Noname manuscript No.
(will be inserted by the editor)

Convolutional Neural Network for Risk Assessment in Polycrystalline Alloy Structures via Ultrasonic Testing

Hassan Alqahtani* · **Asok Ray**

Received: date / Accepted: date

Abstract In the current state of the art of process industries/manufacturing technologies, computer-instrumented and computer-controlled autonomous techniques are necessary for damage diagnosis and prognosis in operating machinery. From this perspective, the paper addresses the issue of fatigue damage that is one of the most commonly encountered sources of degradation in polycrystalline-alloy structures of machinery components. It is possible to conduct in-situ detection & classification of damage as well as an assessment of the remaining service life through ultrasonic measurements of material degradation and their computer-based analysis. In this paper, tools of machine learning (e.g., convolutional neural networks (CNNs)) are applied to synergistic combinations of ultrasonic measurements and images from a confocal microscope (Alicona) to detect and evaluate the risk of fatigue damage. The database of the confocal microscope has been used to calibrate the ultrasonic database and to provide the ground truth for fatigue damage assessment. The results show that both the ultrasonic data and confocal microscope images are capable of classifying the fatigue damage into their respective classes with considerably high accuracy. However, the ultrasonic CNN model yields better accuracy than the confocal microscope CNN model by almost 9%.

Keywords Fatigue damage ; Damage detection , Damage classification ; Ultrasonic; Convolutional neural networks.

H. Alqahtani
Department of Mechanical Engineering, Taibah University, Medina, KSA, 42353
E-mail: hhqahtani@taibahu.edu.sa
Corresponding author

A. Ray
Department of Mechanical Engineering, Pennsylvania State University, University Park, PA 16802.
Department of Mathematics, Pennsylvania State University, University Park, PA 16802
E-mail: axr2@psu.edu

1 Introduction

Ultrasonic testing is a typical nondestructive method that is extensively used for detection and evaluation of defects in mechanical structures. The accuracy and duration of testing should be taken into consideration during the nondestructive testing (NDT) procedure. In this context, it is well known that ultrasonic testing is capable of accurate detection, classification. and characterization of defects in mechanical structures in a timely fashion [1, 2]. However, these techniques depend on an inspector's competence and experience. Usually, during the examination procedure of ultrasonic testing (UT), a human operator often determines the status of tested components by evaluating the signal features, such as the echo shape, amplitude level, and defect position. Hence, defect characterization and identification of relevant and non-relevant flaws using NDT are highly dependent on the (human) inspector's skill [3].

Inspector's errors and, distractions, such as those due to stress, may result in missing an existing defect or incorrectly identifying a false (i.e., nonexisting) defect. Therefore, the defects diagnosed in mechanical structures by conventional NDT are often error-prone and may require an experienced and highly skilled human inspector. To improve the NDT-based evaluation, many industries have focused on building autonomous ultrasonic defect detection & classification systems that could perform detection and classification of defects in mechanical components without direct assistance of human inspectors [4].

The last three decades have seen the evolution of diagnostics of structural damage by usage of artificial intelligence (AI) techniques. Specifically, the tools of pattern recognition have been applied to improve the ability of AI for detection & classification of fatigue damage in mechanical structures, where the pattern of the measurements of a tested component may differ from those of the available records of undamaged and damaged ones. Neural networks (NN) have been widely used for ultrasonic flaw detection and several researches have shown the capability of NN for classification. Thiago et al. [5] evaluated the efficiency and accuracy of artificial intelligence techniques to classify ultrasonic signals, and their model classification performance reaches 93%. Margrave et al. [6] applied several types and configurations of the neural network to detect flaws in steel pipes using ultrasonic signatures. Liu et al. [7] studied classification of crack growth behavior by applying an NN on characteristic values obtained from ultrasonic signals. Sambath et al. [8] built an automatic ultrasonic flaw detection & classification model using NN, and the model performance was improved by applying the wavelet transform. Song et al. [9] used an Intelligent Ultrasonic Evaluation System (IUES) to detect and classify weldment flaws in a real-time fashion. Drai et al. [10] classified volumetric and planar defects by applying NN on extracted features of ultrasonic data in the time domain, frequency domain, and discrete wavelet representations. Seyedtabaii [11] experimented with new intelligent algorithms for classification of weld defects using single fixed-angle ultrasonic probes. Several feature extraction methods were discussed in [12] to classify different defects by the NN-based method. Recently, the usage of large dimensional data is no longer an obstacle

due to the incredible increase in computing power. Several types of research used a fully connected (vanilla) neural network and a convolutional neural network (CNN), to classify defects without extracting any features from the raw data. Chen. Z. and his team applied CNN for fault diagnosis of an automotive five-speed gearbox. Their proposed CNN model provides an excellent fault diagnosis performance, 99%.[13] Although the performance of CNN for damage classification was reasonable well [14], the performance was checked for only the ultrasonic signal without considering a real defect image. Therefore, as an extension to the previously reported research, this paper is a step further to synchronously compare the CNN performance on the ultrasonic echo signal and real defects images. In the authors' previous work [15], the CNN model was developed for crack status classification, where the severity of the crack was classified into three categories: healthy, low-risk, and high risk. The raw data was real damage images that were taken by an Alicona confocal microscope.

Almost all of the above-mentioned research works require feature extraction from raw data by statistical and/or signal processing techniques. Although feature extraction methods are applicable for dimensionality reduction, they involve an exhaustive process, because of the need for careful selection of features that must stay insensitive to the operating conditions. This paper provides a novel approach to quantify the risk of fatigue damage , independently of human involvement, where a convolutional neural network (CNN) model classifies the ultrasonic echo signal into three classes: healthy, low-risk damage, and high-risk damage. The ultrasonic echo signal classification is based on the damage size shown in the confocal microscope image. Therefore, this study provides an accurate risk assessment of fatigue damage using an ultrasonic echo signal.

The major contributions of this paper are delineated as:

1. *Development of a risk assessment and calibration procedure*: The ultrasonic echo signal are synchronized with an image from a confocal microscope, which illustrates the state of health of the tested component.
2. *Construction of a robust classification model*: The ultrasonic signals are classified using convolutional neural networks (CNNs) without using additional signal processing techniques.
3. *Performance comparison*: The CNN model of the ultrasonic signals is compared with another CNN model by using a different database.

The paper is organized into five main sections including the present one. The second section presents a description of the laboratory apparatus that serves as the data generator for validation of the methodology of fatigue-damage detection & classification, proposed in this paper. The third section illustrates the methodology including a strategy of data augmentation, training and testing datasets, and an overview of the convolutional neural networks (CNN). The fourth section shows and discusses the results of experimental validation of the proposed methodology. Finally, the fifth section summarizes and concludes the paper with recommendations for future research.

2 Description of the Experimental Apparatus

The main objective of this investigation is validation of the theoretical results on crack detection at the initiation stage, because a large part of the service life of ductile-alloy structures under medium to high-cycle fatigue is consumed in the crack initiation stage [16].

This section describes the experimental apparatus, as depicted in Figure 1, which is built upon a hydraucally-operated, computer-instrumented and computer-controlled fatigue testing machine[1], equipped with ultrasonic testing (UT) probes[2], a confocal microscope[3], and a digital microscope[4]. The core concept of anomaly detection during the crack initiation stage in this investigation is built upon a synergistic combination of the heterogeneous measurement data, generated from optical images (of the Alicona confocal microscope) and an ensemble of time series from ultrasonic sensors, where the goal is to enhance the performance of the damage-tolerant design, maintenance, and operation of mechanical components of machinery.

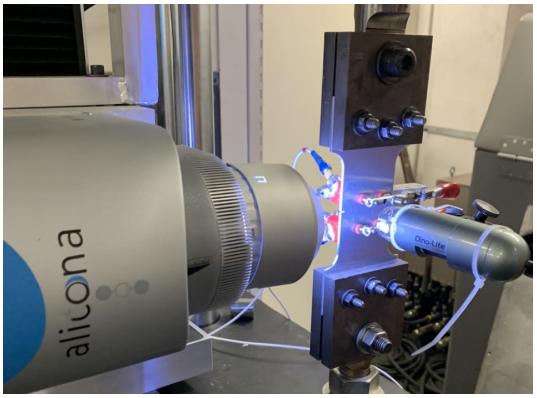


Fig. 1 The experimental apparatus

Twenty-one tests were conducted on the experimental apparatus in a laboratory environment at room temperature to develop an automated monitoring system to identify the fatigue damage properties of polycrystalline alloys. These tests were performed on specimens of 7075-T6 aluminum alloy, where all specimens were tested either for low-cycle or medium-cycle fatigue tests under variable-amplitude and variable-frequency random loading. The dimensions of these specimens are 3 mm thick, 50 mm wide with (1 mm×3.5 mm) slot cut at the edge. The testing of all specimens were performed on tension-tension load

[1] Manufacturer: MTS Systems Corporation, Berlin, NJ, USA
[2] Manufacturer: OLYMPUS, Shinjuku, Tokyo, Japan
[3] Manufacturer: Alicona Imaging GmbH, Dr.-Auner-Strasse 21a, 8074 Raaba/Graz, Austria
[4] Manufacturer: QUESTAR®, New Hope, Pennsylvania, USA

cycles at 60 Hz. The mean load set-point was 8,000 N with a load-amplitude of 3,500 N. The ensemble of information from the three sensors in the apparatus (see Figure 1) have been fused for NDT evaluation to detect the point of crack initiation and to assess the risk of failure in the tested specimens.

2.1 Ultrasonic testing:

An angle-beam transducer has been used in this investigation, which consists of a transmitter and a receiver. The transmitter is used to inject high-frequency acoustic pulses (e.g., 15 MHz ultrasonic waves) into the test specimen and, after their propagation through the test specimen, the waves reach the receiver which is located on the test specimen in the opposite side of the transmitter. The received pulses are influenced by material defects (e.g., anomalous grain boundaries, voids, and inclusions) that may exist on the path of the propagated pulses, but their influence must be limited and stable. On the other hand, these pulses are significantly affected (i.e., attenuated) by the defect growth, because part of the pulses are reflected and not received by the receiver.

2.2 Optical metrology device:

The optical metrology device, which is the Infinite-Focus, Alicona, has been used in this investigation to make 3D surface measurements. The operating principle of the Alicona confocal microscope is that the topographical and color information is generated from variations of the focus, where the small depth of focus of an optical system is combined with vertical scanning; and the range of vertical resolution of the infinite-Focus system is $\sim 20nm$. The Alicona image size is typically $0.4mm \times 0.4mm$, as shown in Figure 2, and each such image has $\sim 4,161,600$ pixels. Thus, the Alicona confocal microscope is able to identify micro-cracks that occur at the crack initiation stage. In addition, All measured surfaces were polished (mirror finish) to clarify the path of the crack initiation.
In this investigation, Alicona measurements have been synchronized with ultrasonic data to provide the ground truth for attenuation of the ultrasonic echo signal .

2.3 Digital microscope:

The risk assessment is quantified from the estimated value of crack length, which is obtained from a magnified image. Digital microscope (DM) images have been taken in loose synchronism with the ultrasonic echo signal and Alicona images. The image resolution of the digital microscope is $\sim 640 \times 480$ pixels, and the range of variable magnification of these images is 10-200X.

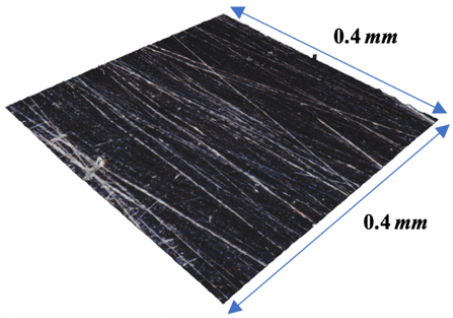


Fig. 2 3D surface generated by the Infinite-Focus device

3 Methodology

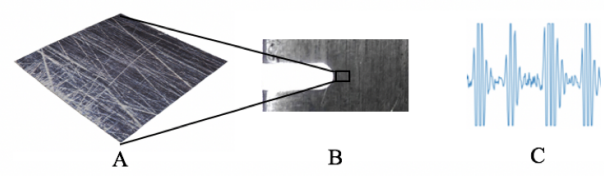
While the procedure of risk assessment is explained in details in previous publications [15,17] of the authors, the criteria of risk assessment are succinctly explained below for completeness of the current paper.

1. All measurements before the onset of the crack on the notch surface of test specimens are considered to be in a healthy (non-risky) state.
2. All crack measurements, before the critical crack length is reached, are considered to be in a low-risk state.
3. All crack measurements, after the critical crack length is reached, are considered to be in a high-risk state.

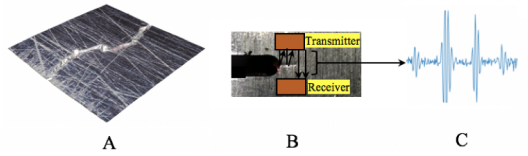
Each of the three plates (a), (b), and (c) in Figure 3 present the following two views of a 3D surface measurement: (i) a side view image by the digital microscope (DM), and (ii) a waveform of the corresponding ultrasonic signal. In the case of images with a crack, the crack is computed from the DM image, as shown in Figure 4, along with the corresponding the ultrasonic echo signal and Alicona images. If the crack length is less than the critical crack length, measured data are characterized to be in the low-risk state; and measured data exceeding the critical crack length are characterized to be in the high-risk state.

3.1 Data augmentation

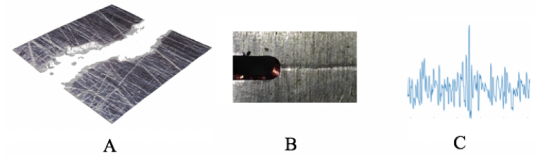
One of the uncertainties types of AI is the epistemic uncertainties which refers to uncertainty caused by lack of data. Therefore, insufficient training data may result in a poor approximation, and thus consequences of insufficient test data are (possibly) optimistic and high-variance estimation of model performance.Usually the causes of uncertainty arise when the training data and test data are Incompatible, [18,19]. Therefore, the amount of data for deep neural networks must be acceptable for both training and testing purposes. In the case of having few data, a data augmentation technique is usually adopted to boost the size of the database [20–22]. In this paper, the size of the original database



(a) Healthy state measurements of a tested specimen



(b) Low-risk state measurements of a tested specimen.



(c) High-risk state measurements of a tested specimen.

Fig. 3 The illustration of damage state of a tested specimen using Alicona measurements (A), digital microscope measurements (B), and the ultrasonic echo signal (C)

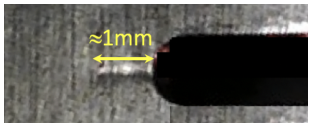


Fig. 4 Estimated Crack length from a digital microscope image

was inadequate. So, the data set has been augmented to build an efficient neural networks (NN) model for prediction and classification. The techniques used for data augmentation were rotation, transition, reflection, and scaling. In the rotation method, every image produced 37 different images. The transition method shifted the original signal image forward by a distinct distance, and it produced 21 images. The third augmentation technique is the reflection, where every signal image generated its reflected image. The last augmentation technique is scaling that produced 31 different scaled images. Figures 5 & 6 show the techniques of the rotation and transition, respectively.

Using the above data augmentation techniques, the number of signal images in the database was boosted from 881 to 80,171. Table 1 shows the data size before and after the augmentation. Table 2 shows the number of signals for each risk state in the augmented database.

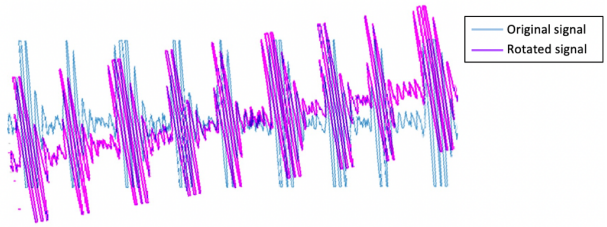


Fig. 5 An example of an augmentation method by rotation

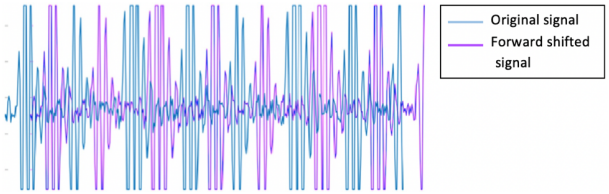


Fig. 6 An example of an augmentation method by transition

Table 1 Techniques for database augmentation.

Augmentation Technique	Original Data Size	Augmented Data Size
Rotation	881	32,597
Scaling	881	27,311
Transition	881	18,501
Reflection	881	1,762
Total		80,171

Table 2 No of signals for each risk state in augmented database

Risk State	No. of signals
Non-Risky	33,943
Low-Risk	29,029
High-Risk	17,199
Total	80,171

3.2 Training and testing datasets

Training data sets refer to the samples of data used to build the models, where the weights and biases of the neural network (NN) models are adjusted during the training process by using these data; thus, the models learn from these data. However, training the models on the actual data and examining their performance must be done on different data sets. In order to evaluate actual performance of the models, each data set is divided into training, validation, and testing categories, where the validation data set is used to provide an unbiased assessment of the respective model. In contrast, the weights and biases of a model are adjusted based on the respective training data set; once the NN model is completely trained by the training and validation data sets, the

testing data set is used to provide an unbiased assessment of the constructed model's performance.

A data set in this paper is split into a training set and a testing set. Furthermore, a part of the training set is further divided into a training subset and a cross-validation subset. The split ratio of the data set in this paper follows the 75/25 rule (i.e., $\sim$ 75% for training and $\sim$ 25% for testing), and then almost 10% of the training dataset is used for the validation purpose [23].

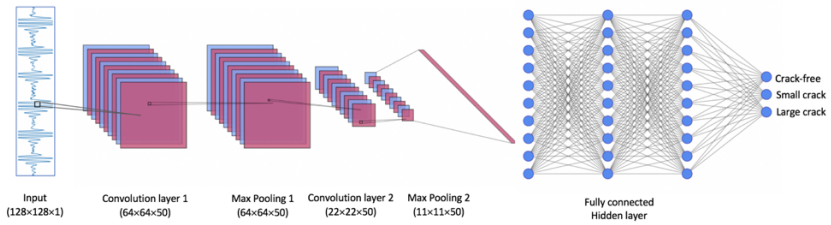
3.3 The Convolutional Neural Network (CNN)

A neural network (NN) is a computational method that is based upon Rosenblatt's perceptron training algorithm that attempts to mimic the logic of a human brain. The NN is based on a collection of nodes and a set of connections that link the neurons layerwise. The simplest feed-forward fully connected neural network is composed of three layers: an input layer, a hidden layer, and an output layer. In essence, an NN works by building connections between the nodes, where every node in the current layer is connected to each node in the previous layer and has an associated weight and a threshold. The node is activated and passing data to the next layer of the network when the output of the node is above the specified threshold, else the node is deactivated and no data are passed to the next layer of the NN. When the NN architecture has at least more than one hidden layer between input and output layers, this NN architecture is poularly known a Deep neural network (DNN).[24].

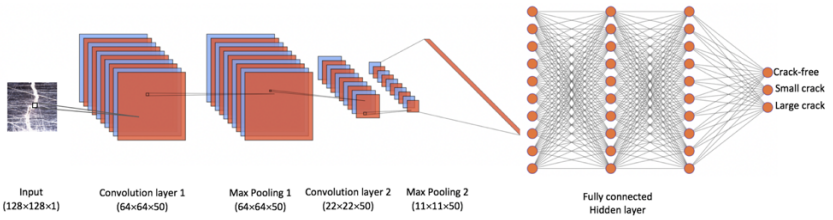
One of the common DNN types is the convolutional neural network that is comprised of two main parts. The first part is feature learning which contains various layers, such as the convolution layer (CL) and pooling layer (PL), while the second part is dedicated to classification which includes a fully connected layer (FCL) and Softmax layer (SML). Different CNN networks can be created by several combinations of these layers. Figures 7 (a) and (b) illustrate a CNN architecture with five layers for an ultrasonic database and an Alicona database, respectively. The input image is of size 128×128, representing height and width respectively. During the feature learning process, the input image flows into couples of convolutional layers with pooling layers such that feature extraction and redundancy reduction occur. The feature maps are the vital part of classification, where the simple features gradually assemble effectively. Then, all the features are merged partially and the resultant features are "flattened" to form a $1D$ column vector as an input to the fully connected layer (FCL). The resultant score for each particular class is converted to probability scores by the softmax layer.

3.3.1 Convolution Layer (CL):

The convolutional layer is one of the essential building blocks of the CNN architecture, which is also the most computation-intensive. A CL consists of several kernels that are learnable filters. Every kernel is spatially small of size



(a) Typical CNN architecture of Ultrasonic signals.



(b) Typical CNN architecture of Alicona images.

Fig. 7 Convolution Neural Network architecture for fatigue damage classification. It consists of 2 convolutional layers with corresponding ReLU activation layers, 2 max pooling layers, 2 fully-connected layers and a softmax output layer.

$h \times w \times n$, where h and w represent the height and width of the filter and n is the channel number of in the input image. Typical options of kernel size are 3×3 or 5×5. During the forward propagation, each kernel performs convolution on the input image across the image size (along width and height) and calculate the dot products between the kernels elements and the elements at any position of the image, this computational process is followed by a nonlinear activation function such as ReLU and sigmoid. Then, $2D$ activation maps (also known as feature maps) are created, where the $2D$ convolution of two signals is defined as:

$$\begin{aligned} h(m,n) &= g(m,n) \star f(m,n) \\ &= \sum_{i=-k}^{k} \sum_{j=-k}^{k} f(i,j) g(m-i, n-i) \end{aligned}$$

where the operator $\star$ is the convolution product of two functions; $h(m,n)$ is the convolved output; $g(m,n)$ is the input image; and $f(m,n)$ is the kernel. Figure 8 illustrates the convolution process; the computed activation maps depend on three hyperparameters: depth, stride, and padding. The number of kernels that are used in the convolution operation determines the depth of the activation maps. Each kernel learns a special pattern of the image such as edges, blobs, and colors. The stride is defined as the number of steps that the kernel is moved in the input image. For example, when the kernel moves one pixel at a time, the stride is one, but it is two when the kernel jumps two

pixels at a time as they are moved around; hence, the size of the feature maps is reduced.

Applying convolution to an input image may lead to loss of information, because the output size is reduced after the convolution operation. Therefore, the padding step is used to control the output size. For example, if the kernel size is 3×3, the input image is padded with zeros around the border [25]. The output size of the convolution layer is calculated in the following way:

$$CL_{size} = \left[\frac{W - K + 2P}{S}\right] + 1 \tag{1}$$

W: The input volume.
K: The Kernel size.
P: The padding.
S: The stride.

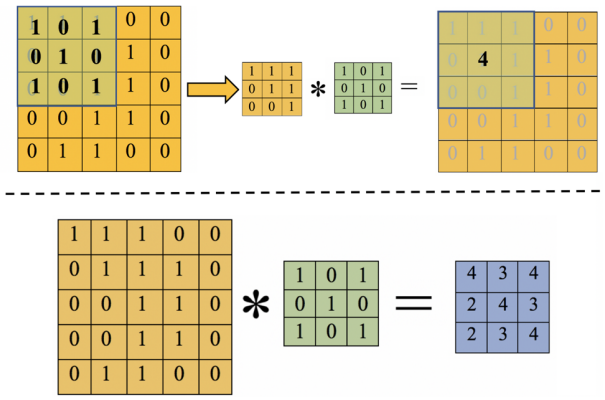


Fig. 8 An example of convolution ($\star$) with the stride length being equql to 1, where the kernel is slid over the entire image to produce a feature map. The stride length is a hyper-parameter which can be used to tune the network.

3.3.2 Pooling layer:

The poolong layers (PLs) are often applied after convolution layers to progressively reduce the spatial size of the respective input images; and hence, the computational cost may increase in the network. The PLs are referred as subsampling or downsampling, and it does not have parameters to learn. By applying a pooling operation, overfitting of the network can be controlled, while hyperparameters of the pooling layer PL indicate the filter size and strides.

Two common pooling techniques are mean-pooling and max-pooling. The mean-pooling takes the average of the matrix, while the max-pooling takes the maximum of matrix [26], where max-pooling is used more commonly than

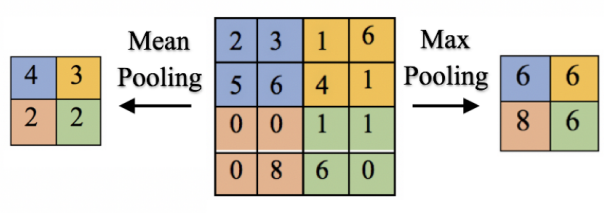


Fig. 9 An example of pooling with max pooling on the right hand side and mean pooling on the left. The pooling process downsamples the image to reduce the number of parameters and thereby reduced computational complexity of the network.

mean-pooling. The concept of the max-pooling technique is that the large number refers to a feature of the input image. In this example, a stride of 2 is selected. That is, the matrix size of the max-pooling is 2×2. It can be seen that the result of the max-pooling operation is downsampling; for example, a 4×4 matrix is downsampled to a matrix of 2×2. In the architecture, a max-pooling layer has been used after each convolution. The first pooling layer has a stride of 3×3 and the second pooling layer has a stride of 2×2.

3.3.3 Fully Connected layer:

The last few layers of the CNN architecture are typically constructed as fully connected layers, where all feature maps achieved at the last convolutional layer are flattened and are associated with the fully-connected layer. The basic idea of the fully connected (FC) layer is to transform the tensor at the output of the convolution and pooling layers into a vector and then several neural network layers are added. A fully connected layer is made up of neurons (perceptrons). As shown in Figure 10, the neuron involves several inputs and produces a single output; x_1, x_2, x_3 are inputs and Y is the output of the single neuron. Each neuron is associated with inputs through a real number called the *weight*, which implies the importance of the respective inputs relative to the output. The output of the neuron is computed by whether the weighted sum $\sum w_i x_i$ is greater than the threshold value, the output is 1, or less than the threshold value, the output is 0. The last layer of FL layers is the Softmax layer, and it is used to turn a vector with real-values [27] into a vector with elements in the range $[0, 1]$, which sum to 1. The softmax function is defined by the equation given below.

$$f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \tag{2}$$

3.3.4 Backpropagation:

Typically, at the first CNN, outputs tend to deviate from the desired output because the initial values of the kernels are arbitrarily chosen; this deviation

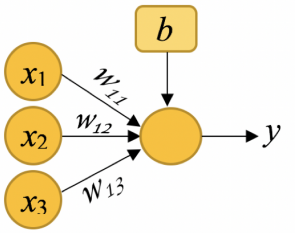


Fig. 10 An artificial neuron.

is computed using loss function [28]. The cross-entropy loss function of CNN architecture is defined as

$$L = -\sum_{i} y_i log(p_i) \tag{3}$$

where L is the loss function, y_i is the desired output and p_i is the probability of the i^{th} class. The main objective of the training phase is to minimize the loss function by readjusting the weights. The weights of CNN architecture are adjusted using a technique called backpropagation, where the loss function back-propagates through the network, and in each layer, the gradient is computed and weights in the filters are updated. The gradient descent is one of the efficient and simplest techniques that are used for updating the weights in the filters [29]. The gradient descent is an optimization algorithm often used to determine gradients in each layer and pass it to the previous layer. After the entire data set is shown to the network, the weights are updated to enhance the performance of the CNN model. Gradient descent can vary in terms of the number of examples used to estimate error; The main three types of gradient descent are:

1. The stochastic gradient descent (SGD): SGD is a gradient descent algorithm that is used for faster convergence of the loss function and updates weights after every sample is shown to the network.
2. Batch gradient descent (BGD): BGD is a gradient descent algorithm that computes the error for each sample in the training dataset, but weights are only updated after all training samples have been tested. One cycle through the entire training sample is known as a training epoch. Hence, the weights are updated at the end of each training epoch for BGD.
3. Mini-batch gradient descent (MBGD); it is a gradient descent algorithm that combines the efficiency of BGD and the robustness of SGD by splitting the training samples into small batches that are used to compute the error and update the weights. For example, if we assume the dataset has 500 images, weights are re-adjusted after a batch of 50 images are shown to the network. Appling one mini-batch through the forward pass and backpropagation is defined as an iteration, while an epoch is defined when all mini-batches (all images) are trained.

In the field of the convolutional neural network, MBGD is the most common implementation of gradient descent methods [30, ?]. This paper uses the mini-batch mode of gradient descent algorithm. By taking the gradient of the loss function (∇L) with respect to the weights, we obtain the update equation.

$$\nabla L = p_i \left(y_i + \sum_{k \neq 1} y_k \right) - y_i \tag{4}$$

where y is one-hot encoded vector for the labels, so $\sum_k y_k = 1$ and $y_i + \sum_{k \neq 1} y_k = 1$. Therefore,

$$\nabla L = p_i - y_i \tag{5}$$

The weights os the network are updates as shown below,

$$W_{ij} = W_{ij} - \alpha \nabla L \tag{6}$$

where, W_{ij} are the filter weights and α is the learning rate.

Table 3 Parameters of convolutional neural network

Layer	**Activation Shape**	**Activation Size**	**#Parameter**
Input	(128,128,1)	16,384	0
Convl_1 (f=5, s=2, p=2)	(64,64,50)	819,200	1,300
Max Pool_1 (s=2)	(64,64,50)	24,200	0
Convl_2 (f=5, s=2, p=2)	(22,22,50)		1300
Max Pool_2 (s=2)	(11,11,50)	6,050	0
FC1	(25,1)	25	151,250
FC2	(3,1)	3	75

4 Results and Discussion

This section presents the experimental results for validation of two convolutional (NN) models, Alicona model and Ultrasonic model. The parameters of our custom CNN model are illustrated in table 3. The model complexities depend on the number of the parameters, where the computation complexity of a CNN model increases as the number of the parameters increases, [31]

4.0.1 The learning curve evaluation

The performance of the Alicona model was computed by training the network on the training database and using the testing database for testing the model performance. Figure 11 shows the learning curve for 10 epochs, where the accuracy of the Alicona model started with approximately 30%. The learning curve reached stability at approximately the 9^{th} epoch. Hence, the process

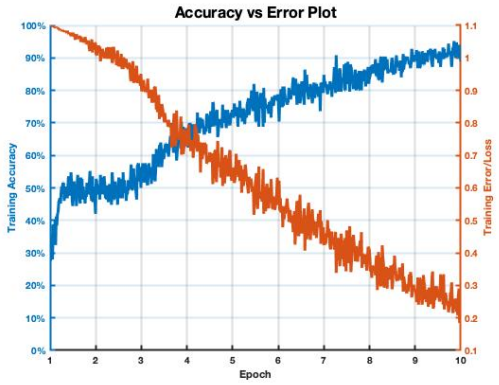


Fig. 11 Learning curve of Alicona model for 10 epochs

was terminated at the 10th epoch due to the stability in the performance. Accuracy of the Alicona model is almost 90%.

Figure 12 illustrates the learning curve of the Ultrasonic model for five epochs, which exhibits a significant improvement in the performance and gives more accurate prediction and classification.. The accuracy profile of the Ultrasonic model starts at approximately 45% and thereafter reaches the saturation point at approximately 3^{rd} epoch and becomes stable onwards, and that is why the process was terminated at the fifth epoch. The steady=stste accuracy of the Ultrasonic model is approximately 98%.

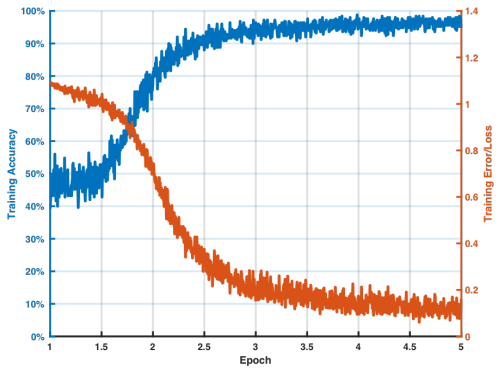


Fig. 12 Learning curve of Ultrasonic model for 5 epochs

4.0.2 Confusion matrices evaluation:

Figures 13 and 14 present the confusion matrices for both training and testing models of Alicona and ultrasonic databases, respectively. the rows of each

of the four matrices correspond to the output class or the predicted class and the columns correspond to the ground truth labels. The diagonal cells correspond to observations that are correctly classified. The off-diagonal cells correspond to incorrectly classified observations. The right-most column of the matrix shows the percentages of all the examples predicted to belong to each class that is correctly and incorrectly classified. These metrics are called the precision and false discovery rate, respectively. The last row of each matrix represents the percentages of all examples belonging to each class, which are correctly (top) and incorrectly bottom) classified. The cell in the bottom right of the plot shows the overall accuracy/ inaccuracy. For example, referring to Figure 14(a), the overall accuracy of Ultrasonic model in training is 97.2%; and referring to Figure 14(b), the overall accuracy of Ultrasonic model in testing is 97.6%.

4.1 Performance comparison:

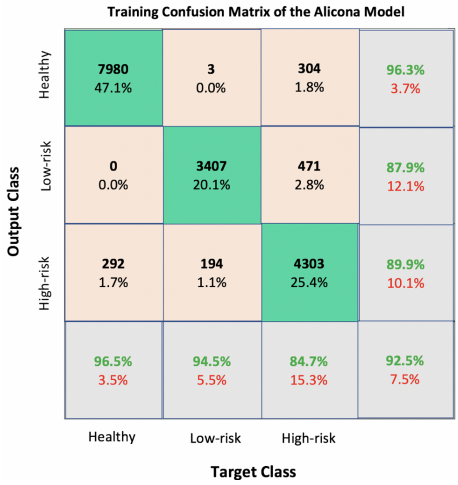
Following Table 4, a comparison of the Alicona model and Ultrasonic model shows that the Ultrasonic model predicts and classifies fatigue cracks at an earlier stage of the model learning process, i.e., at the 3^{rd} epoch while it is at the 9^{th} epoch for the Alicona model.During the training phase, the validation phase, and the testing phase of the Alicona model, the model classification accuracy reaches 89.38 % 91.34% & 92.54, respectively. On the other hand, the Ultrasonic model provides better classification accuracy, 97.20 % for the training model, 96.91 % for the validation model, & 97.56 % for the testing model. In addition, the performance of the Ultrasonic model is better: $\sim 5\%$ for training, $\sim 6.1\%$ for validation, and $\sim 9.1\%$ for testing, as illustrated in Figure 15. The rationale for the improved performance of the CNN model are as follows.

1. Data size: It is well known that having a large dataset is crucial for good performance. The augmented database in Alicona Model is 23,205, while it is 80,171 in the Ultrasonic Model.
2. Complexity of the image: As shown in Figures 2, Alicon image is more complex than the the ultrasonic echo signal . Each pixel of the image is referred to a color depth (0-255), and the color depth of Alicona image is significantly high as compared to ultrasonic signal which is of only two colors.

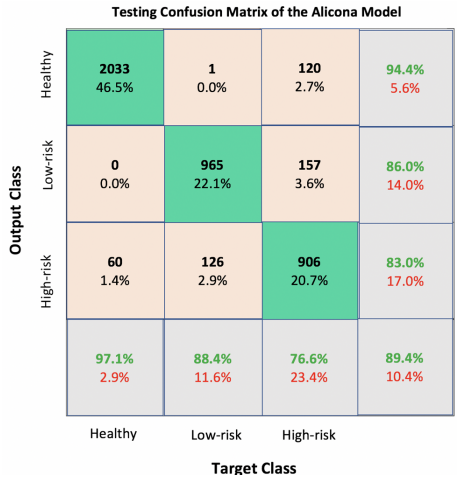
Therefore, features learning in Ultrasonic Model is expected to be more effective than that in the Alicona Model.

5 Summary, Conclusions, and Future Work

This paper has proposed an experimentally validated autonomous technique for detection and classification of fatigue damage in machinery components



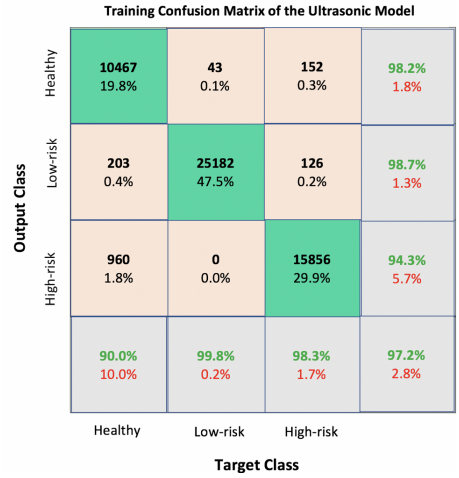
(a) Training of Alicona data for a cracked specimen



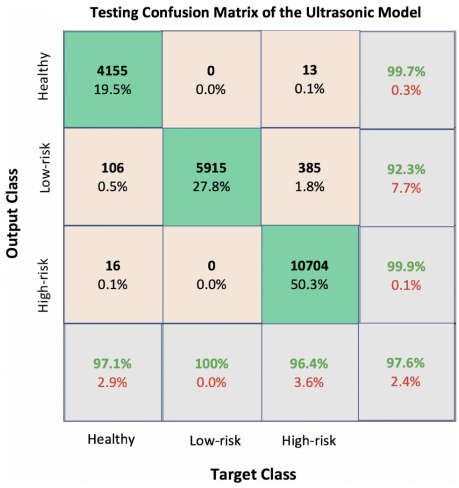
(b) Testing of Alicona model for a cracked specimen.

Fig. 13 Confusion matrices for classification on Alicona model in training and testing phases of CNN architecture, which consists of 2 convolutional layers with corresponding ReLU activation layers, 2 max pooling layers, 2 fully-connected layers and a Softmax output layer.

that are made of ductile materials (e.g., polycrystalline alloys). The main goal here is to create a robust network-based nondestructive testing (NDT) system that provides enhanced performance for damage detection and classification without using feature extraction techniques. Two models for fatigue damage detection and classification are built upon the concept of convolutional neural networks (CNNs). The first model is called the Alicona (confocal microscope) model and the second model is called the Ultrasonic model. It is noted that



(a) Training of Ultrasonic model for a crack-free specimen.



(b)Testing of Ultrasonic model for a cracked specimen.

Fig. 14 Confusion matrices for classification on Ultrasoic data in training and testing phases of CNN architecture, which consists of 2 convolutional layers with corresponding ReLU activation layers, 2 max pooling layers, 2 fully-connected layers and a Softmax output layer.

confocal microscopes are usually available in the laboratory environment only, while ultrasonic probes are available in both field environments (e.g., manufacturing sites) and laboratory sites. However, the Alicona model in the laboratory environment provides a proof of concept of the damage evolution process, while the Ultrasonic model indirectly derives the results from experimental data.

Table 4 The performance accuracy of training database, validation database, and testing database

Network Parameter	Alicona	Ultrasonic
No. of epochs	10	5
Training Accuracy	92.54 %	97.20 %
Validation Accuracy	91.34 %	96.91 %
Testing Accuracy	89.38 %	97.56 %

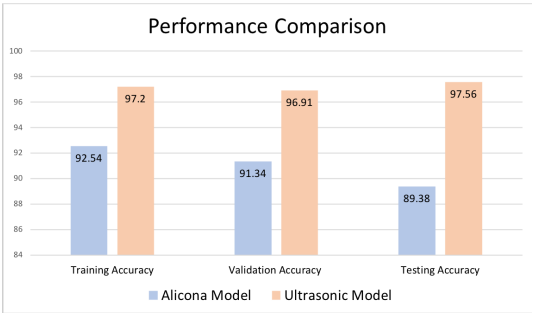


Fig. 15 Performance Comparison of two models in training, validation, and testing

Both Alicona and Ultrasonic data bases have been classified based on the damage status in the side-notch surface of test specimens. The following three damage states are defined to describe the damage state.

- *Free-crack state* that belongs to the non-risky class;
- *Small-crack state* that belongs to the low-risk class;
- *Large-crack state* that belongs to the high-risk class.

In this paper, the Ultrasonic database was synchronized with the Alicona database. Both databases were augmented by rotation, transition, scaling, and reflection. The performance of CNN models, the Alicona model Ultrasonic model, was then evaluated on the augmented database of each model. Results show that the Ultrasonic model performed better than the Alicona model in classification, where the Ultrasonic model gave 97.6 % for the testing database, while the Alicona model gave 89.38 %. The learning curves show that the Ultrasonic model starts to perform well at the 3^{rd} epoch, while the Alicona model was delayed to reach the best performance by six epochs. This delay indicates that the image features of UT are much simpler than the image features of Alicona. Hence, variation in image patterns for each class in the Ultrasonic database can be detected clearly. In general, these findings suggest that image features have a significant effect on detection and classification.

While there are many areas of both theoretical and experimental research that should be undertaken before its commercial application, the following topics are suggested for future research:

1. Developing the risk assessment classification using CNN, such that the risk assessment has multi classes that represent the severity of the damage.

2. Modify the CNN model by selecting the best kernels for features selection and having the optimal CNN structure that predicts and classifies the fatigue damage with high accuracy and in a short time.
3. Building a more realistic CNN model that involves other factors (e.g., environmental effect, structure vibration).

Acknowledgments

The work reported in this paper has been supported in part by U.S. Air Force Office of Scientific Research (AFOSR) under Grant No. FA9550-15-1-0400, and by the U.S. Army Research Office under Grant No. W911NF-20-1-0226. Any opinions, findings, and conclusions in this paper are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

Conflicts of interest

We have no conflicts of interest to disclose.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author, Dr. Hassan Alqahtani, upon reasonable request.

References

1. F. Honarvar, A. Varvani-Farahani, A review of ultrasonic testing applications in additive manufacturing: Defect evaluation, material characterization, and process control, Ultrasonics (2020) 106227.
2. S. Gholizadeh, A review of non-destructive testing methods of composite materials, Procedia Structural Integrity 1 (2016) 50–57.
3. C.-h. Chan, G. K. Pang, Fabric defect detection by fourier analysis, IEEE transactions on Industry Applications 36 (5) (2000) 1267–1276.
4. C. G. Drury, J. Watson, Good practices in visual inspection, Human factors in aviation maintenance-phase nine, progress report, FAA/Human Factors in Aviation Maintenance.@ URL: http://hfskyway. faa. gov (2002).
5. T. M. Nunes, V. H. C. De Albuquerque, J. P. Papa, C. C. Silva, P. G. Normando, E. P. Moura, J. M. R. Tavares, Automatic microstructural characterization and classification using artificial intelligence techniques on ultrasound signals, Expert systems with applications 40 (8) (2013) 3096–3105.
6. F. Margrave, K. Rigas, D. A. Bradley, P. Barrowcliffe, The use of neural networks in ultrasonic flaw detection, Measurement 25 (2) (1999) 143–154.
7. S.-W. Liu, J. H. Huang, J.-C. Sung, C. Lee, Detection of cracks using neural networks and computational mechanics, Computer methods in applied mechanics and engineering 191 (25-26) (2002) 2831–2845.
8. S. Sambath, P. Nagaraj, N. Selvakumar, Automatic defect classification in ultrasonic ndt using artificial intelligence, Journal of nondestructive evaluation 30 (1) (2011) 20–28.

9. S.-J. Song, H.-J. Kim, H. Cho, Development of an intelligent system for ultrasonic flaw classification in weldments, Nuclear Engineering and Design 212 (1-3) (2002) 307–320.
10. R. Drai, M. Khelil, A. Benchaala, Time frequency and wavelet transform applied to selected problems in ultrasonics nde, NDT & e International 35 (8) (2002) 567–572.
11. S. Seyedtabaii, Performance evaluation of neural network based pulse-echo weld defect classifiers, Measurement Science Review 12 (5) (2012).
12. F. P. B. Cruz, G. Johann, K. C. de Oliveira, F. Palú, E. A. da Silva, R. Guirardello, N. C. Pereira, Crambe grain drying: Evaluation of a linear and double resistance driving force model and energetic performance, Renewable and Sustainable Energy Reviews 80 (2017) 1–8.
13. Z. Chen, K. Gryllias, W. Li, Mechanical fault diagnosis using convolutional neural networks and extreme learning machine, Mechanical systems and signal processing 133 (2019) 106272.
14. N. Munir, H.-J. Kim, J. Park, S.-J. Song, S.-S. Kang, Convolutional neural network for ultrasonic weldment flaw classification in noisy conditions, Ultrasonics 94 (2019) 74–81.
15. H. Alqahtani, S. Bharadwaj, A. Ray, Classification of fatigue crack damage in polycrystalline alloy structures using convolutional neural networks, Engineering Failure Analysis 119 (2021) 104908.
16. E. Keller, A. Ray, Real-time health monitoring of mechanical structures, Structural Health Monitoring 2 (3) (2003) 191–203.
17. H. Alqahtani, A. Ray, Neural network-based automated assessment of fatigue damage in mechanical structures, Machines 8 (4) (2020) 85.
18. M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (2021) 243–297.
19. R. Feng, D. Grana, N. Balling, Uncertainty quantification in fault detection using convolutional neural networks, Geophysics 86 (3) (2021) M41–M48.
20. C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of Big Data 6 (1) (2019) 1–48.
21. X. Cui, V. Goel, B. Kingsbury, Data augmentation for deep neural network acoustic modeling, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (9) (2015) 1469–1477.
22. H. Nishizaki, Data augmentation and feature extraction using variational autoencoder for acoustic modeling, in: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2017, pp. 1222–1227.
23. Q. H. Nguyen, H.-B. Ly, L. S. Ho, N. Al-Ansari, H. V. Le, V. Q. Tran, I. Prakash, B. T. Pham, Influence of data splitting on performance of machine learning models in prediction of shear strength of soil, Mathematical Problems in Engineering 2021 (2021).
24. O. Moselhi, T. Hegazy, P. Fazio, Neural networks as tools in construction, Journal of construction engineering and management 117 (4) (1991) 606–625.
25. M. Sarıgül, B. M. Ozyildirim, M. Avci, Differential convolutional neural network, Neural Networks 116 (2019) 279–287.
26. D. Scherer, A. Müller, S. Behnke, Evaluation of pooling operations in convolutional architectures for object recognition, in: International conference on artificial neural networks, Springer, 2010, pp. 92–101.
27. M. Jogin, M. Madhulika, G. Divya, R. Meghana, S. Apoorva, et al., Feature extraction using convolution neural networks (cnn) and deep learning, in: 2018 3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT), IEEE, 2018, pp. 2319–2323.
28. M. Martinez, R. Stiefelhagen, Taming the cross entropy loss, in: German Conference on Pattern Recognition, Springer, 2018, pp. 628–637.
29. Y. A. LeCun, L. Bottou, G. B. Orr, K.-R. Müller, Efficient backprop, in: Neural networks: Tricks of the trade, Springer, 2012, pp. 9–48.
30. P. Gou, J. Yu, A nonlinear ann equalizer with mini-batch gradient descent in 40gbaud pam-8 im/dd system, Optical Fiber Technology 46 (2018) 113–117.
31. P. Maji, R. Mullins, On the reduction of computational complexity of deep convolutional neural networks, Entropy 20 (4) (2018) 305.