SHM
Structural Health Monitoring

Original Manuscript

Structural Health Monitoring
2022, Vol. 0(0) 1–18
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14759217211073503
journals.sagepub.com/home/shm
SAGE

Automatic defect detection for ultrasonic wave propagation imaging method using spatio-temporal convolution neural networks

Jiaxing Ye and Nobuyuki Toyama

Abstract

Ultrasonic wave propagation imaging enables the detection of anomalies in various structures; hence, it has been applied as one of the promising techniques for damage identification in structural health monitoring (SHM). The interpretation of imaging data is vital to SHM; however, it relies significantly on expert subjective judgment, rendering the results vulnerable to human errors. Recent advances in the field of computer vision arising from the adoption of deep neural networks have resulted in new perspectives for substituting human roles in laborious data interpretation tasks. This paper presents an effective learning architecture that can characterize the ultrasonic wave propagation videos for automatic non-destructive inspection. The main contribution is threefold: 1. To the best of our knowledge, this is the first study to leverage video content analysis techniques to exploit ultrasonic wave propagation image series. Previous approaches that focused on the still wavefield images are likely to lose critical temporal information, thereby resulting in an inferior performance. 2. We devise a model that progressively aggregates both temporal and spatial information encoded in multiple adjacent snapshots of ultrasonic wave propagation motions for efficient data analysis. We presented the details regarding the system implementation and critical parameter settings. 3. The proposed approach is validated through extensive experimental comparisons with other state-of-the-art computer vision techniques on a real dataset which is publicly available. We hope that this study will encourage further investigations into video-based non-destructive data interpretation, not limited to ultrasonic signals.

Keywords

Non-destructive testing, ultrasonic wave propagation imaging, computer vision, video content analysis, spatiotemporal convolutions

Introduction

Ultrasonic inspection is a well-established technique that has been extensively applied in various non-destructive testing (NDT) and structural health monitoring (SHM) applications for decades. Because of its favorable long-range diagnostic capability, this method is extensively applied in detecting multiple defects such as cracks, delamination, and fatigue-based losses in various structures. In general, an electrical pulser is employed to generate an ultrasonic signal that propagates through inspection objects in the form of elastic waves, allowing the volume of the material between the transmitting and receiving transducers to be inspected. Once flaw/damage is encountered, part of the wave energy is reflected back to the surface of the structure. These defects can be discerned by investigating the waveforms.[1] As a versatile damage detection technique, ultrasonic inspection offers several favorable merits, such as high sensitivity to most material damages and proficiency in extracting defect location and size specifications. However, direct interpretation of complicated waveforms can be challenging due to their multimodal characteristics.

Visualization techniques, which render a series of snapshots of ultrasonic wave propagation, have been introduced to alleviate difficulties in ultrasonic inspection data

National Metrology Institute of Japan, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan

Corresponding author:
Jiaxing Ye, National Metrology Institute of Japan, National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan.
Email: jiaxing.you@aist.go.jp

interpretation. The resultant image sequence significantly facilitates the understanding of the wave propagation mechanism as well as the detection of wave scattering. Therefore, damage identification can be performed efficiently by analyzing the visualized scattered waves induced by defects instead of exploiting complicated waveforms.[2,3] A typical ultrasonic imaging inspection system comprises three components: a pulsed laser that generates thermoelastic ultrasonic waves, a transducer that receives ultrasonic signals, and a computer that stores the signals and visualizes wave propagation. In addition, image post-processing is commonly applied to enhance the signal-to-noise ratio of inspection wavefields, thereby facilitating visual inspection.

Damage detection lies at the heart of SHM and the goal is to identify the damage with high accuracy and efficiency. In the past decade, there is an increasing interest on the automatic interpretation of inspection data for structural health monitoring (SHM) in the field of civil engineering.[4,5] In respect to minimizing the labor force in ultrasonic inspection data interpretation, a majority of the effort is offloaded to the subject, in which advanced signal processing and machine learning techniques are primarily involved.

The initial studies toward wavefield imaging data analysis were performed by using conventional signal processing tools. For instance, Fourier analysis and wavelet packet transform had been employed to investigate the behavior of temporal and spatial frequency components of wavefield for damage detection.[6] Soon the focus of wavefield analysis shifted to quantifying wave propagation. It had been acknowledged that the ultrasonic wavefield is difficult to interpret, mainly due to multiple modes and reflections constructively and destructively interfering.[3] To eliminate the incident wavefield, which conveys no information regarding damage identification while retaining the critical damage-induced patterns, some efficient numerical methods had been proposed, such as wavenumber filtering methods[7] and root mean square (RMS) method.[8] Furthermore, derived from the wavenumber method, advanced methods had been validated for defect detection and quantification in composite.[9] More lately, weighted RMS and wavenumber filtering had been further investigated for detection and localization of barely visible identified damages (BVID) and characterize the severity in carbon-fiber-reinforced polymer composite structures.[10]

In the last few years, the impact of deep learning has been widespread across diverse research fields, including the area of structural health monitoring.[11] The application of the latest deep learning models to interpret the ultrasonic wavefield data for damage detection and condition assessment has recently gained considerable attention.[12–14] Before we dive into the details of technical review, we present a brief comparison between wavefield signal processing and machine learning based schemes for ultrasonic wavefield image pattern investigation.

In general, wavefield data processing techniques are developed under (relatively) explicit/ideal assumptions, such as the physical properties of ultrasonic waves and the homogeneous media had been thoroughly characterized.[3] It can be understood as a top-down approach, in which the physical properties are represented by using well-formed mathematics formulas and subsequently being applied for wavefield data analysis. Because the prior knowledge (physics of wave propagation) had been carefully exploited, wavefield signal processing methods are data efficient. They do not rely on massive datasets to extract the underlying data structure for damage identification. However, the initial assumption may not hold when dealing with the real data, making the methods incapable of generating favorable assessment results.

On the contrary, the deep learning models handle the ultrasonic wavefield pattern classification problem in an alternative manner. They formulate the problem of damage identification from wavefield data as a cognitive task. The models would exhaust all efforts to establish a mapping (using function composition) between visual patterns and corresponding condition labels. According to the universal approximation theorem,[15] with a powerful design of stacking layers of nonlinear transformations and numerous tunable parameters, deep neural networks can exploit a hypothesis space of sufficient flexibility and complexity.[16] Furthermore, stochastic optimization is employed to search the optimal functions in the hypothesis space with high efficiency, which can build mapping relationships specified by the training data (e.g., wavefield images and the corresponding condition indexes).[17] Deep learning takes a bottom-up approach that combines subtle visual cues distilled from massive wavefield data to give rise to a more comprehensive system with high complexity and completeness for the wavefield data assessment. However, since the model performance depends purely on its own previous experience, its ability is bounded by how many annotated samples were available for training the model.

Overall, both approaches have the same objective: to minimize human efforts and reduce errors in ultrasonic inspection signal interpretation but adopt different strategies. The wavefield signal processing methods, which treat a signal as the physical support of information, endeavor to find provably correct and optimal solutions for damage identification and assessment with consideration of the properties of ultrasonic waves. In contrast, deep learning approaches tackle the cognitive task of damage identification from wavefield data more directly by establishing the mapping from the input wavefield images to the corresponding condition labels. And the models are more tolerant of uncertainty, partial loss in wavefield data, and approximation, which bring new perspectives and possibilities for ultrasonic wavefield interpretation. In the long run, the two types of methodologies could be combined and contribute complementarily to automatic ultrasonic wavefield data interpretation with high precision.

For the development of statistical machine learning system for ultrasonic wavefield pattern classification, in an early attempt, the artificial neural network (ANN) with a single hidden layer was evaluated to classify ultrasonic waveforms with crack, porosity, and inclusion defects.[18] As for the statistical pattern classification, since the performance of ANNs is limited by the computation power and size of a dataset, some researchers have adopted simpler approaches, such as principal component analysis (PCA) and self-organizing maps (SOM).[19] More recently, efficient machine learning algorithms, such as support vector machines (SVMs), have been evaluated for flaw-induced ultrasonic wave detection,[20] and the random forest algorithm[21] was employed to detect the extent of internal damage due to rebar corrosion. Although these models reported high classification accuracy, the evaluation was confined to small scale datasets, and their performances were proved to be difficult to extend to unseen cases.

With the rapid development of computer hardware, machine learning models have become more effective. In recent years, deep neural networks (also known as deep learning) have consistently demonstrated state-of-the-art performances in various cognitive tasks.[22] The shift in applying deep learning for ultrasonic imaging data analysis is in progress.[23] Convolutional neural networks (CNNs) have been employed to characterize the wavelet spectrogram of ultrasonic waveforms to classify voids and delamination in carbon fiber composite materials.[24] Furthermore, CNNs have been employed to analyze ultrasonic wavefield images to detect multiple types of defects in planar specimens.[13] It is noteworthy that CNNs, owing to their large number of learnable parameters, have been acknowledged to be data hungry, that is, their performance relies significantly on the size of the training data. As in ultrasonic testing, samples from real or manufactured flaws are expensive, thereby resulting in insufficient experimental data. Hence, simulation approaches have been devised to generate large datasets to facilitate CNN model training.[25,26] In summary, recent systems for ultrasonic testing data analysis are typically constructed using deep neural networks, and the primary targets are A-scan waveforms or wavefield images. Problems pertaining to small sample sizes are expected to continue in the application of CNNs for ultrasonic data analysis.

The aim of this study is to provide a novel perspective for characterizing ultrasonic wave propagation imaging (UWPI) data using video content analysis techniques. This is inspired by practices in ultrasonic wavefield testing, that is, an inspector watches a "movie" of wave motion snapshots to monitor the manner in which waves propagate and then analyze whether an obstacle (flaw) is present in their path.[2] Although a single snapshot of the ultrasonic wavefield provides discriminant information for flaw detection, it is evident that video-based assessment is more efficient and accurate; therefore, we propose devising a spatio-temporal convolutional neural network that can mimic human ability in understanding UWPI videos. An overview of the proposed system is presented in Figure 1.

Video understanding, which is a core problem in the field of computer vision, is an established research topic and has attracted significant attention in recent decades. Although substantial progress has been reported, the existing models may not be suitable for UWPI data analysis due to three main reasons: First, the UWPI dataset obtained from real specimens is small compared with massive datasets used for genetic action recognition,[27] causing the complicated models to overfit the data easily. In other words, because the model parameters are numerous whereas the available training samples are limited, the model tends to perform extremely well on the data used for training but cannot yield correct predictions when handling unseen data. Multiple methods can be used to solve this issue; herein, we present our solution to the problem. Second, action recognition models are primarily focused on dynamic human actions and gestures,[28] whereas UWPI exhibits a relative structured pattern of wave propagation and scattering. For instance, in a continuous medium, the behavior of ultrasonic waves propagation can be approximated by combination of longitudinal waves, transverse waves, and surface waves. Such essential property implies that simple neural network architectures might be sufficient to describe the normal wave propagation. Third, the central problem of video content understanding is the aggregation of multi-frame information. We revisit this fundamental principle and empirically validate multiple CNN architectures for efficient spatio–temporal information characterization. Furthermore, a few questions remain: How does additional temporal information affect UWPI pattern analysis? What is the overall performance improvement afforded? In this study, we designed a series of experiments to answer these questions. In summary, the contributions of this study are as follows:

*. We propose a novel scheme that adopts a video content analysis technique for UWPI data investigation. It differs from previous approaches that focus on still wavefield images. Our approach is advantageous for exploiting rich motion information for better ultrasonic wave propagation pattern modeling.

*.In the field of computer vision, spatio–temporal convolutional neural networks, that is, three-dimensional (3D) CNNs, have demonstrated state-of-the-art performance for video understanding.[27] In this study, we redesign them to adapt to UWPI video analysis.

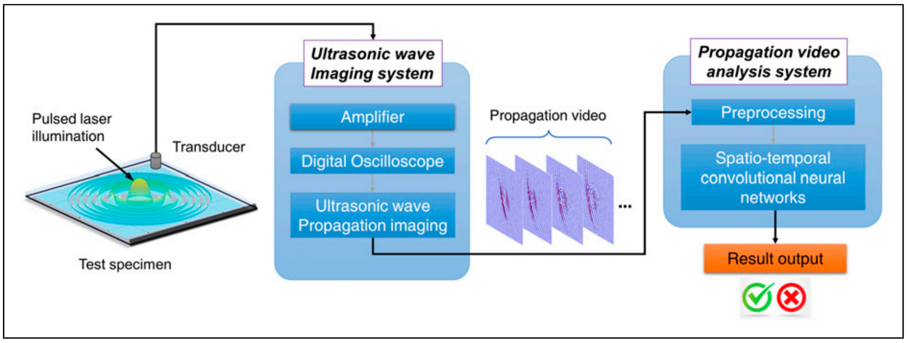
Pulsed laser illumination
Transducer
Test specimen
Ultrasonic wave Imaging system
Amplifier
Digital Oscilloscope
Ultrasonic wave Propagation imaging
Propagation video
...
Propagation video analysis system
Preprocessing
Spatio-temporal convolutional neural networks
Result output
✓ ✗

Figure 1. The processing diagram of the proposed ultrasonic wave propagation video analysis system.

Table 1. Overview of papers using machine learning techniques for ultrasonic non-destructive tests.

Input signal	Year	Signal representation	Pattern classification scheme
1-D waveform	2002[31]	Ultrasonic A-scan signal with matching pursuit coefficients	ANN with RBF activation functions
	2013[19]	Ultrasonic A-scan signal	Self-organizing maps (SOM) classifiers
	2015[20]	Ultrasonic A-scan signal	Support vector machines (SVM)
	2017[32]	Ultrasonic A-scan signal with wavelet transformed features	Principal component analysis (PCA)
	2017[33]	Ultrasonic A-scan signal	Dictionary learning with K-SVD
	2020[23]	Ultrasonic A-scan signal	Autoencoder (AE)
	2021[26]	A-scan signal with adjacent concatenation and virtual flaw augmentation	Deep convolutional neural network
2-D image	1996[18]	Ultrasonic A-scan signal with PCA pre-processing	Artificial neural network (ANN) with 3 layers
	1997[6]	Ultrasonic B-scan image	Wavelet packet transform
	2005[34]	Phase information extracted from TOFD images	Cross-correlation coefficient
	2007[35]	Co-occurrence based matrix features	Multilayer neural-fuzzy network
	2011[36]	Ultrasonic B-sacn image with time-frequency representation	4-Layer artificial neural networks
	2016[37]	Ultrasonic B-scan image	Sparse deconvolution method
	2017[38]	Ultrasonic B-scan image with Hilbert-Huang transform (HHT)	Cross-correlation
	2018[13]	Ultrasonic wavefield images	Deep convolutional neural network with residual module
	2020[25]	Ultrasonic plane wave imaging data and simulation data	Deep convolutional neural network
	2021[39]	Ultrasonic wavefield images	Latest deep convolutional neutral networks
	2021[40]	Ultrasonic wavefield images with adaptive wavenumber filtering and RMS feature extraction	Fully convolutional network (FCN)

Through extensive investigations, we determined the optimal network design for this task.

*. The proposed video-based analysis approach is validated through experimental comparisons for a practical task, that is, to discern flaw-induced patterns from ultrasonic wave propagation motions. Various state-of-the-art computer vision techniques are added to the comparison list, including the conventional approaches composed of feature extraction with statistical classifiers and the latest approaches based on neural networks, such as ResNet[29] and DenseNet,[30] which even delivered human-level visual object recognition performance. By comparing the results on real data, it is demonstrated the proposed video based UWPI data analysis approach outperformed all other methods by a significant margin.

The remainder of this paper is organized as follows. We first present a comprehensive review of the machine learning techniques employed for understanding ultrasonic data. Then, the following section describes the proposed approach for UWPI video analysis, and several variant designs for the learning architectures are discussed. Subsequently, we present the experimental validation, including an introduction to the inspection method and real ultrasonic imaging test dataset, parameter settings for spatio-temporal pattern modeling, and comprehensive comparison results. We conclude this study by presenting a summary and discussion in the final section.

Related works

Over the last 5 years, machine learning-enabled ultrasonic inspection data interpretation emerged as an active research topic and a wide variety of research progresses had been reported. In this section we summarize the remarkable progress achieved in the field. Table 1 presents a concise review of the recent progress in this field. We categorized the methods in terms of ultrasonic signal representation and machine learning algorithms.

The table reveals the technical trend of devising machine learning systems for automatic ultrasonic data investigation. Advanced machine learning approaches such as SVMs[20] and random forests[21] have been discussed for ultrasonic signal analysis. It was confirmed that novel machine learning/pattern recognition techniques can contribute substantially to ultrasonic data interpretation. This finding is further intensified owing to the remarkable success of deep learning models. Based on the universal approximation theorem,[15] deep neural network architectures can be used to approximate any complex function $f(\cdot)$ that maps attributes $\mathbf{x}$ to the desired output y. Within the context of ultrasonic testing, $\mathbf{x}$ can be the ultrasonic wavefield image, and y can be the corresponding condition index, which is a numerical code with 1 denoting normal, and 0 indicating a flaw pattern. Neural networks fully utilize deep stacking architectures to establish a highly nonlinear relationship between the two spectrums.[17] Although deep neural networks have been extensively discussed, the most complex neural network design applied hitherto comprised seven convolution layers.[13] This is attributed primarily to ultrasonic inspection datasets being limited to small scales, thereby hindering the complicated network design from achieving a significant performance gain.

Second, in terms of ultrasonic data representation, computerized pattern analysis systems typically involve two input modalities: ultrasonic A-scan waveforms and B-scan wavefield images. Accordingly, different signal processing techniques have been adopted, such as Fourier and wavelet transforms,[32] which have been introduced to process A-scan waves. Furthermore, statistical co-occurrence features were investigated using ultrasound B-scan images.[35] Moreover, advanced ultrasonic testing techniques, such as ultrasonic wave propagation imaging (i.e., wavefield imaging), can provide a series of snapshots (video) depicting ultrasonic wave propagation through the test piece, thereby easing defect inspection significantly. However, to the best of our knowledge, video-based machine learning approaches for ultrasonic wavefield data analysis have not been investigated.

The analysis above motivated us to conduct further studies. We discovered that previous studies focused on ultrasonic signals in the representations of A-scan waveforms and two-dimensional wavefield images. In this study, we considered an alternative scheme to perform UWPI video analysis. We assumed that both spatial and temporal information are essential for wavefield motion pattern modeling. The rich discriminant conveyed in a spatiotemporal manner can provide further advancements.

The proposed approach

The aim of this study is to develop an efficient machine learning system to discern flaw-induced patterns in an ultrasonic wave propagation video. Figure 2 illustrates an example for ease of understanding. A sequence of ultrasonic wavefield snapshots was used as the input to the machine learning system. The corresponding condition indexes were binary values indicating the presence/absence of flaw-induced patterns and were the desired outputs. In this section, we introduce the proposed computerized UWPI data analysis system as follows: First, we provide a general scheme for video understanding based on spatio-temporal feature learning. Subsequently, we explain our customization of the learning architecture to achieve an efficient UWPI video investigation.

Video understanding problem and 3D convolution-based learning scheme

In computer vision, video understanding refers to the capability in automatically analyzing videos to extract knowledge/information for the detection and recognition of meaningful events. Video content analysis is primarily performed by investigating two crucial and complementary aspects: appearance and dynamics. Inspired by the remarkable success of CNNs in image recognition, CNNs have been extended to video classification in recent studies. Because video data can be regarded as 3D signals (height, width, and time), a new type of deep neural network, that is, spatiotemporal CNNs, was investigated jointly to exploit the spatial appearance and temporal motions exhibited in adjacent frames.[41] This method is also known as 3D CNNs, as convolution is applied to 3D data. The key aspect of 3D CNN for video understanding is that human engineers do not design the feature layers. Instead, they are learned from massive training data (raw pixels) using a general-purpose learning procedure. By utilizing multiple levels of spatiotemporal features learned using 3D CNNs, remarkable progress has been achieved in video understanding.[42] In this section, we present learning approaches based on 3D CNNs for UWPI video data analysis.

We denote the ultrasonic wavefield video data, which comprise a series of wavefield images, as 3D tensors $V \in \mathbb{R}^{T \times H \times W}$, where T, H, and W are the number of images, height, and width, respectively. In 3D CNNs, a series of 3D convolution operators (i.e., kernels) are applied to the video data, which slide along three directions (T, H, and W) to automatically exploit both the visual appearance of waves and dynamic propagation motions in a hierarchical manner. Figure 3 Shows the flowchart of the 3D CNN model for UWPI video analysis. The model is defined by $f(\cdot)$ which defines a function composition involving a series of computations as follows:

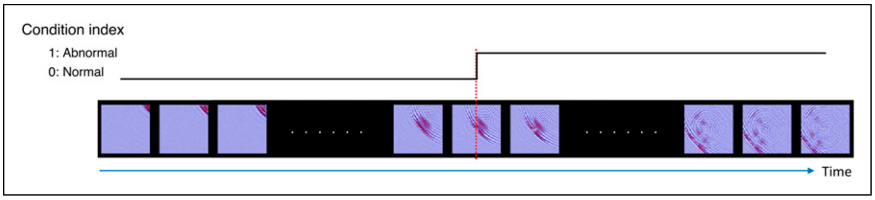

Figure 2. The ultrasonic wave propagation imaging (ultrasonic wave propagation imaging) snapshots and corresponding condition indexes.

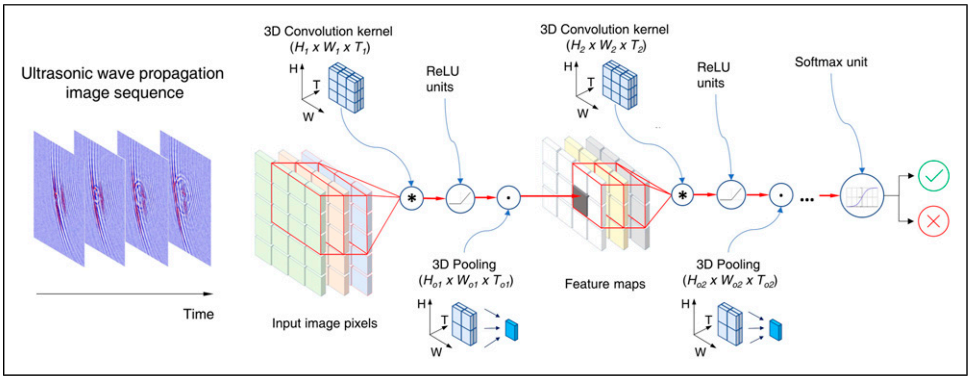

Figure 3. The flowchart of spatio-temporal convolutional neural networks for ultrasonic wave propagation imaging video analysis.

$$\tilde{\boldsymbol{y}} = f\left(\boldsymbol{V}; \boldsymbol{W}^{1,\ldots,K}, b^{1,\ldots,K}\right)$$
$$= \sigma\left(\sigma\left(\cdots \sigma\left(\boldsymbol{V}\boldsymbol{W}^{1} + b^{1}\right) \cdots + \right)\boldsymbol{W}^{K} + b^{K}\right) \quad (1)$$

where $\boldsymbol{W}^{1,\ldots,K}$ and $b^{1,\ldots,K}$ are the learnable weights of all the convolution filters and the biases vectors that operate at the l-th layer, respectively. The network is organizedinto groups of units known as layers $(k = 1,\ldots,K)$ and the mechanism of information propagation between layers, for example, from the k-1 to k layers, complies with the same principal as follows:

$$\boldsymbol{h}^{(k)} = \sigma\left(\boldsymbol{a}^{(k)}\right), \text{where} \quad \boldsymbol{a}^{(k)} = \boldsymbol{b}^{(k)} + \boldsymbol{W}^{(k)}\boldsymbol{h}^{(k-1)} \quad (2)$$

Here, $\sigma(\cdot)$ is the activation function, which generally performs element-wise nonlinear filtering to prevent the model from collapsing into a linear model. The activation function that we employed in this study is a rectified linear unit (ReLU), defined as $\sigma(\tau,\boldsymbol{a}) = \max\ (0,\boldsymbol{a})$ which has been validated for large-scale visual learning problems. The subsequent operator–3D pooling performs down-sampling by segmenting the input into cuboidal pooling regions and computing each region's maximum. The objective for applying 3D pooling is twofold, that is, (i) to downscale the data size, and (ii) to introduce some level of translation invariance to feature extraction, which is quite important for visual feature extraction. The composition of such computations enables extremely complex functions to be learned. $\tilde{\boldsymbol{y}}$ is the prediction output generated by $f(\cdot)$, which approximates the interpretation results of human inspectors, such as normal and abnormal judgments are denoted as 0 and 1, respectively.

Among the various problems encountered in deep learning architecture design, the most difficult is neural network training, that is, to automatically adjust tens of millions of parameters in a systematic way to establish an efficient and robust mapping between the input and output when the model is fed with raw data and the corresponding label pairs $\{\boldsymbol{V}_n, \boldsymbol{y}_n\}$. We present a concise overview of 3D CNN model training algorithm in Table 2.

In the algorithm, $\nabla_{\tilde{\boldsymbol{y}}}\boldsymbol{J}$ is the derivative induced by the loss of training data, and η is the learning rate that governs the network update step/speed. A regularization term is typically added to the prediction error function to prevent the model overfitting, which is expressed as $\Omega(\boldsymbol{\theta})$. We denote all the learnable parameters with $\boldsymbol{\theta}$, and stochastic optimization was performed to tune all the parameters to minimize the objective function $\boldsymbol{J}$. For simplicity, we used the stochastic gradient descent (SGD) optimization scheme in the algorithm, although modified solvers are available such as adaptive moment estimation (ADAM) and stochastic gradient descent with momentum; more details can be referred to reference[17].

In this section, we present the 3D CNN learning model employed for UWPI video analysis. We abbreviate it as ST-Net-3D, which represents spatio-temporal neural network using 3D convolutions. In the following sections,

Table 2. Algorithm for training neural networks.

```
Algorithm 1: Train Neural Networks ($V_n$, $y_n$, W, b)
Initialization: W, b, $\eta$
For n = 1,2,…,N
  do perform forward propagation $\widehat{y}_n = f(V_n, W, b)$
Compute the prediction loss $J = L(y_n, \widehat{y}_n) + \lambda\Omega(\theta)$
Compute the gradient on the output layer $g^{(K)} \leftarrow \nabla_{\widehat{y}} J$
Compute the gradients at each layer by back-propagation: For $k = K, K - 1, \ldots, 1$
$g^{(k)} \leftarrow \nabla_{a(k)} J = g \dot{o} \sigma'(a^{(k)})$
$\nabla_{b(k)} J = g^{(k)} + \lambda \nabla_{b(k)} \Omega(\theta)$; $\nabla_{w(k)} J = g^{(k)} h^{(k-1)T} + \lambda \nabla_{w(k)} \Omega(\theta)$
$g^{k-1} \leftarrow \nabla_{h(k-1)} J = W^{(k)T} g^{(k)}$
  Update all the parameters via SGD $\theta \leftarrow \theta - \eta \cdot g$
Return (W, b)
```

beginning with the basic architecture, we discuss how we devise the ST-Net-3D to achieve better performance.

Temporal information fusion scheme

The central problem in video analysis is the modeling of the temporal dimension. Ultrasonic wave propagation video classification involves clips containing several contiguous frames in time. Therefore, we can establish the connection of the deep neural network in the time dimension to learn the spatio-temporal features through neural network training. Although the connections can be established via multiple methods, we herein present three major designs: late score fusion, early feature fusion, and hierarchical feature fusion.

Single-frame prediction: We employed single frame-based prediction as a baseline to understand the significance of static appearance for ultrasonic wave propagation image analysis. In detail, individual ultrasonic wavefield images were as input in the scheme, and a series of advanced two-dimensional (2D) CNN models were adopted to distinguish flaw-induced patterns from normal cases. Moreover, because temporal information were not considered, the results can be used to justify the contribution of spatio-temporal information for the task.

Late score fusion: After building the baseline model, we devised a temporal information aggregation scheme. A direct solution can be late fusion, which performs majority voting over the single-frame predication results generated by a 2D CNN model. Although straightforward in design, the method is error tolerant, thereby enabling a successful prediction (i.e., as if the failure predictions are less than half of the total frames in an input video clip).

Early feature fusion: The early feature fusion scheme combines information across neighboring frames immediately at the pixel level. It is implemented by modifying the filters on the first 2D convolutional layer in the single-frame model by extending them to 3D convolutions, by which temporal information can be characterized. All the early and direct connectivity to the raw pixel from adjacent frames allows the network to precisely model local motions, such as direction and speed.

Hierarchical feature fusion: The hierarchical fusion method aggregates both temporal and spatial information throughout the network such that higher layers can access more global information progressively in both dimensions. It is implemented by extending the connectivity along the time axis of all layers and computing 3D convolutions to compute activations. We explain this method using an example, as follows: Assume that we are addressing a short video clip containing five frames. In the model we used, the first 3D convolutional layer was set to apply local temporal filtering with a timespan of three frames. We shift the convolution operator along time axis by one frame at each time. By setting valid convolution parameters, the third convolutional layer can access the spatio-temporal information across all five input frames. The valid convolution parameters mentioned here include the kernel size K_T, padding length P and striding step S. The desired output volume size was obtained by adjusting the three parameters using the following formula: $D_{out} = (D_{in} - K_T + 2P)/S + 1$, where D_{in} and D_{out} are the input/output dimensions, respectively. This principle also applies to other dimensions of the imaging data such as height (H) and width (W). Specifically, we set the 3D convolution operators at first, second, and third layers with parameters $\{K_T = 3, P = 0, S = 1\}$, $\{K_T = 1, P = 0, S = 1\}$; $\{K_T = 3, P = 0, S = 1\}$, respectively. Compared with other methods, the hierarchical feature fusion scheme introduces more nonlinear transformations, thereby enabling the model to represent more complex relationships between the input data and corresponding labels. Figure 4 shows an overview of the temporal information fusion schemes which clearly demonstrates the difference.

3D convolution kernel decomposition

Although 3D CNNs are suitable for creating hierarchical representations of spatio-temporal patterns for video understanding, they contain far more parameters than 2D CNNs because of the additional kernel dimension,

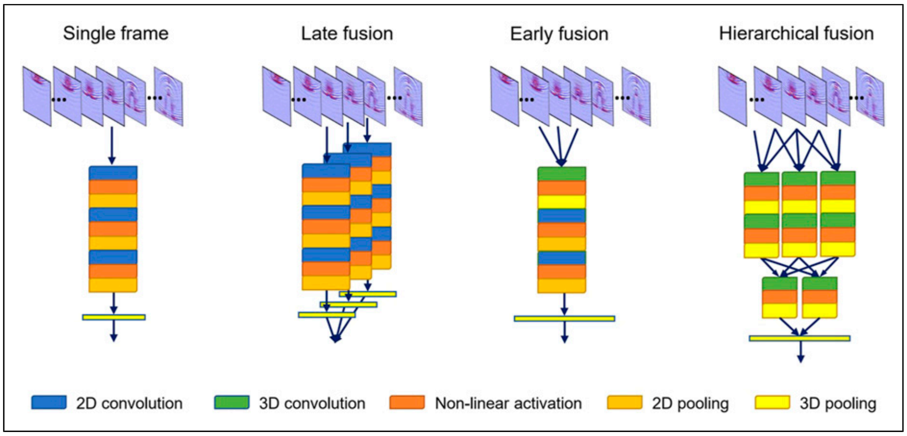

Figure 4. Demonstration of three fusion rules for ultrasonic wave propagation video analysis.

rendering them more difficult to train.[37] It can be understood easily from a computational perspective, that is, the 3D convolution with size of $t \times d \times d$ has complexity $\mathcal{O}$ (tdd) which is much higher compared with the case of 2D $d \times d$ convolution kernel that costs $\mathcal{O}$ (dd). To mitigate this issue and facilitate the efficient modeling of ultrasonic wave propagation video, we employed an approach to decompose 3D convolution to a 2D spatial convolution, followed by a one-dimensional (1D) temporal convolution. The method has been discussed in previous studies and proven to be effective for accelerating the deep network learning while retaining favorable performance.[43] The rationale behind the decomposition is that the 2D spatial appearance and 1D temporal dynamics in videos always exhibit different characteristics. Compared with modeling the spatial and temporal information simultaneously, it may be more efficient and feasible to model the two-way information separately. In this design, we replaced the 3D convolution kernel with size of $t \times d \times d$ by using two consecutive kernels with size of $1 \times d \times d$ and $t \times 1 \times 1$, respectively. The decomposition kernel is hereafter denoted by (2+1)D convolution. By performing (2+1)D convolution, the computation complexity can be reduced from $\mathcal{O}$ (tdd) to $\mathcal{O}$ (t) + $\mathcal{O}$ (dd). An illustration of the decomposition process is presented in Figure 5. The 2D convolution kernel contains nine (3×3) learnable parameters, whereas the 3D convolution kernel and the (2+1)D kernel contain 27 (3×3×3) and 12 (3×3+3×1) parameters, respectively. In addition to high efficiency, the kernel decomposition can facilitate the optimization, thereby necessitating fewer data is required to achieve a lower prediction error.

Table 3 presents the proposed spatio-temporal neural network architecture for UWPI video analysis and we denoted it as ST-Net-(2+1)D. Hierarchical fusion method was adopted to aggregate both temporal and spatial information throughout the network. To speed up the model training and prevent overfitting, we substitute the (2+1)D convolution for 3D convolution operators. Batch normalization and ReLU activation units were applied after the convolutions. We employ Batch normalization to mitigate the change in means and variances of the inputs to internal layers during training. By reducing these unwanted distribution shifts, batch normalization can speed up network training and produce more reliable models. Moreover, it has been deemed also that batch normalization offers a regularization effect, which performed comparably to the dropout method.[17] Subsequently, the Rectified Linear Unit (ReLU) operator is adopted to generate sparse feature representations for robust pattern recognition. For simplicity, we omitted the ReLU processes in Tab 2. 3D pooling was adopted to downsize the feature map and enhance the translation invariance. Additionally, we show the key specifications of the proposed spatio-temporal network, including dimensions of the filters, number of learnable parameters, and size of output feature maps in the order of height, width, and time. The series of convolutions culminated in a global spatio-temporal pooling layer yielded a *long* feature vector. Then a dropout layer was implemented to enhance the neural networks stability and robustness, where a certain set of features were disregarded at random.[17] Finally, the resultant vector was fed into a fully connected layer, which outputs the class-wise probabilities (condition index) through a softmax function. Cross-entropy loss is employed to assess the model performance and to facilitate gradient-based network training.

Experimental validation

In this section, we evaluate the proposed approach based on an extensive experimental comparison using a real dataset. We first present the ultrasonic inspection dataset, parameter settings for neural network training, and evaluation metrics.

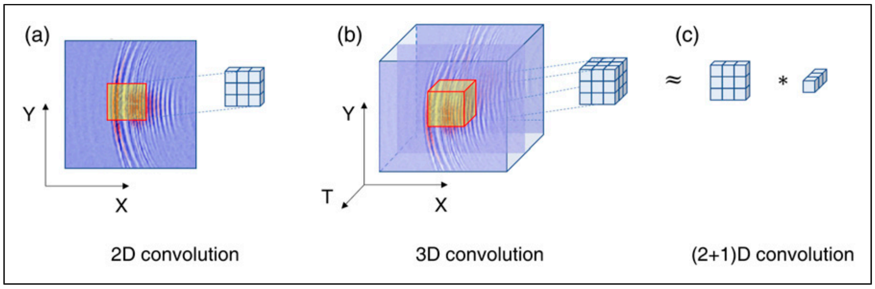

Figure 5. Convolution kernel comparison for wavefield image feature learning.

Table 3. Summary of the proposed ST-Net-(2+1)D architectures.

Layer #	Operations	*Learnable parameters*	*Output feature map*
(2+1)D conv1	5 × 5 × 1 × 3, 64, stride [2, 2, 1], padding [0, 0, 1] 1 × 1 × 3, 64	(5 × 5 × 1 × 3 + 1) × 64 (1 × 1 × 3 × 64+1) × 64	62 × 62 × 7 × 64
BatchNoram1	—	(1 × 1 × 1 + 1 × 1 × 1) × 64	62 × 62 × 7 × 64
pool1	2 × 2 × 1, stride [2, 2, 1]	—	31 × 31 × 7 × 64
(2+1)D conv2	3 × 3 × 1, 64, stride [2, 2, 1], padding [1, 1, 0] 1 × 1 × 3, 64, padding [0,0,1]	(3 × 3 × 1 × 64 + 1) × 64 (1 × 1 × 3 × 64 + 1) × 64	16 × 16 × 7 × 64
BatchNorm2	—	(1 × 1 × 1 + 1 × 1 × 1) × 64	16 × 16 × 7 × 164
pool2	2 × 2 × 2, stride [2,2,1]	—	8 × 8 × 6 × 64
(2+1)D conv3	3 × 3 × 1, 96 1 × 1 × 3, 96	(3 × 3 × 1 × 64 + 1) × 64 (1 × 1 × 3 × 96 + 1)	6 × 6 × 4 × 96
BatchNoram3	—	(1 × 1 × 1 + 1 × 1 × 1) × 96	6 × 6 × 4 × 96
(2+1)D conv4	3 × 3 × 1, 128, padding [0,0,1] 1 × 1 × 3, 96	(3 × 3 × 1 × 96 + 1) × 128 (1 × 1 × 3 128 + 1) × 128	4 × 4 × 4 × 128
BatchNoram4	—	(1 × 1 × 1 + 1 × 1 × 1) × 128	4 × 4 × 4 × 128
(2+1)D conv5	3 × 3 × 1, 128 1 × 1 × 3, 128	(3 × 3 × 2 × 128 + 1) × 128 (1 × 1 × 3 × 128 + 1) × 128	2 × 2 × 2 × 128
BatchNoram5	—	(1 × 1 × 1 + 1 × 1 × 1) × 128	2 × 2 × 2 × 128
FC layer	—	(1024 + 1) × 2	1 × 1 × 1 × 2

Subsequently, the proposed approach and other state-of-the-art methods are compared.

Dataset

To assess the performance of computer vision algorithms for automatic ultrasonic wave propagation image analysis, we employed the system developed in our laboratory to establish the database.[44] The system comprised three components: A pulsed laser scan unit that generates thermoelastic ultrasonic waves, a transducer attached to the surface of a specimen that collects ultrasonic waves propagating through the specimens, a computer that stores the amplified signals and visualizes wave propagation.

In the evaluation, we addressed the problem of flaw detection from stainless-steel plate specimens. Because the appearance of the UWPI data varied significantly owing to the flaw shape, size, and depth, we prepared a series of specimens with intended flaws to include more propagation patterns. Two types of defects were implemented: drilled-holes with diameters ranging from 1 to 5 mm, and slits with lengths ranging from 3 to 10 mm. Table 4 introduces the flaw conditions of the 17 specimens used for dataset development, and Figure 6 shows the flaw specifications. A laser scan was performed on the central region of the 3-mm-thick specimens measuring 100 mm 100 mm (green zone in Figure 6) on the steel plates. Notably, the transducers incident angle governed the propagation direction of the ultrasonic waves; hence, it was fundamental in detecting slit flaws. To rate the robustness of the computer vision system to incident angle selections, we obtained ultrasonic inspection images with different incident angles varying from 0 to 90° with 22.5 degree interval. Table 5 presents the key parameters applied for laser ultrasonic imaging inspection.

Based on the settings above, we captured 49 UWPI videos from 17 specimens and then prepared one defect-free specimen to obtain the wave propagation data from normal sample. Altogether, the dataset contained 50 full-scan

Table 4. Summary of the specimen specifications.

Specimen #	Flaw type	Depth	Transducer side	Defect Size (mm)
1–3	Hole	Penetrated	Front	ϕ 1, ϕ 3, ϕ 5
4–6	Hole	1.5 mm	Front	ϕ 1, ϕ 3, ϕ 5
7–9	Hole	1.5 mm	Back	ϕ 1, ϕ 3, ϕ 5
10–11	Slit	Penetrated	Front	5, 10
12–14	Slit	1.5 mm	Front	3, 5, 10
15–17	Slit	1.5 mm	Back	3, 5, 10

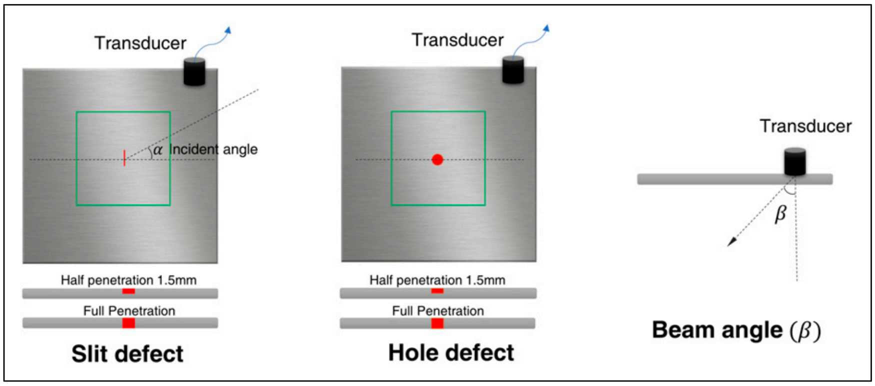

Figure 6. Flaws implemented on specimens and transducer placement.

Table 5. Summary of the experimental setting of inspection device.

Probe frequency	Beam angle β (°)	Pulse repetition	Incident angle α (°)
1 MHz	90	500 Hz	0, 22.5, 45, 67.5, 90

videos with 7004 wavefield snapshots with resolution of 128 × 128. We manually labeled all the images with binary condition indexes: 1 represents observable defective-induced scattering patterns, and 0 indicates normal ultrasonic wave propagation. The dataset had been used to evaluate CNN models for *still* wavefield image pattern analysis, and it had been made available for public access.[39] Previously, the binary condition indexes were assigned in a frame-wise manner, and in this work, we further process the wavefield images and the corresponding condition labels to facilitate video-based classification. In detail, we trimmed the full-scan videos to multiple short clips, and each clip consists of five contiguous frames with one overlapping frame. There are mainly three reasons to trim the videos: (1) We can obtain a sufficient number of video snippets, which enable spatio-temporal neural network training. (2) Although short, the trimmed clips can provide rich wave propagation dynamics for learning discriminant feature representations for defect detection. (3) The trimming scheme can retain high annotation accuracy to the resultant short clips in the time domain. As a result, we obtained 1681 annotated UWPI video clips. Figure 7 demonstrates the video clip trimming scheme from UWPI full scan data. The three examples represented the cases of a normal specimen and defective specimens with slit-shaped and hole-shaped flaws. Moreover, the plots reflect the effect of key parameters of incident angles.

The defective patterns can be discerned by focusing on the flaw-induced scattered wave radiation. However, it is not feasible to define the scattering waves using explicit programming. The significance of the CNN-based learning scheme is that owing to multiple levels of feature extraction, the raw pixel input can be systematically grouped to describe high-level and meaningful patterns by automatic feature discovery from the example dataset. It has been acknowledged that deep neural networks so far are data inefficient. That is, thousands, or even millions of training examples are required to be fed to the model to learn the efficient feature representations for a specific cognitive task. Data augmentation is regarded as a prospective process for mitigating the small sample size issue; it helps in increasing the dataset size and thus reduces overfitting. As for

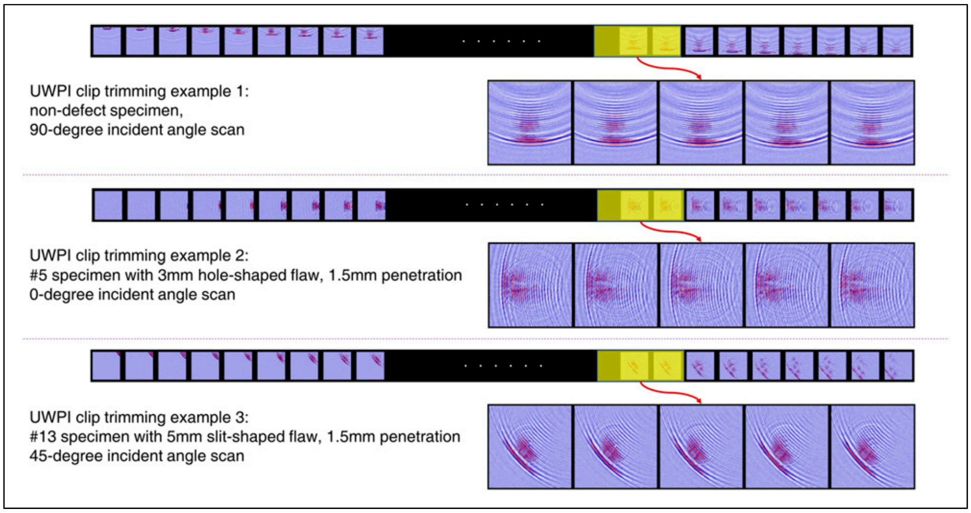
UWPI clip trimming example 1:
non-defect specimen,
90-degree incident angle scan
UWPI clip trimming example 2:
#5 specimen with 3mm hole-shaped flaw, 1.5mm penetration
0-degree incident angle scan
UWPI clip trimming example 3:
#13 specimen with 5mm slit-shaped flaw, 1.5mm penetration
45-degree incident angle scan

Figure 7. Demonstration of ultrasonic wave propagation imaging video clips trimming with three examples.

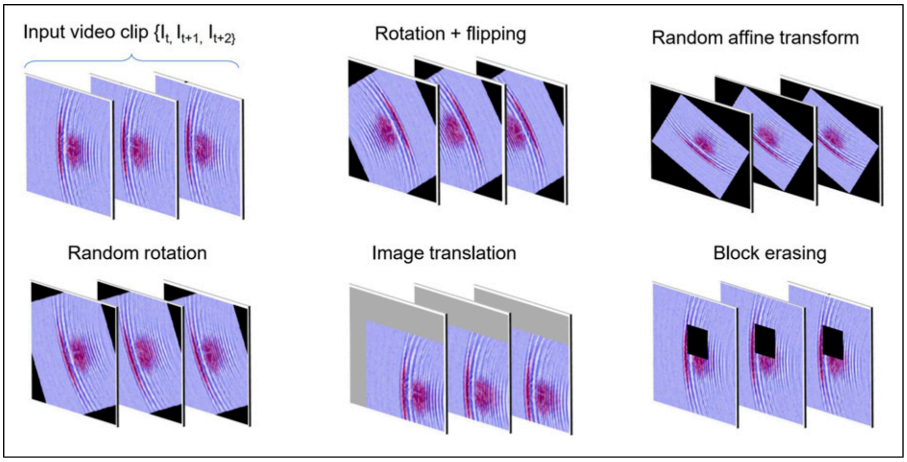

Figure 8. Augmented video clips by geometry transformation.

computer vision tasks, geometry transforms are the most applied data augmentation methods, including rotation, horizontal/vertical axis flipping, translation, and affine transform. Besides those four basic transforms, we introduce a block erasing scheme for dataset augmentation, randomly selecting a rectangle region in a wavefield image and erasing its pixels with zeros values. The method was proven to be effective in reducing the risk of over-fitting and making the model robust to occlusion.[45] The augmentation operation was implemented in the UWPI video clips such that the same geometry transform was applied to all frames. Figure 8 shows an illustration of the video augmentation schemes. The neural networks that have learned augmented data are assumed to generate robust feature representations to the corresponding visual appearance changes. However, the current augmentation scheme is incapable of describing real variations in ultrasonic wavefield patterns caused by different settings of an inspection system, such as transducer locations, incident angles, and target specimen flaw conditions. Recently, some progress had been made to devising new approaches of doing physically realistic data augmentation for non-destructive testing, such as generating synthetic ultrasonic inspection data by fusion of real and physics-model simulation data.[26] Realistic data augmentation will be an essential research topic for building reliable machine learning models for ultrasonic inspection data interpretation without gathering an absurd amount of training data.

At the evaluation stage, we adopted the leave-one-specimen-out protocol, that is, at each iteration, we selected one stainless-steel specimen with defects, and the UWPI data generated from it were assigned as testing data. All clips captured from other specimens were segregated into training and validation sets on an 80%-20% ratio.

Meanwhile, video clips obtained from non-flaw specimens were randomly inserted into the training/validation/testing sets with a fraction of 65%-10%–25%. At the training stage, we randomly selected 30% of the augmented clips and added them to the training set. Besides, the reference 2D CNN models, that is, AlexNet, ResNet, and DenseNet, require the input images to be of the size 224 × 224, we upscale the UWPI dataset from original resolution (128 × 128) to 224 × 224 to fulfill the requirement.

Parameter settings

In this section, we present the configuration for training the machine learning models. The dataset was rescaled to the range of [0, 1] with min-max normalization to ease model training. To perform PCA feature extraction, we need to determine the number of principal components. To this end, we evaluated three cases of using 50, 100, 200 components, which presented explained variances of 97.31%, 98.10%, and 98.76%, respectively. Because each parameter renders a sufficient amount of explained variance, we tune the hyper parameter using validation set via cross validation. For Histogram of Oriented Gradients (HoG)[46] feature extraction, we adopt default settings with cell size of [8, 8] and 9 bins of orientation histogram. For training the SVM model with RBF kernel, both C and gamma parameters need to be optimized simultaneously. We define the following parameter grid for optimization: we set C with log-uniform distribution of [0.1, 1, 10, 100] and gamma of [0.001, 0.1, 1, 10]. Two hyper-parameters were tuned with the validation set through leave-one-specimen-out cross-validation as described in the previous section. For evaluation of random forest classifier, although there had been some works that demonstrated using 128 trees is sufficient for most cases and the best depth is 5–8 splits.[47] Since the specific optimal values may exist depending on the application, we performed cross-validation on random forest consisting of 50, 100, and 200 trees with maximum depth parameters of 4, 6, and 8 in this evaluation.

For the neural network training, batch size was set to 64, which refers to the number of training video clips processed at each iteration. We set the epoch number to 10, that is, the entire dataset was passed forward and backward through the neural network 10 times. We adopted the SGD optimizer with an initial learning rate of 1×10^{-3} to train the networks from the beginning. During the training, the learning rate was reduced by half by every two epochs. Such weight decay scheduling has been empirically observed to facilitate both model optimization and prediction power generalization.[32] We set the dropout ratio to 0.3, which can suppress model overfitting. It is noteworthy that to compare the early fusion and hierarchical fusion schemes, we prepared two spatio-temporal neural networks both consisted of five convolution layers. The difference is that early fusion adopted spatio-temporal convolution at only the first layer, while hierarchical fusion employs spatio-temporal convolutions at each convolution layer. By controlling the model depth, a valid and fair comparison can be performed.

Evaluation metric

The evaluation metrics, which are integral to an appropriate assessment, are described herein. By comparing the prediction results with the ground truth labels, we can derive four descriptive statistics, as follows:

1. True positive (TP): the number of video clips exhibiting flaw-induced patterns is correctly discerned.
2. True negative (TN): the number of normal wave propagation video clips classified as non-defects.
3. False positive (FP): the number of normal wave propagation video clips is incorrectly assessed to contain a defect.
4. False negative (FN): the number of video clips exhibiting flaw-induced patterns is incorrectly classified as no-defect.

We further extract four metrics of Accuracy (ψ), Recall (Re), precision (Pr), and F-score(γ)

$$\psi = \frac{TP + TN}{TP + FP + TN + FN}, \quad Pr = \frac{TP}{TP + FP}, \\ Re = \frac{TP}{TP + FN}, \quad \gamma = 2 \cdot \frac{Pr \cdot Re}{Pr + Re}. \tag{3}$$

Because the validation was performed using the leave-one-specimen-out method, we can extract the four metric statistics denoted as $\{\psi_p, Pr_p, Re_p, \gamma_p\}, p = [1, 2, \ldots, 17]$, where p denotes the p-th specimens. To investigate the overall performance, we extracted the average scores among all 17 specimens. The resultant statistics were expressed as $\{\overline{\psi}, \overline{Pr}, \overline{Re}, \overline{\gamma}\}$ and these four metrics were employed to evaluate the models. It is noteworthy that the F-score is the harmonic mean of the precision and recall, and it has been applied as a representative indicator for model assessment.

Results comparison

The objective of the experimental validation is threefold: First, we would like to justify the contribution of spatio-temporal information for UWPI video modeling. Second, after confirmation of the significance of spatio-temporal information, a problem arises: how to fuse temporal and spatial information efficiently. An empirical comparison has been performed to determine the optimal scheme. Third, 3D convolution filter decomposition could be regarded as an efficient way to ease spatio-temporal nerual network training. By conducting experiments on real data, we intend to understand the manner in which 3D kernel decomposition facilitates the spatio-temporal feature characterization of ultrasonic wave propagation patterns.

We first tested the conventional computer vision approaches composed of visual feature extraction and statistical classification algorithms. At the feature extraction stage, two dimension reduction techniques of principal component analysis (PCA) and uniform manifold approximation and projection (UMAP)[48] methods were employed to extract low-dimensional features from ultrasonic wavefield images. In the experiments, we standardize the wavefield image data to have a mean of 0 and a variance of 1, and the latent feature dimension is empirically set to 100. Besides, we evaluate the HoG feature to characterize incident traveling waves and flaw-induced anomalous waves. Three classification algorithms Naive Bayes classifier, support vector machines (SVM) with RBF kernel and random forest (RF)[50] were adopted to perform content-based classification. It is noteworthy that SVM and RF classifiers played dominant roles in machine learning before the era of deep learning. Therefore, we add them to the comparison list. Then, we evaluated a series of representative still image recognition CNN models as baselines, namely AlexNet,[49] ResNet,[29] and DenseNet.[30] It is noteworthy that we adopted those off-the-shelf models as reference baseline without any manual tweaking. Moreover, we added the most recent USresNet[13] for comparison, which is equipped with a dedicated network design for ultrasonic wavefield image classification. The results of reference methods are presented in the upper half of Table 6. As shown, DenseNet and USresNet were the top performers in addressing still wavefield images, and both yielded F-scores exceeding 95%. Two findings were revealed from the results: First, the advanced CNN models with a deeper depth can usually achieve better performances, and the results of AlexNet, ResNet, and DenseNet obtained in this application were consistent with their performances in universal image recognition.[30] Second, for a specific application, the neural network architecture can be customized to reduce the model complexity while retaining high performances. We discovered that the seven-layer USresNet achieved comparable performance to that of DenseNet with 120 layers. This is primarily attributed to the limitation of the data size. We further applied the late fusion method to the 2D CNN-based models. In detail, the 2D CNNs were applied to recognize each frame of a video clip and then average the predictions at the whole video clip level. Although simple, the late fusion scheme achieved slight performance improvements as an initial attempt to incorporate temporal information.

Next, we evaluated the core part, that is, the spatio-temporal neural networks for the task. In Table 6, ST-Net-3D denoted that approach that it is based on 3D convolution filters for combining three-way spatial and temporal information for UWPI data investigation. In detail, we evaluated two schemes for time-domain information characterization: early fusion and hierarchical fusion. Based on the results shown in Table 6, ST-Net-3D achieved superior performance gain in terms of the F-score. By considering spatio-temporal information, even the simple ST-Net-3D scheme using early fusion with a five-layer design achieved an average F-score of 96.35%, which outperformed the 120-layer DenseNet model (95.85%). The results comparison also demonstrated the effectiveness of the hierarchical spatio-temporal feature aggregation scheme using deep-stacking neural networks. Hence, spatio-temporal information is validated to be essential for understanding UWPI videos.

We subsequently investigated the effects of 3D convolution filter decomposition on the motion modeling of ultrasonic wave propagation. The method, denoted as ST-Net-(2+1)D, involves a method to approximate the 3D filter using two sequential filters, which address spatial and temporal information individually. The ST-Net-(2+1)D attained the highest F-score of 97.20% compared with those of other methods in Table 6. Moreover, we compared the model complexity regarding the number of learnable parameters (Million) of neural networks and model size (MB) in Table 7. The proposed ST-Net-(2+1)D is a rather lightweight model that requires less computational resources. It could be a desirable feature because it frees up memory and enables real-time operations in a real deployment scenario.

The factors contributing to this result are as follows: For a small dataset, a relatively simpler model is preferred over complex ones. In this study, our dataset comprised less than 1700 clip samples, which is regarded to be small for deep neural network training. The experimental results validated that the 3D filter decomposition can accelerate model training and improve performance.

We comprehensively analyzed the consistency of the wave propagation video classification performance for all specimens. This was performed because the defect-induced ultrasonic wave scattering pattern can vary significantly due to the flaw properties. An efficient computerized UWPI data analysis system should be robust to all these variations and always generate reliable predictions. To demonstrate the stability of the models, we present boxplots that describe the distribution of F-scores for all specimens. A full comparison is shown in Figure 9. In the boxplot, the central part is known as the interquartile range (IQR) box, which ranges between the first and third quartiles (Q3–Q1) of F-scores, and the red line indicates the median F-score, and the red "+" represents outliers that are below $Q1$-1.5× IQR or above $Q3$+1.5× IQR. Figure 9 shows the variability or dispersion of the F-scores of each method over the dataset. The comparison shows that the spatio-temporal CNNs generally performed better than the 2D CNN-based methods that focused on still images wavefield appearance. Moreover, the comparison results validated that the proposed UWPI

Table 6. Performance of the spatio-temporal neural networks for ultrasonic wave propagation imaging data analysis compared with the state-of-the-art methods.

	Model depth	Temporal aggregation	Accuracy $\overline{\psi}(\%)$	Recall $\overline{Re}(\%)$	Precision $\overline{Pr}(\%)$	F-score $\overline{\gamma}(\%)$
UMAP[48]+Naive Bayes	—	—	68.53	71.62	68.32	69.63
		Late fusion	69.20	71.45	68.69	69.72
UMAP+SVM	—	—	69.05	76.81	67.71	71.60
		Late fusion	69.79	76.77	68.20	71.83
UMAP+RF	—	—	68.83	76.44	67.69	71.41
		Late fusion	69.53	76.32	68.15	71.58
PCA+Naive Bayes	—	—	82.00	82.25	83.24	80.84
		Late fusion	81.01	79.36	83.36	78.64
PCA+SVM	—	—	91.19	96.33	88.23	91.68
		Late fusion	91.72	96.59	88.77	92.08
PCA+RF	—	—	88.30	96.08	83.95	89.15
		Late fusion	88.46	96.47	83.63	89.13
HoG+Naive Bayes	—	—	83.77	86.62	83.97	84.81
		Late fusion	83.37	90.58	82.91	85.45
HoG+SVM	—	—	90.80	92.70	89.27	90.95
		Late fusion	90.91	92.83	89.53	91.34
HoG+RF	—	—	91.44	92.13	91.77	92.02
		Late fusion	91.53	92.09	91.92	92.14
AlexNet, 2012[49]	5	—	86.96	94.88	67.70	77.87
		Late fusion	86.70	94.80	67.35	77.51
ResNet, 2016[29]	17	—	95.36	91.43	92.11	90.95
		Late fusion	96.14	91.54	94.46	92.20
DenseNet, 2017[30]	120	—	97.35	95.82	95.21	95.33
		Late fusion	97.62	95.68	96.37	95.85
UsresNet, 2018[13]	7	—	95.59	96.77	94.81	95.61
		Late fusion	95.58	96.85	94.64	95.50
ST-net-3D	5	Early fusion	96.29	97.39	95.30	**96.35**
ST-net-3D	5	Early fusion	96.67	96.51	97.09	**96.68**
ST-net-(2+1)D	5	Early fusion	97.19	96.81	97.73	**97.20**

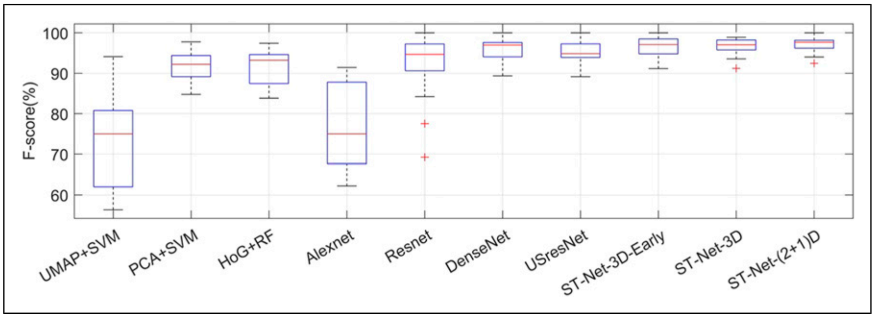

Figure 9. Specimen-wise classification accuracy comparison.

video analysis system achieved superior accuracy and favorable performance consistency.

Furthermore, we investigate the automatic defect detection performances with respect to flaw types because an efficient defect detection system is required to render high performance for various types of defects. As shown in Table 4, specimen #1 ~#9 and #10 ~ #17 contain hole-shaped and slit-shaped flaws, respectively. We calculated the average evaluation metric scores between the two sets of specimens to evaluate the flaw type-wise performance. The resultant statistics for hole-shaped defects and slit-like defects were expressed as $\{\overline{\psi_{\text{hole}}}, \overline{Re_{\text{hole}}}, \overline{Pr_{\text{hole}}}, \overline{\gamma_{\text{hole}}}\}$ and $\{\overline{\psi_{\text{slit}}}, \overline{Re_{\text{slit}}}, \overline{Pr_{slit}}, \overline{\gamma_{slit}}\}$ and we summarized all comparison results in Table 8 and Table 9. Through extensive comparison with other methods, the proposed ST-Net-(2+1)D achieved superior defect detection performance in terms of

Table 7. Complexity comparison of automatic defect detection models for ultrasonic wave propagation imaging method.

	AlexNet, 2012	ResNet, 2016	DenseNet, 2017	ST-Net-(2+1)D
Parameters(M)	36.53	6.39	7.98	0.51
Model Size(MB)	157.02	139.41	126.01	14.41

Table 8. Detailed detection performance comparison for hole-shaped flaws.

	UMAP + SVM	PCA + SVM	*AlexNet,* 2012	*ResNet,* 2016	*DenseNet,* 2017	*UsresNet,* 2018	*ST-Net-*(2+1)D
Acc. $\overline{\psi}_{\text{hole}}(\%)$	69.55	91.28	85.92	94.12	97.10	96.63	**97.91**
Prec. $\overline{Pr}_{\text{hole}}(\%)$	68.13	88.15	64.91	90.73	94.86	97.14	**99.08**
Rec. $\overline{Re}_{\text{hole}}(\%)$	76.78	96.56	93.45	87.07	94.84	96.64	**96.94**
F1$\overline{\gamma}_{\text{hole}}(\%)$	71.79	91.71	75.62	87.55	94.59	96.78	**97.97**

Table 9. Detailed detection performance comparison for slit-shaped flaws.

	UMAP + SVM	PCA + SVM	*AlexNet,* 2012	*ResNet,* 2016	*DenseNet,* 2017	*UsresNet,* 2018	*ST-Net-*(2+1)D
Acc. $\overline{\psi}_{\text{slit}}(\%)$	69.29	91.63	88.13	96.76	**97.63**	94.42	96.54
Prec. $\overline{Pr}_{\text{slit}}(\%)$	67.79	88.84	70.83	93.67	95.60	92.19	**96.20**
Rec. $\overline{Re}_{\text{slit}}(\%)$	76.79	96.35	96.50	96.33	**96.93**	96.91	96.67
F1$\overline{\gamma}_{\text{slit}}(\%)$	71.64	92.05	80.40	94.78	96.16	94.28	**96.33**

high detection precision and low miss detection rates for both slit-shaped and hole-shaped flaws.

The proposed ST-Net-(2+1)D learning architecture is designated to learn multiple levels of representations to amplify aspects of the input wavefield snapshots that are important for flaw-induced pattern discrimination and suppress irrelevant variations. To demonstrate the effectiveness of feature learning, we perform the latent space analysis by exploiting the bottleneck features of using t-Distributed Stochastic Neighbor Embedding (t-SNE) method. The bottleneck features is defined as the last output feature map (BatchNorm5, Tab. 3) before the fully-connected (FC) layers, which can be regarded as the highest level feature representation rendered by ST-Net-(2+1)D with respect to the flaw detection task. On the other hand, t-SNE has been acknowledged as an essential method to interpret the functionality of neural networks by exploiting how it transforms the high-dimensional raw (image) data to a low-dimensional space is preferable for the respective task. In Figure 10, we illustrate both the two-dimensional principal component analysis (PCA) features (left) extracted from raw UWPI snapshots and the t-SNE visualization of ST-Net-(2+1)D bottleneck features (right). The bottleneck feature visualization shows that in the latent space, the two classes—defective and normal cases, appear in two clusters which is ideal for a binary classification task. In contrast, the PCA features were incapable of capturing critical visual features for flaw-induced pattern discrimination. The latent space analysis manifests that the non-linear hierarchical feature representation learned by ST-Net-(2+1)D can greatly facilitate discrimination of the defective patterns from UWPI data.

Finally, we conducted the over-fitting analysis by investigating learning curves. Overfitting refers to the case where model performance on the training dataset is improved at the cost of worse performance on data not seen during training. It is regarded as one of the most common problems for training deep neural networks because they have far more parameters than training samples. We employed a batch of techniques to avoid overfitting, including simplifying the model by 3D convolution kernel decomposition, increasing the dataset with data augmentation, and applying batch normalization and dropout in the ST-Net architecture. This section presents the learning curves with the metrics of training loss and prediction accuracy to demonstrate model performance. Figure 11 showed the three learning curves of training loss for USresNet, DenseNet, and ST-Net-(2+1)D models. As the training continued, we can observe that the loss decreased to a point of stability, and validation loss exhibited a similar trend while maintaining a small gap with the training loss. By examining the two learning curves, we can confirm that the model training had proceeded successfully. The same conclusion can also be inferred from Figure 12 that presented the learning curves of prediction accuracy.

Conclusion and future works

The aim of this study was to develop a video content analysis system that can mimic the ability of the human eye

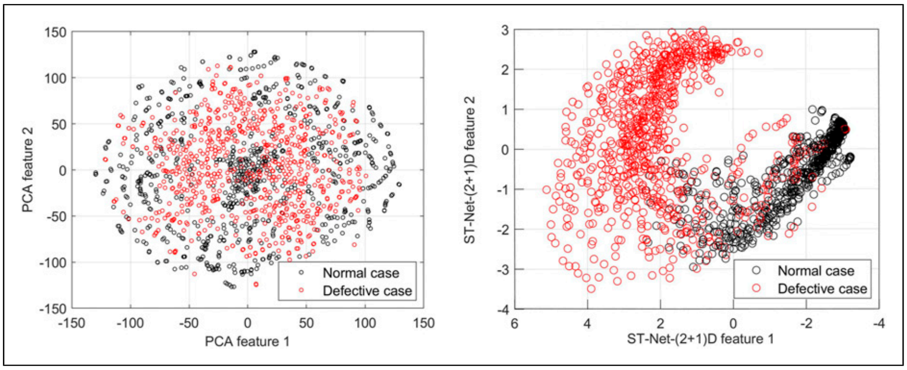

Figure 10. Latent space comparison: raw wavefield image PCA features (left) and ST-Net-(2+1)D bottle neck features with t-SNE visualization (right).

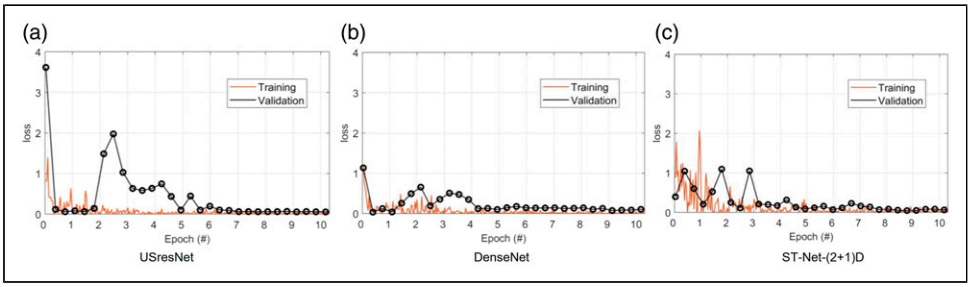

Figure 11. Learning curves comparison: Model training loss.

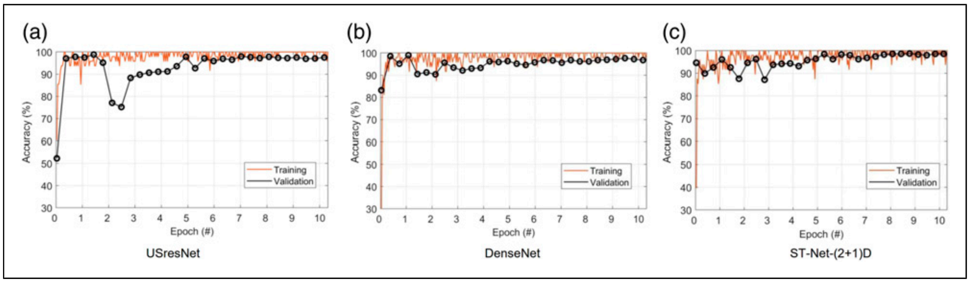

Figure 12. Learning curves comparison: Defect detection accuracy.

for watching a movie of ultrasonic testing wave motion resulting from a scanning pulsed laser and for understanding the manner in which waves propagate and interact with any obstacle (flaw) in their path. First, we prepared an ultrasonic wavefield imaging dataset that included more than 1600 annotated clips for model assessment. Subsequently, we presented the idea and backbone architecture of spatio-temporal CNNs for generic video understanding. Regarding ultrasonic wave motion analysis, the fundamental problem was the method to use time-domain information. Hence, we investigated three approaches to incorporate temporal features for better wave propagation modeling. Moreover, to reduce the complexity of spatio-temporal learning neural networking and facilitate model training, we factorized thes patio-temporal (3D) convolution filters using coupled 2D and 1D convolution operators. Because we were using a small-scale dataset, it can be beneficial to adopt a light-weight model with fewer learnable parameters. Subsequently, we performed an extensive evaluation of representative image classification methods. The results validated that the proposed video-based deep learning system outperformed all conventional approaches. In addition, to obtain a better understanding of the model performance consistency, we analyzed its behavior on all specimens with different flaws. The results confirmed that the proposed method achieved superior performance in terms of both defect pattern classification accuracy and model stability. We hope that our analysis will inspire new research that harness the potential efficiency of video-based analysis for automatic NDT data interpretation, not limited to ultrasonic wave propagation data.

The proposed method was validated for automatic damage detection, which is one of the essential tasks in SHM. Our future work will be as follows: First, an elaborate study will be conducted to obtain more information about the defects present in the wavefield imaging data. The current system can only determine whether a defect exists in the wavefield image and cannot provide further details, such as location and size. Second, we will expand our dataset by using new specimens with other kinds of defects. It can be anticipated that with the extended dataset, both the defect detection performance and generalization ability to defect multiple damages of spatio-temporal neural networks would be improved. Moreover, it has been widely acknowledged that there is a need to go further than just damage detection to the assessment of performance and life of a structure. From this perspective, the proposed methodology has the potential to serve as a compelling anomalies detector in ultrasonic wavefield imaging inspection for a wide range of SHM applications. The rapid progress in hardware platforms for deep learning and communication technologies will soon become feasible to deploy the proposed method on site. Validating the method in the context of practical SHM will be one of the focuses of our future research.

Acknowledgments

This article is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Jiaxing Ye https://orcid.org/0000-0002-6680-1201
Nobuyuki Toyama https://orcid.org/0000-0002-9657-891X

References

1. Schmerr LW. *Fundamentals of Ultrasonic Nondestructive Evaluation*. New York, NY: Springer, 2016.
2. Yashiro S, Takatsubo J, Miyauchi H, et al. A novel technique for visualizing ultrasonic waves in general solid media by pulsed laser scan. *NDT E Int* 2008; 41(2): 137–144.
3. Michaels JE. Ultrasonic wavefield imaging: research tool or emerging nde method? In: 43rd Annual Review of Progress in Quantitative Nondestructive Evaluation AIP Conference Proceedings, Atlanta, Georgia, 17–22 July, 1806, p. 020001.
4. Yeager M, Gregory B, Key C, et al. On using robust mahalanobis distance estimations for feature discrimination in a damage detection scenario. *Struct Health Monit* 2019; 18(1): 245–253.
5. Wang R, Chencho, An S, et al. Deep residual network framework for structural health monitoring. *Struct Health Monit* 2021; 20(4): 1443–1461.
6. Robini MC, Magnin IE, Benoit-Cattin H, et al. Two-dimensional ultrasonic flaw detection based on the wavelet packet transform. *IEEE Trans Ultrason Ferroelectrics Frequency Control* 1997; 44(6): 1382–1394.
7. Kudela P, Radzieński M and Ostachowicz W. Identification of cracks in thin-walled structures by means of wavenumber filtering. *Mech Syst Signal Process* 2015; 50-51: 456–466.
8. Radzienski M, Krawczuk M, ´ Zak A, et al. Application of rms for damage detection by guided elastic waves. In: Journal of Physics: Conference Series,Volume 305, Oxford, United Kingdom, IOP Publishing, 11–13 July 2011.
9. Mesnil O, Yan H, Ruzzene M, et al. Fast wavenumber measurement for accurate and automatic location and quantification of defect in composite. *Struct Health Monit* 2016; 15(2): 223–234.
10. Dafydd I and Sharif Khodaei Z. Analysis of barely visible impact damage severity with ultrasonic guided lamb waves. *Struct Health Monit* 2020; 19(4): 1104–1122, p. 012085.
11. Bao Y and Li H. Machine learning paradigm for structural health monitoring. *Struct Health Monit* 2021; 20(4): 1353–1372.
12. Virkkunen I and Koskinen T. Flaw detection in ultrasonic data using deep learning.In: Baltica XI 2019: International Conference on Life Management and Maintenance for Power Plants, Helsinki, 11–13 June 2019. VTT Technical Research Centre of Finland.
13. Ye J, Ito S and Toyama N. Computerized ultrasonic imaging inspection: From shallow to deep learning. *Sensors* 2018; 18(11): 3820.
14. Ye J and Toyama N. Usimgaist: ultrasonic inspection image dataset for non-destructive evaluation. Available at: https://sites.google.com/site/yejiaxingweb/usimgaist (acessesd 2021).
15. Leshno M, Lin VY, Pinkus A, et al. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks* 1993; 6(6): 861–867.
16. Lu Z, Pu H, Wang F, et al. The expressive power of neural networks: a view from the width. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NY, USA, 4 December 2017. pp. 6232–6240.
17. Goodfellow I, Bengio Y, Courville A, et al. *Deep Learning*, Cambridge: MIT press Cambridge, 2016.
18. Lawson SW and Parker GA. Automatic detection of defects in industrial ultrasound images using a neural network. In: Vision Systems: Applications, volume 2786, pp. 37–47, International Society for Optics and Photonics. SPIE, 1996.
19. Guarneri GA, Junior FN and de Arruda LVR. Comparative evaluation of artificial neural networks models to classify weld flaws using pulse-echo ultrasonic signals. In: 22nd International Congress of Mechanical Engineering, Brazil, 3–7 November 2013, pp. 2126–2134.

20. Virupakshappa K and Oruklu E. Ultrasonic flaw detection using support vector machine classification. In: 2015 IEEE International Ultrasonics Symposium (IUS), Taipei, Taiwan, 21–24 October 2015, pp. 1–4.
21. Chun P-j, Ujike I, Mishima K, et al. Random forest-based evaluation technique for internal damage in reinforced concrete featuring multiple nondestructive testing results. *Construction Building Mater* 2020; 253: 119238.
22. LeCun Y, Bengio Y and Hinton G. Deep learning. *Nature* 2015; 521(7553): 436–444.
23. Munir N, Park J, Kim H-J, et al. Performance enhancement of convolutional neural network for ultrasonic flaw classification by adopting autoencoder. *NDT E Int* 2020; 111: 102218.
24. Meng M, Chua YJ, Wouterson E, et al. Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks. *Neurocomputing* 2017; 257: 128–135.
25. Pyle RJ, Bevan RL, Hughes RR, et al. Deep learning for ultrasonic crack characterization in nde. *IEEE Trans Ultrason Ferroelectrics, Frequency Control* 2021; 68: 1854–1865.
26. Virkkunen I, Koskinen T, Jessen-Juhler O, et al. Augmented ultrasonic data for machine learning. *J Nondestructive Eval* 2021; 40(1): 1–11.
27. Zhu Y, Li X, Liu C, et al. A comprehensive study of deep video action recognition. arXiv preprint arXiv:201206567 2020.
28. Wang H, Kläser A, Schmid C, et al. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *Int Journal Computer Vision* 2013; 103(1): 60–79.
29. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778.
30. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017, pp. 4700–4708.
31. Drai R, Khelil M and Benchaala A. Time frequency and wavelet transform applied to selected problems in ultrasonics nde. *NDT E Int* 2002; 35(8): 567–572.
32. Cruz FC, Simas Filho EF, Albuquerque MCS, et al. Efficient feature selection for neural network based detection of flaws in steel welded joints using ultrasound testing. *Ultrasonics* 2017; 73: 1–8.
33. Alguri KS, Chia CC and Harley JB. Model-driven, wavefield baseline subtraction for damage visualization using dictionary learning. In: Structural Health Monitoring, Stanford, CA, 12–14 September 2017.
34. Zahran O and Al-Nuaimy W. Utilising phase relationships for automatic weld flaw categorisation in time-of-flight diffraction images. In: 11th International Conference on Fracture (ICF), Turin, Italy, pp. 20–25.
35. Shitole CSN, Zahran O and Al-Nuaimy W. Combining fuzzy logic and neural networks in classification of weld defects using ultrasonic time-of-flight diffraction. *Insight - Nondestructive Test Condition Monit* 2007; 49(2): 79–82.
36. Sambath S, Nagaraj P and Selvakumar N. Automatic defect classification in ultrasonic ndt using artificial intelligence. *J Nondestructive Evaluation* 2011; 30(1): 20–28.
37. Jin H, Yang K, Wu S, et al. Sparse deconvolution method for ultrasound images based on automatic estimation of reference signals. *Ultrasonics* 2016; 67: 1–8.
38. Tian Y, Maitra R, Meeker WQ, et al. A statistical framework for improved automatic flaw detection in nondestructive evaluation images. *Technometrics* 2017; 59(2): 247–261.
39. Ye J and Toyama N. Benchmarking deep learning models for automatic ultrasonic imaging inspection. *IEEE Access* 2021; 9: 36986–36994. DOI: 10.1109/ACCESS.2021.3062860.
40. Ijjeh AA, Ullah S and Kudela P. Full wavefield processing by using fcn for delamination detection. *Mech Syst Signal Process* 2021; 153: 107537.
41. Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7–13 December 2015, pp. 4489–4497.
42. Sun L, Jia K, Yeung DY, et al. Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7–13 December 2015, pp. 4597–4605.
43. Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018, pp. 6450–6459.
44. Toyama N, Yamamoto T, Urabe K, et al. Ultrasonic inspection of adhesively bonded CFRP/aluminum joints using pulsed laser scanning. *Adv Compos Mater* 2019; 28(1): 27–35.
45. Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, 7–12 February 2020, volume 34. pp. 13001–13008.
46. Dalal N and Triggs B. Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005, pp. 886–893.
47. Oshiro TM, Perez PS and Baranauskas JA. How many trees in a random forest? In: International workshop on machine learning and data mining in pattern recognition, Berlin, Germany, 13 July 2012, Springer, pp. 154–168.
48. McInnes L, Healy J, Saul N, et al. Umap: uniform manifold approximation and projection. *J Open Source Softw* 2018; 3(29): 861.
49. Krizhevsky A, Sutskever I and Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Information Processing Systems* 2012; 25: 1097–1105.
50. Murphy KP. *Machine Learning: A Probabilistic Perspective*. MA: MIT press, 2012.