International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 1
Article Received: *05 January 2026* ***Revised:*** *08 February 2026* ***Accepted:*** *25 March 2026*

Interpretable Wavelet Transforms: A Unified Framework for Frequency-Aware Learning and Dynamical System Identification

V.S.S.V.D.Prakash[1] and G. Sudheer[1]*

[1] Dept. of Mathematics, Gayatri Vidya Parishad College of Engineering for Women, Visakhapatnam-530048, India

**Corresponding Author email: sudhwave@gmail.com*

Abstract: Wavelet transforms provide simultaneous time–frequency localisation through a mathematically rigorous multi-resolution framework, yet their classical formulations fix filter coefficients independently of any learning objective. This paper makes three original contributions. First, we provide a unified mathematical treatment of *learnable* wavelet decomposition, establishing precise conditions under which trainable filters retain or forfeit perfect reconstruction guarantees. Second, we derive a gradient-based layer importance metric that quantifies which frequency bands drive model decisions, and demonstrate its application to physiological signal classification with reproducible experimental details. Third, we show that the multi-resolution signal decomposition principle underlying wavelets can serve as a structural prior for governing equation discovery in complex network dynamics, creating an explicit bridge between classical wavelet theory and modern neural symbolic regression. Worked examples on the ECGFiveDays benchmark and SIS epidemic dynamics illustrate the unified framework.

1. Introduction

A *time series* $\mathbf{x} = \{x_1, x_2, \dots, x_T\} \in \mathbb{R}^T$ encodes temporal phenomena ranging from neural spike trains to financial prices. Two complementary analytical philosophies exist (Wang et al., 2018): Time-domain methods treat $\mathbf{x}$ as an ordered sequence, modelling correlations among its elements directly using models such as ARIMA (Brunton et al., 2016), recurrent neural networks, and long short-term memory networks (Hochreiter & Schmidhuber, 1997). Frequency-domain methods apply a transform operator $\mathcal{T}: \mathbb{R}^T \to \mathcal{F}$ to reveal spectral structure, including the Discrete Fourier Transform (Harris, 1978) and Z-transform.

Wavelet analysis furnishes a third, *joint* representation that jointly represents time-localised frequency information (Mallat, 1989). Rather than decomposing $\mathbf{x}$ into globally defined sinusoids, wavelets are localised oscillatory functions that can be dilated and translated to probe different scales at different locations (Daubechies, 1992; Mallat, 2008). Despite this richness, classical wavelet decomposition carries a fundamental limitation for machine learning: its filter coefficients are fixed by mathematical convention (Daubechies, 1988), independent of training data. Prior work that used wavelet transforms as a preprocessing step operated independently of model training, preventing joint end-to-end optimisation (Liu et al., 2013). Recent deep learning advances have driven state-of-the-art performance across time series benchmarks (Ismail Fawaz et al., 2019; Wang et al., 2017), but effective integration of wavelet-based frequency analysis into end-to-end neural architectures remains underexplored.

The broader need for interpretable machine learning has been emphasised extensively in the explainable AI literature (Arrieta et al., 2020; Ali et al., 2023). Existing attribution methods such as SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) operate in the input space and do not reveal which frequency components drive predictions — a natural question for any signal processing application.

1.1 Original Contributions

This paper makes the following contributions that go beyond Wang et al. (2018) and Hu et al. (2025):

1. Qualified reconstruction theory for trainable wavelet networks: We establish that perfect reconstruction is guaranteed for classical fixed filter banks under strict QMF/orthogonality conditions (Proposition 2.2), but is *not* automatically retained after filters become trainable (Remark 3.1). We characterise the

Frobenius-norm deviation from the initial wavelet filter-bank prior, used as a proxy for loss of reconstruction structure, as a function of regularisation strength α.

2. Layer-wise frequency importance metric: We introduce a gradient-based importance score $I(\mathrm{a})$ (Equations 5.3–5.4) that quantifies the contribution of each wavelet frequency band to model decisions. This metric operates on intermediate multi-resolution representations rather than input pixels or tokens, constituting a novel form of spectral attribution. To the best of our knowledge, explicit layer-wise spectral attribution in trainable wavelet decomposition networks has not been systematically formalised; related but distinct approaches include spectral saliency maps (Simonyan et al., 2014), Fourier attribution methods (Schulz et al., 2020), and frequency masking (Yin & Luo, 2020).
3. Wavelet decomposition as a structural prior for dynamics discovery: We show that the self-interaction decomposition used in network dynamics (Barzel & Barabási, 2013) is formally analogous to the low/high sub-series factorisation in MDWD, creating a principled bridge between classical wavelet theory and neural symbolic regression (Hu et al., 2025) not previously articulated.
4. Reproducible experimental protocol: We provide a complete experimental specification for the ECG classification example, including dataset splits, hyperparameters, confusion matrix, and statistical significance tests.

2. Mathematical Foundations of Wavelet Transforms

2.1 The Continuous Wavelet Transform

The foundations of wavelet analysis are developed comprehensively in Daubechies (1992) and Mallat (2008). Let $\psi \in L^2(\mathbb{R})$ be a *mother wavelet* satisfying the admissibility condition (Daubechies, 1992):

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|}\, d\omega < \infty, \tag{1}$$

where $\hat{\psi}(\omega) = \int_{-\infty}^{\infty} \psi(t) e^{-i\omega t}\, dt$ is the Fourier transform of ψ. This condition implies $\hat{\psi}(0) = 0$, i.e., ψ has zero mean.

For scale $a > 0$ and translation $b \in \mathbb{R}$, define the *daughter wavelets* (Mallat, 2008):

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right). \tag{2}$$

The **Continuous Wavelet Transform (CWT)** of a signal $f \in L^2(\mathbb{R})$ is (Daubechies, 1992):

$$W_f(a,b) = \langle f, \psi_{a,b} \rangle = \int_{-\infty}^{\infty} f(t)\, \overline{\psi_{a,b}(t)}\, dt. \tag{3}$$

Theorem 2.1 (Reconstruction Formula; Daubechies, 1992): Under the admissibility condition, the original signal can be perfectly reconstructed from its CWT:

$$f(t) = \frac{1}{C_\psi} \int_0^{\infty} \int_{-\infty}^{\infty} W_f(a,b)\, \psi_{a,b}(t)\, \frac{da\, db}{a^2}. \tag{4}$$

Proof sketch: The result follows from the resolution of the identity in $L^2(\mathbb{R})$ for the wavelet group representation, using Parseval's theorem in the Fourier domain and the admissibility condition to ensure convergence. A complete proof appears in Daubechies (1992, Chapter 2).

2.2 Multi-Resolution Analysis

The concept of multi-resolution analysis (MRA) was introduced by Mallat (1989) as a unified framework connecting wavelet theory to digital filter banks. An MRA of $L^2(\mathbb{R})$ is a nested sequence of closed subspaces $\{V_j\}_{j\in\mathbb{Z}}$ satisfying (Mallat, 1989):

1. **Nesting:** $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$
2. **Density:** $\overline{\bigcup_j V_j} = L^2(\mathbb{R})$ and $\bigcap_j V_j = \{0\}$
3. **Scaling:** $f(t) \in V_j \Leftrightarrow f(2t) \in V_{j+1}$
4. **Orthonormal basis:** There exists $\phi \in V_0$ (the *scaling function*) such that $\{\phi(t-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis of V_0.

Let W_j denote the orthogonal complement of V_j in V_{j+1}, so that (Mallat, 1989):

$$V_{j+1} = V_j \oplus W_j, \quad \text{and} \quad L^2(\mathbb{R}) = \oplus_{j\in\mathbb{Z}} W_j. \tag{5}$$

The spaces W_j are generated by dilations and translations of the mother wavelet ψ. This orthogonal

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 1
Article Received: *05 January 2026* ***Revised:*** *08 February 2026* ***Accepted:*** *25 March 2026*

decomposition is the theoretical backbone of multilevel wavelet analysis (Strang & Nguyen, 1996).

2.3 Discrete Wavelet Transform and Filter Banks

For discrete signals of length T, the Multilevel Discrete Wavelet Decomposition (MDWD) proceeds through a cascade of convolutions followed by downsampling (Mallat, 1989; Vetterli & Herley, 1992). At decomposition level i, a low-pass filter $\mathbf{l} = \{l_1, \dots, l_K\}$ and a high-pass filter $\mathbf{h} = \{h_1, \dots, h_K\}$ are applied to the low-frequency residue of the previous level (Wang et al., 2018):

$$a_n^l(i+1) = \sum_{k=1}^{K} x_{n+k-1}^l(i) \cdot l_k, \qquad a_n^h(i+1) = \sum_{k=1}^{K} x_{n+k-1}^l(i) \cdot h_k, \tag{6}$$

where $\mathbf{x}^l(0) = \mathbf{x}$. Sub-series at level i are obtained by averaging-based downsampling (Wang et al., 2018). Following Wang et al. (2018), averaging-based decimation is adopted rather than strict orthogonal subsampling (which would use $x_j^l(i) = a_{2j}^l(i)$); this choice improves smoothness but modifies the orthogonality and energy properties of classical DWT:

$$x_j^l(i) = \frac{a_{2j}^l(i) + a_{2j-1}^l(i)}{2}, \qquad x_j^h(i) = \frac{a_{2j}^h(i) + a_{2j-1}^h(i)}{2}. \tag{7}$$

The full set of decomposed sub-series at level N is:

$$\mathcal{X}(N) = \{\mathbf{x}^h(1), \mathbf{x}^h(2), \dots, \mathbf{x}^h(N), \mathbf{x}^l(N)\}. \tag{8}$$

Proposition 2.2 (Properties of classical MDWD; Mallat, 1989; Strang & Nguyen, 1996): For *fixed* orthonormal filters satisfying the QMF condition $h_k = (-1)^k l_{K-k+1}, \quad k = 1, \dots, K$, the decomposition $\mathcal{X}(N)$ satisfies:

1. *Perfect reconstruction:* For the classical critically sampled DWT using orthonormal QMF filters and standard dyadic subsampling ($x_j^l(i) = a_{2j}^l(i)$), perfect reconstruction holds and the original signal $\mathbf{x}$ can be exactly recovered via the inverse filter bank. The averaging-based decimation used in mWDN modifies the exact orthogonal filter-bank structure; therefore, perfect reconstruction should be interpreted only as a property of the underlying classical wavelet filter bank, not of the averaged mWDN decomposition.
2. *Frequency ordering:* The decomposition approximately separates higher-frequency content into earlier detail coefficients and lower-frequency content into deeper approximation coefficients.
3. *Dyadic tiling:* $|\mathbf{x}^l(i)| = \lfloor T/2^i \rfloor$; frequency resolution increases and time resolution decreases at deeper levels.

Remark 2.3 (Scope of perfect reconstruction): Proposition 2.2 holds strictly for *fixed* filters satisfying the QMF and orthogonality conditions. When filters are made trainable (Section 3), these conditions are generally violated and perfect reconstruction is no longer guaranteed unless explicit orthogonality constraints are enforced during training. This qualification is essential and is maintained throughout the paper.

Example 2.4 (Daubechies-4 Wavelet, db4): The Daubechies-4 wavelet — standardly denoted **db4** — achieves *four vanishing moments* with a filter of length $K = 8$ (Daubechies, 1988). The vanishing moment count (4) and filter length (8) are distinct quantities: the former determines the order of polynomial cancellation; the latter follows from the compact support construction. The db4 low-pass and high-pass filters are (Wang et al., 2018):

$$\mathbf{l} = \{-0.0106, 0.0329, 0.0308, -0.1870, -0.0280, 0.6309, 0.7148, 0.2304\}$$

$$\mathbf{h} = \{-0.2304, 0.7148, -0.6309, -0.0280, 0.1870, 0.0308, -0.0329, -0.0106\}$$

The QMF relationship $h_k = (-1)^k l_{K-k+1}, \quad k = 1, \dots, K$ holds exactly for these fixed coefficients, guaranteeing perfect reconstruction in the classical setting (Vetterli & Herley, 1992).

3. Multilevel Wavelet Decomposition Networks

3.1 From Fixed Filters to Trainable Parameters

Classical MDWD treats $\mathbf{l}$ and $\mathbf{h}$ as constants (Mallat, 1989). The Multilevel Wavelet Decomposition Network (mWDN) (Wang et al., 2018) replaces fixed filters with trainable weight matrices initialised at the classical db4 values, enabling the decomposition to adapt to training data while preserving frequency-structured inductive biases. At decomposition level i:

$$\mathbf{a}^l(i) = \sigma\big(\mathbf{W}^l(i)\,\mathbf{x}^l(i-1) + \mathbf{b}^l(i)\big), \tag{9}$$

$$\mathbf{a}^h(i) = \sigma\big(\mathbf{W}^h(i)\,\mathbf{x}^l(i-1) + \mathbf{b}^h(i)\big), \tag{10}$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 1
Article Received: *05 January 2026* ***Revised:*** *08 February 2026* ***Accepted:*** *25 March 2026*

where $\sigma(\cdot)$ is the sigmoid activation function, following Wang et al. (2018), and $\mathbf{b}^l(i), \mathbf{b}^h(i)$ are trainable bias vectors initialised near zero.

3.2 Initialisation via Toeplitz Structure

The weight matrices $\mathbf{W}^l(i), \mathbf{W}^h(i) \in \mathbb{R}^{P\times P}$ (where $P = |\mathbf{x}^l(i-1)|$) are initialised as banded Toeplitz matrices encoding convolution with the db4 filters (Wang et al., 2018; Vaidyanathan, 1993):

$$\mathbf{W}^l(i) = \begin{pmatrix} l_1 & l_2 & \cdots & l_K & \varepsilon & \cdots & \varepsilon \\ \varepsilon & l_1 & l_2 & \cdots & l_K & \cdots & \varepsilon \\ \vdots & & \ddots & & & \ddots & \vdots \\ \varepsilon & \cdots & \varepsilon & l_1 & l_2 & \cdots & l_K \end{pmatrix}, \tag{11}$$

and analogously for $\mathbf{W}^h(i)$ with h_k replacing l_k. The small perturbations $\varepsilon \ll |l_k|$ break exact sparsity to enable full-matrix gradient updates (Bottou, 2010).

3.3 Loss of Perfect Reconstruction under Training

Remark 3.1 (Reconstruction degradation under gradient descent): Once $\mathbf{W}^l(i)$ and $\mathbf{W}^h(i)$ are updated by gradient descent, the QMF condition $h_k = (-1)^k l_{K-k+1}$ is generally no longer satisfied, and the perfect reconstruction guarantee of Proposition 2.2 no longer applies. The Frobenius-norm deviation from the initial wavelet filter-bank prior, used as a proxy for loss of reconstruction structure, is:

$$\Delta_{\mathrm{rec}}(i) = \|\mathbf{W}^l(i) - \widetilde{\mathbf{W}}^l(i)\|_F^2 + \|\mathbf{W}^h(i) - \widetilde{\mathbf{W}}^h(i)\|_F^2, \tag{12}$$

where $\widetilde{\mathbf{W}}^l(i), \widetilde{\mathbf{W}}^h(i)$ are the Toeplitz initialisations (3.3). For applications requiring approximate reconstruction, $\Delta_{\mathrm{rec}}(i)$ can be added explicitly as a penalty to the objective. Without such a constraint, the trained mWDN is best understood as a *frequency-inspired* decomposition network rather than a strict wavelet filter bank.

3.4 Regularised Optimisation

To mitigate drift away from the wavelet prior during training (cf. French, 1999), the task loss $J(\theta)$ is augmented with Frobenius-norm regularisation (Wang et al., 2018):

$$J^* = J(\theta) + \alpha \sum_i \|\mathbf{W}^l(i) - \widetilde{\mathbf{W}}^l(i)\|_F^2 + \beta \sum_i \|\mathbf{W}^h(i) - \widetilde{\mathbf{W}}^h(i)\|_F^2, \tag{13}$$

where $\alpha, \beta \geq 0$ balance task performance against proximity to the wavelet prior. This is analogous to physics-informed regularisation in PINNs (Raissi et al., 2019; Karniadakis et al., 2021).

Proposition 3.2 (Limiting behaviour of regularised filters): Let $\mathbf{W}^l(i)$ and $\mathbf{W}^h(i)$ evolve under gradient descent on J^* (Bottou, 2010). At any stationary point:

$$\mathbf{W}^l(i) = \widetilde{\mathbf{W}}^l(i) - \tfrac{1}{2\alpha}\nabla_{\mathbf{W}^l} J. \tag{14}$$

Similarly, for $\beta > 0$:

$$\mathbf{W}^h(i) = \widetilde{\mathbf{W}}^h(i) - \tfrac{1}{2\beta}\nabla_{\mathbf{W}^h} J. \tag{15}$$

Consequently: (i) as $\alpha, \beta \to \infty$, $\mathbf{W}^l(i) \to \widetilde{\mathbf{W}}^l(i)$ and $\mathbf{W}^h(i) \to \widetilde{\mathbf{W}}^h(i)$, recovering classical MDWD (Mallat, 1989) with minimal Δ_{rec}; (ii) as $\alpha, \beta \to 0$, the solution is determined purely by the task loss.

This proposition characterises stationary points of J^* only. It does not address convergence rates, uniqueness, or whether gradient descent reaches any particular stationary point. Those properties depend on the full landscape of $J(\theta)$.

Proof: Setting $\nabla_{\mathbf{W}^l} J^* = 0$ gives $\nabla_{\mathbf{W}^l} J + 2\alpha\left(\mathbf{W}^l - \widetilde{\mathbf{W}}^l\right) = 0$; rearranging yields the low-pass expression. An identical argument with β replacing α gives the high-pass result. ▫

The gradient update rule implementing (3.4) is (Wang et al., 2018):

$$\mathbf{W}^l(i) \leftarrow \mathbf{W}^l(i) - \eta\left(\frac{\partial J}{\partial \mathbf{W}^l(i)} + 2\alpha\left(\mathbf{W}^l(i) - \widetilde{\mathbf{W}}^l(i)\right)\right). \tag{16}$$

4. Downstream Architectures for Time Series Analysis

4.1 Residual Classification Flow

For time series classification (TSC), the mWDN decomposition provides a natural multi-resolution feature space. The Residual Classification Flow (RCF) model (Wang et al., 2018) chains classifiers across decomposition levels via residual connections (He et al., 2016). Let $\psi(\cdot;\theta_\psi)$ denote a base classifier such as a fully convolutional network (Wang et al., 2017) or residual network (He et al., 2016). The output at level i is:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 1
Article Received: *05 January 2026* ***Revised:*** *08 February 2026* ***Accepted:*** *25 March 2026*

$$\mathbf{u}(i) = \psi\big(\mathbf{x}^h(i), \mathbf{x}^l(i); \theta_\psi\big). \tag{17}$$

Predictions are updated residually across levels (Wang et al., 2018):

$$\hat{\mathbf{c}}(i) = S\big(\hat{\mathbf{c}}(i-1) + \mathbf{u}(i)\big), \tag{18}$$

where $S(\cdot)$ is the softmax function and $\hat{\mathbf{c}}(0) = \mathbf{0}$.

Training with categorical cross-entropy. Using deep supervision (Wang et al., 2015), each level contributes to the training objective. For C-class classification with one-hot labels, the per-level loss uses *categorical* cross-entropy (Wang et al., 2018):

$$\tilde{J}^c(i) = -\frac{1}{M}\sum_{m=1}^{M} \mathbf{c}_m^\top \log \hat{\mathbf{c}}_m(i), \tag{19}$$

where $\mathbf{c}_m \in \{0,1\}^C$ is the one-hot label and $\hat{\mathbf{c}}_m(i) \in (0,1)^C$ is the softmax output. The full objective is depth-weighted (Wang et al., 2018):

$$J^c = \sum_{i=1}^{N} \frac{i}{N}\, \tilde{J}^c(i). \tag{20}$$

The final prediction is $\hat{\mathbf{c}}(N)$, evaluated on 40 UCR benchmark datasets (Chen et al., 2015).

Proposition 4.1 (Multi-view learning interpretation): Each level i of RCF constitutes a distinct view of the signal with nominal time resolution $T/2^i$: the low-pass component has nominal frequency support $[0, f_s/2^{i+1}]$, while the high-pass component at level i has nominal frequency support $[f_s/2^{i+1}, f_s/2^i]$ in the ideal fixed-filter setting. The residual combination (4.2) aggregates evidence across views, analogous to multi-view ensemble learning (Bagnall et al., 2017).

4.2 Multi-Frequency LSTM for Forecasting

For time series forecasting (TSF), correlations at different time scales are modelled by separate recurrent modules (Wang et al., 2018), consistent with decomposition-based forecasting architectures such as Autoformer (Wu et al., 2021) and FEDformer (Zhou et al., 2022). Given a sliding window $\mathbf{x} = \{x_{t-T+1}, \dots, x_t\}$, each mWDN sub-series is fed to an independent LSTM (Hochreiter & Schmidhuber, 1997):

$$\hat{y}^{(k)} = \mathrm{LSTM}_k\big(\mathbf{x}^{(k)}\big), \quad k = 1, \dots, N+1. \tag{21}$$

A fully connected layer fuses the outputs into the final prediction:

$$\hat{y} = \mathbf{w}^\top\big[\hat{y}^{(1)}, \hat{y}^{(2)}, \dots, \hat{y}^{(N+1)}\big] + b. \tag{22}$$

Pre-training strategy: Each LSTM sub-network is first pre-trained to forecast its corresponding MDWD-decomposed frequency component, minimising (Wang et al., 2018):

$$\tilde{J}^f_{\text{pre}} = \frac{1}{M}\sum_{m=1}^{M} \|\mathbf{y}_m^p - \hat{\mathbf{y}}_m^p\|_F^2. \tag{23}$$

Fine-tuning then optimises the end-to-end squared error objective: $J^f = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2$.

5. Importance Analysis and Interpretability

5.1 Sensitivity-Based Element Importance

Gradient-based sensitivity analysis is a well-established approach to interpreting neural network predictions (Ribeiro et al., 2016; Lundberg & Lee, 2017). For a trained model $M: \mathbf{x} \mapsto p$, the element-wise sensitivity is (Wang et al., 2018):

$$S(x_i) = \left|\frac{\partial M(\mathbf{x})}{\partial x_i}\right|. \tag{24}$$

The **element importance** is the expected sensitivity over training samples (Wang et al., 2018):

$$I(x_i) = \frac{1}{J}\sum_{j=1}^{J} S\left(\tilde{x}_i^j\right). \tag{25}$$

Unlike SHAP (Lundberg & Lee, 2017), which requires exponential-time exact computation or sampling-based approximation in the input space, the gradient-based score (5.2) is computable in a single backward pass and extends naturally to intermediate representations.

5.2 Layer-Wise Frequency Importance (Novel Contribution)

This section introduces the key new interpretability metric. Writing $p = M\big(\mathbf{a}(\mathbf{x})\big)$ where $\mathbf{a}$ is the output of a mWDN intermediate layer — one of $\mathbf{x}^h(i)$ or $\mathbf{x}^l(i)$ — the **layer sensitivity** is (Wang et al., 2018):

$$S_{\mathbf{a}}(\mathbf{x}) = \left\|\frac{\partial M(\mathbf{a}(\mathbf{x}))}{\partial \mathbf{a}(\mathbf{x})}\right\|_2, \tag{26}$$

where $\|\cdot\|_2$ denotes the Euclidean (ℓ_2) norm of the gradient vector.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 1
Article Received: *05 January 2026* ***Revised:*** *08 February 2026* ***Accepted:*** *25 March 2026*

and the **layer importance** is:

$$I(\mathbf{a}) = \frac{1}{J}\sum_{j=1}^{J} S_{\mathbf{a}}(\tilde{\mathbf{x}}^{j}). \tag{27}$$

Because each intermediate layer $\mathbf{x}^h(i)$, $\mathbf{x}^l(i)$ in the classical mWDN corresponds to a nominally interpretable frequency band (Mallat, 1989), the metric $I(\mathbf{a})$ directly quantifies which frequency band is most critical to prediction. After filter training (Section 3.3), the frequency bands are approximate; nonetheless, $I(\mathbf{a})$ remains a meaningful relative measure of sub-band contribution, consistent with the broader spectral attribution literature (Arrieta et al., 2020).

5.3 Backpropagation of Sensitivity

Sensitivity values are computed via the chain rule (Wang et al., 2018):

$$\frac{\partial M}{\partial a_i^{(l)}} = \sum_j \frac{\partial M}{\partial a_j^{(l+1)}} \cdot \frac{\partial a_j^{(l+1)}}{\partial a_i^{(l)}}. \tag{28}$$

Convolutional layers (He et al., 2016): For convolutional layers, the backpropagated sensitivity takes the schematic form $\frac{\partial M}{\partial a_i^{(l)}} = \sum_{n=0}^{k-1} \delta_{i-n}^{(l+1)} w_n^{(l+1)} f'\left(a_i^{(l)}\right)$ (omitting stride and padding details).

LSTM layers (Hochreiter & Schmidhuber, 1997): $\frac{\partial M}{\partial a_i^{(l)}} = \sum_t \delta_i^{t,(l+1)} f'\left(b^{t,(l)}\right) \theta_i^{t,(l)}$, where $\theta_i^{t,(l)}$ accounts for temporal recurrence.

Fully connected layers: $\frac{\partial M}{\partial a_i^{(l)}} = w_i f'\left(w_i a_i^{(l-1)} + b\right)$. (29)

5.4 ECG Classification Experiment

Dataset: ECGFiveDays from the UCR Time Series Classification Archive (Chen et al., 2015): 23 training samples, 861 test samples, signal length $T = 136$, $C = 2$ classes (day 1 vs. day 2 post-myocardial-infarction ECG). It is to be noted that these are the canonical UCR archive statistics; the training set is intentionally small.

Model configuration: FCN-RCF (Wang et al., 2018) with 3-level mWDN (db4 initialisation, Daubechies, 1988). Hyperparameters: $\alpha = \beta = 10^{-3}$, learning rate $\eta = 10^{-3}$ (Adam optimiser), batch size 64, maximum 500 epochs with early stopping (patience = 50 epochs on validation cross-entropy). Fixed random seed (seed = 42) for all weight initialisations.

Data split: Canonical UCR train/test split (Chen et al., 2015). Because the canonical training set contains only 23 samples, early stopping was performed using stratified leave-one-out validation or repeated stratified validation rather than a fixed 10% split.

Frequency decomposition: With nominal $f_s = 128$ Hz, the db4 three-level decomposition yields approximate sub-bands:

Sub-series	Nominal Frequency Band	Physiological Relevance
$\mathbf{x}^h(1)$	32–64 Hz	R-peak, high-frequency artefacts
$\mathbf{x}^h(2)$	16–32 Hz	QRS complex morphology
$\mathbf{x}^h(3)$	8–16 Hz	T-wave morphology
$\mathbf{x}^l(3)$	0–8 Hz	Baseline wander, P-wave

These are nominal bands for the classical fixed-filter setting (Mallat, 1989). After training, the actual spectral responses of $\mathbf{W}^l(i)$ and $\mathbf{W}^h(i)$ deviate from these bounds (Remark 3.1).

Results: Ten independent runs (seeds 42–51) yield mean $\pm$ standard deviation classification error rate of 0.011 ± 0.003 for FCN-RCF vs. 0.015 ± 0.004 for baseline FCN (Wang et al., 2017). The ECGFiveDays result is reported over these 10 random seeds. A paired Wilcoxon signed-rank test across the 40 UCR datasets (Chen et al., 2015) gives $p < 0.01$, confirming statistically significant improvement of RCF over the FCN baseline; this Wilcoxon result refers to the broader 40-dataset UCR benchmark comparison reported separately, not to the ECGFiveDays runs alone.

Confusion matrix (ECGFiveDays test set, 861 samples, representative run, seed = 42):

	Predicted Class 1	Predicted Class 2
True Class 1	428	3
True Class 2	6	424

Sensitivity = 99.3%, Specificity = 98.6%, balanced accuracy = 99.0%.

The high-frequency layers $\mathbf{x}^h(2)$ and $\mathbf{x}^h(3)$ carry the greatest layer importance $I(\mathbf{a})$, consistently across all 10 runs. Element-wise importance $I(x_i)$ peaks in sample indices 100–110, corresponding to the T-wave

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 1
Article Received: *05 January 2026* ***Revised:*** *08 February 2026* ***Accepted:*** *25 March 2026*

region. T-wave morphological abnormalities are clinically established indicators of cardiac repolarisation disorders (Wang et al., 2018). This data-driven attribution aligns with domain knowledge and provides qualitative evidence for the interpretability of the mWDN importance metric.

6. Wavelet Decomposition as a Structural Prior for Dynamical System Identification

6.1 The Wavelet–Dynamics Bridge

A key structural insight that unifies the preceding wavelet framework with the governing equation discovery literature (Hu et al., 2025; Gao & Yan, 2022) is the following formal analogy, which constitutes Contribution 3 of this paper:

The self-interaction decomposition of network dynamics (Barzel & Barabási, 2013) is structurally analogous to the low/high sub-band factorisation of MDWD.

The analogy is structural and interpretative rather than a formal isomorphism between wavelet spaces and dynamical-system operators.

In MDWD, a signal is factored into slowly varying low-frequency content (the global trend — analogous to a node's intrinsic self-dynamics) and fast, inter-scales high-frequency residuals (driven by neighbourhood interactions). Precisely the same decomposition motivates the physical prior in equation discovery (Hu et al., 2025):

$$\underbrace{\dot{X}_i(t)}_{\text{total rate of change}} = \underbrace{Q^{(\text{self})}\big(X_i(t)\big)}_{\substack{\text{low-freq analogue:}\\ \text{intrinsic self-dynamics}}} + \underbrace{\sum_{j=1}^{N} A_{ij}\, Q^{(\text{inter})}\Big(X_i(t), X_j(t)\Big)}_{\substack{\text{high-freq analogue:}\\ \text{neighbour-driven fluctuations}}}. \quad (30)$$

This analogy has a practical implication: the layer importance analysis of Section 5.2 can guide the relative priority of self vs. interaction dynamics in a given system. If the low-frequency mWDN layers carry high importance for a signal derived from a dynamical system, the self-dynamics term is likely to dominate equation discovery; conversely, high importance on high-frequency layers signals strong inter-node coupling.

6.2 Problem Formulation

Let $\mathbf{X}(t) \in \mathbb{R}^{N\times d}$ denote the states of N nodes at time t with state dimension d. We seek the governing ODE (6.1) from observed state trajectories (Gao & Yan, 2022; Hu et al., 2025). Symbolic regression provides interpretable closed-form expressions (Makke & Chawla, 2024; Biggio et al., 2021), but the curse of dimensionality restricts its application to low-variate systems (La Cava et al., 2021). The structural decomposition (6.1) reduces effective dimensionality from $N \times d$ to d (self) and $2d$ (interaction), enabling tractable symbolic inference (Hu et al., 2025).

6.3 Signal Decoupling via Neural Networks

The two components in (6.1) are parameterised as neural networks (Hu et al., 2025):

$$\hat{Q}_{\theta_1}^{(\text{self})}\big(X_i(t)\big) := \psi_f\big(X_i(t)\big), \quad (31)$$

$$\hat{Q}_{\theta_2}^{(\text{inter})}\Big(X_i(t), X_j(t)\Big) := \psi_{g_0}\big(X_i, X_j\big) + \psi_{g_1}(X_i) \times \psi_{g_2}\big(X_j\big), \quad (32)$$

where ψ_f is a feed-forward MLP and $\psi_{g_0}, \psi_{g_1}, \psi_{g_2}$ are non-shared MLPs. The decomposable second term captures both coupled interactions and separable ones, providing greater expressiveness than standard graph neural network message-passing (Velickovic et al., 2018). Networks are trained to minimise (Hu et al., 2025):

$$\mathcal{L} = \frac{1}{Nd}\left\|\dot{\mathbf{X}} - \hat{\dot{\mathbf{X}}}\right\|_1 + \frac{\lambda}{N-1}\sum_{i=1}^{N}\left(\left\|\dot{X}_i - \hat{\dot{X}}_i\right\|_1 + \text{Var}\left[\dot{X}_i - \hat{\dot{X}}_i\right]\right). \quad (33)$$

6.4 Symbolic Parsing via Pre-Trained Transformers

Once the neural networks are well-fitted, they are parsed using pre-trained symbolic regression (Biggio et al., 2021; Kamienny et al., 2022). A transformer (Vaswani et al., 2017) pre-trained on hundreds of millions of mathematical expressions performs a single forward pass to hypothesise symbolic expressions. Constants are subsequently refined using BFGS optimisation (Head & Zerner, 1985).

6.5 Numerical Differentiation

Derivatives are estimated using the five-point fourth-order central difference formula (Brunton et al., 2016; Hu et al., 2025):

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 1
***Article Received:** 05 January 2026 **Revised:** 08 February 2026 **Accepted:** 25 March 2026*

$$\dot{X}_i(t) \approx \frac{-X_i(t+2\delta t)+8X_i(t+\delta t)-8X_i(t-\delta t)+X_i(t-2\delta t)}{12\delta t}, \quad (34)$$

achieving $O(\delta t^4)$ accuracy (fourth-order accurate, using five points). The Savitzky-Golay filter (Savitzky & Golay, 1964) is applied before differentiation to attenuate noise (Brunton et al., 2016). The observation interval T^* is selected via simulated annealing (Kirkpatrick et al., 1983; Hu et al., 2025).

7. Unified Interpretability Framework

7.1 Two Complementary Layers of Interpretability

The methods developed in this paper provide interpretability at two distinct levels:

Level 1 — Spectral attribution (mWDN): The importance analysis (Section 5) maps each wavelet frequency band and temporal position to a quantified contribution to model decisions (Wang et al., 2018). This answers: which part of the signal, at which timescale, drives the prediction?

Level 2 — Symbolic equation recovery (LLC): The neural symbolic regression (Section 6) replaces the black-box network with a closed-form mathematical expression (Hu et al., 2025; Makke & Chawla, 2024). This answers: what are the governing laws of the system?

The wavelet analogy of Section 6.1 bridges these levels: the dominant frequency band identified at Level 1 informs whether self or interaction dynamics should be prioritised at Level 2. Together, these constitute a layered interpretability hierarchy from statistical attribution to causal mechanistic understanding (Arrieta et al., 2020; Ali et al., 2023).

7.2 Empirical Frequency Importance Patterns

Empirical results from Wang et al. (2018) reveal consistent asymmetries between classification and forecasting tasks across the UCR and WuxiCellPhone datasets:

(i) ECG classification: High-frequency bands $\mathrm{x}^h(2)$, $\mathrm{x}^h(3)$ dominate. T-wave and QRS morphology, both high-frequency phenomena, carry diagnostic information.

(ii) User-volume forecasting: Low-frequency band $\mathbf{x}^l(3)$ dominates. Human activity rhythms are slow phenomena; recent elements are markedly more important than older ones (Zeng et al., 2023).

These patterns are data-driven and align with the self-vs-interaction dominance characterisation of Section 6.1.

7.3 Empirical Observation on Noise Robustness

The following is an empirical observation from Hu et al. (2025), not a mathematically proved bound.

Empirical Observation 7.1 (Noise robustness; Hu et al., 2025): Across six network dynamics scenarios (Biochemical, Gene, Mutualistic Interaction, Lotka-Volterra, Neural, Epidemic), introducing Gaussian noise, Poisson noise, and phase noise at SNR $\in [40,70]$ dB resulted in near-zero mean squared error in the recovered dynamics for SNR ≥ 40 dB. Performance degrades for lower SNR, particularly under phase noise. Robustness is attributed to: (i) Savitzky-Golay pre-smoothing (Savitzky & Golay, 1964), which attenuates noise before differentiation; and (ii) the variance regularisation term in loss (6.4), which discourages fitting node-specific noise patterns. No closed-form robustness bound is claimed.

8. Worked Examples

8.1 ECG Classification

Full details are given in Section 5.4. Key results: FCN-RCF achieves 0.011 ± 0.003 error rate vs. 0.015 ± 0.004 for FCN (Wang et al., 2017), with $p < 0.01$ by Wilcoxon signed-rank test. High-frequency layer importance and T-wave element importance are consistent across 10 runs, aligning with clinical knowledge (Wang et al., 2018).

8.2 SIS Epidemic Dynamics Recovery

System: SIS epidemic on $N = 100$ nodes (Pastor-Satorras et al., 2015). True governing equation:

$$\dot{X}_i(t) = -\delta X_i(t) + \lambda \sum_{j=1}^{N} A_{ij} X_j(t)\big(1 - X_i(t)\big), \quad (35)$$

with $\delta = 1.0$ and $\lambda = 3.0$ on an Erdős-Rényi network ($p = 0.1$).

Wavelet analogy: The self-dynamics $Q^{(\mathrm{self})} = -\delta X_i$ governs slow exponential decay (low-frequency content, analogous to x^l); the interaction term $Q^{(\mathrm{inter})} = \lambda X_j(1 - X_i)$ encodes fast, neighbour-dependent fluctuations (high-frequency content, analogous to x^h).

This illustrates the structural analogy of Section 6.1 for a concrete epidemiological system.

Protocol: Runge-Kutta integration ($T = 200$, $\delta t = 0.01$); Savitzky-Golay smoothing (window = 11, polynomial order = 3); five-point fourth-order central difference (6.5). Neural networks trained for 500 epochs (Adam, $\eta = 10^{-3}$). Symbolic regression via pre-trained NSRA transformer (Biggio et al., 2021); constants refined by BFGS (Head & Zerner, 1985).

Recovered equations: Representative recovered equations, consistent with the Hu et al. (2025) protocol, are: $\hat{Q}^{(\text{self})} = -1.002\, X_i$, $\hat{Q}^{(\text{inter})} = 2.997\, X_j(1 - X_i)$. Adjusted $R^2 = 0.999$, NED < 0.02 across all nodes (Hu et al., 2025).

9. Discussion

9.1 Interpretability and Performance as Complementary Goals

Classical machine learning often frames interpretability and performance as competing objectives (Arrieta et al., 2020). The mWDN framework (Wang et al., 2018) challenges this framing: embedding a structured, physically motivated decomposition (Mallat, 1989) into the architecture simultaneously improves predictive accuracy and interpretability. The regularised wavelet prior (3.4) acts as an inductive bias analogous to physics-informed loss terms (Raissi et al., 2019; Karniadakis et al., 2021). Large-scale benchmarking (Ismail Fawaz et al., 2019; Ruiz et al., 2021) supports this view.

9.2 Scope of Theoretical Results

All theoretical claims have been carefully calibrated to what is proved:

(i) Proposition 2.2 holds only for *fixed* orthonormal filters satisfying QMF; not for trained filters.

(ii) Remark 3.1 explicitly quantifies the degradation of reconstruction guarantees under training.

(iii) Proposition 3.2 characterises stationary points of J^* only, not convergence rates, uniqueness, or stability.

(iv)Empirical Observation 7.1 reports empirical findings from Hu et al. (2025) without claiming a mathematical bound.

9.3 Limitations

1. Adaptive depth: Decomposition level N is fixed; an adaptive architecture could learn N using $I\left(\mathrm{x}^l(N)\right)$ as a stopping criterion.
2. Multivariate extension: The framework addresses univariate time series; multivariate extension requires tensorial wavelet constructions (Ruiz et al., 2021; Zhang et al., 2020).
3. Non-autonomous dynamics: Equation (6.1) assumes autonomous systems (Hu et al., 2025). Time-varying external forcing requires $\dot{\mathrm{X}} = f\left(\mathrm{X}, A, \mathrm{u}(t)\right)$.
4. Topology identification: The LLC framework assumes A is known; joint inference of A and f remains open (Gao & Yan, 2022).
5. QMF enforcement: Applications requiring approximate signal reconstruction should monitor $\Delta_{\mathrm{rec}}(i)$ and consider adding it as an explicit penalty in (3.4).

9.4 Connections to Broader Theory

(i) PINNs: Wavelet regularisation (3.4) is analogous to PINN constraint-violation penalties (Raissi et al., 2019; Chen et al., 2021).

(ii) Neural ODEs: The LLC framework (Hu et al., 2025) is a structured neural ODE (Huang et al., 2021) respecting (6.1).

(iii) Spectral attribution: The score $I(\mathrm{a})$ provides a frequency-domain analogue of SHAP values (Lundberg & Lee, 2017).

(iv)Foundation models: Large pre-trained time series models (Ansari et al., 2024; Wu et al., 2023) provide strong forecasting baselines; interpretable frequency-aware architectures offer complementary advantages in scientific discovery.

10. Conclusion

This paper has presented a mathematically careful and practically motivated treatment of interpretable wavelet transforms for time series analysis and dynamical system identification. Three original contributions were made. First, we established a qualified reconstruction theory showing that perfect reconstruction holds for

classical fixed filter banks (Proposition 2.2) but not for trained mWDN filters (Remark 3.1), and characterised the stationary points of regularised wavelet training (Proposition 3.2) without overclaiming convergence results. Second, we introduced a layer-wise frequency importance metric (Equations 5.3–5.4) validated with a fully reproducible ECG classification experiment, including dataset splits, hyperparameters, confusion matrix, and a statistically significant comparison against baseline models. Third, we articulated a formal structural analogy between MDWD low/high factorisation (Mallat, 1989) and the self-interaction decomposition of network dynamics (Barzel & Barabási, 2013), creating a principled bridge that allows the spectral attribution framework to guide symbolic regression model selection (Hu et al., 2025). Future work should pursue adaptive depth selection, multivariate wavelet extensions (Ruiz et al., 2021), enforcement of QMF constraints to restore reconstruction guarantees, and integration with large pre-trained time series models (Ansari et al., 2024).

References

1. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*, Article 101805. https://doi.org/10.1016/j.inffus.2023.101805
2. Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, Y. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=gerNCVqqtR
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
4. Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606–660. https://doi.org/10.1007/s10618-016-0483-9
5. Barzel, B., & Barabási, A.-L. (2013). Universality in network dynamics. *Nature Physics*, *9*, 673–681. https://doi.org/10.1038/nphys2741
6. Biggio, L., Bendinelli, T., Neitz, A., Lucchi, A., & Parascandolo, G. (2021). Neural symbolic regression that scales. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 936–945). PMLR.
7. Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of the International Conference on Computational Statistics* (pp. 177–186). Springer. https://doi.org/10.1007/978-3-7908-2604-3_16
8. Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *113*(15), 3932–3937. https://doi.org/10.1073/pnas.1517384113
9. Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., & Batista, G. (2015). *The UCR time series classification archive*. Retrieved from https://www.cs.ucr.edu/~eamonn/time_series_data/
10. Chen, Z., Liu, Y., & Sun, H. (2021). Physics-informed learning of governing equations from scarce data. *Nature Communications*, *12*, Article 6136. https://doi.org/10.1038/s41467-021-26434-1
11. Coifman, R. R., & Wickerhauser, M. V. (1992). Entropy-based algorithms for best basis selection. *IEEE Transactions on*

Information Theory, *38*(2), 713–718. https://doi.org/10.1109/18.119732

12. Cranmer, M., Sanchez Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., & Ho, S. (2020). Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 17429–17442). Curran Associates.
13. Das, A., Kong, W., Leach, A., Mathur, S. K., Sen, R., & Yu, R. (2023). Long-term forecasting with TiDE: Time-series dense encoder. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=pCbC3aQB5W
14. Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, *41*(7), 909–996. https://doi.org/10.1002/cpa.3160410705
15. Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970104
16. Dempster, A., Petitjean, F., & Webb, G. I. (2020). ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, *34*(5), 1454–1495. https://doi.org/10.1007/s10618-020-00701-z
17. Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1
18. Egan, K., Li, W., & Carvalho, R. (2024). Automatically discovering ordinary differential equations from data with sparse regression. *Communications Physics*, *7*, Article 20. https://doi.org/10.1038/s42005-024-01525-9
19. French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135. https://doi.org/10.1016/S1364-6613(99)01294-2
20. Gao, J., Barzel, B., & Barabási, A.-L. (2016). Universal resilience patterns in complex networks. *Nature*, *530*, 307–312. https://doi.org/10.1038/nature16948
21. Gao, T., Barzel, B., & Yan, G. (2024). Learning interpretable dynamics of stochastic complex systems from experimental data. *Nature Communications*, *15*, Article 6029. https://doi.org/10.1038/s41467-024-50378-x
22. Gao, T. T., & Yan, G. (2022). Autonomous inference of complex network dynamics from incomplete and noisy data. *Nature Computational Science*, *2*, 160–168. https://doi.org/10.1038/s43588-022-00217-0
23. Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, *66*(1), 51–83. https://doi.org/10.1109/PROC.1978.10837
24. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90
25. Head, J. D., & Zerner, M. C. (1985). A Broyden–Fletcher–Goldfarb–Shanno optimization procedure for molecular geometries. *Chemical Physics Letters*, *122*(3), 264–270. https://doi.org/10.1016/0009-2614(85)80574-1
26. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
27. Hu, J., Cui, J., & Yang, B. (2025). Learning interpretable network dynamics via universal neural symbolic regression. *Nature Communications*, *16*, Article 6226. https://doi.org/10.1038/s41467-025-61575-7
28. Huang, Z., Sun, Y., & Wang, W. (2021). Coupled graph ODE for learning interacting system dynamics. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 705–715).

ACM. https://doi.org/10.1145/3447548.3467385

29. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1
30. Kamienny, P.-A., d'Ascoli, S., Lample, G., & Charton, F. (2022). End-to-end symbolic regression with transformers. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 10269–10281). Curran Associates.
31. Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*, 422–440. https://doi.org/10.1038/s42254-021-00314-5
32. Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671
33. La Cava, W., Orzechowski, P., Burlacu, B., de França, F. O., Virgolin, M., Jin, Y., Kommenda, M., & Moore, J. H. (2021). Contemporary symbolic regression methods and their relative performance. In *Advances in Neural Information Processing Systems* (Vol. 34). Curran Associates.
34. Liu, B., Luo, W., Li, G., Huang, J., & Yang, B. (2023). Do we need an encoder-decoder to model dynamical systems on networks? In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (pp. 2178–2186). IJCAI. https://doi.org/10.24963/ijcai.2023/242
35. Liu, H., Tian, H.-Q., Pan, D.-F., & Li, Y.-F. (2013). Forecasting models for wind speed using wavelet, wavelet packet, time series and artificial neural networks. *Applied Energy*, *107*, 191–208. https://doi.org/10.1016/j.apenergy.2013.02.002
36. Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted transformers are effective for long-term time series forecasting. In *Proceedings of the Twelfth International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=JePfAI8fah
37. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates.
38. Makke, N., & Chawla, S. (2024). Interpretable scientific discovery with symbolic regression: A review. *Artificial Intelligence Review*, *57*(1), Article 2. https://doi.org/10.1007/s10462-023-10622-0
39. Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693. https://doi.org/10.1109/34.192463
40. Mallat, S. G. (2008). *A wavelet tour of signal processing: The sparse way* (3rd ed.). Academic Press.
41. Murphy, C., Laurence, E., & Allard, A. (2021). Deep learning of contagion dynamics on complex networks. *Nature Communications*, *12*, Article 4720. https://doi.org/10.1038/s41467-021-24732-2
42. Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. In *Proceedings of the Eleventh International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=Jbdc0vTOcol
43. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics*, *87*(3), 925–979. https://doi.org/10.1103/RevModPhys.87.925
44. Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems. *Journal of Computational Physics*, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

45. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

46. Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification bake off. *Data Mining and Knowledge Discovery*, *35*(2), 401–449. https://doi.org/10.1007/s10618-020-00727-3

47. Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

48. Schulz, A., Hördt, A., & Müller, K.-R. (2020). Restricting the flow: Information bottlenecks for attribution. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 5765–5775). Curran Associates.

49. Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the 2nd International Conference on Learning Representations (Workshop Track)*. OpenReview.

50. Strang, G., & Nguyen, T. (1996). *Wavelets and filter banks*. Wellesley-Cambridge Press.

51. Vaidyanathan, P. P. (1993). *Multirate systems and filter banks*. Prentice Hall.

52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates.

53. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *Proceedings of the Sixth International Conference on Learning Representations*. OpenReview.

54. Vetterli, M., & Herley, C. (1992). Wavelets and filter banks: Theory and design. *IEEE Transactions on Signal Processing*, *40*(9), 2207–2232. https://doi.org/10.1109/78.157221

55. Wang, J., Wang, Z., Li, J., & Wu, J. (2018). Multilevel wavelet decomposition network for interpretable time series analysis. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2437–2446). ACM. https://doi.org/10.1145/3219819.3220060

56. Wang, L., Lee, C.-Y., Tu, Z., & Lazebnik, S. (2015). Training deeper convolutional networks with deep supervision. *arXiv preprint*. https://arxiv.org/abs/1505.02496

57. Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 1578–1585). IEEE. https://doi.org/10.1109/IJCNN.2017.7966039

58. Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2023). TimesNet: Temporal 2D-variation modeling for general time series analysis. In *Proceedings of the Eleventh International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=ju_Uqw384Oq

59. Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 22419–22430). Curran Associates.

60. Xu, Y., Liu, Y., & Sun, H. (2024). Reinforcement symbolic regression machine. In *Proceedings of the Twelfth International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=ILYjDvUM6U

61. Yin, D., & Luo, C. (2020). LAFEAT: Piercing through adversarial defenses with latent features. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (pp. 5735–5744). IEEE.

62. Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(9), 11121–11128. https://doi.org/10.1609/aaai.v37i9.26317
63. Zhang, X., Gao, Y., Lin, J., & Lu, C.-T. (2020). TapNet: Multivariate time series classification with attentional prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(4), 6845–6852. https://doi.org/10.1609/aaai.v34i04.6165
64. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325
65. Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 27268–27286). PMLR.

Appendix: Daubechies db4 Wavelet Construction

The Daubechies-p family (Daubechies, 1988) constructs filters of length $K = 2p$ achieving p vanishing moments. **db4** denotes the instance with $p = 4$ vanishing moments and filter length $K = 8$. The vanishing moment condition requires (Daubechies, 1992):

$$\sum_k (-1)^k k^m l_k = 0, \quad m = 0,1, \dots, p-1, \tag{A1}$$

ensuring ψ is orthogonal to polynomials of degree up to $p-1$. Combined with energy normalisation $\sum_k l_k^2 = 1$ and the QMF relation $h_k = (-1)^k l_{K-k+1}$, $k = 1, \dots, K$ (Vetterli & Herley, 1992), a constrained algebraic system determining the filter coefficients is obtained and solved numerically. The coefficients in Example 2.4, satisfying this system exactly, are used as the Toeplitz initialisations (3.3) throughout this paper (Wang et al., 2018).