Learnable Continuous Wavelet Transform Network

Masaharu Yamamoto
System AI R&D Department, Corporate Laboratory,
Toshiba Corporation
Kawasaki, Japan
masaharu2.yamamoto@toshiba.co.jp

Shigeru Maya
System AI R&D Department, Corporate Laboratory,
Toshiba Corporation
Kawasaki, Japan
shigeru1.maya@toshiba.co.jp

Abstract

Wavelet transform is widely used to analyze non-stationary data whose frequencies change exponentially over time. In recent years, there have been attempts to use the continuous wavelet coefficients obtained by the continuous wavelet transform as features for machine learning. Mother wavelets are usually chosen from limited collections of functions. To obtain mother wavelets suitable for a specific machine learning task, we propose a novel method to flexibly construct a mother wavelet with a compact support in the Fourier domain using a neural network. We call the model Learnable Continuous Wavelet transform Network (LCWnet). We show that LCWnet significantly improves the performance of classification tasks compared with existing mother wavelets in three datasets.

CCS Concepts

• **Computing methodologies** → Neural networks.

Keywords

Time series, Wavelet, Neural networks

ACM Reference Format:
Masaharu Yamamoto and Shigeru Maya. 2025. Learnable Continuous Wavelet Transform Network. In *Proceedings of The 11th Mining and Learning from Time Series Workshop: From Classical Methods to LLMs (SIGKDD MILETS Workshop '25).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Time-series data is widely observed in various domains, such as industry [22], speech recognition [7], medicine [26], wind speed prediction [16], and geology [11]. With the increase in data obtained by attaching sensors to various devices, there has been an emphasis on the importance of incorporating signal processing techniques, which have been extensively researched from both theoretical and practical perspectives, into machine learning. These techniques are essential for extracting meaningful patterns from raw data and enhancing model interpretability and performance.

Among the signal processing techniques, wavelet transform is particularly useful for non-stationary data that change exponentially in frequency, such as acoustic data and vibration data. Wavelet transform provides time-frequency representation of the data convolving a function called mother wavelet while changing its scale. Wavelet transform can be either discrete or continuous [6]. Discrete transform is an efficient computation for obtaining wavelet coefficients and is used for information compression and feature extraction. Continuous transform gives wavelet coefficients for all scales. The wavelet coefficient, which is obtained by wavelet transform, depends on the mother wavelet.

It has been reported that using continuous wavelet coefficients to extract features of machine learning can improve the performance of some tasks [14]. For instance, when diagnosing arrhythmia by analyzing electrocardiogram heartbeats, it is effective to use *continuous wavelet transform* (CWT) to extract the time-frequency characteristics of the electrocardiogram [27]. The *continuous wavelet coefficients* (CWC), which is obtained by CWT, can be treated as a two-dimensional feature and classified using a *convolutional neural network* (CNN). On the other hand, the discrete wavelet transform offers high time resolution for high frequencies and low time resolution for low frequencies. Therefore, unlike the continuous wavelet transform, the discrete wavelet transform cannot be computed at a constant resolution and thus cannot be directly applied to calculations requiring two-dimensional feature analysis.

Mother wavelets must be selected from functions that satisfy the so-called admissible condition. One criterion for choosing it is that the wavelet coefficients should be as sparse as possible to enhance information compression and the interpretability of visualization [20]. Qualitatively, a sparse representation can be achieved by using a mother wavelet that matches the shape of the time-series data [23]. In the quantitative method, the mother wavelet is selected based on the entropy [5] or the minimum description length of the wavelet coefficients [9]. However, there are expected to be more appropriate functions for mother wavelets when using wavelet coefficients as features for machine learning because mother wavelets are selected from a limited collection of functions in these methods.

Learning the mother wavelet from data is a promising method to improve performance. However, this poses a significant challenge because the mother wavelet must satisfy the admissible condition. If the CWT is performed with a function that fails to meet this condition, the resulting transform may not accurately represent the time-frequency characteristics of the data. Additionally, the range of scales to be calculated could become excessively large, leading to practical difficulties. These issues highlight the need for a learnable and theoretically sound approach that ensures both computational feasibility and signal fidelity.

To address these challenges, we propose a Learnable Continuous Wavelet Transform using a neural network, named LCWnet. LCWnet is a mother wavelet whose CWC is suitable for a specific

SIGKDD MILETS Workshop '25, Toronto, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

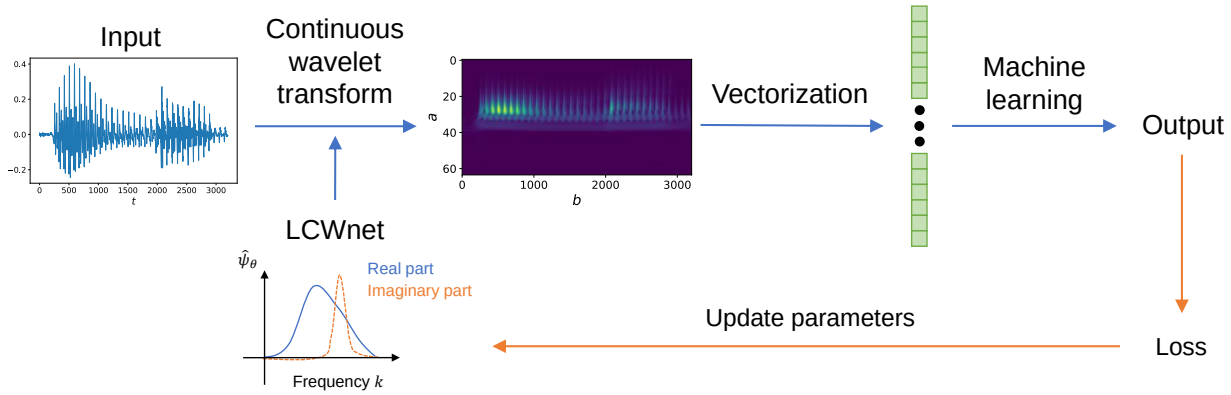


Figure 1: Overview of mother wavelets trained using LCWnet. LCWnet consists of real and imaginary parts, each of them has a compact support in the Fourier domain. LCWNet transforms input data to CWC. After vectorized, CWC is input to a machine learning model as a feature. The parameters of LCWnet are updated by the output of the model. Blue arrows and orange arrows represent inference and learning processes, respectively.

task, and trained from the output of the subsequent machine learning model (Figure 1). It has a compact support in the Fourier domain in order to stable learning and ensure that it satisfy the admissible condition.

Our contributions are as follows:

- We propose a novel framework where the mother wavelet can learn from data.
- It is proven that LCWnet satisfies the admissibility condition, thereby establishing itself as a mother wavelet.
- LCWnet achieves the highest performance in binary classification experiments compared with existing mother wavelets.

2 Related works

In this section, we describe existing researches from three perspectives. The first is learning discrete wavelet transform from the viewpoint of finite-length filter. The second is research on optimizing the parameters of conventional mother wavelets. The third is the relationship between operator learning and CWT when regarded as an operator.

2.1 Discrete wavelet transform.

In discrete wavelet transform, mother wavelets are represented as filters, and the time-series data can be decomposed at scales that are powers of 2. *Multi resolution analysis* (MRA) [18] is a time-efficient method for performing discrete wavelet transform. In recent years, numerous efforts have been made to utilize the discrete wavelet coefficients learned from the data for machine learning tasks. For instance, each level of the MRA filters are constructed as a one-dimensional CNN, and learn from the data to obtain a sparse representation [19]. In this method, the filter might not necessarily satisfy the construction condition of the mother wavelet because there is no restriction on the filter. Furthermore, because this method represents the mother wavelet as a filter, it cannot be directly extended to the continuous wavelet transform.

2.2 Existing continuous mother wavelets.

There have been attempts to train existing mother wavelets by adjusting their parameters. The parameters of Shannon wavelet filter has been learned for speech recognition [24]. The center frequency and width parameters of the Morlet wavelet have been learned to classify gravitational waves [25]. However, the mother wavelets learned by these methods are limited to existing ones, and they are not necessarily suitable for machine learning tasks.

2.3 Operator learning.

Learning CWT differs from the discrete transform, which involves learning a finite number of filters. CWT should be treated as an operator that maps one-dimensional time-series function to a two-dimensional image. Neural networks have been employed to learn operators for a considerable period [4]. Various effective models have been proposed based on DeepONet [17], which was proposed as a method for finding solutions to partial differential equations. In particular, Fourier neural operator (FNO) [15] is a method that transforms functions by repeatedly using Fourier layers that transform the input functions by convolutions. Learning the mother wavelet for continuous wavelet transform is similar to FNO because both involve convolutions. However, since FNO learns the integral kernel without imposing any restrictions, it cannot consistently learn functions that satisfy the admissible condition. Therefore, learning the CWT requires a new method in which the convolution kernel function always satisfies the admissibility condition.

3 Methodology

In this section, we describe the preliminary of wavelet transform to formulate the definition of mother wavelets and numerical calculation to implement CWT. Following an examination of the necessary conditions for the mother wavelet, we propose LCWNet based on these considerations.

3.1 Preliminary

The mathematical formulation of *continuous wavelet transform* (CWT) is presented as follows [6]. CWT provides a time-frequency representation of a function by convolving a mother wavelet, at different scales. Complex wavelet transform, where the mother wavelet takes complex values, is used to extract the phase information of the time-series data. To perform complex wavelet transform, complex mother wavelets ψ are chosen from the following subspace $\mathcal{W}$ of $\mathbb{C}$-valued square-integrable functions,

$$\mathcal{W} = \left\{ \psi \in L_2(\mathbb{R}) \middle| \int_{-\infty}^{\infty} dk \frac{|\mathcal{F}[\psi](k)|^2}{|k|} < \infty \right\}, \tag{1}$$

where $\mathcal{F}$ represents the Fourier transform,

$$\hat{\psi}(k) = \mathcal{F}[\psi](k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx \psi(x) e^{-\sqrt{-1}kx}. \tag{2}$$

We also use $\hat{\cdot}$ to represent functions in the Fourier domain. Mother wavelets must satisfy the condition in Equation (1), which is called the admissible condition. Functions to be wavelet transformed are assumed to be square integrable. For a function $f \in L_2(\mathbb{R})$, *continuous wavelet coefficients* (CWC) using a complex mother wavelet ψ is obtained by

$$T_{\psi}^{\mathrm{wav}} f(a,b) = |a|^{-1/2} \int_{-\infty}^{\infty} dx f(x) \psi\left(\frac{x-b}{a}\right)^*, \tag{3}$$

where the parameters a and b represent the scale and the translation of the mother wavelet, respectively, and the asterisk $*$ denotes the complex conjugate.

Equation (3) can also be expressed using the Fourier transform $\mathcal{F}$ and its inverse transform $\mathcal{F}^{-1}$ as follows,

$$T_{\psi}^{\mathrm{wav}} f(a,b) = \sqrt{a}\mathcal{F}^{-1}[\mathcal{F}[f](\cdot)\mathcal{F}[\psi]^*(a\cdot)](b), \tag{4}$$

where $\cdot$ represents the argument of the function. The inverse Fourier transform in Equation (4) is obviously written as follows,

$$\begin{aligned} &\mathcal{F}^{-1}[\mathcal{F}[f](\cdot)\mathcal{F}[\psi]^*(a\cdot)](b) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk \mathcal{F}[f](k)\mathcal{F}[\psi]^*(ak) e^{\sqrt{-1}kb}. \end{aligned} \tag{5}$$

It is well-established that the wavelet transform is invertible. Specifically, when f is a real-valued function and the support of the mother wavelet in the Fourier domain $\mathcal{F}[\psi]$ is $[0,\infty)$, the inverse transform weakly converges as follows,

$$f = 2C_{\psi}^{-1} \int_0^{\infty} \frac{da}{a^2} \int_{-\infty}^{\infty} db \mathrm{Re}\left[T_{\psi}^{\mathrm{wav}} f(a,b) \psi^{a,b}\right], \tag{6}$$

where

$$C_{\psi} = 2\pi \int_0^{\infty} dk |k|^{-1} |\hat{\psi}(k)|^2, \tag{7}$$

$$\psi^{a,b}(x) = |a|^{-1/2} \psi\left(\frac{x-b}{a}\right). \tag{8}$$

This expression can be interpreted that $T_{\psi}^{\mathrm{wav}} f$ has the all information of the original function f at scales $a > 0$.

3.2 Problem setting

The numerical computation of the CWC can be computed by discretizing Equation (4). For time series data $\boldsymbol{f} = (f_0, \ldots, f_{L-1}) \in \mathbb{R}^L$ of length L obtained at a sampling frequency q_{freq} and a complex mother wavelet ψ and discretized S scales $\boldsymbol{a} = (a_0, \ldots, a_{S-1}) \in \mathbb{R}^S$, the complex continuous wavelet transform $T_{\psi}^{\mathrm{wav}} : \mathbb{R}^L \to \mathbb{C}^{S \times L}$ is computed by the following calculation for $s = 0, \ldots, S-1$,

$$T_{\psi}^{\mathrm{wav}} \boldsymbol{f}_{(s,l)} = \sqrt{a_s} \mathcal{F}_{\mathrm{disc}}^{-1}\left[\mathcal{F}_{\mathrm{disc}}[\boldsymbol{f}] \odot \Psi_{a_s}\right]_l, \tag{9}$$

where $\odot$ represents the point-wise product of vectors, F_{disc} and F_{disc}^{-1} are discrete Fourier transform and its inverse transform defined as

$$\mathcal{F}_{\mathrm{disc}}[\boldsymbol{f}]_k = \sum_{l=0}^{L-1} f_l e^{\frac{-2\sqrt{-1}\pi kl}{L}}, \tag{10}$$

$$\mathcal{F}_{\mathrm{disc}}^{-1}[\boldsymbol{f}]_l = \frac{1}{L} \sum_{k=0}^{L-1} f_k e^{\frac{2\sqrt{-1}\pi kl}{L}}, \tag{11}$$

and Ψ_{a_s} is a discretized form of $\mathcal{F}[\psi]^*(a_s \cdot)$ defined as

$$(\Psi_{a_s})_l = \mathcal{F}[\psi]^*(a_s q_l) \text{ for } l = 0, \ldots, L-1, \tag{12}$$

where

$$\boldsymbol{q} = \begin{cases} \frac{q_{\mathrm{freq}}}{L}(0, 1, \ldots, \frac{L-1}{2}, -\frac{L-1}{2}, \ldots, -1) & \text{if } L \text{ is odd,} \\ \frac{q_{\mathrm{freq}}}{L}(0, 1, \ldots, \frac{L}{2}-1, -\frac{L}{2}, \ldots, -1) & \text{otherwise,} \end{cases} \tag{13}$$

is a vector representing frequencies obtained by discrete Fourier transform.

CWC is numerically computed in Equation (9). This calculation has the advantage of computational efficiency compared to original Equation(3) because it does not require integration, and can be calculated simultaneously for all time l.

While the discrete wavelet transform is generally computed such that the scales are halved, the continuous wavelet transform allows for specifying scales at any resolution. Consequently, $\mathcal{F}[\psi]^*$ must be computed at any scale, requiring the mother wavelet to be defined not only at a finite points but at any point.

In this paper, we address the classification of time series data. Throughout the task, the CWC is treated as a feature input for machine learning classification. We consider a framework to obtain the mother wavelet that maximizes performance.

3.3 Constructing a learnable mother wavelet

We now describe the configuration of LCWnet. For a function to qualify as a mother wavelet, it must satisfy the admissible condition, which can be interpreted as the following three necessary conditions:

(1) In the integral of the admissible condition, it must converge to zero at the origin faster than the order $O(k^{1/2})$ to avoid divergence.
(2) To prevent divergence at infinity, it must decay at infinity with an order of $O(k^{-\alpha})$ for any $\alpha > 0$.
(3) It must avoid divergence over the integration interval.

Based on this consideration, we introduce the LCWnet to construct the mother wavelet which have a compact support in the Fourier domain as illustrated in Figure 2. First, we prepare a *multilayer perceptron* MLP_{θ} parameterized by θ with one-dimensional

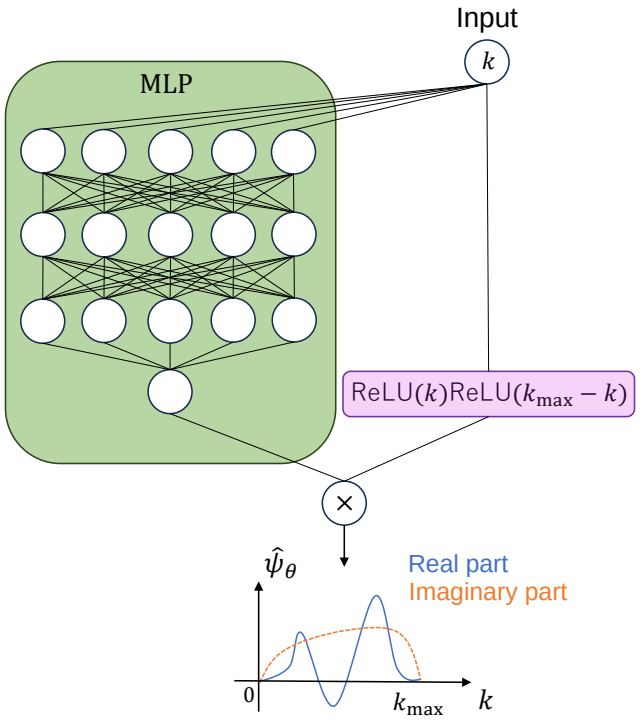


Figure 2: The architecture of LCWnet.

input and output layers. The activation function of the MLP is linear for the final layer, and nonlinear continuous functions over $\mathbb{R}$ such as *Rectified Linear Unit* (ReLU) or logistic sigmoid function for the other layers. This MLP is assumed to have finite layers and finite units for each layer in order to be a continuous function. Next, we construct the real part $\hat{\psi}^{(r)}_{\theta_r}$ and the imaginary part $\hat{\psi}^{(i)}_{\theta_i}$ of the complex mother wavelet in the Fourier domain as follows,

$$\hat{\psi}^{(r)}_{\theta_r}(k) = \mathrm{MLP}_{\theta_r}(k)\mathrm{ReLU}(k)\mathrm{ReLU}(k_{\max} - k), \tag{14}$$

$$\hat{\psi}^{(i)}_{\theta_i}(k) = \mathrm{MLP}_{\theta_i}(k)\mathrm{ReLU}(k)\mathrm{ReLU}(k_{\max} - k), \tag{15}$$

where $k_{\max}$ is a hyperparameter that limits the domain of the mother wavelet and, θ_r and θ_i are the parameters of real and imaginary mother wavelets. ReLU represents

$$\mathrm{ReLU}(k) = \begin{cases} k & \text{if } k \geq 0, \\ 0 & \text{if } k < 0. \end{cases} \tag{16}$$

Using these, we construct the complex mother wavelet defined in the Fourier domain as follows,

$$\hat{\psi}_\theta(k) = \hat{\psi}^{(r)}_{\theta_r}(k) + \sqrt{-1}\hat{\psi}^{(i)}_{\theta_i}(k). \tag{17}$$

Then, the constructed function satisfies the admissible condition.

Proposition 3.1. *$\hat{\psi}_\theta$ in Equation (17) is a mother wavelet. In other words, this function satisfies the admissible condition.*

Proof. We prove that the output of the function satisfies the admissible condition. The integral in Equation (1) can be decomposed of its real part and imaginary part as follows,

$$\begin{aligned} &\int_{-\infty}^{\infty} dk \frac{|\hat{\psi}_\theta(k)|^2}{|k|} \\ &= \int_{-\infty}^{\infty} dk \frac{|\hat{\psi}^{(r)}_{\theta_r}(k)|^2}{|k|} + \int_{-\infty}^{\infty} dk \frac{|\hat{\psi}^{(i)}_{\theta_i}(k)|^2}{|k|}. \end{aligned} \tag{18}$$

It is enough to show that only the real part is bounded because the real part and the imaginary part have the same structure.

$$\begin{aligned} &\int_{-\infty}^{\infty} dk \frac{|\hat{\psi}^{(r)}_{\theta_r}(k)|^2}{|k|} \\ &= \int_0^{k_{\max}} dk \left|\mathrm{MLP}_{\theta_r}(k)\right|^2 k(k_{\max} - k)^2, \\ &< k_{\max}^3 \int_0^{k_{\max}} dk \left|\mathrm{MLP}_{\theta_r}(k)\right|^2. \end{aligned} \tag{19}$$

The MLPs are assumed to be obtained by repeating affine transforms and nonlinear continuous transforms a finite number of times. Since a function obtained by composition of a finite number of continuous functions is also continuous, $|(\mathrm{MLP}_{\theta_r})|$ is continuous and takes the maximum value M in $[0, k_{\max}]$ by the maximum value principle. Therefore, the left hand side of Equation (19) is bounded above by $k_{\max}^4 M^2$. Hence, the function satisfies the admissible condition. □

To calculate the CWC using LCWnet, simply replace $\mathcal{F}[\psi]^*$ in Equation (12) with LCWnet and use Equation (9). LCWnet takes values only in the positive frequency, inverse wavelet transform can be conducted in Equation (6). This means the positive scales of CWC transformed by LCWnet contain all information of the data. Therefore, we consider only the positive scales.

Having a compact support not only satisfies the admissible condition, but also stabilizes the learning. We calculate the CWC for the prepared scales during the learning process. As the shape of the function changes during learning, the activation position of the CWC shifts in the scale direction This might cause the CWC to move outside the scales at which we calculate. However, if mother wavelets have a compact support, this shift can be prevented because the range where CWC can shift is limited.

3.4 Learning algorithm

Here, we provide a detailed description of the LCWnet learning framework illustrated in Figure 1. We consider K class prediction model $C : \mathbb{C}^{S\times L} \to \mathbb{R}^K$ to output the probability that the time-series data belong to each class from its CWC. To train both the mother wavelet and the machine learning model simultaneously by gradient descent method, the model is assumed to be differentiable.

Let $\mathcal{D} = \{(\boldsymbol{f}_i, \boldsymbol{l}_i)\}_i$ be the N pairs of time-series data $\boldsymbol{f}_i \in \mathbb{R}^L$ and its one-hot label $\boldsymbol{l}_i \in \{0,1\}^K$ for $i = 1, \ldots, N$. One-hot label represents the element corresponding to the class of the data takes 1 and others take 0. Then the loss function $\mathcal{L}$ is described as follows,

$$\mathcal{L} = \mathbb{E}_{(\boldsymbol{f},\boldsymbol{l})\sim\mathcal{D}}[\mathrm{CrossEntropy}(C(T^{\mathrm{wav}}_\psi \boldsymbol{f}), \boldsymbol{l})], \tag{20}$$

where $\mathbb{E}_{(\boldsymbol{f},\boldsymbol{l})\sim\mathcal{D}}$ represents data average and CrossEntropy is defined as

$$\mathrm{CrossEntropy}(\boldsymbol{p}, \boldsymbol{l}) = \sum_{k=1}^{K} l_k \log p_k. \tag{21}$$

In this paper, we employ a simple linear separator as the subsequent machine learning model to examine the differences in the impact of CWC features. Specifically, the linear separator C is implemented using linear regression and softmax. This selection is made to ensure that the prediction performance is not predominantly influenced by the complexity of the subsequent model, thereby highlighting the contribution of the input CWC.

The following transformations are applied before the linear separator. Firstly, since the linear separator accepts vectors as input, the CWC is rearranged from a second-order tensor into a one-dimensional form. Secondly, given that the CWC is a complex valued vector, it must be converted to be real values. This can be achieved by either concatenating the real and imaginary parts of the vector or by taking the absolute value. In this paper, we use the absolute value to conduct robust inference. This is because the absolute value of the CWC represents the amplitude of the time-frequency components of the data, eliminating the influence of phase differences. Finally, LCWnet does not normalize the mother wavelet so that its L2 norm becomes 1, resulting in the CWC potentially taking on extremely large or small values. This makes training the subsequent linear separator difficult. To stabilize learning and also reduce training time, we utilize layer normalization [2] to standardize the vectorized CWC. After these transformations, the CWC is input into a one-layer fully connected layer, which is treated as a linear separator.

4 Experiments

We demonstrated through numerical experiments that the proposed method can provide features more suitable for linear separation compared to existing mother wavelets. The data were selected based on the following two criteria: The first is that the time series data should exhibit frequency variations to leverage the properties of the wavelet transform. The second is that the dataset should be large enough to reflect the learning effect of the mother wavelet. Based on these criteria, we selected two speech dataset, TIMIT [12] and LIBRISPEECH [21], and machine sounds dataset MIMII [10]. For each of these datasets, we performed binary classification tasks, which can be accomplished by linear separation. To ensure a fair evaluation, we assessed performance using multiple metrics, including accuracy and the F1 score. The F1 score, which is the harmonic mean of precision and recall, is particularly useful for checking biased predictions, even with imbalanced data. Since our method aims to select an appropriate mother wavelet tailored to the data and task, we benchmarked it against existing mother wavelets that do not involve learning.

4.1 Datasets

The TIMIT corpus includes speech utterances from 630 speakers across eight regions of the United States. Each person reads ten sentences. We used gender information as labels for the binary classification task. The training and test data were given in the dataset. We created the validation data by extracting one-fifth of the entire training data.

The LIBRISPEECH dataset contains over 1000 hours of English audiobook recordings. The training, development, and test data were given in the dataset. Additionally, there are separate development and test datasets for both clean and other speech. In this paper, the development data was used for validation. We trained the models by 100 hours of clean speech training data (clean-100) and 500 hours of other speech data (other-500), each of them was evaluated by clean data and other data, respectively. We used gender information as labels for the binary classification task as well as TIMIT.

MIMII dataset includes machine sounds from four types of machines: valve, pump, slider, and fan, with labels for normal and abnormal sounds. Artificially collected noises were added to the dataset in three levels so that the signal-to-noise ratio (SNR) becomes 6dB, 0dB and -6dB. In this experiments 6dB was used because low SNR was difficult for linear separation. We used the label of normal or abnormal condition for the classification task.

All three datasets were recorded at 16kHz. Non-speech periods were removed from the TIMIT and LIBRISPEECH dataset in advance. During training, a random selection of mini-batch was taken for each epoch, and a time series length of 3200 was randomly extracted and used as input data for the binary classification task. Similarly, for the validation and test data, a time series length of 3200 was extracted and some portion of them were fixed to be used for evaluation.

4.2 Existing mother wavelets

Various mother wavelets have been proposed for continuous wavelet transform. We used complex Mexican hat wavelet, Gabor wavelet, and Shannon wavelet as benchmarks. These functions in the Fourier domain are shown in Figure 3.

The Mexican hat wavelet is introduced as the second derivative of the Gaussian function. The positive frequency part of the Mexican hat wavelet is called the complex Mexican hat function $\hat{\psi}_{\text{Mexican}}$ [1].

$$\hat{\psi}_{\text{Mexican}}(k) \propto \begin{cases} k^2 e^{-k^2/2} & \text{for } k \geq 0, \\ 0 & \text{for } k < 0. \end{cases} \tag{22}$$

Gabor wavelet is a function that exponentially decays a triangular wave [8]. It has parameters that represent the center frequency k_0 and the width w.

$$\hat{\psi}_{\text{Gabor}}(k) \propto e^{-(k-k_0)^2 w^2}. \tag{23}$$

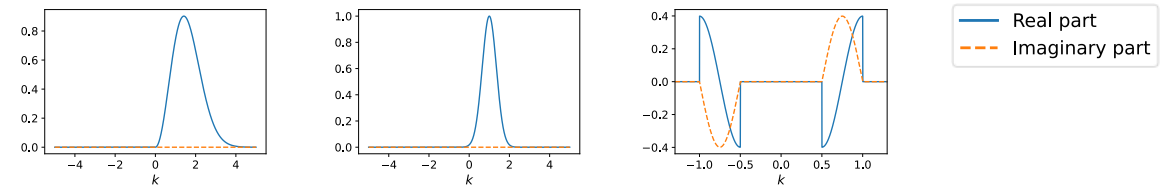
Shannon wavelet is a function that has a compact support in the Fourier domain [3]. It has a role as a band-restriction function.

$$\hat{\psi}_{\text{Shannon}}(k) \propto e^{-2\pi\sqrt{-1}k}(\Pi(2k) + \Pi(-2k)), \tag{24}$$

where Π is a gate function representing $\Pi(k) = 1$ for $1 < k < 2$, and $\Pi(k) = 0$ for others,

4.3 Experimental settings

Cross-entropy was used as the loss function. Additionally, we added the L2 norm of the coefficients in the linear model to the loss function to prevent overfitting. The regularization parameter was set to 10^{-2} for other-500 dataset of LIBRISPEECH and 10^{-4} for other datasets. Regularization was not applied to the weights of the MLP in the mother wavelet, because their weights do not directly affect prediction performance.



(a) Complex Mexican hat (b) Gabor (c) Shannon

Figure 3: Function shapes of existing mother wavelets in the Fourier domain.

Mother wavelet	Metric	TIMIT	LIBRISPEECH	
			clean-100	other-500
Mexicanhat	Accuracy	0.9598	0.8511	0.7839
	F1 score	0.9713	0.8499	0.7528
Shannon	Accuracy	0.9564	0.8579	0.8000
	F1 score	0.9689	0.8533	0.7668
Gabor	Accuracy	0.9577	0.8579	0.7861
	F1 score	0.9568	0.8553	0.7486
LCWnet	Accuracy	**0.9610**	**0.8598**	**0.8028**
	F1 score	**0.9721**	**0.8622**	**0.7741**

Table 1: Experimental results on two human voice datasets, TIMIT and LIBRISPEECH. The highest score for each dataset is highlighted in bold, and the second highest score is underlined.

Mother Wavelet	Metric	Fan	Pump	Slider	Valve
Mexicanhat	Accuracy	0.8886	0.9374	0.8545	0.8856
	F1 score	0.7776	0.6262	0.5392	0.0000
Shannon	Accuracy	0.9165	0.9488	0.8980	0.8863
	F1 score	0.8371	0.7231	0.7167	0.0285
Gabor	Accuracy	0.9290	0.9526	0.8881	0.8854
	F1 score	0.8630	0.7440	0.6810	0.0070
LCWnet	Accuracy	**0.9315**	**0.9535**	**0.9046**	**0.8871**
	F1 score	**0.8670**	**0.7497**	**0.7363**	**0.0338**

Table 2: Experimental results on MIMII for 6dB SNR for each machine. The highest score for each dataset is highlighted in bold, and the second highest score is underlined.

The number of scales S was set to 64. The minimum scale a_0 was $2^{k_{\max}L/q_{\text{freq}}}$ and the maximum scale a_{S-1} was $2^{10^{-4}L/q_{\text{freq}}}$. The intermediate scales were set to have equal intervals over power of 2.

The range of explored hyperparameters is as follows: The maximum frequency $k_{\max}$ of LCWnet is [1, 4, 16]. For the Gabor wavelet, the central frequency k_0 was examined from [2, 4, 8] with a width $w = 1$. The Shannon wavelet and Mexican hat wavelet do not have function-specific hyperparameters.

The Adam optimizer [13] was used to train all models. The MLP to construct real and imaginary parts of the mother wavelets has a one-dimensional input and output layer, three 128-dimensional intermediate layers. The number of epochs was set to 50,000, with a mini-batch size of 2,024. Model selection was performed using early stopping, meaning the performance was monitored every 500 epochs during the training and selected the model with the highest performance.

All the experiments were performed using one NVIDIA A100 40GB GPU. The real time required for training one model was approximately 80 minutes for TIMIT and 120 minutes for LIBRISPEECH and MIMII.

4.4 Results

From Table 1, the proposed method achieved the highest performance in terms of both accuracy and F1 score on human speech data from the TIMIT and LIBRISPEECH datasets. Figures 4 (a) and (b) show the mother wavelets learned from TIMIT and clean-100 of LIBRISPEECH take similar shapes. The real part forms a small peak near $k = 0$, returns to zero, and then forms a larger peak. The imaginary gradually increase away from $k = 0$. Their shapes are considered similar because this shape is useful for determining gender from clean human speech data.

Other-500 of LIBRISPEECH is difficult to learn from due to the diversity in pronunciation. Consequently, the model trained on this data exhibited lower classification performance compared to the other two datasets. The shape of mother wavelet derived from this data, as illustrated in Figure 4 (c), was different from the others. This indicates that, even for the same task, different mother wavelets may be learned depending on the data.

Table 2 presents the performance results from the experiments using the MIMII. The proposed method achieved the highest performance on this dataset as well. The valve data records acoustics during opening and closing, where timing information is crucial for anomaly detection. However, due to the loss of such timing information in the linear separation, the F1 scores for all methods were close to zero. For other machines, frequency characteristics related to abnormal sounds were important, enabling successful learning.

As shown in the Figure 5, the mother wavelets learned from the MIMII take different shapes compared to those learned from the other speech datasets. This indicates that useful mother wavelets differ depending on the task and data, such as detecting anomalies in machine sounds versus determining gender from human speech. LCWnet can learn the shapes of functions with multiple peaks, not just unimodal functions, because the mother wavelets trained from fan and pump data resemble a period of a sine wave in the imaginary part.

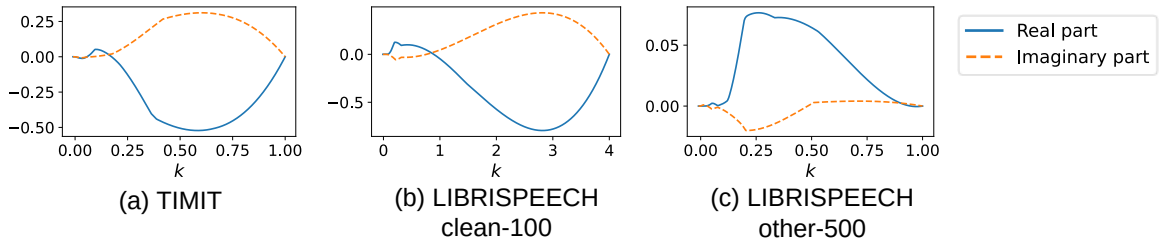


Figure 4: Mother wavelets in the Fourier domain learned from TIMIT and LIBRISPEECH. The maximum frequency was selected 1, 4, and 1 for (a), (b), and (c), respectively.

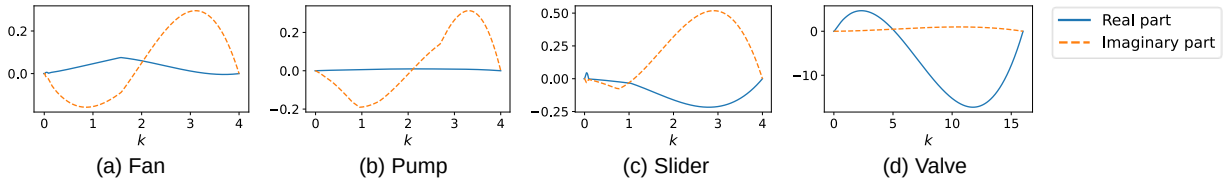


Figure 5: Mother wavelets in the Fourier domain learned from MIMII. The maximum frequency was selected 4, 4, 4, and 16 for (a), (b), (c), and (d), respectively.

Throughout all experiments, it was demonstrated that the proposed method could utilize highly performant functions without the need to select a specific mother wavelet. Additionally, it was confirmed that the learned mother wavelets tend to have similar shapes when dealing with the same type of task.

5 Discussion and conclusion

In this paper, we proposed LCWnet, a novel framework for flexibly obtaining mother wavelets suitable for a specific machine learning task. In classification experiments using human voice and machinery sound datasets, we confirmed that LCWnet provides useful CWCs compared to existing mother wavelets. LCWNet can be directly applied to tasks other than linear separation.

The limitation of the LCWnet method is that the mother wavelet is restricted to functions with compact support in the Fourier domain. However, many mother wavelets without a compact support, such as the Mexican hat or Gabor wavelets, exponentially decay. Since LCWnet can approximate exponential decay in the range of the support, the difference between exponential decay and reaching zero is negligible. Therefore, it is considered that having a compact support does not lead to performance degradation.

CWT using LCWnet can be performed regardless of the sampling rate of the time-series data because it is an operator. In future work, we consider our method can maintain its performance under various sampling rates if the subsequent model does not depend on sampling rate. Furthermore, it is possible to use a lighter neural network to construct the mother wavelets by knowledge distillation because the mother wavelet is just a one-dimensional function.

References

[1] Paul S Addison, James N Watson, and Tong Feng. 2002. Low-oscillation complex wavelets. *Journal of Sound and Vibration* 254, 4 (2002), 733–762.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. arXiv:1607.06450 [stat.ML] https://arxiv.org/abs/1607.06450

[3] Carlo Cattani. 2008. Shannon wavelets theory. *Mathematical Problems in Engineering* (2008).

[4] Tianping Chen and Hong Chen. 1995. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks* 6, 4 (1995), 911–917.

[5] Ronald R Coifman and M Victor Wickerhauser. 1992. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory* 38, 2 (1992), 713–718.

[6] Ingrid Daubechies. 1992. *Ten lectures on wavelets*. SIAM.

[7] Aditya Dutt and Paul Gader. 2023. Wavelet Multiresolution Analysis Based Speech Emotion Recognition System Using 1D CNN LSTM Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

[8] Dennis Gabor. 1946. Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-part III: radio and communication engineering* 93, 26 (1946), 429–441.

[9] Effrina Yanti Hamid, Redy Mardiana, and Zenichio Kawasaki. 2001. Wavelet-based compression of power disturbances using the minimum description length criterion. In *Proceedings of 2001 Power Engineering Society Summer Meeting*, Vol. 3. IEEE, 1772–1777.

[10] Purohit Harsh, Tanabe Ryo, Ichige Kenji, Endo Takashi, Nikaido Yuki, Suefusa Kaori, and Kawaguchi Yohei. 2019. MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection. In *Proceedings of the 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*. 209.

[11] Ali Heidari and Noorollah Majidi. 2021. Earthquake acceleration analysis using wavelet method. *Earthquake Engineering and Engineering Vibration* 20 (2021), 113–126.

[12] S. Garofolo John, F. Lamel Lori, M. Fisher William, G. Fiscus Jonathan, S. Pallett David, L. Dahlgren Nancy, and Zue Victor. 1993. DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM.

[13] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *the 3rd International Conference on Learning Representations, Conference Track Proceedings*. http://arxiv.org/abs/1412.6980

[14] Hyeon Kyu Lee and Young-Seok Choi. 2019. Application of continuous wavelet transform and convolutional neural network in decoding motor imagery brain-computer interface. *Entropy* 21, 12 (2019), 1199.

[15] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, Anima Anandkumar, et al. 2020. Fourier Neural Operator for Parametric Partial Differential Equations. In *the 8th International Conference on Learning Representations*.

[16] Hui Liu, Xiwei Mi, and Yanfei Li. 2018. Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network. *Energy Conversion and Management* 156 (2018), 498–514.

[17] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. 2021. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence* 3, 3 (2021), 218–229.

[18] Stephane G Mallat. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 7 (1989), 674–693.

[19] Gabriel Michau, Gaetan Frusque, and Olga Fink. 2022. Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. *Proceedings of the National Academy of Sciences* 119, 8 (2022), e2106598119.

[20] Wai Keng Ngui, M Salman Leong, Lim Meng Hee, and Ahmed M Abdelrhman. 2013. Wavelet analysis: mother wavelet selection methods. *Applied Mechanics and Materials* 393 (2013), 953–958.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *Proceedings of the 40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5206–5210.

[22] Aditi B Patil, Jitendra A Gaikwad, and Jayant V Kulkarni. 2016. Bearing fault diagnosis using discrete wavelet transform and artificial neural network. In *Proceedings of the 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 399–405.

[23] J Rafiee, MA Rafiee, N Prause, and Marco P Schoen. 2011. Wavelet basis functions in biomedical signal processing. *Expert systems with Applications* 38, 5 (2011), 6190–6201.

[24] Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 1021–1028.

[25] Jason Stock and Chuck Anderson. 2022. Trainable Wavelet Neural Network for Non-Stationary Signals. arXiv:2205.03355 [cs.LG] https://arxiv.org/abs/2205.03355

[26] RK Tripathy and U Rajendra Acharya. 2018. Use of features from RR-time series and EEG signals for automated classification of sleep stages in deep neural network framework. *Biocybernetics and Biomedical Engineering* 38, 4 (2018), 890–902.

[27] Tao Wang, Changhua Lu, Yining Sun, Mei Yang, Chun Liu, and Chunsheng Ou. 2021. Automatic ECG classification using continuous wavelet transform and convolutional neural network. *Entropy* 23, 1 (2021), 119.