Deep learning-based identification of characteristic regions for picosecond ultrasonics metrology

Jing Min, Xiuguo Chen, Zhongyu Wang, Jing Hu, Yong Sun, Zirong Tang, Shiyuan Liu

School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China

ARTICLE INFO

Keywords:
Picosecond ultrasonics
Characteristic region identification
Deep learning
Signal segmentation

ABSTRACT

Picosecond ultrasonics (PU) is an ideal metrology technique for opaque nanostructures in the manufacture of integrated circuits (IC). The identification algorithm for the characteristic signals has a significant impact on the measurement precision and throughput. In this work, deep learning is applied to the identification of characteristic regions in PU signals, and the method is evaluated on single-layer nanofilm specimens, where data augmentation is achieved by adding four types of typical noises separately to experimental signals. Three deep learning models with different architectures are developed, trained, and compared to do the identification. Among the three models, the hybrid model shows superior overall performance in terms of the noise impact and identification speed. The analyzing results also show that the deep learning-based method is feasible to be applied to PU metrology for IC fabrication, especially the higher precision under noise compared to traditional methods, with the acceptable time cost.

1. Introduction

Picosecond ultrasonics (PU) [1–3] is a non-destructive characterization method utilizing high-frequency ultrasound excited by an ultrashort laser pulse typically ≤ 1 ps in duration [4–6]. As such a laser pulse incident on an opaque solid medium, part of the laser energy is absorbed within the optical penetration depth (typically a few tens of nanometers in metals), thereafter a local stress is set up by several possible mechanisms and launches the ultrasound [5,7]. Since the ultrasound is usually at a frequency up to hundreds of GHz, the PU method achieves a nanoscale resolution [5,8]. Meanwhile, due to the strong acoustic attenuation at such a high frequency, the propagation depth of the ultrasound is commonly in the submicron range [3,9]. Well suited for the characterization of opaque nanofilms, the PU method has become one of the critical metrology tools in integrated circuits (IC) fabrication [8,10–12].

In PU measurement, the process of ultrasound generation and propagation, in the order of picoseconds, is measured through the pump–probe technique to obtain a time-domain curve containing characteristic regions. According to the features of the characteristic region signals, such as the position, shape, amplitude, etc., the properties such as thickness [13–15], adhesion quality [16–19], and density [20,21] of the measured specimen can be analyzed. As a result, the identification of the characteristic regions is an essential procedure for PU signal analysis. Based on PU theories, some fundamental features of the characteristic signals can be learned, so it is not difficult to conduct artificial identification. However, for industrial applications, automatic signal processing is necessary, which requires an algorithm to identify the characteristic regions quickly and accurately. Embodying all the fundamental features we have learned into the identification algorithm is still a challenge. Due to the insufficient utilization of the fundamental features, conventional identification algorithms suffer from some shortcomings. Concretely, the Findpeaks-based algorithms merely utilize the peak feature of the characteristic signals, one of the spatial features, then easily interfered by the peaks of noise to present poor robustness. Another class of traditional algorithms relies on the correlation between the temporal positions of the characteristic signals and the known information about the specimen, therefore tends to present low accuracy due to the inevitable errors in the specimen information. With the trend towards more complex device structures and increasing demands on measurement precision in the semiconductor industry [22,23], conventional algorithms will exhibit more limitations.

For target identification algorithms, machine learning has offered a new horizon. Through lots of data, machine learning models can learn some regularities about the target, which are originally learned by humans through theories or experiences [24], such as the fundamental
features of PU characteristic signals mentioned above. Thus, there is no need to consider how to embody the regularities into the algorithm when exploiting machine learning to do the target identification. To implement this strategy in ultrasonic techniques, Saniie et al. [25] trained a neural network with the signals processed by split-spectrum processing, enabling it to locate the target echo in the time-domain ultrasonic measurement signal mixed with clutter. Additionally, Wang et al. [26] proposed to use the frames separated from the measurement signal to train the neural network to distinguish between the target echo and non-target regions. In these studies, the raw signals need to undergo some pre-processings before being fed into the neural network. These pre-processings are actually the feature extractions for the signals and are critical to the performance of machine learning algorithms.

Recently, deep learning models have realized end-to-end signal processing due to their powerful capabilities of feature extraction, and have been widely applied in the tasks related to target identification, such as object detection [27] and semantic segmentation [28]. In the field of photo-induced ultrasonics, there have been some relevant studies [29–31], but focusing on the processing of two-dimensional ultrasound images. Compared to these studies, the applications of deep learning in the processing of electrocardiogram (ECG) signals [32–38], which are one-dimensional time-domain signals containing multiple characteristic regions, are more enlightening for the identification algorithm for the characteristic regions in PU signals. It is noted that deep learning models based on the convolutional neural network (CNN) [32–34], the recurrent neural network (RNN) [35], and the hybrid of these two [36–38] have been trained to be able to identify the characteristic regions in ECG signals by category. Nevertheless, the diversity of specimens, the impacts of noise, and the demands on time consumption in IC fabrication make it challenging to migrate the deep learning-based method to PU metrology.

Focusing on the automatic identification of the characteristic regions in PU signals, this work aims to develop a deep learning-based method with trained models to do the identification. Three deep learning models, respectively based on CNN, RNN, and CNN-RNN, are designed and adopted here for comparison study. PU experiments and data augmentation are carried out to prepare the dataset for the training, validation, and test of the models. Then, a reasonable training strategy is made to train the three models. The feasibility of the deep learning-based method is finally judged from the performances of the trained models. By analyzing the identification result under noise and the time consumption, the better model for PU metrology can be selected.

2. Picosecond ultrasonics experiment

The PU experimental setup built in this work, based on conventional pump–probe technique, is shown in Fig. 1(a). The light source was a 920 nm femtosecond fiber laser generating ~ 100 fs optical pulses with a repetition rate of 80 MHz. The laser output was split into two parts: pump beam (~90% of total output) for exciting ultrasound in specimens and probe beam (~10% of total output) for detecting changes in the surface properties of specimens. To monitor the specimen properties at different moments after excitation, an optical delay line was set on the path of the probe beam to control the time delay between the pump and probe pulses before reaching the specimen. After being remerged by a beam splitter, the pump and probe beams were focused onto the specimen surface within spots of ~ 20 μm in diameter through an objective lens. The pump energy incident on the specimen is ~ 1.5 nJ per pulse. The two beams reflected by the specimen became parallel again through a collimating lens. The reflected probe pulse carrying the specimen

Fig. 1. (a) PU experimental setup. (b) Denoised measured signals of 200-nm AlCu, W, and Mo specimens, respectively.
information corresponding to a specific moment entered the photodetector (an amplifier-equipped photodiode) with a power of ~ 3 mW, and the reflected pump beam was blocked by a dump.

In addition to the basic structure described above, cross-polarization method and lock-in amplification scheme were applied to the experimental setup to ensure a relatively high signal-to-noise ratio (SNR). Passing through a \( \lambda/2 \) wave plate, the pump was in an orthogonal polarization state to the probe, so the scattered pump light could be removed from the reflected probe beam by placing a polarizer before the photodetector. An acousto-optic modulator was placed on the pump path to modulate the pump beam at a frequency of 2 MHz, and the lock-in amplifier used the modulation frequency as a reference to extract and amplify the 2 MHz components in the photodetector output.

During the measurement, as the motorized translation stage of the optical delay line moves, transient intensity change \( \Delta I \) of the reflected probe pulse measured by the lock-in amplifier varies with the time delay. Using a computer to sequentially record \( \Delta I \) under each delay time and dividing it by \( I \) (intensity of the reflected probe pulse in the absence of pump) for normalization, the curve of \( \Delta I / I \) versus the delay time was got. In fact, the value \( \Delta I / I \) is equal to the normalized transient reflectance change \( \Delta R / R \). It can be taken as, the transient reflectance change \( \Delta R \) is measured, and also the reference \( R \). The reflectivity of the specimen corresponds to the \( \Delta R / R \), depending on the wavelength \( \lambda \) and the time \( t \). Even if the lock-in detection is used to reduce noise, the noise at 2 MHz still exists in the signal. And the SNR of the raw signal curve is \( \sim 35 \) dB, with the SNR defined by the formula:

\[
SNR = 10\log_{10} \left( \frac{P_{\text{signal}}}{P_{\text{noise}}} \right)
\]

where \( P_{\text{signal}} \) and \( P_{\text{noise}} \) stand for the power of the pure signal and noise, respectively. To further improve the SNR, a combination algorithm of coherent averaging and wavelet filtering was used to reduce the noise in the raw curve.

PU experiments were performed on single-layer metal nanofilms to obtain the measurement signals for study. These specimens included thin films of AlCu, W, and Mo, with nominal thicknesses of 100 nm, 200 nm, 500 nm, and 1000 nm, prepared on silicon substrates covered with a 20 nm AlN buffer layer by magnetron sputtering. At the laser wavelength of 920 nm, all the specimens have sufficient absorption and conversion of the light to generate relatively strong PU signals. Fig. 1(b) displays the measured signal curves after noise reduction, respectively from the 200-nm specimens of AlCu, W, and Mo. The mainbody of these PU signals consists of three kinds of regions, the zero, the echoes, and the thermal decay. The moment PU excitation happens, the value \( \Delta R / R \) changes drastically and this part of the measured curve is the zero. The ultrasound excited at the specimen surface partially reflects at the interfaces, where acoustic impedance mismatch exists, during its longitudinal propagation. As the ultrasound returns to the specimen surface, \( \Delta R / R \) fluctuates slightly, corresponding to the echoes in the curve. The reciprocating travel of the ultrasound inside the film results in successive echoes. After the zero region, the curve outside the echoes is the thermal decay. Here the study of characteristic region identification is for the PU measurement of the specimen properties which are characterized by ultrasounds, such as thickness, adhesion quality, and density, rather than thermal properties. Therefore, in this paper, the zero and the echoes are the characteristic signals, which need to be identified for specimen properties measurement, while the thermal decay belongs to the non-characteristic region.

As can be seen in Fig. 1(b), there are significant differences among specimens of different materials in terms of the shape and amplitude of

Fig. 2. Architecture of the 1D U-Net model.
the characteristic signals. On the other hand, the characteristic signals from different specimens share some fundamental features. For example, the zero always appears as a large spike with a very steep left side and a relatively gentle right side, and the echo always appears as a small and short oscillation. The successive echoes arising from the same film layer appear at a fixed time interval $\Delta t$ (also the interval between the zero and the first echo), which is determined by the layer thickness $d$ and the longitudinal sound velocity $v_s$ by the following formula:

$$\Delta t = \frac{2 \times d}{v_s} \quad (2)$$

Accordingly, the conventional identification algorithms for the characteristic signals in PU measurements mostly follow two lines of thought, one is to find peaks to locate the characteristic signals, another is to calculate the $\Delta t$ by Eq. (2) with the known nominal thickness to determine the characteristic regions. However, the former is highly susceptible to noise, especially when the characteristic signal is weak (e.g., the echoes in the measured curve of the 200-nm Mo specimen in Fig. 1(b)). With regard to the latter, the nominal thickness deduced from the manufacturing process of the film usually deviates from the true thickness, leading to inaccurate identification of the characteristic regions. These shortcomings of the conventional algorithms are essentially attributed to the very limited utilization of the fundamental features of the characteristic signals. Thus we propose to use the deep learning-based identification algorithm, which can adequately grasp and utilize the fundamental features to achieve accurate identification.

3. Deep learning models for characteristic region identification

The PU measurement signal is essentially a time series, and can also be seen as a row of pixels (i.e. a one-dimensional image) with each data of the series treated as the pixel brightness. Therefore, either the RNN, commonly used to process time series, or the CNN, commonly used to process images, can be applied to the characteristic region identification for PU signals. Considering the industrial application, in addition to the nature of the PU signal, the impact of noise and the identification speed should be taken into account when selecting the deep learning model. Due to the different ways of feature extraction, the RNN and CNN are expected to differ in identification performance, especially the identification effect under noise and the identification speed. Thus, three different models, including a CNN-based model, an RNN-based model, and a hybrid model based on CNN and RNN, were adopted in this work for the performance comparison, and then the best characteristic region identification model for PU metrology could be concluded.

The three models, essentially belonging to segmentation models, can segment the PU signal into several regions by category to achieve characteristic region identification. Here the region categories include zero, echo, and background, of which the background stands for the non-characteristic region. The input to the models is the one-dimensional array transformed from the PU signal, here its length is set to 2400. With the number of region categories being 3, the output of the models is required to be of size $2400 \times 3$. The three values in the second dimension of the model output represent the predicted probabilities that the corresponding-position element in the input belongs to the three categories respectively. The category with the highest probability is the model prediction result for the element.

3.1. CNN-based model

The CNN-based model we used is a 1D U-Net model. For PU metrology, a larger dataset means a larger consumption of time and cost. The U-Net, a typical semantic segmentation model, performs well on small datasets of the order of hundreds or even tens [39], which is friendly to PU metrology. As the model input is one-dimensional, all the convolution, pooling, and upsampling layers in the U-Net model need to be 1D mode. After being fed into the model, the one-dimensional input is processed as a single feature map.

As shown in Fig. 2, both the contracting and the expanding path of the 1D U-Net model consist of three blocks. Each block in the contracting path contains two convolutions and one max pooling. After the operations of the contracting block, the number of feature maps is enlarged and the length is shortened. As the feature maps enter the expanding block, they first undergo an upsampling and a convolution to become the same size (length $\times$ number) as the feature maps at the symmetric position in the contracting path. Then these two groups of feature maps are concatenated into one group to undergo another two convolutions,
thus one expansion is completed. At the end of the model, the feature maps from the last expanding block are convolved to be the model output of size $2400 \times 3$.

Among these operations included in the 1D U-Net model, convolution is the only one that involves trainable parameters, and its mathematical expression can be described as:

$$y_{ij} = w_i \cdot x_j + b_i$$  \hspace{1cm} (3)

where $w_i$ and $b_i$ represent the weight matrix and bias of the i-th kernel at this convolution layer, respectively; $x_j$ stands for the j-th local region in the input of this convolution layer, and $y_{ij}$ is the j-th element in the output corresponding to the i-th kernel.

Since the width of the characteristic signal is not very small compared to the length of the entire PU signal, relatively larger convolutional kernels were used to extract graphic features effectively. Specifically, a kernel size of 27 was chosen for most of the convolution layers in the model, instead of the usual size of 3 \cite{34,39}. Enlarging the kernel size leads to an increase in the number of model parameters, which may cause overfitting. Therefore, some dropout layers were added to the model to avoid overfitting.

### 3.2. RNN-based model

As mentioned previously, each element in the input array needs to get the corresponding prediction result, which means that when using an RNN-based model to do this task, only one element can be fed into the RNN layer at each time step. Thus, the $2400 \times 1$ input, which needs to take 2400 time steps to be processed, is a quite long time series to the model. To avoid the gradient vanishing and gradient exploding problems that occur in ordinary RNNs when processing long time series \cite{40}, long short-term memory (LSTM) cells were chosen rather than RNN cells when constructing the model. The operational architecture of the LSTM cell is shown in the local enlargement of Fig. 3, and the corresponding mathematical expression is as follows:

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$  \hspace{1cm} (4)

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$  \hspace{1cm} (5)

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$  \hspace{1cm} (6)

$$c_t = c_{t-1} f_t + i_t \tanh(w_c[h_{t-1}, x_t] + b_c)$$  \hspace{1cm} (7)

$$h_t = o_t \tanh(c_t)$$  \hspace{1cm} (8)

where $x_t$ is the input of the LSTM cell at time step $t$; $f_t, i_t, o_t$, and $c_t$ are the values of the forget gate, input gate, output gate, and cell state in the LSTM cell at time step $t$, respectively; $w$ and $b$ are the weight matrix and bias of the corresponding position, respectively; $h_t$ is not only the hidden state value but also the output of the LSTM cell at time step $t$.

For the accurate identification of the intervening characteristic signal (e.g., the first echo located between the zero and the second echo), comprehensive utilization of the preceding and the following information in the time series is usually required. Accordingly, the bidirectional LSTM structure, which involves two sets of LSTM cells that process the input time series in the forward and backward directions respectively, was used here.

---

**Fig. 4.** Architecture of the CNN-BLSTM model.

---

**Fig. 5.** Schematic diagram of the data augmentation method. (a) Four different types of artificially designed noise. (b) Measured signal of a 500-nm AlCu specimen. (c) Deformed signals obtained by adding the noise to the original measured signal.
The final architecture of the model based on the bidirectional LSTM structure, i.e. the bidirectional LSTM model, is shown in Fig. 3. There are two bidirectional LSTM layers in the model, and each layer contains 512 LSTM cells with 256 LSTM cells in each direction. The architecture and these hyperparameters were chosen to ensure that the model could sufficiently understand and express the information contained in the long sequence of 2400 time steps. The bidirectional LSTM layer produces an output at each time step, and all the outputs form a new time series to be fed into the next layer. At the end of the model, a time-distributed dense layer transforms the output of the second bidirectional LSTM layer at each time step into a three-classified result, i.e. a $1 \times 3$ array, and then the model output of size $2400 \times 3$ is get. In order to

![Fig. 6. Training processes of three deep learning models: (a) accuracy and (b) CWIOU versus epoch time.](image)

![Fig. 7. Output results of three models for the measured signal of a 100-nm AlCu specimen.](image)

<table>
<thead>
<tr>
<th>Segmentation model</th>
<th>Training set</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>CWIOU (%)</td>
<td>Accuracy (%)</td>
</tr>
<tr>
<td>1D U-Net</td>
<td>98.67</td>
<td>94.92</td>
<td>98.53</td>
</tr>
<tr>
<td>Bidirectional LSTM</td>
<td>98.95</td>
<td>95.93</td>
<td>98.48</td>
</tr>
<tr>
<td>CNN-BLSTM</td>
<td>99.32</td>
<td>97.30</td>
<td>99.05</td>
</tr>
</tbody>
</table>
4.1. Data augmentation

PU signals.

abundance of the dataset, which is critical for the generalization described in Section 2, so data augmentation is required to ensure the model with a certain ability to identify the characteristic regions in the trainable parameters in the model are determined and equip the strategy need to be developed for the model training. After training, all dataset containing abundant data samples, and a reasonable training performance of the trained model. In this work, data augmentation was implemented by adding noise to the original measured signals, and the basic principle is shown in Fig. 5. Adopting this method of data augmentation is beneficial to the robustness of the trained model in real measurement environments. According to the actual application scenarios, four types of noise as shown in Fig. 5(a) were designed. The objects simulated by these types of noise and the properties of the noise

type1: Lower-amplitude white noise. This simulation corresponds to the random noise induced by electronic components such as the photodetector. And the SNR was set to be \( \sim 45 \) dB, similar to the noise level of the coherent averaged signal with 2 \( \sim 3 \) averages (acquisition time 2 \( \sim 3 \) s).

type2: Higher-amplitude white noise. Same purpose as type 1, but with a lower SNR of \( \sim 35 \) dB, approximately the noise level of the raw signal without noise reduction.

type3: Low-frequency sine noise. This kind of noise is designed to simulate the case that the output of the light source is unstable. Here we set the sine's period to be \( \sim 5 \) times the echo width and its amplitude to be close to the echo amplitude.

type4: Impulse noise. Abrupt airflow turbulence or shaking of the optical component holder may cause a sudden intensity fluctuation in the pump or probe, eventually leading to such a type of noise. The impulse was designed to be close in size (width and amplitude) to the first echo and was positioned between the zero and the first echo.

Adding the four types of noise separately to an original signal (e.g., the curve shown in Fig. 5(b)), then four derived signals (e.g., the curves shown in Fig. 5(c)) were acquired. All the original signals and the derived signals were transformed into one-dimensional arrays by ignoring the abscissa information, and then resized to the specified length of 2400. For each element in the array, we assigned an individual label to it according to the region it was in, and the rule was: all the elements in the zero region (e.g., the green area in Fig. 5(b)) were labeled as 'zero', those in the echo region (e.g., the red areas in Fig. 5(b)) were labeled as 'echo', and the rest were all labeled as 'background'. An array and a complete set of labels corresponding to it, together form a data sample of the deep learning dataset.

4.2. Critical functions for model training

When developing the training strategy, two functions need to be especially focused on, namely the loss function and the evaluation metrics function. During the training process, all the trainable parameters in the model are updated by backpropagation algorithm. As the loss, a measure of how far off the model output is from the actual result, is the base for the backpropagation algorithm, the choice of loss function will affect the updating direction of the model parameters. While the evaluation metrics function can help us to select the best group of parameters for the model performance.

Since the characteristic region identification in this paper is achieved by semantic segmentation, the model training requires pixel-level loss calculation, which is usually realized by the categorical cross-entropy loss function. Moreover, the PU signal suffers from category imbalance, i.e. the data points in the zero and echo regions are much fewer than those in the background region. To prevent the trained model from being biased towards the background category, a weighted categorical cross-entropy loss function was adopted here, and is given by.

![Fig. 8. CWIOUs of three models on different datasets.](image-url)
Loss = \sum_{i=1}^{M} \sum_{j=1}^{N} w_j \times y_{ij} \times \log(\hat{y}_{ij}) \tag{9}

w_j = \frac{N}{n_j} \times \frac{1}{\sum_{k=1}^{C} n_k} \tag{10}

where \(M\) is the length of the model input, \(N\) is the number of region categories, \(w\) stands for the specific weight for the corresponding category, \(y\) represents the true probability, \(\hat{y}\) represents the predicted probability, and \(n\) represents the number of elements belonging to the corresponding category in the input.

The intersection over union (IOU) is often chosen as the evaluation metric for segmentation models. Considering that the background region is not the thing of interest in this work, we proposed the characteristics-only weighted IOU (CWIOU) to be the evaluation metric. The function is presented as follows:

\[ CWIOU = \sum_{i=1}^{C} w'_i \times \frac{|T_i \cap P_i|}{|T_i \cup P_i|} \tag{11} \]

\[ w'_i = \frac{T_i}{\sum_{j=1}^{C} T_j} \tag{12} \]

where \(C\) refers to the number of characteristic region categories, \(w'\) stands for the specific weight for the corresponding category, \(T\) represents the collection of the elements truly belonging to the corresponding category in the input, and \(P\) represents the collection of the predicted results matching the corresponding category.

4.3. Implementation of model training

The dataset acquired by the method stated in Section 4.1 consisted of 420 samples and was divided into three sets, which were used for the training, validation, and test of the models, respectively. 276 samples

Fig. 9. Echo identification results of three models for (a) the original signal, (b) the type1 signal, and (c) the type2 signal of a 1000-nm AlCu specimen.
were assigned to the training set and the rest were divided equally between the validation and test sets. What’s more, the three sets were designed to have the same distribution of samples. The establishment and training of the three models described in Section 3 were implemented in Python 3.8, using TensorFlow 2.4. All the models were trained under the same strategy. The Adam optimizer [41] was used to carry out the backpropagation algorithm, and the initial learning rate was set to 0.001. In addition to the CWIOU, accuracy was also observed as the evaluation metric. After 500 epochs, the model training was terminated, and the group of model parameters with the highest CWIOU was set to 0.001. In addition to the CWIOU, accuracy was also observed as the evaluation metric. After 500 epochs, the model training was terminated, and the group of model parameters with the highest CWIOU on the validation set was finally kept. From the training processes shown in Fig. 6, it can be found that all the three models had reached stability in both accuracy and CWIOU within 500 epochs. Comparing the evaluation metric curves on the training and validation sets, no serious overfitting is found to occur in any of the three models. Although the accuracy and CWIOU curves look similar overall, there may be differences in local trends. Taking the accuracy and CWIOU values of the bidirectional LSTM model on the validation set as examples, in the first 10 epochs, the maximum points of the two curves appear at different epochs. This reveals that not the higher the accuracy, the higher the CWIOU. Since the proportion of the background is large in the PU signals, when the model performs well on the background category and poorly on the zero and echo categories, there will be a relatively high accuracy and a low CWIOU. Therefore, it’s not sensible to use accuracy as the only evaluation metric in this work.

5. Results and discussion

After getting the trained models, we first assessed the feasibility of the proposed deep learning-based method according to the general identification effects of the trained models. Then, the impacts of different types of noise on identification and the identification speeds of the models were investigated for assessing and comparing the overall performances of the three trained models.

5.1. Feasibility of the deep learning-based identification

The general identification effect of the trained model contains two aspects: the performance on dataset and the concrete identification results for individual test signals. The performances of the trained models on the dataset are shown in Table 1. It can be seen that on the training, validation, and test sets, the CNN-BLSTM model always shows the highest accuracy and CWIOU among the three models. And even the 1D U-Net model, which performs relatively poorly, achieves an accuracy of 98.00% and a CWIOU of 92.88% on the test set. Another point worth noting is that the difference in CWIOU between models is more pronounced than that in accuracy, which thereby indicates that CWIOU is a better metric than accuracy for comparing the performances of different models. Fig. 7 presents the output results of the trained models for an original measured signal from the test set, concretely, the measured signal of a 100-nm AlCu specimen by the PU experiment described in Section 2. The output result includes three probability curves, which are generated from the three 2400 × 1 arrays of the 2400 × 3 model output, corresponding to the zero, echo, and background categories, respectively. As can be seen in Fig. 7, all the three models exhibit good identification for this measured signal, specifically in the following aspects. First, all the critical features (e.g., extreme points and turning points) of the characteristic signals are included in the predicted characteristic regions. Second, no obvious misidentified regions appear. Last, no unidentifiable regions appear. As for the differences in evaluation metrics among the models, it results from the performance differences in determining the boundaries of regions.

5.2. Impacts of different types of noise on identification

According to Section 4.1, the dataset consists of five kinds of signals, namely the original signals without artificially designed noise, and the type1 ~ type4 signals containing the type1 ~ type4 noise, respectively. By analyzing the identification results for the five kinds of signals in the test set, the impacts of different types of noise on identification can be investigated. It should be pointed out that the validation set, which didn’t participate in the update of the trainable parameters during the model training, was also used in this investigation to expand the samples and thus reduce the influence of odd samples on the investigation result. The validation and test sets were merged and then divided into five datasets according to the five signal categories mentioned above. The performances of the three models on the five datasets are shown in Fig. 8.

By comparing the results for different datasets in Fig. 8, some preliminary information about the impact of noise can be obtained. The noise in both the type1 and type2 datasets is white noise, and the comparison of the results for the original, type1, and type2 datasets shows that the higher the noise amplitude, the greater the impact on identification. The differences between the results for the type3 and original datasets indicate that the low-frequency sine noise has a strong impact on the bidirectional LSTM model while a tiny impact on the 1D U-Net model. Comparing the results for the type4 and original datasets, it can be seen that only the 1D U-Net model is significantly affected by the impulse noise. Besides, it should be noted that on four datasets except for the type3 dataset, the CNN-BLSTM model owns the highest CWIOU among the three models, and even on the type3 dataset, it
achieves a CWIOU of 91.53%.

In the following, the specific impacts of noise are analyzed with the model prediction results for individual signals. Due to the highly distinct features of the zero signal, for the three models, the identification of the zero region is hardly affected by noise. Therefore, only the echo identification results are plotted here for analysis.

Fig. 9 displays the echo identification results for three signals from the original, type1, and type2 datasets, respectively. For the longitudinal contrast, the three signals from the same specimen were purposely selected. By observing Fig. 9 laterally, it can be found that on the three signals, the 1D U-Net model always performs worst and the CNN-BLSTM model performs best, which is consistent with the regularity presented on datasets. By observing Fig. 9 longitudinally, we can see that as the white noise amplitude increases, the identification effect of only the 1D U-Net model gets worse significantly. And a detailed look reveals that the deterioration of the effect is mainly embodied in the identification of the second echo region.

The shape of the small-amplitude echo, such as the second echo in Fig. 9, is easily masked by white noise. In this case, it is necessary to utilize the periodicity of the echoes to improve the identification effect. The periodicity is closer to the temporal feature, for which the RNN has a stronger extraction capability than the CNN. As a result, when the characteristic signal is weak, the 1D U-Net model, a CNN-based model, is more affected by white noise, compared to the other two models.

The echo identification results for an individual type3 signal, shown in Fig. 10, present the same regularity as the CWIOUs on the type3 dataset shown in Fig. 8. The low-frequency sine noise interferes with the temporal feature extraction of the echoes due to its own periodicity, and hardly affects the echo shape. In this case, the echo shape, closer to the spatial feature, needs to be utilized for accurate identification. In terms of spatial feature extraction, the CNN is far superior to the RNN. Consequently, under the type3 noise, the 1D U-Net model is able to maintain good identification, while the bidirectional LSTM model, an RNN-based model, cannot. As for the CNN-BLSTM model, although the

Fig. 11. Echo identification results of three models for the type4 signals of a 1000-nm W specimen. The impulse noises were set at three different positions: (a) at a distance from the echo; (b) adjacent to the echo; (c) overlapping partially with the echo.
Fig. 11 (a) and (b), we can find that when the impulse noise is at a distance from the echo or adjacent to the echo, all the three models identify echoes well. But in the case of the impulse noise overlapping partially with the echo, as shown in Fig. 11(c), the 1D U-Net model incorrectly identifies the overlap as the echo region, while the other two models do not. In order to explore whether such a misidentification is related to the echo shape, the echo identification results of the 1D U-Net model for the type4 signals, with the above overlap phenomenon, of the specimens of different materials are presented together in Fig. 12 for observation.

From Fig. 12, it can be seen that regardless of the echo shape, such a misidentification of the 1D U-Net model always occurs when the impulse noise overlaps partially with the echo. This misidentification results in a non-coincidence of the maximum points of the predicted and the true echo regions, as shown in Fig. 12. In film thickness measurements based on PU, after the characteristic region identification, the Findpeaks algorithm is used to determine the precise positions of the characteristic signals to obtain the Δt in Eq. (2). Then the film thickness d can be calculated with the Δt and the known sound velocity v_s. Accordingly, the ~ 20 ps deviation between the two maximum points, shown in the insets of Fig. 12, will lead to a film thickness measurement error of ~ 50 nm, which is unacceptable in engineering applications.

For the three models, there are significant differences in terms of the impacts of noise on identification. The 1D U-Net model is mainly affected by the high-amplitude white noise and the impulse noise overlapping partially with the echo. The bidirectional LSTM model performs worse mainly when there is low-frequency noise in the signal. As for the CNN-BLSTM model, it is always able to identify echoes well under these types of noise designed in this work, and in most cases, it owns the highest evaluation metrics among the three models. Overall, of the three models, the CNN-BLSTM model has the best identification effect under noise.

The identification results under noise also reveal the benefit of the deep learning-based identification method over traditional methods: better robustness to noise, thus ensuring the precision of measurement. For traditional methods, to avoid the disturbance of noise peaks, sufficient noise reduction must be performed in advance. However, the noise with similar frequency to the characteristic signal, such as the type4 noise (impulse noise similar to echo), is hard to be removed. While, even under this type of noise, some deep learning models, such as the CNN-BLSTM model, can identify the characteristic signals accurately.

5.3. Inference time

PU is commonly applied in in-line measurement scenarios where high speed is required. Therefore, when selecting the characteristic region identification algorithm for PU metrology, the time cost of the algorithm should be considered. For the deep learning-based algorithm proposed in this paper, the time cost is mainly determined by the inference time of the model, i.e. the time it takes for the model to compute the input into the corresponding output. In practical applications of PU metrology, the identification algorithm usually needs to process only one signal at one time, rather than a batch of signals. Thus, the inference time for a single signal is what we care about here, and the comparison among the results of the three models is shown in Fig. 13. The inference calculations of the models were carried out on an Intel Core 2.1 GHz (i7-12700F) CPU.

According to Fig. 13, the 1D U-Net model owns the shortest inference time among the three models, due to the fact that it involves much fewer operations to perform one inference than the other two models. Since the complexity of the CNN-BLSTM model is lower than that of the bidirectional LSTM model, the inference time of the CNN-BLSTM model is slightly shorter than that of the bidirectional LSTM model. The actual time cost for single measurement is typically ~ 4 s, compared to which the inference time of 0.66 s of the CNN-BLSTM model is not that long, but still much longer than that of the 1D U-Net model. Yet, in addition to the typical structure followed by the CNN-BLSTM model, there are many other ways to combine CNN and RNN. So, for the hybrid model, it is
In this work, we investigated a deep learning-based method for the automatic identification of the characteristic regions in PU signals, realizing an application of deep learning in PU metrology. Through pump–probe PU experiments on nanofilms of AlCu, W, and Mo, the original dataset for deep learning was acquired, and data augmentation was achieved by adding noises. Based on the nature of the PU signal, three deep learning models, i.e., 1D U-Net, bidirectional LSTM, and CNN–BLSTM models, were designed, trained, and compared to do the identification studies. Comparison among the three models in terms of identification effects under four types of noise and identification speeds shows that the hybrid model based on CNN and RNN, such as the CNN–BLSTM model, has superior overall performance. The results also demonstrate that deep learning indeed provides an effective solution to the identification improvement, with better robustness to noise than traditional methods and acceptable time consumption. With the accumulation of data samples, the hybrid model is expected to be applied to PU metrology in IC fabrication with complex structures.

Funding

This work was funded by the National Natural Science Foundation of China (Grant No. 52022034, 52130504, 62175075); National Key Research and Development Plan of China (2022YFF0709104); Key Research and Development Plan of Hubei Province (2020A008).

CRediT authorship contribution statement


Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data are not publicly available at this time but may be obtained from the authors upon reasonable request.

References


