Enhancing generalizability of a machine learning model for infrared thermographic defect detection by using 3d numerical modeling

Arsenii Chulkov
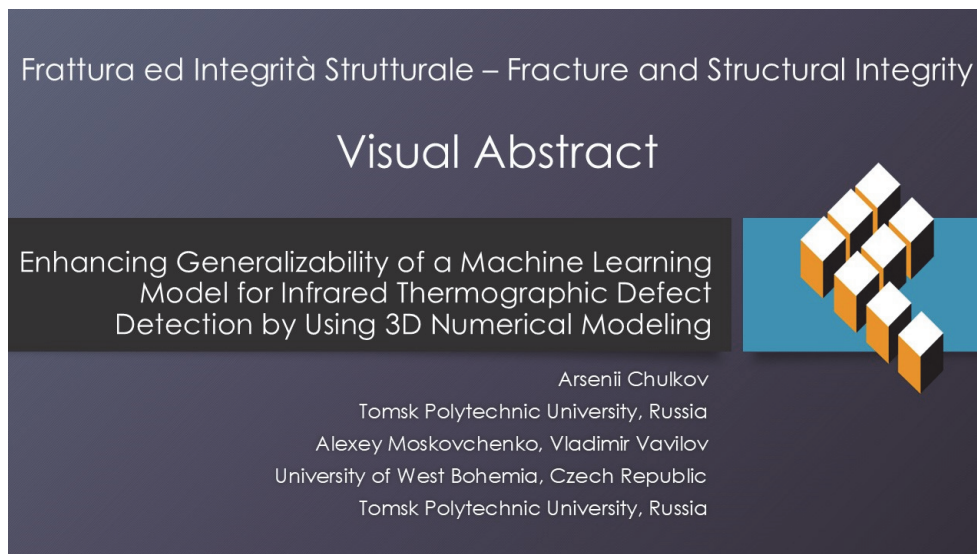*Tomsk Polytechnic University, Russia*
chulkovao@tpu.ru, http://orcid.org/0000-0003-1226-0013

Alexey Moskovchenko
University of West Bohemia, Czech Republic
alexeym@ntc.zcu.cz, http://orcid.org/0000-0002-2813-2529

Vladimir Vavilov
Tomsk Polytechnic University, Russia
vavilov@tpu.ru, http://orcid.org/ 0000-0002-9828-7374

Frattura ed Integrità Strutturale – Fracture and Structural Integrity

Visual Abstract

Enhancing Generalizability of a Machine Learning Model for Infrared Thermographic Defect Detection by Using 3D Numerical Modeling

Arsenii Chulkov
Tomsk Polytechnic University, Russia
Alexey Moskovchenko, Vladimir Vavilov
University of West Bohemia, Czech Republic
Tomsk Polytechnic University, Russia

INTRODUCTION

Infrared (IR) thermography is a method of non-destructive testing (NDT) based on the analysis of thermal patterns on the surface of objects under test by using thermal imagers [1]. Thermal stimulation of objects and subsequent analysis of temperature distributions allow detecting structural defects and thermal anomalies in various materials. Due to its simplicity, non-contact nature of testing and capacity to swiftly assess large areas, IR thermographic NDT has become

widespread in the aerospace industry [2, 3], power production and civil engineering [4–6], including evaluation of composite materials (carbon and glass fiber plastics) [7]. However, IR thermography possesses some drawbacks, such as a diffusive nature of heat conduction, sensitivity to environmental conditions, need for properly-trained operators, etc.

Flash thermography (FT) involves the use of powerful heat sources (e.g. flash tubes and lasers) generating short optical pulses which can be considered as Dirac pulses [8]. Due to short observation times, flash thermography provides high-resolution thermal images facilitating the detection of small-size defects in materials and structures.

The integration of automation and machine learning into IR thermography has significantly expanded potentials of this technique [9–11]. In 2016, Khodayar et al. outlined the use of artificial intelligence as the "2050-horizon" in IR thermographic NDT [12]. The state-of-the-art and recent improvements in Convolutional Neural Networks (CNN) was presented in 2018 by Jiuxiang Gu et al. [13]. In the recent review paper, Yunze He et al. stated that the rapid development of deep learning makes IR machine vision and IR thermographic NDT more intelligent thus contributing to broadening applications of these techniques [14].

In active IR thermographic NDT of materials, the principles of deep learning can be helpful in solving inverse heat conduction problems, i.e. performing defect characterization of hidden defects that is a permanent challenge in thermal NDT (TNDT). For example, Yousefi et al. showed that CNNs can serve as an unsupervised extractor of defect features in IR NDT [15]. Haiyi Wu et al. proposed the deep learning model based on CNNs with a U-shape architecture to predict the heterogeneous distribution of circle-shaped fillers in composites [16]. In pulsed TNDT, a couple of different CNNs were investigated by Qiang Fang et al. with experiments being fulfilled on a series of academic test samples with bottom hole defects and Teflon inserts [17]. The same team used a finite-element model to calculate defect responses in carbon fiber reinforced polymer (CFRP) to be further used for determining defect depth by means of a new technique, which employed the so-called Gated Recurrent Units [18].

To summarize the above-mentioned, one may state that machine learning algorithms, being trained on appropriate datasets, can autonomously analyze IR thermographic data, enhance defect detection and make trust-worthy decisions. However, a problem of generalizability remains one of the most challenging while using artificial intelligence approaches. The respective neural network models often prove their efficiency only under specific conditions, i.e. if they are trained on particular training setups and sample datasets [19, 20].

This study was motivated by the fact that the machine learning models trained on datasets with fixed parameters yield limited defect detection and characterization capabilities. The results obtained provide a useful scientific contribution to the field of defect detection using IR data and machine learning. First, it presents a comprehensive evaluation of the generalizability of machine learning models trained on datasets with varying degrees of parameter variability. By systematically manipulating numerical model parameters, such as defect depth, material thermal conductivity and sample thickness, this study provides a detailed understanding of how these factors influence model performance. This approach offers valuable insights into the optimal design of training datasets, highlighting the need for a balanced data variability to enhance model robustness without compromising accuracy. Secondly, the introduction of multiple test datasets, each with distinct unseen parameter variations, represents a novel methodology for assessing model generalizability. This rigorous testing framework goes beyond conventional validation approaches by simulating real-world scenarios where defects and material properties may differ significantly from those used in training. This aspect of the study demonstrates the practical applicability of the proposed machine learning models, showcasing their potential to reliably detect defects in diverse and unpredictable environments.

The outline of the paper is as follows. First, the theory of FT and basic processing approaches will shortly be introduced. Next, a couple of training datasets used for machine learning will be developed by means of advanced 3D numerical modeling. Then, these datasets will be used for evaluating efficiency of a particular Gaussian Support Vector Machine (SVM) model in characterizing defect parameters. The robustness of the suggested learning machine model toward noise of an additive and multiplicative nature will be explored. Finally, some data processing algorithms will be analyzed to demonstrate that the use of Thermographic Signal Reconstruction (TSR) and Temperature Contrast significantly improve the model efficiency.

THEORY

F T is based on applying a brief heat pulse onto a material under examination followed by measurement of material temperature response by means of an IR camera [21]. Typically, such heat pulses last only few milliseconds, and the analysis focuses solely on the material thermal response following the pulse, i.e. at the cooling stage of the thermal process. The time-temperature responses at sample surface points are then subjected to processing in order to extract meaningful information on subsurface defects.

Classical heat conduction solutions were exhaustively summarized by Carslaw and Jaeger [22]. The solutions for pulsed, continuous or harmonic heating of an adiabatic semi-infinite body and slab are often used. For example, the surface response of a semi-infinite body toward Dirac-pulse heating is given by a simple formula [23]:

$$\frac{T(t)}{Q_o} = \frac{1}{\sqrt{\pi \rho C K t}} = \frac{1}{e\sqrt{\pi t}} \qquad (1)$$

where $T(t)$ represents the temperature on the sample surface at the time t, Q_o stands for the energy of the heat pulse, while ρ, C, K and e denote the material density, specific heat capacity, thermal conductivity and thermal effusivity ($e = \sqrt{\rho C K}$), respectively.

A region containing a delamination-like defect may be considered as a slab of the thickness d. The pulse response of a slab can be expressed as [24]:

$$T(t) = \frac{Q_o}{e\sqrt{\pi t}}[1 + 2\sum_{n=1}^{\infty} R^n e^{-\frac{(nd)^2}{\alpha t}}] \qquad (2)$$

where α is the thermal diffusivity, R is the thermal reflection coefficient describing the effusivity contrast between two materials at the defect boundary, and n is a summation index. The temperature contrast between defect and non-defect areas can be obtained by subtracting Eq. (1) from Eq. (2):

$$\Delta T(t) = \frac{2Q_o}{e\sqrt{\pi t}}[\sum_{n=1}^{\infty} R^n e^{-\frac{(nd)^2}{\alpha t}}] \qquad (3)$$

Nonetheless, practical procedures of TNDT are often aggravated by significant noise/clutter, non-uniform heating, variations in material thermal properties and other complexities. Consequently, the applicability of robust but simple analytical methods may be limited in practical TNDT problems.

Thermographic data processing techniques
In TNDT, similarly to many other inspection techniques, the data analysis focuses on evaluating various forms of signal contrasts [25]. These contrasts underline difference between signals in defect and defect-free areas. TNDT employs some definitions of contrasts used as figures of merit, including but not limited to absolute contrast (AC), running and normalized contrasts (NC).

Absolute thermal contrast involves calculating the temperature difference between defect (*d*) and non-defect (*nd*) areas. However, a significant limitation of both AC and NC is the necessity to identify a non-defect, or sound, area. This requirement poses a challenge in data processing, in particular, when locations of defects are not *a priori* known.

In [26], an automated method for identifying a reference zone was introduced being based on the determination of a minimal value of the integral involving the $T\sqrt{t}$ function. Since a reference (non-defect) area is determined, the dimensionless running contrast can be obtained as follows:

$$C^{run} = \frac{T_d - T_{nd}}{T_{nd}} \qquad (4)$$

The Pulse Phase Thermography (PPT) technique, initially proposed by Maldague and Marinetti [27], combines advantages of pulsed and thermal wave TNDT. In fact, any form of thermal stimulation, be it a flash or pulse of a certain duration, can be represented as a combination of harmonic thermal waves, therefore, it is fruitful to examine the propagation of individual waves within a solid material and their interaction with structural inhomogeneities, or defects.

The process involves monitoring the surface temperature with an IR camera after the heat pulse was delivered onto sample surface. Subsequently, the discrete Fourier transform is applied to the $T(t)$ data, resulting in calculation of signal phases as

a function of frequency $\varphi(f)$. The detection of subsurface defects is based on observing the phase difference $\Delta\varphi$ between defect and non-defect areas that are identified in phase images (phasegrams).

TSR is a well-established method proposed by Shepard for processing pixel-based temperature evolutions [28, 29]. The technique is based on the polynomial fitting of experimental temperature data in the Log-Log scale. The fitting procedure effectively replaces a raw set of temperature data with a series representing polynomial coefficients. Such approach facilitates reconstruction of initial thermographic data to effectively discriminate defect and non-defect areas. Subsequently, the first and second derivatives of logarithmic temperature are analyzed thus contributing to enhancing the signal-to-noise ratio and producing sharp images of subsurface defects. Also, the derivative analysis is efficient in characterizing defect depth.

Machine learning

Machine learning techniques have gained popularity in IR thermography due to their ability to automatically learn and adapt from data. They can enhance defect detection by recognizing subtle patterns, extracting features and making decisions based on a learned knowledge [30].

In supervised learning, machine learning models are trained on labeled datasets containing examples of both defect and non-defect thermographic data. These models learn to distinguish between the two input classes enabling them to identify defects in new, unlabeled data.

Machine learning models can be applied to defect detection as classification tasks (e.g., identifying whether an image contains a defect) or regression tasks (e.g., estimating the size or depth of defects).

In this study, the emphasis is made on the binary classification task and pixel-by-pixel analysis of temperature evolutions that allows classification of points in thermographic images as belonging to either defect or non-defect areas.

Some relatively simple models, such as SVM and Bagged Trees, were chosen to focus on how the variability in the training datasets influences the model performance. By using these models, it became possible to systematically analyze the impact of dataset variability on defect detection accuracy without the added complexity of more advanced and computationally-intensive algorithms. It is worth mentioning that these models have also demonstrated good performance in IR thermography applications in plentiful previous studies [31, 32, 33].

The SVM model has been chosen because it has demonstrated a good performance in defect classification when processing raw temperature data. Its theoretical foundation is rooted in the concept of finding an optimal hyperplane in a high-dimensional feature space to best separate data points belonging to different classes [34]. The SVM concept is to identify a hyperplane that maximizes the margin, which represents the distance between the hyperplane and the nearest data points (called support vectors) from each class. This margin maximization not only leads to a better generalization but also improves the model robustness toward outliers.

SVMs are powerful machine learning models that excel in finding optimal decision boundaries for both linearly and non-linearly separable data. They are widely used in various applications including image and text classification, anomaly detection, etc., thanks to their robust theoretical foundation and versatility.

Ensemble machine learning models aggregate predictions from multiple base models to create a final prediction. The Bagged Trees ensemble method involves training multiple decision trees on different subsets of training data and combining their predictions. This approach significantly reduces overfitting, a common challenge in machine learning, especially when dealing with varied and noisy data. Bagged Trees have shown effectiveness in numerous studies, providing robust and reliable results across different applications.

By choosing the models above, this study aimed to ensure transparency and interpretability in analyzing the crucial role of dataset composition in machine learning-based defect detection systems. The focus on dataset variability is essential for understanding and improving generalizability and accuracy of defect detection methods.

NUMERICAL SIMULATION

In this study, the viability of training defect detection models by using IR thermographic data derived from numerical simulations will be assessed. The accent is made on a thorough examination of the impact of model parameters and dataset size on the performance of the models, which are applied to previously unobserved data. To achieve this objective, some multiple sets of training and testing datasets, each characterized by different model parameters, were generated. These models were designed by using the ThermoCalc-3D software, a specialized tool developed by Tomsk Polytechnic University for simulating heat transfer processes in solid materials with defects by using the finite difference method.

The numerical simulation of 3D heat conduction problems yields temporal evolutions at all surface points of a solid body subjected to uniform or uneven heating. The model used represents a rectangular plate containing air-filled parallelepiped-like defects. A visual presentation of the model is shown in Fig. 1. The example of calculated temperature distributions is given in Fig. 2.
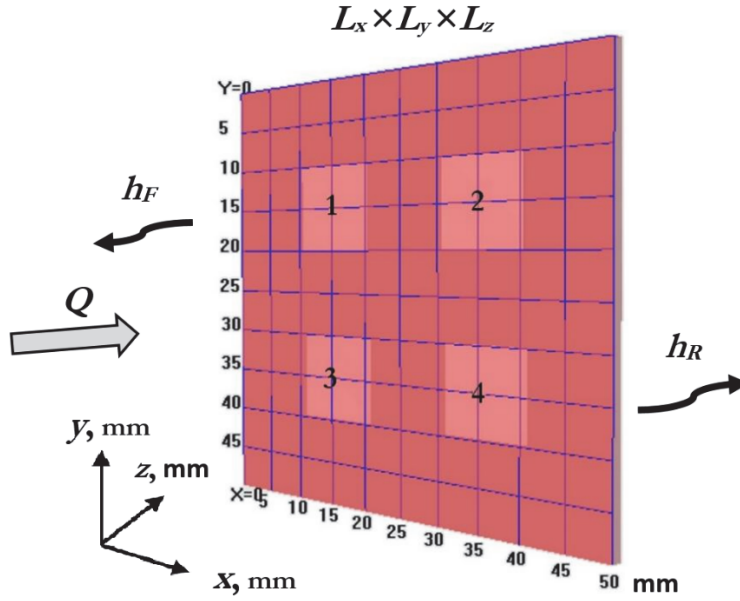

Figure 1 : TNDT 3D numerical model.

The general mathematical formulation of the 3D model of non-adiabatic heat conduction in a multilayer body with defects accepted in ThermoCalc-3D is as follows (Fig. 1):

$$\frac{\partial T_i(x,y,z,\tau)}{\partial \tau} = \alpha_i^x \cdot \frac{\partial^2 T_i(x,y,z,\tau)}{\partial x^2} + \alpha_i^y \cdot \frac{\partial^2 T_i(x,y,z,\tau)}{\partial y^2} + \alpha_i^z \cdot \frac{\partial^2 T_i(x,y,z,\tau)}{\partial z^2} \tag{5}$$

$$i = 1 \div 76 \quad (36 \; layers + 40 \; defects)$$
$$T_i(\tau = 0) = T_{in} \tag{6}$$

$$-K_1^z \cdot \frac{\partial T_1(x,y,z=0,\tau)}{\partial z} = Q(x,y,\tau) - h_F \cdot [T_1(x,y,z,\tau) - T_{amb}] \tag{7}$$

$$K_m^z \cdot \frac{\partial T_3(x,y,z=L_z,\tau)}{\partial z} = -h_R \cdot [T_3(x,y,z,\tau) - T_{amb}] \tag{8}$$

$$\frac{\partial T_i(x,y,z,\tau)}{\partial x} = 0 \; \text{for} \; x = 0, \, y = 0 - L_y; \quad x = L_x, \, y = 0 - L_y$$
$$\frac{\partial T_i(x,y,z,\tau)}{\partial y} = 0 \; \text{for} \; y = 0, \, x = 0 - L_x; \quad y = L_y, \, x = 0 - L_x \tag{9}$$

$$T_i(x,y,z,\tau) = T_{i\pm1}(x,y,z,\tau) \; \text{and} \; K_i^{q_j} \cdot \frac{\partial T_i(x,y,z,\tau)}{\partial q_j} = K_{i\pm1}^{g_j} \cdot \frac{\partial T_{i\pm1}(x,y,z,\tau)}{\partial q_j} \tag{10}$$

Here: T_i is the temperature in the i-th region counted from the initial object temperature (i =1-36 corresponds to 36 specimen layers, i =37-76 corresponds to 40 defects); T_{in} is the specimen initial temperature; $\alpha_i^{q_j}$, $K_i^{q_j}$ are the thermal diffusivity and the thermal conductivity in the i-th region by the coordinate q_j; x, y, z are the Cartesian coordinates; q_j is one of the Cartesian coordinates x, y or z (j =1-3); τ is the time; $Q(x, y, \tau)$ is the power density of the absorbed heat flux that, in a general case, varies in both time and space; h_F, h_R are the heat exchange coefficients on a front and rear surface respectively; these coefficients combine both the radiation and convection phenomenon; m is the number of layers (m =36), T_{amb} is the ambient temperature; L_x, L_y, L_z are the specimen dimensions.

Eqn. (5) is the 3D parabolic equation of heat conduction; Eqn. (6) is the initial condition; Eqn. (7) is the boundary condition on a front surface (heating and cooling); Eqn. (8) is the boundary condition on a rear surface (cooling only); Eqns. (9) are the adiabatic conditions on side surfaces by the coordinates x and y; Eqns. (10) are the temperature and heat flux continuity conditions on the boundaries between layers and between layers and defects.

Note that the ThermoCalc-3D allows modeling a 36-layer plate containing up to 40 parallelepiped-like defects. In this study, a classical 1-layer plate with 4 defects was modeled, see Fig. 1. By using a numerical grid including up to several million nodes, ThermoCalc-3D ensures accuracy of calculating non-defect temperatures under 0.5% and defect temperatures under 3% compared to known 1D analytical solutions.

The following model parameters were chosen: plate lateral size 50 x 50 mm, number of numerical grid steps by X, Y, Z – axes 50 x 50 x 100, lateral size of defects 10 x 10 mm; defect thermal properties (air): $k = 0.07$ W·m^{-1}·K^{-1}, $\rho = 1.3$ kg·m^{-3}, $C = 928.4$ J·kg^{-1}·K^{-1}, heat time 0.02 s (square pulse), time step 0.02 s, number of collected frames 250, ambient and initial temperature 0°C, and spatial distribution of the heat pulse is Gaussian:

$$Q = Q_0 e^{-\sigma_x (x_o - x)^2 - \sigma_y (y_o - y)^2} \tag{11}$$

where σ_x, σ_y - coefficients of spatial distribution of heat pulse, m^{-2}, $x_o = y_o = 25$ mm - coordinates of heat source center (sample center).

Some model parameters presented in Tab. 1 varied to produce different datasets.

The Train 1-6 datasets include changeable model parameters, such as material thermal properties, specimen thickness and heating power. Although not all combinations of the parameters were calculated, the total number of datasets used for training reached 63.

The first training dataset (Train 1) represented a particular numerical model with thermal properties corresponding to those of CFRP. Train 2 incorporated variations in heat pulse energy and spatial distribution. In the Train 3 dataset, sample thickness, heat pulse power and spatial distribution varied. As mentioned above, not all combinations of input data were calculated but each parameter value occurred at least once. For instance, for a sample thickness of 1 mm, the heat power was set at 200,000 W/m², and the spatial distribution coefficients were 50 m^{-2}. Train 4 introduced variations in thermal conductivity, heat power and spatial distribution, while in the Train 5 dataset, there were three combinations of sample thickness, two combinations of heat capacity and density, as well as variations in thermal conductivity, heat power and spatial distribution. The Train 6 dataset was the most comprehensive including wide-range variations in sample thickness, thermal properties and heat power.

Furthermore, in order to evaluate the training datasets, three additional datasets of varying complexity were calculated. Test 1 represented the simplest variant comprising a 1 mm-thick plate (similar to the training models) but with slight differences in thermal conductivity, defect depth and heat power, to compare to the first training dataset. The second Test 2 dataset differed from the respective training model by defect depth and material thermal properties that corresponded to 1 mm-thick glass fiber and polyamide composites. The Test 3 dataset was most complex additionally including varying material thickness.

Each pixel of the calculated IR images was categorized as related to either defect or defect-free area. The obtained 3D IR thermographic sequences were then transformed into 2D matrices of feature vectors for model training and testing purposes. Each feature vector encapsulated the temperature evolution of an individual pixel. Fig. 2 represents the examples of the simulated thermograms for the Train 1 model.

| Dataset	Sample thickness, mm	Density, kg·m⁻³, and heat capacity, J·kg⁻¹·K⁻¹	Conductivity in Z-direction, W·m⁻¹·K⁻¹	Defect depth, mm	Defect thickness, mm	Heat pulse spatial distribution σ_x, σ_y, m⁻²	Maximum heat pulse density, W·m⁻²
Train 1	1	500, 2462	0.64	0.2 0.4 0.6 0.7	0.05	100, 50	200 000
Train 2	1	500, 2462	0.64	0.2 0.4 0.6 0.7	0.05	50, 50 100, 100	150 000 200 000 250 000 300 000 350 000 400 000
Train 3	1 2 3 4 5 10 15	500, 2462	0.64	0.2 0.4 0.6 0.7	0.05	50, 50 100, 100	150 000 200 000 250 000 300 000 350 000 400 000
Train 4	1	500, 2462	0.1 0.3 0.5 0.7 0.9 1.1	0.2 0.4 0.6 0.7	0.05	50, 50 100, 100	150 000 200 000 250 000 300 000 350 000 400 000
Train 5	1 3 5	500, 2462 2100, 800	0.1 0.3 0.5 0.7 0.9	0.2 0.4 0.6 0.7	0.05 0.1	50, 50 100, 100	150 000 200 000 250 000 300 000 350 000 400 000
Train 6	1 3 5 7 10	500, 2462 800, 2100 800, 3000 900, 4000	0.1 0.3 0.5 0.9 1.2 1.5	0.2 0.4 0.6 0.7	0.05 0.1	50, 50 100, 100	150 000 200 000 250 000 300 000 350 000 400 000
Test 1	1	500, 2462	0.65	0.15 0.35 0.55	0.05	70, 70	280 000
			0.8	0.25 0.45 0.65	0.05	70, 70	270 000
Test 2	1	1240, 1800	0.13	0.15 0.35 0.55	0.1	70, 70	270 000
	1	1900, 1200	0.3	0.15 0.35 0.55	0.05	70, 70	230 000
Test 3	5	1240, 1800	0.13	0.15 0.35 0.55	0.1	70, 70	270 000
	3	1900, 1200	0.13	0.2 0.4 0.6	0.1	70, 70	290 000
	1.5	1240, 1800	0.2	0.1 0.3 0.5	0.1	70, 70	290 000
	2	1900, 1200	0.3	0.2 0.4 0.6	0.1	70, 70	290 000

Table 1: Numerical model parameters.
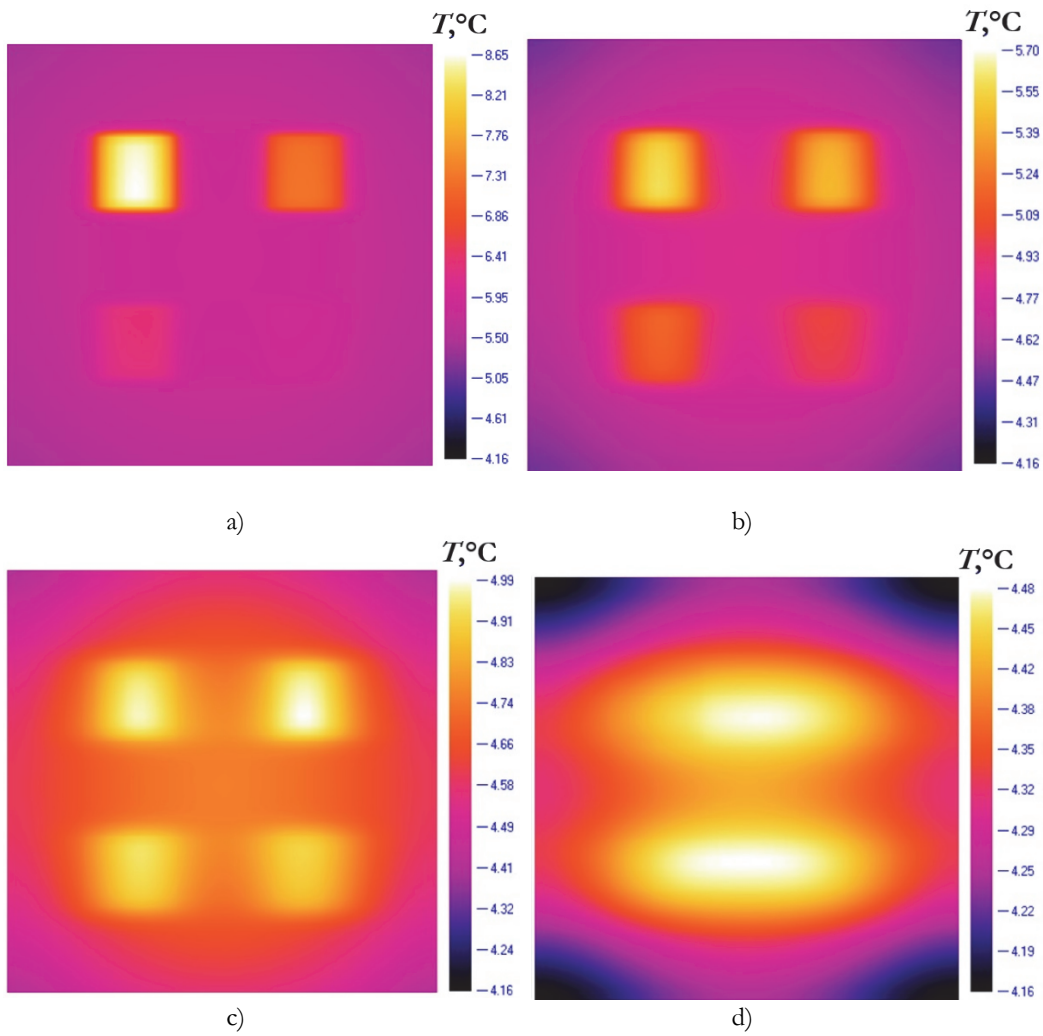
Figure 2: Simulated thermograms (Train 1 dataset) at times 0.3 s (a), 0.7 s (b), 1.5 s (c) and 5 s after heating.

The temperature evolutions at defect and non-defect points are presented in Fig. 3 in the Log-Log scale. The blue curves represent temperature evolutions over the defect centers. The red curves represent non-defect temperature evolutions at the corner points, in the middle of the sample and between the defects.
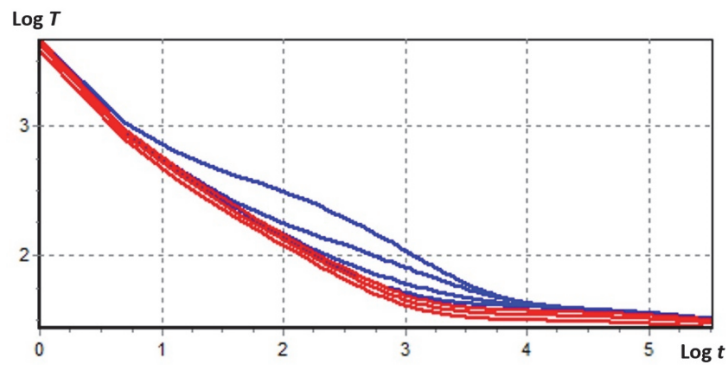

Figure 3: Modelled temperature evolutions at defect (blue) and non-defect (red) points (Log T *vs.* Log t).

EVALUATION OF DATASETS AND MODELS PERFORMANCE

A Gaussian SVM and Bagged Trees models were trained using the six datasets (Train 1-6), and their performance was assessed by using the validation data and three distinct test datasets (Test 1-3).

The validation data was used to evaluate the model at the training phase. This data was not used in the very training process, but it helped to tune hyperparameters and prevent overfitting. The models were trained by a 5-folds cross-validation scheme. This means that each train dataset was randomly divided into 5 subsets of approximately equal size. Each subset represented a fold. The model was trained and evaluated 5 times, each time using a different fold as the validation set and the remaining 4 folds served as the training set. After having performed training on the training set, the model performance was evaluated on the validation set (the remaining fold) to estimate how well it responded to unseen data.

Finally, the performance metrics (True Positive Rate-TPR and True Negative Rate-TNR) obtained from 5 iterations (5 validation sets) were averaged to provide an overall assessment of the model performance. Then models were trained on all train datasets and evaluated by test datasets.

To summarize, the validation data was used during the training phase to adjust model parameters or hyperparameters and also prevent overfitting. This can indirectly influence the performance of the model, while the test data provided an independent evaluation of the final model performance on new, unseen data.

The performance of two machine learning models, namely, SVM and Bagged Trees Ensemble, was evaluated by analyzing different training and testing datasets. The used evaluation metrics included Sensitivity (True Positive Rate, TPR), Specificity (True Negative Rate, TNR), and Precision (Positive Predictive Value, PPV). These metrics were chosen to give a comprehensive understanding of the model performance in detecting defects. The results are presented in Tabs. 2 and 3, illustrating the model performance across various datasets. Since in many cases the TPR was small, some models classified all data as corresponding to defects, leading to TPR and TNR values of 100 and 0, respectively. In these cases, the tables indicate zero values, thus indicating that the model is not appropriate.

It is important to note that Accuracy is not a representative metric in this context because it does not account for the imbalance between defect and defect-free cases. Instead, the metrics like TPR, TNR, and Precision (PPV) provide a clearer picture of model performance in detecting defects.

	Sensitivity (TPR)/ Specificity (TNR)/Precision (PPV) %					
Dataset	Train 1	Train 2	Train 3	Train 4	Train 5	Train 6
Validation	**99.8**/100/99.9	**78.4**/99.9/93.2	**45.9**/100/84.5	**96.4**/100/99.1	**87.2**/100/96.9	**69**/99.2/92.7
Test 1	100/**0**/100*	**98**/99.5/99.3	**87.3**/99.7/95.9	**0**/100/79.7	**86.6**/100/95.6	**62.5**/100/88.7
Test 2	100/**0**/100	100/**0**/100	100/**33.9**/100	**0**/100/79.7	**82.5**/99.6/95.7	**96.5**/97.1/99.1
Test 3	100/**0**/100	100/**0**/100	100/**0**/100	**0**/100/79.7	**75.9**/90/91.7	**84**/80.3/93.6

Table 2: SVM Model performance for different datasets.

	Sensitivity (TPR)/ Specificity (TNR)/Precision (PPV) %					
Dataset	Train 1	Train 2	Train 3	Train 4	Train 5	Train 6
Validation	99.5/99.8/99.8	98.9/99.9/99.6	99/99.8/99.6	99.8/100/99.9	99.4/100/99.9	96.9/99.4/99.2
Test 1	100/0/100	26.5/68.8/73.2	30.4/67.2/74	0/100/79.7	55.6/52.5/77.7	37/57.1/72.8
Test 2	100/1.6/100	100/0/100	100/0/100	0/100/79.7	50/57.9/82	100/6.4/100
Test 3	100/0/100	100/0/100	100/0/100	0/100/79.7	82.6/40.9/87.4	87.7/40/90.5

Table 3: Bagged Trees Ensemble model performance for different datasets.

The tables allow evaluating model performance by using different training datasets. The model trained on the results of the single simulation with the fixed parameters (Train 1) failed on all test datasets.

The introduction of heat power variations in the model yielded the best results for the Test 1 dataset with slight differences in the parameters, but it failed to identify defects in more complex cases (Test 2 and 3).

Surprisingly, introducing variations in plate thickness (Train 3) from 1 to 15 mm produced an adverse effect, causing the model to perform worse even on the validation dataset.

Similarly, the introduction of variations in thermal conductivity (Train 4) did not improve model performance not only on Test 1 dataset 1 but also failed to identify defects in Test 2 and 3 datasets. This shows that varying only thermal conductivity is not sufficient to generalize the model onto different materials, see Train 2.

The most promising results were achieved by training the model on the Train 5 dataset, which included slight variations in sample thickness (from 1 to 5 mm) and thermal conductivity (from 0.2 to 0.7 W·m^{-1}·K^{-1}). Additionally, it introduced the combination of density and heat capacity not present in the test data. This model effectively identified most of defects across all test datasets.

However, the use of the training dataset with a greater variability in model parameters (Train 6) resulted in a worse performance for validation and Test 1 datasets.

Tab. 2 demonstrates that increasing the size and variation of the training data can negatively affect the results obtained on both the validation data and the data not very different from the training one (Test 1).

The Bagged Trees Ensemble model shows poor performance for Test 1, Test 2 and Test 3 with Train 1, Train 2, and Train 3, reflecting a high false positive rate (Specificity at 0% in most cases). The model performs better with Train 5 and Train 6 indicating that less variability in training data improves the generalizability.

The SVM model struggles with overfitting when trained on highly variable datasets, as evidenced by the poor Specificity in certain test cases.

The Bagged Trees Ensemble model, while showing robust validation performance, also encounters the problem of overfitting, especially with training datasets of high variability.

Both models benefit from training datasets with controlled variability (Train 5 and Train 6) thus enhancing their generalization ability in respect to unseen test data.

The performance generally improves with the variability of numerical model parameters in training sets 1 through 5. However, the performance decreases with too much variability, as shown by Train 6. This demonstrates the need to find an optimal balance in training data variability to achieve the best model performance.

To conclude, the proposed machine learning models has proven to be efficient in the case of varying heat power and spatial distribution when applied to materials with similar thermal properties and thickness. Nevertheless, to enhance the model generalizability by involving different materials and thicknesses, training data parameters are carefully to be chosen. For example, excessive variability in the training data may compromise model performance producing worse evaluation results and failing to improve overall generalizability.

Learning curve evaluation

In this section, a comprehensive evaluation of the SVM machine learning model performance is presented through the analysis of learning curves. Learning curves are crucial diagnostic tools, which illustrate the model learning process by plotting the training and validation errors against different training set sizes. Learning curves provide valuable insights into the model performance and behavior during training. They help to understand how well the model is being learned from the data and whether it generalizes well to new, unseen data. Even with different training and testing datasets aimed at evaluating generalizability, learning curves can still offer meaningful information. The particular learning curves have been constructed using two distinct training datasets: one with a higher variability in properties and parameters (Train 6) and another with a lower variability (Train 5), see Fig. 4. This comparison between the two training datasets highlights the impact of dataset variability on the model performance thus illustrating the differences between a sufficiently comprehensive dataset and one that might be an overly variable.

It is important stressing that a training curve shows how the model performance evolves on the training data as the number of training examples increases. In its turn, the validation curve shows how the model performance evolves on the validation or test data. If the training accuracy is high but the validation accuracy is low, the model is likely overfitting. If both the training and validation accuracies are low, the model is likely underfitting.

While considering training error curves, in the case of the lower variability dataset (Train 5), the training error starts as high as 21.97% but then drops sharply to 2.23%. This suggests that the model learns more quickly when the variability in the training data is lower. In the case of the higher variability dataset (Train 6), the training error decreases steadily from 18.06% to 5.98%, showing a gradual improvement as more data is used for training.

Validation error curves were first evaluated on Test 1 set. In the case of the lower variability, the validation error starts at 25.33% but shows a significant drop to 3.21% with some fluctuations. It suggests the better performance on this dataset, especially with higher training sizes. In the case of the higher variability, the validation error decreases from 25.33% to 9.50% showing good generalizability to this set.

The results obtained on the Test 2 set show that, in the case of the Lower Variability, the validation error starts at 20.28% and decreases to 3.55% with slight fluctuations. The model performs well but not as consistently as with the higher variability data.
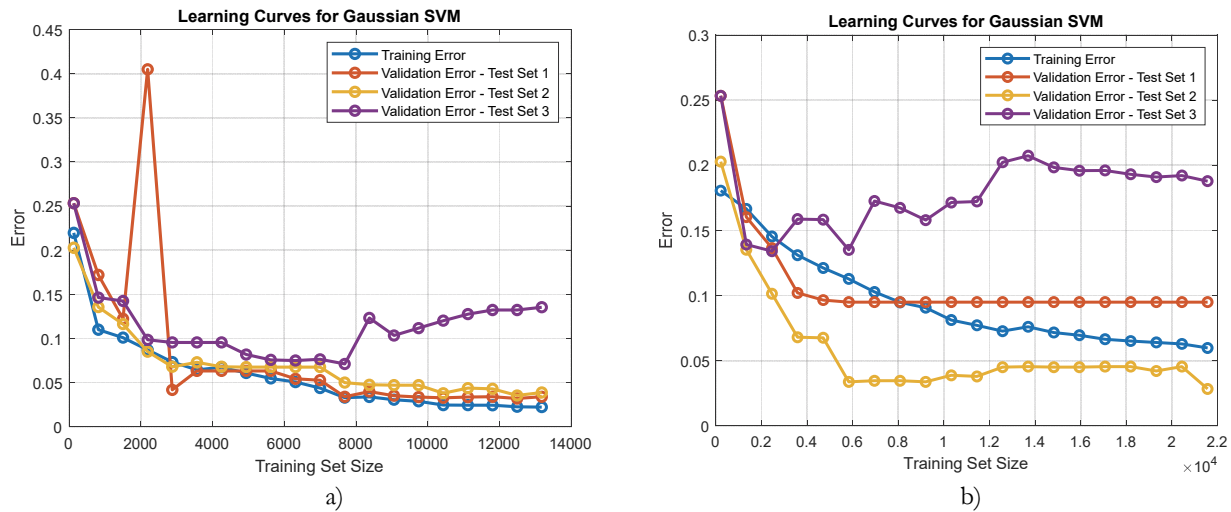

Figure 4: Learning curves for Gaussian SVM model trained on Train 5 (a) and Train 6 (b) datasets.

While considering the higher variability of the data, the validation error decreases from 20.28% to 2.84% indicating good generalization to different thermal properties and defect depths.

Finally, the Test 3 set was evaluated to reveal that, in the case of the lower variability, the validation error starts at 25.33% and diminishes to around 13.55% with a better consistency than in the case of the higher variability data but still showing significant fluctuations. If the data is characterized by the higher variability, the validation error fluctuates between 13.92% and 20.23% thus showing that the model struggles with the diverse conditions in this dataset.

The analysis of model performance shows that the model trained on the lower variability dataset shows the steeper decrease in the training error, indicating faster learning. The validation errors for the lower variability datasets are generally lower, suggesting the better performance and generalizability.

Considering generalizability leads to the following conclusions. The model trained on the higher variability dataset performs well on the Test 1 and 2 sets but struggles with the diverse conditions in the Test 3 set. The model trained on the lower variability dataset performs better on all three validation sets with errors being lower and more consistent. Training with lower variability data may help the model to learn more quickly and perform better on similar validation sets.

For datasets with higher variability, more sophisticated models or additional training data may be required to improve generalizability.

Overall, the model trained on the lower variability dataset shows better performance and generalizability indicating that reducing variability in training data may lead to more robust models.

Assessing Robustness to Noise

Tab. 4 demonstrates the resistance of the model toward noise of two types. Additive and multiplicative Gaussian-type noise with varying standard deviations (STD) was introduced into the Test 2 dataset, and the performance of the optimal model (trained on the Train 5 dataset) was subsequently evaluated. It is worth reminding that additive noise is conditioned by background thermal reflections and ultimately represents a random noise of an IR detector; in the most IR imagers this kind of noise can be assumed 0.01-0.1°C. Additive noise is added to temperature evolutions recorded in TNDT tests. In its turn, multiplicative noise is mainly determined by material surface clutter, such as natural inhomogeneities in absorptivity/emissivity, and it is proportional to the sample excess temperature; the minimum amplitude of multiplicative noise is about 2-4% for black body-like materials [23].

For additive noise with a standard deviation of up to 0.2°C and multiplicative noise up to 2%, the model quality metrics have demonstrated only marginal reductions. With the additive noise increased to 0.7°C, the True Positive Rate (TPR) has revealed a minor decline, while the Negative Predictive Value (NPV) has demonstrated a more significant reduction thus indicating an increase in false positive indications. It is interesting that the introduction of the additive noise with STD=1°C has resulted in the slightly higher TPR but significantly diminished the NPV down to 49.3%. This can be explained by the

complicated combinations of defect and noise signals, which may appear in some particular cases. It seems that, with noise added, some "defect" pixels not found in the raw data may be correctly identified. Respectively, the TPR may grow up, but the number of false positive indications also increases. The results in Tab. 4 show that additive noise has corrupted the model performance more than multiplicative noise. This can probably be explained by relatively low amplitudes of the multiplicative noise (not higher than 4%).

STD of additive noise, °C	STD of multiplicative noise, %	Sensitivity (TPR), %	Specificity (TNR), %	Precision (Positive Predictive Value, PPV), %	Negative Predictive Value (NPV), %
0.1	0	82.3	99.6	95.7	98.3
0.2	0	81.7	99.0	95.5	95.4
0.5	0	80.5	94.1	95.0	77.6
0.7	0	80.5	88.1	94.7	63.2
1	0	82.7	78.3	94.7	49.3
0	2	81.7	99.3	95.5	96.8
0.2	2	81.1	98.5	95.3	93.1
0.5	2	80.7	92.3	94.7	72.7
0	4	79.6	97.0	94.9	87.2
0.2	4	79.6	96.4	94.9	85.1
0.5	4	79.6	92.4	94.7	72.6

Table 4: Model resistance toward noise.

Thermographic data processing

This section explores efficiency of some known data processing algorithms, namely, Thermographic Signal Reconstruction (TSR), Pulse Phase Thermography (Fourier transform), and Temperature Contrast. Tab. 5 shows the quality metrics of the model (the minimum value between TPR and TNR), which was trained on the Train 5 dataset processed by using the above-mentioned algorithms.

Dataset	Efficiency			
	Raw temperature data	Fourier phase	TSR, 1st derivative	Contrast
Validation	94.4	89.6	97.0	99.6
Test 1	86.6	87.5	87.5	99.9
Test 2	82.5	50.0	99.2	98.1
Test 3	75.9	37.5	80.8	98.4

Table 5: Data processing efficiency (Machine Learning model).

The table illustrates that the use of Fourier phasegrams as input images has surprisingly corrupted the model performance making it inappropriate for detecting defects in the Test 2, 3 datasets. On the contrary, the use of the first derivative and contrast data has led to a notable enhancement of the model quality. For example, in the case of contrast, the sensitivity values consistently surpassed 98% across all test datasets.

Fig. 5 provides the illustrations to the model efficiency while using various types of the training models, which are applied to one of the sequences related to the Test 3 dataset.

Fig. 5a shows that the deepest defect was not detected when using raw temperature data. The same results but with some noisy indications and more distorted defect patterns were provided by the model where the raw data was corrupted by the Gaussian noise (Fig. 5b). Finally, all defects were detected when the machine learning model was trained on the contrast data (Fig. 5c).
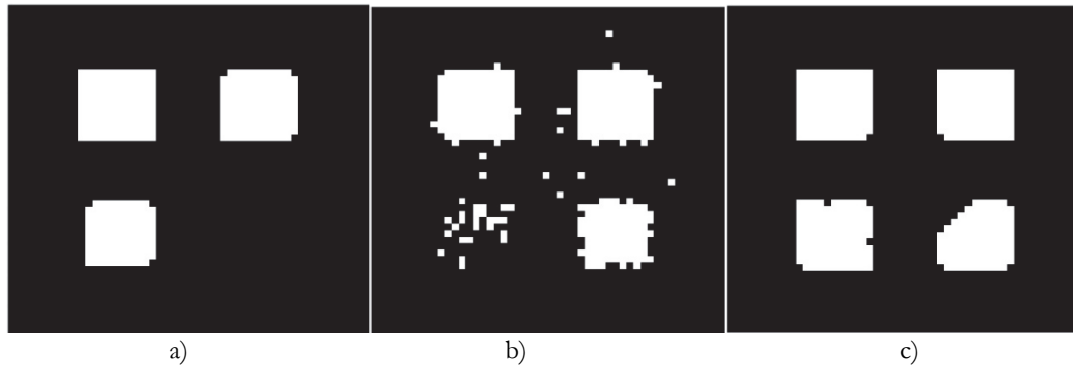
Figure 5: Applying machine learning model: a) raw data training; b) raw data training, Gaussian noise (STD=0.5 °C); c) contrast data training.

Discussion on merits and limitations

This study provides a detailed analysis of how variability in training datasets impacts the performance of machine learning models for defect detection using IR thermographic data. By incorporating different training and testing datasets, this research systematically evaluates model generalizability and robustness.

The choice of relatively simple yet effective models, such as SVM and Bagged Trees Ensemble, allows a clear understanding of how dataset variability influences model performance. These models have demonstrated effectiveness in other studies, thus reinforcing their suitability for particular applications.

The study design, which implements multiple training and testing scenarios, facilitates the understanding of the model generalizability. By testing the accepted models on unseen datasets, this research assesses how well the models can adapt to new data, that is crucial for practical deployment.

However, some limitations of the technique under discussion should be mentioned. Model Complexity: while the use of simple models like SVMs and Bagged Trees Ensemble allows for the clear analysis, it may limit the exploration of more complex relationships within the data; advanced models such as deep learning algorithms could potentially capture more intricate patterns, but they are not considered in this study. Dataset Limitations: the study relies on numerically simulated datasets, which, while controlled, may not fully capture the complexity and variability of real data; future work could include experimental data to further validate the findings. Overfitting Concerns: although cross-validation was used to mitigate overfitting, the performance of the models on highly variable datasets (Train 6) indicates potential overfitting; this suggests that while the models perform well on less variable datasets, their robustness on more complex datasets could be improved.

CONCLUSION

In this study, the possibility of enhancing defect evaluation in IR thermographic NDT through the application of Machine Learning Models has been explored. The suggested SVM and Bagged Trees Ensemble models were trained on the data derived from numerical simulations. The number of model parameters, including material thermal properties, specimen thickness and heating parameters, were analyzed in order to evaluate how general can be a model to be used in machine learning.

It was demonstrated that the models trained on datasets with fixed parameters yielded limited defect detection capabilities. Introducing variations in heating parameters proved to be promising for detecting defects with minor parameter differences, but it appeared unsuccessful in cases that are more complicated. It is worth noting that the introducing of variations in specimen thickness and thermal conductivity worsened the model performance. The Train 5 dataset, which included subtle variations in specimen thickness, thermal conductivity, as well as various combinations of material density and heat capacity, provided the best results and a noticeable ability to identify defects in all test datasets.

Furthermore, the model robustness in regard to noise was explored to demonstrate its ability to withstand additive and multiplicative random noise with a standard deviation up to 0.5 °C for additive and 2% for multiplicative noise. However, with noise greater than the above-mentioned thresholds, the model performance deteriorated increasing false negative indications.

The potentials of some known techniques of thermographic data processing, such as TSR, Fast Fourier Transform and Temperature Contrast, were examined. While the efficiency of Pulse Phase Thermography (Fourier transform) was surprisingly low, the use of the first derivatives (TSR) and contrast data significantly improved the model efficiency. In particular, the use of temperature contrast data ensured sensitivity (TPR) better than 98% across all test datasets.

In conclusion, this study has revealed that machine-learning models exhibit a substantial potential for enhancing defect detection in IR thermographic NDT. However, further expanding results onto different materials and sample thicknesses requires careful selection of training data parameters, as excessive variability in the training data may lead to worsened results. Additionally, by performing proper data processing, in particular, determining temperature contrast, one may significantly enhance model performance. A deeper insight in this research area is a topic of further research.

REFERENCES

[1] Maldague, X. (2001). Theory and Practice of Infrared Technology for Nondestructive Testing, New York, Wiley-Interscience.

[2] Deane, S., Avdelidis, N.P., Ibarra-Castanedo, C., Zhang, H., Nezhad, H.Y., Williamson, A.A., Mackley, T., Davis, M., J., Maldague X. and Tsourdos, A. (2019). Application of NDT thermographic imaging of aerospace structures, Infr. Phys. Technol., 97, pp. 456-466. DOI: 10.1016/j.infrared.2019.02.002.

[3] Avdelidis, N.P., Almond, D.P., Dobbinson, A., Hawtin, B.C., Ibarra-Castanedo, C. and Maldague, X. (2004). Aircraft composites assessment by means of transient thermal NDT, Prog. Aerosp. Sci., 40, pp. 143–62. DOI: 10.1016/j.paerosci.2004.03.001.

[4] Yang, R., He, Y. and Zhang, H. (2016). Progress and trends in nondestructive testing and evaluation for wind turbine composite blade, Renew Sustain. Energy Rev., 60, pp. 1225–50. DOI: 10.1016/j.rser.2016.02.026.

[5] Yang, B., Zhang, L., Zhang, W. and Ai, Y. (2013). Non-destructive testing of wind turbine blades using an infrared thermography: A review. Intern. Conf. Mater. for Renew. Energy and Environ., Chengdu, China, pp. 407-410. DOI: 10.1109/ICMREE.2013.6893694

[6] Sfarra, S., Cicone, A., Yousefi, B., Perilli, S., Robol, L. and Maldague, X.P.V. (2022). Maximizing the detection of thermal imprints in civil engineering composites via numerical and thermographic results pre-processed by a groundbreaking mathematical approach, Int. J. Therm. Sci., 177, pp. 107553. DOI: 10.1016/j.ijthermalsci.2022.107553.

[7] Gholizadeh, S. A review of non-destructive testing methods of composite materials. (2016). Procedia. Struct. Integr., 1, pp. 50–57. DOI: 10.1016/j.prostr.2016.02.008.

[8] Shepard, S.M. (2007). Flash thermography of aerospace composites. IV Conferencia Panamericana de END, Buenos Aires, Argentina.

[9] Liu, Y. and Bao, Y. (2022). Review on automated condition assessment of pipelines with machine learning, Adv. Eng. Informatics, 53, pp. 101687. DOI: 10.1016/j.aei.2022.101687.

[10] Niccolai, A., Caputo, D., Chieco, L, Grimaccia, F. and Mussetta, M. (2021). Machine learning-based detection technique for NDT in industrial manufacturing, Mathematics, 9, p. 1251. DOI: 10.3390/math9111251.

[11] Westphal, E. and Seitz, H. (2021). A machine learning method for defect detection and visualization in selective laser sintering based on convolutional neural networks, Addit. Manuf., 41, p. 101965. DOI: 10.1016/j.addma.2021.101965.

[12] Khodayar F., Sajasi S. and Maldague X. (2016). Infrared thermography and NDT: 2050 horizon, Quantit. InfraRed Thermogr. J., 13(2), p. 210-231.

[13] Jiuxiang Gu, Zhenhuan Wang, Kuen J., Lianyang Ma, Shahroudy A., Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai and Tsuhan Chen. Recent advances in convolutional neural networks. (2018). Pattern recognition, 77, pp. 354-377.

[14] Yunze He, Baoyuan Deng, Hongjing Wang, Liang Cheng, Ke Zhou, Siyan Cai and Ciampa, F. (2021). Infrared machine vision and infrared thermography with deep learning: A review. Infr. Phys. & Technol., 116, p. 103754.

[15] Yousefi, B., Kalhor, D., Usamentiaga, R., Lei, L., Ibarra-Castanedo, C. and Maldague, X.P.V. (2018). Application of deep learning in infrared nondestructive testing. Conf. 2018 Quantitative Infrared Thermography, pp. 97-105. DOI: 10.21611/qirt.2018.p27.

[16] Haiyi Wu, Hongwei Zhang, Guoqing Hu and Rui Qiao. Deep learning based reconstruction of the structure of heterogeneous composites from their temperature fields. (April 2020). AIP Advances, 10(4), p. 045037. DOI: 10.1063/5.0004631

[17] Qiang Fang, Ibarra-Castanedo, C., Garrido, I., Yuxia Duan and Maldague, X. Automatic detection and identification of defect by deep learning algorithms from pulsed thermography data. (2023). Sensors, 23(9), p. 4444.

DOI: 10.3390/s23094444.

[18] Fang, Q., Mamoudan, F., Ibarra Castanedo, C. and Maldague, X. (2020). Defect depth estimation in infrared thermography with deep learning. Virtual: 3rd International Symposium on Structural Health Monitoring and Nondestructive Testing 25-26 Nov 2020, Quebec, Canada. e-Journal of Nondestructive Testing, 25(12). https://www.ndt.net/?id=25550.

[19] Alhammad, M., Avdelidis, N.P., Ibarra-Castanedo, C., Torbali, M.E., Genest, M., Zhang, H., Zolotas, A. and Maldague, X.P.V. (2022). Automated impact damage detection technique for composites based on thermographic image processing and machine learning classification, Sensors, 22, p. 9031. DOI: 10.3390/s22239031.

[20] Chulkov, A.O., Nesteruk, D.A., Vavilov, V.P., Moskovchenko, A.I., Saeed, N. and Omar M. (2019). Optimizing input data for training an artificial neural network used for evaluating defect depth in infrared thermographic nondestructive testing, Infr. Phys. Technol., 102(3), p. 103047. DOI: 10.1016/j.infrared.2019.103047.

[21] Marinetti, S., Plotnikov, Y.A., Winfree, W.P. and Braggiotti, A. (1999). Pulse phase thermography for defect detection and visualization. In: Proc. SPIE Nondestructive Evaluation of Aging Aircraft, Airports, and Aerospace Hardware III, 3586, pp. 230–238. DOI: 10.1117/12.339890.

[22] Carslaw, H.S. and Jaeger, J.C. (1959). Conduction of heat in solids, Oxford, Oxford Univ. Press., UK.

[23] Vavilov, V.P. and Burleigh, D.D. (2015). Review of pulsed thermal NDT: Physical principles, theory and data processing. NDT & E Int., 73, pp. 28–52. DOI: 10.1016/j.ndteint.2015.03.003.

[24] Almond, D.P. and Patel, P.M. (1996). Photothermal science and techniques. Chapman & Hall Series in Accounting and Finance, Physics and its applications, 10, p. 241.

[25] Vavilov, V.P. (2007). Pulsed thermal NDT of materials: Back to the basics, Nondestruct. Test. Eval., 22, pp. 177–197. DOI: 10.1080/10589750701448407.

[26] Moskovchenko, A., Švantner, M., Muzika, L., Skála, J., Pereira, C.M.C. and Das, S. (2023). The apparent effusivity method for normalized thermal contrast evaluation in infrared thermographic testing. Infr. Phys. Technol., 134, p.104931. DOI: 10.1016/j.infrared.2023.104931.

[27] Maldague, X. and Marinetti, S. (1996). Pulse phase infrared thermography, J. Appl Phys., 79, p. 2694–2698. DOI: 10.1063/1.362662.

[28] Shepard, S.M. (2007). Automated processing of thermographic derivatives for quality assurance, Opt. Eng., 46, p. 051008. DOI: 10.1117/1.2741274.

[29] Oswald-Tranta, B. Comparative study of thermal contrast and contrast in thermal signal derivatives in pulse thermography. (2017). NDT & E Int., 91, pp. 36–46. DOI: 10.1016/j.ndteint.2017.06.005.

[30] Caggiano, A., Zhang, J., Alfieri, V., Caiazzo, F., Gao, R. and Teti, R. (2019). Machine learning-based image processing for on-line defect recognition in additive manufacturing. CIRP Ann., 68, pp. 451–454. DOI: 10.1016/j.cirp.2019.03.021.

[31] Moskovchenko, A. and Švantner, M. Thermographic data processing and feature extracting approaches for machine learning-based defect detection. (Oct. 2023). Eng. Proc., 51(1), p. 5. DOI: 10.3390/engproc2023051005.

[32] Alhammad, M., Avdelidis, N.P., Ibarra-Castanedo, C., Torbali, E., Genets, M., Hai Zhang, Zolotas, A. and Maldague X.P.V. Automated impact damage detection technique for composites based on thermographic image processing and machine learning classification. (2022). Sensors, 22(23), p. 9031. DOI: 10.3390/s22239031.

[33] Alhammad, M., Avdelidis, N.P., Ibarra-Castanedo, C., Maldague, X., Zolotas A., Torbali, E. and Genest, M. (2024). Multi-label classification algorithms for composite materials under infrared thermography testing. Quantit. InfraRed Thermogr. J., 21(1), p. 3-29. DOI: 10.1080/17686733.2022.2126638.

[34] Pisner, D.A. and Schnyer, D.M. Support vector machine. Mach. Learn., Elsevier; 2020, pp. 101–121. DOI: 10.1016/B978-0-12-815739-8.00006-7.