

Continuous wavelet transform based variable projection networks

Attila Miklós Ámon¹
Eötvös Loránd University
Department of Numerical Analysis¹
1117, Budapest Pázmány Péter sétány 1/C
Email: ze3vjn@inf.elte.hu

Kristian Fenech²
Eötvös Loránd University
Department of Artificial Intelligence²
1117, Budapest Pázmány Péter sétány 1/C
Email: fenech@inf.elte.hu

Péter Kovács¹
Eötvös Loránd University
Department of Numerical Analysis
1117, Budapest Pázmány Péter sétány 1/C
Email: kovika@inf.elte.hu

Tamás Dózsa^{1,3,4}
Széchenyi István University
and HUN-REN Institute for Computer Science and Control
Vehicle Industry Research Center³
Systems and Control Laboratory⁴
1111 Budapest, Kende u. 13-17.
Email: dozs.tamas@hun-ren.sztaki.hu

Abstract—A novel feature extraction layer using the continuous wavelet transform is developed for use with neural networks. The learned parameters (scales and shifts in time) of the proposed wavelet layer can be interpreted by humans. In addition, using the concept of adaptive orthogonal projections to approximate wavelet coefficients, a sparse representation of the input is achieved. The proposed method is applicable with any analytic mother wavelet. As a case study, the application of bearing fault detection is considered to demonstrate the efficiency of the method, using a benchmark dataset. The results clearly show that the proposed model can be used to solve fault detection problems achieving near perfect classification accuracy. Qualitative and quantitative comparisons are made to a previous wavelet based neural network architecture. It is shown that the proposed method matches or outperforms previous approaches while using significantly less parameters. This sparsity improves model performance from the point of view of interpretability as well as model complexity.

I. INTRODUCTION

In traditional supervised learning methods input signals are subjected to feature extraction transformations before the classification or regression problem is solved. These types of transformations try to capture meaningful information from the input while reducing signal dimension. The feature transformations are usually based on intuition or some prior knowledge about the signals. Thus, in order to capture meaningful features, extensive domain knowledge about the problem is required. In addition, for traditional methods the feature extraction transformation is performed separately from the machine learning model responsible for the classification of the extracted features. Despite this rigid approach, traditional feature extraction schemes often retain a high degree of interpretability and have been shown to increase model performance [1]. Examples of such static feature extraction schemes include principal component analysis (PCA) [2], the discrete Fourier transform (see e.g. [3]) and various adaptive orthogonal transformations [4], [5].

More recently, neural network architectures relying on convolution layers to automatically find meaningful features have become popular. These machine learning (ML) models train the parameters of the feature extraction transformations (usually weights of a discrete convolution kernel) *together* with the parameters of the underlying ML model. The use of convolutional neural networks led to many novel highly efficient applications especially in image processing (see e.g. [6]). Automatic feature extraction allowed for the construction of end-to-end neural network based applications, and (at least from the point of view of performance metrics) reduced the need for domain knowledge. Interestingly, despite this property being one of the main advantages of deep learning, simultaneously it is also one of the most frequently cited limitations [7].

High performing convolutional models usually contain a large number of parameters, which are very difficult to interpret. This makes the use of deep learning problematic in a number of domains, such as medical signal processing and autonomous driving, where a degree of explainability is often required [7], [8]. These problems led to intensive research of explainable artificial intelligence (XAI) methods with the following purposes (see e.g. [8]).

- Explain the behavior of previously trained deep learning models. These types of approaches, known are usually referred to as Post-Hoc methods.
- Construct so-called *transparent* neural network models, where the learned parameters retain meaningful information.

Post-hoc approaches include layer-wise relevance propagation (LRP) [9] and integrated gradients (IG) [10], which have been successfully applied in a number of disciplines, for example medical signal processing [11].

Transparent methods often try to incorporate classical math-

emational transformations with interpretable parameters into ML architectures. Such ML models are also frequently referred to as model driven methods. In [12] and [13] the concepts of variable projection neural networks (VP-Net) and variable projection support vector machines (VP-SVM) were developed. These modified ML models propose the use of adaptive orthogonal projection operators instead of convolution layers for automatic feature extraction. If these transformations are chosen correctly (exploiting the characteristics of the problem at hand), then the learned parameters will have exact physical meaning (see e.g. [5] and [13]).

Another recent result used a variation of the continuous wavelet transform to automatically extract features in a neural network model known as WaveletKernelNet (WKN) [14]. The continuous wavelet transform [15] has been successfully used to solve a large number of signal processing problems (see e.g. [16], [17]). Wavelet coefficients are highly interpretable, as they provide time localised information about the scale (closely related to frequency) profile of signals. In [14] the connection between the wavelet transform and convolution operators was exploited to replace convolution layers in deep neural models with layers calculating wavelet coefficients.

In this study, a continuous wavelet transform layer is proposed, which, instead of discrete convolution, uses variable projection [4] as a way of approximating the wavelet coefficients. The resulting transparent neural network architecture will be referred to as continuous wavelet transform based variable projection network (CWT-VP). CWT-VP learns signal representations which rely on *only a few relevant wavelet coefficients*, thus providing a sparse representation of the input. The proposed layer allows for the construction of much simplified model architectures (compared to e.g. [14]), while improving interpretability. The example problem of bearing fault detection is investigated, where it is shown that in addition to the above mentioned benefits, the proposed model also outperforms similar methods in terms of performance metrics.

II. CONTINUOUS WAVELET TRANSFORM

Suppose $f, \psi \in L_2(\mathbb{R})$. The *wavelet transform* of f with respect to ψ is defined as

$$W_\psi f(x, s) := |s|^{-1/2} \int_{\mathbb{R}} f(t) \overline{\psi(s^{-1} \cdot (t - x))} dt, \quad (1)$$

where $(x, s \in \mathbb{R}, s \neq 0)$ and $\overline{\psi(t)}$ denotes the complex conjugate of $\psi(t)$. The wavelet transform is thus a function of translation x and dilation s parameters and is determined by the *wavelet* (also mother wavelet or analyzing wavelet) ψ (see e.g. [3] or [15]). For fixed x and s the value $W_\psi(x, s)$ will also be referred to as a wavelet coefficient. Henceforth, the dilated and translated mother wavelet $\psi(s^{-1} \cdot (t - x))$ will be denoted by $\psi_{x,s}(t)$. Notice, that for a fixed dilation s , one has

$$W_\psi f(x, s) = (f * \psi_{\cdot, s})(x),$$

where $*$ denotes the convolution operation. This observation was exploited in [14], where neural network layers producing wavelet coefficients were introduced using convolutions. In this work, instead of exploiting the relationship between wavelet transform and convolutions, another interpretation of wavelet coefficients is considered. Namely,

$$W_\psi f(x, s) := \langle f, \psi_{x,s} \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the $L_2(\mathbb{R})$ inner product (see e.g. [3]). Thus, the wavelet transform of f can be interpreted as the coordinates of the projection of f onto the subspace spanned by all dilations and translations of the mother wavelet ψ [15].

In applications, several algorithms have been developed to calculate the wavelet coefficients (1). It is important to note, that in applications, one only has access to $\mathbf{f} \in \mathbb{R}^N$, an N -point sampling of f :

$$\begin{aligned} \mathbf{f}_i &:= f(t_i), \\ (t_0, \dots, t_{N-1} \in \mathbb{R}, t_j &\neq t_k (j \neq k), N \in \mathbb{N}). \end{aligned} \quad (2)$$

This means that in practice, the coefficients (1) may only be approximated. If, for a fixed mother wavelet ψ a set of discrete dilation s_k and translation x_j parameters exist such that the function system $\psi_{s_k, x_j}(k, j \in \mathbb{Z})$ forms an orthogonal system in $L_2(\mathbb{R})$, then ψ is referred to as an orthogonal wavelet [3]. The discrete wavelet transform (DWT) uses orthogonal wavelets and can efficiently calculate wavelet coefficients through the iterative application of digital filters (see e.g. [18]). The construction of orthogonal wavelets however is difficult, especially if the resulting mother wavelet should depend on learnable parameters. Furthermore, discrete wavelet coefficients are not translation invariant, which makes their application difficult when analysing time dependent signals.

For this reason, the so-called continuous wavelet transform (CWT) (see e.g. [15]) is also widely used. CWT has the advantage of retaining the above mentioned translation invariant property, however usually signal reconstruction requires a redundant representation. That is, $M \gg N$ wavelet coefficients are usually needed to invert the transformation. Here, N denotes the number of sampling points of f . CWT approximates the coefficients (1) using numerical methods. In turn, CWT is not restricted to a predefined set of s_k and x_j parameters, therefore it can give a finer representation of signal details. Another big advantage of CWT is that it can be used with non-orthogonal wavelets, which usually exhibit a number of desirable traits such as smoothness and symmetry. Several popular CWT implementations exist (see e.g. [15] and [19]). In this work, methods are constructed, where features are extracted from the input signals by calculating only a few CWT-like approximations of the wavelet coefficients. The proposed methods learn which wavelet coefficients (given by $(s, x) \in \mathbb{R} \setminus \{0\} \times \mathbb{R}$) best capture information about the signal needed to solve an underlying classification problem.

III. CWT-VP NETWORKS

Suppose that the input data is available in the form given in Eq. (2). Feature extraction schemes usually aim to represent $\mathbf{f} \in \mathbb{R}^N$ with $n \ll N$ points. This can be achieved for example by the so-called variable projection operator [4]

$$\mathbf{f} \approx P_{\Psi(\boldsymbol{\eta})} \mathbf{f} = \Psi(\boldsymbol{\eta}) (\Psi(\boldsymbol{\eta})^+ \mathbf{f}), \quad (3)$$

where $\Psi(\boldsymbol{\eta}) \in \mathbb{R}^{N \times n}$, $\boldsymbol{\eta} \in \mathbb{R}^m$ ($m \in \mathbb{N}$) and $\Psi(\boldsymbol{\eta})^+$ is the Moore-Penrose pseudoinverse of $\Psi(\boldsymbol{\eta})$. Eq. (3) defines the projection of $\mathbf{f}$ onto the column space of $\Psi(\boldsymbol{\eta})$. The columns of this matrix are usually chosen as discrete samplings of an $L_2(\mathbb{R})$ basis that depends analytically on the parameters $\boldsymbol{\eta}$. For example, this framework has been successfully used with the orthogonal Hermite function system to extract features from biological signals [20], [21].

Suppose that the derivatives $\partial \Psi(\boldsymbol{\eta}) / \partial \boldsymbol{\eta}$ exist. Under this assumption, Golub and Pereyra gave an explicit formula for the Jacobian of $\Psi(\boldsymbol{\eta})^+$ in [22]. Using this formula it is possible to calculate the Jacobian of the projection $P_{\Psi(\boldsymbol{\eta})}$ from Eq. (3) analytically. This idea was used by Kovács et al. in [12] to construct VP-Net, a neural network architecture which included layers implementing projections of the type (3) for feature extraction. Later this idea was extended to support vector machines in [13].

The main contribution of this work is the approximation of the wavelet coefficients (1) using the variable projection technique to acquire CWT layers. In particular, let $m = n$ and consider the parameter vector $\boldsymbol{\eta} := (s_1, x_1, s_2, x_2, \dots, s_n, x_n) \in \mathbb{R}^{2n}$, with $s_k > 0$, ($k = 1, \dots, n$). If the matrix $\Psi(\boldsymbol{\eta})$ is constructed such that

$$\Psi(\boldsymbol{\eta})_{k,j} = \psi_{s_j, x_j}(t_k) = \frac{1}{\sqrt{s_j}} \psi \left(\frac{t_k - x_j}{s_j} \right),$$

then the j -th component of $\mathbf{w} := \Psi(\boldsymbol{\eta})^+ \mathbf{f} \in \mathbb{C}^n$ approximates $W_\psi(s_j, x_j)$ defined in Eq. (1).

The proposed variable projection continuous wavelet transform layer (VP-CWT) can now be introduced as:

$$L^\eta : \mathbb{R}^N \rightarrow \mathbb{R}^n, \quad L^\eta(\mathbf{f}) := \Psi(\boldsymbol{\eta})^+ \mathbf{f}. \quad (4)$$

Consider a neural network, whose first layer performs the automatic continuous wavelet transform based feature extraction step given (4). In order to be able to train the parameters $\boldsymbol{\eta}$ together with the model parameters it suffices to assume (see [12] and [13]) that the derivatives $\partial \psi_{s_j, x_j} / \partial s_j$ and $\partial \psi_{s_j, x_j} / \partial x_j$ exist for every $j = 1, \dots, n$. In this work, the Morlet wavelet was considered. These wavelets are given by

$$\psi_{s,x}(t) := \frac{1}{\sqrt{s}} \cdot e^{-\left(\frac{t-x}{s}\right)^2/2} \cdot \cos \left(\omega_0 \cdot \frac{t-x}{s} \right), \quad (5)$$

for $s > 0, x, \omega_0 \in \mathbb{R}$. The needed partial derivatives can be expressed as $\frac{\partial \psi_{s,x}(t)}{\partial s} = -\left(\frac{1}{2s}\psi_{s,x}(t) + a(t)\frac{\partial a(t)}{\partial s}\right)\psi_{s,x}(t) - \omega_0 s^{-1/2} \frac{\partial a(t)}{\partial s} e^{-a^2(t)/2} \sin(\omega_0 a(t))$ and $\frac{\partial \psi_{s,x}(t)}{\partial x} =$

$$-a(t)\frac{\partial a(t)}{\partial x}\psi_{s,x}(t) - s^{-1/2}\omega_0\frac{\partial a(t)}{\partial x}e^{-a^2(t)/2}\sin(\omega_0 a(t)),$$

where

$$a(t) = a_{s,x}(t) := \frac{t-x}{s}, \quad \frac{\partial a}{\partial s}(t) = -\frac{t-x}{s^2}, \quad \frac{\partial a}{\partial x}(t) = -\frac{1}{s}.$$

It is important to mention that the proposed method is applicable with any other mother wavelet as long as the above mentioned partial derivatives exist.

Consider a labeled dataset $\mathcal{S} := (\mathbf{f}_i, y_i)_{i=1}^q$ ($q \in \mathbb{N}$), where $\mathbf{f}_i \in \mathbb{R}^N$ and $y_i \in \{0, 1\}$. Although, for simplicity, a binary classification problem is considered here, it should be noted that the proposed method can be applied to multiclass and regression problems in a straightforward manner. In [12], the term

$$J_{VP}(\boldsymbol{\eta}) := \frac{\alpha}{q} \sum_{i=1}^q \frac{\|\mathbf{f}_i - P_{\Psi(\boldsymbol{\eta})} \mathbf{f}_i\|_2^2}{\|\mathbf{f}_i\|_2^2},$$

was introduced for use with variable projection based neural networks, where the data examples $\mathbf{f}_i$ come from the dataset $\mathcal{S}$, $\alpha > 0$ is a hyperparameter and L^η denotes the VP-CWT layer (4). Denote a fully connected neural network employing the proposed VP-CWT layer by $NN_\theta : \mathbb{R}^N \rightarrow \{0, 1\}$, where the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^M$ contains every parameter of the network (including the $\boldsymbol{\eta}$ parameters of the VP-CWT layer (4)). Suppose that NN_θ is trained with respect to the binary cross entropy loss function on $\mathcal{S}$:

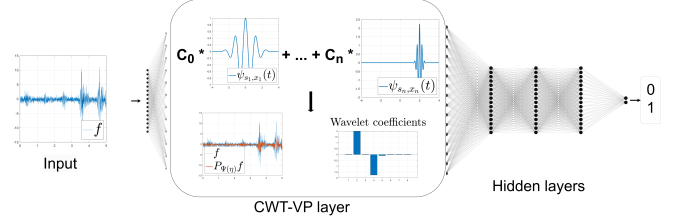


Fig. 1. The proposed continuous wavelet transform architecture.

$$J_{BCE}(\boldsymbol{\theta}) := -\frac{1}{q} \sum_{i=1}^q (y_i \log NN_\theta(\mathbf{f}_i) + (1 - y_i) \cdot \log(1 - NN_\theta(\mathbf{f}_i))), \quad ((\mathbf{f}_i, y_i) \in \mathcal{S}).$$

Then, as suggested for neural networks using variable projection layers in [12], the regularized loss function

$$J(\boldsymbol{\theta}) := J_{BCE}(\boldsymbol{\theta}) + J_{VP}(\boldsymbol{\eta}), \quad (6)$$

should be used to avoid the problem of vanishing and exploding gradients in the variable projection layer. The gradient of J_{VP} with respect to the wavelet coefficient parameters can be analytically calculated using the formulas given in [12]. A schematic representation of a neural network using the proposed variable projection based continuous wavelet transform is given in figure 1.

IV. COMPARISON TO THE STATE OF THE ART

As mentioned in section II, in [14] a neural network architecture employing the continuous wavelet transform for feature extraction has already been proposed. In that work, the relationship between wavelet coefficients (1) and convolutions was exploited. In particular, the WaveletKernelNet (WKN) network proposed in [14] uses the (discrete) convolutions

$$W_\psi f(\cdot, s_i) \approx \mathbf{w} := \mathbf{f} * \psi_{s_i, x_i} \quad (7)$$

to approximate wavelet coefficients. The WKN model learns the dilation $s_i > 0$ and initial time shift $x_i \in \mathbb{R}$ parameters for $i = 1, \dots, n$. Then, for each (s_i, x_i) pair, wavelet coefficients were approximated for all possible translations using the above convolution operation. In turn, for each parameter pair, the input signal $\mathbf{f}$ was represented by M approximated wavelet coefficients, where M denotes the dimension of $\mathbf{w}$ in (7). If a total of $n \in \mathbb{N}$ wavelets are used, this yields $M \cdot n$ wavelet coefficients to represent the input signals. For some applications, this high resolution signal representation may be necessary (see e.g. [14]). In section V however, a real life fault detection problem is presented, where the proposed VP-CWT network outperforms the WKN approach despite relying on fewer wavelet coefficients. As described in section III, the VP-CWT layer represents input signals with exactly n wavelet coefficients. This sparse representation allows for better interpretability of the learned parameters, as well as a simplified network model. In fact, VP-CWT layers do not need to be followed by pooling operations to reduce the output features' dimension, which provides simpler model architecture and improves explainability.

V. EXPERIMENTS

In the maintenance of mechanical systems, it is important to predict failures of elements as early as possible to protect the other components or the whole system from significant damage. In this section, the proposed CWT-VP neural network is used to solve a fault detection problem for bearing components. In addition, a comparison is made between the proposed method and WKN [14]. Acceleration signals needed to train the different models were acquired from the MFPT challenge open source rolling element bearing fault database [23].

A rolling bearing facilitates the transfer of forces between stationary and rotating elements through rolling elements. The primary advantage of a rolling bearing lies in its reduction of rolling friction compared to sliding friction. Figure 2 shows the structural components of a typical bearing, illustrating common faults that may occur within its assembly.

The MFPT Challenge dataset comprises 23 measurements gathered from machines operating under diverse fault conditions. These measurements consist of vibration acceleration signals. The initial 20 datasets originate from a bearing test rig, including 3 measurements representing good conditions, 3 with outer race faults under constant load, 7 with outer race faults under varying loads, and 7 with inner race faults under different loads. The remaining 3 measurements are

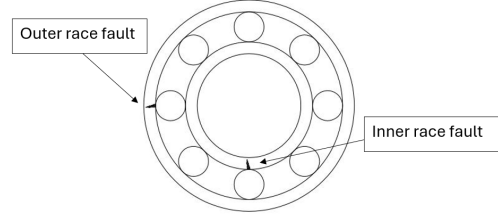


Fig. 2. Rolling element bearing cross-section with faults.

derived from real-world machines: an oil pump bearing, an intermediate speed bearing, and a planet bearing. The fault locations in the datasets are unknown.

For these experiments, data was obtained under good (baseline) conditions with a load of 270 lbs, along with inner and outer race conditions data collected under a similar load of 250 lbs.

Set	Layer	# of cfs.	Mean Acc.	Highest	Lowest
baseline-outer	CWT-VP	32	96%	100%	88.89%
	WKN	124160	92.89%	100%	64.45%
baseline-inner	CWT-VP	24	96.89%	100%	91.11%
	WKN	124160	95.11%	100%	75.56%
inner-outer	CWT-VP	32	96.67%	100%	90%
	WKN	124160	96%	100%	80%

TABLE I
CLASSIFICATION RESULTS OF THE BEARING DATASET

Three sets of data were created for training and testing. The preprocessing steps suggested in [24] were employed. In total 150 normal, 75 inner fault and 75 outer fault data examples were obtained. Each set contained two types of data: baseline (good condition) and outer race fault, baseline and inner race fault, and inner and outer race fault, respectively. The training was conducted using the best training parameters for each test setup. The datasets were split into 80% for training and 20% for testing with 5-fold cross-validation, which consisted of stratified folds to preserve the percentage of samples for each class.

Table I contains classification accuracy results for each of the above mentioned test scenarios. It is clear that both the proposed CWT-VP and WKN architectures could differentiate between the classes with high accuracy. The CWT-VP model consisted of a single wavelet layer (4) followed by three linear layers with 50 neurons each. ReLU activation functions were applied between these layers, with a sigmoid function used for binary classification. In order to measure the difference between the two wavelet layers, the WKN model used the same architecture, except for the first layer, where the cwt layer suggested in [14] was used. The implementation [25] was used to construct the WKN model. The hyperparameters of both models were tuned using a trial and error approach, with

the best achieved results displayed in table I for each model and each experiment. The WKN model seemed to achieve the best performance when used with the hyperparameter setup suggested in [14]. In this setup, a total of 124160 wavelet coefficients were calculated for each input signal, using a total of 64 learnable dilation and translation parameters. In contrast, the best performing VP-CWT models only required 24 – 32 wavelet coefficients to achieve similar results. This sparsity allows for more interpretability. In addition, the VP-CWT approach seemed less prone to overfitting achieving around 90% accuracy on the worst fold for each experiment. In summary, the proposed model performed well during the conducted experiments, in some cases significantly outperforming the WKN model despite relying on considerable sparser signal representations.

VI. CONCLUSION

In this study a novel variable projection based numerical approximation of the continuous wavelet transformation was proposed for use with neural networks. The proposed transformation can be used to represent input signals with only a few wavelet coefficients, thus the learned parameters describing the coefficients contain interpretable information. In fact, the VP-CWT features represent signal components at specific scales and translations learned from the data, carrying relevant information for the subsequent classifier. The use of the variable projection formulation allows for the construction of transparent neural network architectures, where the parameters of the VP-CWT layer are trained together with the weights of the underlying ML model.

Several concrete objectives will be pursued to continue this research. These include the adaptation of the proposed VP-CWT layer for non-neural network based models [13], as well as considering a wider array of mother wavelets. In addition, other applications (e.g. medical signal processing) should also be considered.

ACKNOWLEDGMENT

Project no. C1748701, K146721, TKP2021-NVA-29 have been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under NVKDP-2021, K_23 "OTKA", and TKP2021-NVA-03 funding schemes, respectively. Support by the the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. P. Kovács was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

REFERENCES

- [1] I. Guyon and A. Elisseeff, *An Introduction to Feature Extraction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–25.
- [2] M. Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 100, 2022.
- [3] C. Gasquet and P. Witomski, *Fourier analysis and applications: filtering, numerical computation, wavelets*. Springer Science & Business Media, 2013, vol. 30.
- [4] G. Golub and V. Pereyra, "Separable nonlinear least squares: the variable projection method and its applications," *Inverse problems*, vol. 19, no. 2, p. R1, 2003.
- [5] T. Dózsa, J. Radó, J. Volk, Á. Kisari, A. Soumelidis, and P. Kovács, "Road abnormality detection using piezoresistive force sensors and adaptive signal models," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.
- [6] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.
- [7] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital signal processing*, vol. 73, pp. 1–15, 2018.
- [8] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
- [9] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS one*, vol. 10, no. 7, p. e0130140, 2015.
- [10] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.
- [11] T. Bender, J. M. Beinecke, D. Krefting, C. Müller, H. Dathe, T. Seidler, N. Spicher, and A.-C. Hauschild, "Analysis of a deep learning model for 12-lead ecg classification reveals learned features similar to diagnostic criteria," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2023.
- [12] P. Kovács, G. Bognár, C. Huber, and M. Huemer, "Vpnet: variable projection networks," *International Journal of Neural Systems*, vol. 32, no. 01, p. 2150054, 2022.
- [13] T. Dózsa, F. Deuschle, B. Cornelis, and P. Kovács, "Variable projection support vector machines and some applications using adaptive hermite expansions," *International Journal of Neural Systems*, vol. 34, no. 1, 2024.
- [14] T. Li, Z. Zhao, C. Sun, L. Cheng, X. Chen, R. Yan, and R. X. Gao, "Waveletkernelnet: An interpretable deep neural network for industrial intelligent diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 4, pp. 2302–2312, 2021.
- [15] D. Jordan, R. Miksad, and E. Powers, "Implementation of the continuous wavelet transform for digital time series analysis," *Review of scientific instruments*, vol. 68, no. 3, pp. 1484–1494, 1997.
- [16] H. Goelz, R. D. Jones, and P. J. Bones, "Continuous wavelet transform for the detection and classification of epileptiform activity in the eeg," in *Proceedings of the First Joint BMES/EMBS Conference. 1999 IEEE Engineering in Medicine and Biology 21st Annual Conference and the 1999 Annual Fall Meeting of the Biomedical Engineering Society (Cat. N*, vol. 2. IEEE, 1999, pp. 941–vol.
- [17] A. R. Ismail and S. S. Asfour, "Continuous wavelet transform application to emg signals during human gait," in *Conference Record of Thirty-Second Asilomar Conference on Signals, Systems and Computers (Cat. No. 98CH36284)*, vol. 1. IEEE, 1998, pp. 325–329.
- [18] D. Sundararajan, *Discrete wavelet transform: a signal processing approach*. John Wiley & Sons, 2016.
- [19] M. Unser, A. Aldroubi, and S. J. Schiff, "Fast implementation of the continuous wavelet transform with integer scales," *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3519–3523, 1994.
- [20] A. Sandryhaila, S. Saba, M. Puschel, and J. Kovacevic, "Efficient compression of qrs complexes using hermite expansion," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 947–955, 2011.
- [21] T. Dózsa and P. Kovács, "Ecg signal compression using adaptive hermite functions," in *ICT Innovations 2015: Emerging Technologies for Better Living 7*. Springer, 2016, pp. 245–254.
- [22] G. H. Golub and V. Pereyra, "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate," *SIAM Journal on numerical analysis*, vol. 10, no. 2, pp. 413–432, 1973.
- [23] E. Bechhoefer, "Condition based maintenance fault database for testing diagnostics and prognostic algorithms," <https://www.mfpt.org/fault-datasets/>, 2013.
- [24] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—a tutorial," *Mechanical systems and signal processing*, vol. 25, no. 2, pp. 485–520, 2011.
- [25] L. Tianfu, "Waveletkernelnet," <https://github.com/HazeDT/WaveletKernelNet/>, 2020.