NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 16 (2026), 1766941
Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences
Araştırma makalesi / Research article
www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

# Effect of wavelet family selection on transformer health index prediction

## Transformatör sağlık endeksi tahmininde dalgacık ailesi seçiminin etkisi

**Kübra Nur Akpınar[1,*], Seçil Genç[2], Barış Çavuş[3]**

[1] *Marmara University, Vocational School of Technical Sciences, Department of Electrical and Energy, 34865, İstanbul,Türkiye*
[2,3] *Ondokuz Mayıs University, Electrical and Electronics Engineering Department, 55200, Samsun, Türkiye*

**Abstract**

In transmission systems, power transformers are key high-value components, and their consistent operation is fundamental for maintaining grid stability and achieving cost-effective performance. The Transformer Health Index (THI) integrates key diagnostic parameters including dissolved gas analysis, water content, oil quality indicators, and power factor providing essential insights for asset condition assessment and investment planning. In this study, THI prediction is conducted using the Random Forest algorithm, recognized in literature for its high predictive accuracy for transformer applications, in combination with data preprocessing and filtering techniques applied to transformer dataset. For the first time, to the best of our knowledge in the THI prediction literature, various wavelet families are systematically compared at the preprocessing stage to examine their influence on predictive accuracy. The results show that the Symlet-2 configuration consistently outperformed other families in both filtered and non-filtered datasets, while Coiflet-3 and Coiflet-5 achieved higher efficiency through dimensionality reduction but with an accuracy decrease of approximately 0.09–0.10 in $R^2$ compared to Symlet-2. The findings demonstrate that the choice of wavelet family in the preprocessing phase directly impacts feature selection outcomes and model performance, offering valuable guidance for the development of high-accuracy transformer condition assessment frameworks.

**Keywords:** Transformer Health Index, Wavelet families, Pearson correlation, Random Forest, Condition assessment.

**Öz**

Güç transformatörleri, iletim şebekelerinin kritik ve yüksek maliyetli bileşenleri olup, güvenilir çalışmaları şebeke kararlılığı ve ekonomik verimlilik açısından büyük önem taşımaktadır. Transformatör Sağlık Endeksi (TSE) çözünmüş gaz analizi, yağ kalitesi göstergeleri, su içeriği ve güç faktörü gibi temel teşhis parametrelerini bir araya getirerek, varlık durumu değerlendirmesi ve yatırım planlaması için önemli bilgiler sunmaktadır. Bu çalışmada, literatürde transformatör uygulamalarında yüksek tahmin doğruluğu ile bilinen Rastgele Orman algoritması; transformatör veri kümesi üzerinde, veri ön işleme ve filtreleme teknikleri ile kullanılarak TSE tahmini gerçekleştirilmiştir. Yazarların bilgisine göre, literatürde ilk kez, TSE tahmininde ön işleme aşamasında farklı dalgacık aileleri sistematik olarak karşılaştırılmış ve tahmin doğruluğu üzerindeki etkileri incelenmiştir. Sonuçlar, Symlet-2 konfigürasyonunun hem filtreli hem de filtresiz veri setlerinde diğer ailelere kıyasla tutarlı biçimde en yüksek performansı sağladığını, Coiflet-3 ve Coiflet-5'in ise boyut indirgeme yoluyla daha yüksek hesaplama verimliliği elde ederken Symlet-2'ye kıyasla $R^2$ değerinde yaklaşık 0,09–0,10'luk bir düşüş yaşadığı görülmüştür. Bulgular, önişleme aşamasında seçilen dalgacık ailesinin hem özellik seçimi sonuçlarını hem de model performansını doğrudan etkilediğini ortaya koymakta olup yüksek doğruluklu transformatör durumu değerlendirme yaklaşımlarının geliştirilmesine yönelik değerli çıkarımlar sunmaktadır.

**Anahtar kelimeler:** Transformatör sağlık endeksi, Dalgacık aileleri, Pearson korelasyonu, Rastgele Orman, Durum izleme

## 1 Introduction

As high-value and essential components of power systems, power transformers must operate continuously and reliably to maintain the safety, stability, and efficiency of the network. [1]. Unexpected failures can cause prolonged outages, high replacement costs, and severe cascading impacts on the grid. Consequently, modern asset management strategies increasingly depend on accurate diagnostics and prognostics to optimize maintenance schedules, extend service life, and prevent severe breakdowns [2–4]. The Transformer Health Index (THI) has emerged as an effective calculation method, integrating diagnostic inputs such as dissolved gas analysis (DGA), water content, oil quality assessments, and insulation degradation indicators into a single quantitative value [3, 4]. This enables utility operators to prioritize maintenance, assess operational risks, and plan asset replacement in a structured, data-driven manner.

THI estimation techniques have evolved from rule-based and weighted scoring systems to advanced computational approaches, including statistical models, fuzzy logic, and machine learning algorithms encompassing Random Forest (RF), Support Vector Machines (SVM), k-Nearest Neighbors (kNN), Artificial Neural Networks (ANN), and

* Sorumlu yazar / Corresponding author, e-posta / e-mail: kubraaakpinar@gmail.com (K. N. Akpınar)
Geliş / Received: 16.08.2025 Kabul / Accepted: 08.04.2026 Yayımlanma / Published: 14.04.2026
doi: 10.28948/ngumuh.1766941

Convolutional Neural Networks (CNN) [5–10]. Hybrid and optimization frameworks like LightGBM with Robust Expectation-Maximization (EM) for missing data imputation have reported error reductions exceeding 70% [11], while deep generative and multi-task LSTM-GRU models have been deployed within digital twin frameworks to jointly predict THI and remaining useful life [4,12]. In addition, feature selection techniques have been shown to significantly reduce computational cost without compromising prediction accuracy [10,13]. Despite these advances most studies process measurements with limited attention to signal preprocessing especially for small and medium size datasets [14].

Wavelet analysis has demonstrated strong potential in feature extraction in multi engineering domains such as mechanical fault detection [15,16], hydrological forecasting [17], and wind power prediction [18]. In [19], the mother wavelet Daubechies 1 (db1) was adopted for its strong performance in detecting sudden signal transitions and its applicability in differential protection. While power transformer related studies have predominantly focused on the Daubechies family, the influence of wavelet choice on simulation results underscores this selection. However, alternative families such as Symlets, Coiflets, Biorthogonal, Reverse Biorthogonal, and Haar have received limited attention [20]. This study addresses that gap by systematically comparing multiple wavelet families for THI estimation and integrating a Pearson correlation based adaptive thresholding method for feature selection, followed by performance evaluation using the RF model. In contrast to earlier studies that relied solely on DGA data and a single wavelet family, the proposed framework incorporates additional datasets and evaluates both the widely used Daubechies and other wavelet families. Model performance is assessed through Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and the coefficient of determination ($R^2$) metrics. By integrating multi-family wavelet preprocessing with adaptive feature selection, this work provides a scalable and noise-tolerant THI prediction method, addressing a clear gap in the transformer asset management literature. This study contributes to the literature by (i) examining different configurations within a selected wavelet family at the preprocessing stage, (ii) providing one of the first systematic evaluations of wavelet family effects on THI prediction accuracy, (iii) feature selection with Pearson–Kneedle integration, and (iv) demonstrating how preprocessing decisions directly affect predictive performance and asset management decisions.

This paper is organized as follows: Section 2 introduces the dataset and preprocessing procedures including wavelet-based denoising. Section 3 presents the feature selection approach and the RF-based prediction model. Section 4 reports the experimental results and comparative performance analysis. Section 5 discusses the results for transformer asset management. Finally, Section 6 concludes the study and outlines future research and limitations.

## 2 Materials and methods

The methodology employed in this study for THI prediction comprises feature extraction with wavelet families, feature selection with filter (Pearson) based method, model training with RF, and performance evaluation, as illustrated in Figure 1. All analyses and model implementations were performed in MATLAB.

The process begins with importing a publicly available transformer dataset to guarantee reproducibility and facilitate comparison with prior research. In the wavelet feature extraction stage, multiple mother wavelet families including Daubechies, Symlets, Coiflets, Biorthogonal, Reverse Biorthogonal, and Haar are applied to capture a range of time–frequency localization properties. This step enhances signal quality by reducing noise while preserving diagnostically relevant features.

Subsequently, feature selection is performed using the Pearson correlation coefficient in combination with the Kneedle algorithm. This hybrid method adaptively determines the correlation threshold, enabling the most informative features. The dataset is divided into 70% for training and 30% for testing, with robustness evaluated via a 5-fold cross-validation procedure.

The RF model which is selected for its robustness, ability to handle nonlinear relationships, and proven performance in THI prediction is trained on the selected features. Model performance is then evaluated using RMSE, MAE, MAPE, and $R^2$, providing a comprehensive view of both absolute and relative prediction accuracy. Finally, a comparative analysis of wavelet families quantifies their impact on model performance, ensuring that the most effective preprocessing configuration is identified for THI estimation under realistic operating conditions.

### *2.1 Transformer data*

The dataset used for THI prediction was obtained from a publicly available platform [21]. Public datasets are widely used in power transformer health assessment research to ensure reproducibility across different modelling approaches [3,5,9]. Unlike datasets containing only DGA measurements, this dataset includes additional diagnostic parameters relevant to transformer condition assessment, enabling a more comprehensive evaluation of asset health [2,10].

It comprises 470 records and 14 variables: Hydrogen, Oxygen, Nitrogen, Methane, CO, $CO_2$, Ethylene, Ethane, Acetylene, Dibenzyl Disulfide (DBDS), Water Content, Power Factor, Interfacial Voltage (InterfacialV), Dielectric Rigidity, and THI. DGA related gases such as Hydrogen, Methane, and Ethylene are widely used to detect thermal and electrical faults [3,14]. Parameters like DBDS associated with copper corrosion and insulation aging along with power factor and InterfacialV, provide critical insights into oil quality and insulation condition [13]. Dielectric rigidity and water content are directly linked to insulation breakdown risk offering essential information for preventive maintenance [17].

In contrast to many studies in the literature, where datasets are often restricted to DGA or a small subset of oil quality parameters, the chosen dataset combines chemical,

electrical, and physical indicators into a single framework [3,10,13,14,17]. This diversity enables the proposed methodology to assess the relative importance of a broader range of features and to evaluate the impact of multi-family wavelet denoising on THI prediction performance. Although the dataset is tabular, the samples are ordered according to THI and remaining life. This ordering introduces a degradation-based axis, allowing wavelet denoising to be applied in a physically meaningful manner.

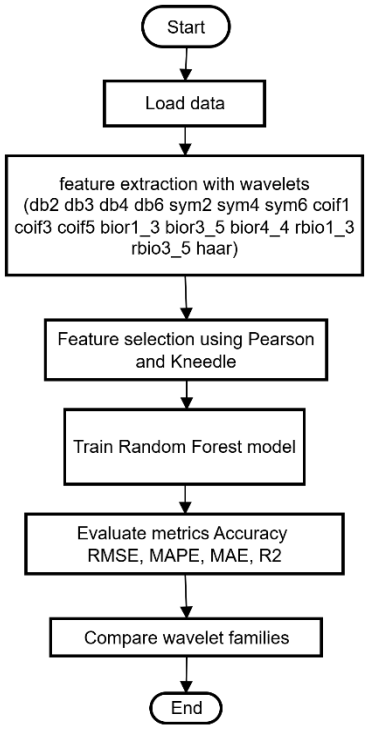


**Figure 1.** Flowchart of the THI prediction method

### *2.2 Preprocessing: Multi-Family wavelet denoising*

In signal processing and analysis, different wavelet families offer distinct mathematical structures and functional characteristics, making them suitable for diverse applications. For instance, Symlets provide near-symmetric waveforms that reduce phase distortion, Coiflets offer higher-order vanishing moments for improved frequency localization, and Haar wavelets enable computationally efficient [22]. The choice of mother wavelet has been shown to significantly affect predictive accuracy in various domains, such as financial time series forecasting and optical imaging noise reduction [23-24].

To enhance the quality of input features and suppress measurement noise prior to THI prediction, this study adopts a multi-family wavelet denoising approach. The methodology involves a systematic assessment of various mother wavelet families with distinct time–frequency localization characteristics, seeking to preserve diagnostically important signal features to enhance the robustness and generalizability of the ensuing machine learning models.

The Discrete Wavelet Transform (DWT) is a multi-resolution analysis technique that decomposes an ordered feature sequence $x[n]$ into approximation and detail coefficients through low-pass and high-pass filtering, followed by down-sampling by a factor of two [25, 26]. At decomposition level $j$, the approximation coefficients $A_j[n]$ and detail coefficients $D_j[n]$ are obtained in Equation (1) and (2).

$$A_j[n] = \sum_k h[k-2n]\, A_{j-1}[k] \tag{1}$$

$$D_j[n] = \sum_k g[k-2n]\, A_{j-1}[k] \tag{2}$$

Here, $A_{j-1}[k]$ are the approximation coefficients from the previous level. $h$ is the low-pass filter associated with the scaling function $\varphi(t)$, and $g$ is the high-pass filter associated with the wavelet function $\psi(t)$. In this paper the index n does not represent time; it denotes the ordering of samples along the transformer degradation axis.

Figure 2 illustrates the multi-level wavelet decomposition process, where the original signal $S$ is iteratively split into approximation ($cAj$) and detail ($cDj$) coefficients at each level [26].

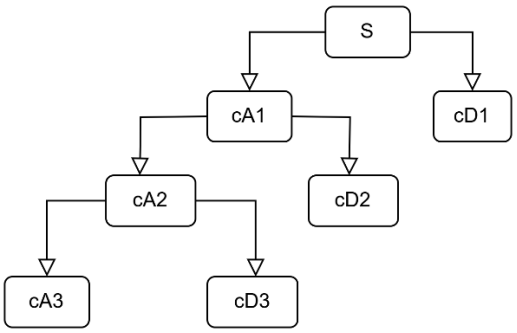


**Figure 2.** Wavelet Decomposition (c: coefficients) [26]

The following formulation is provided for theoretical completeness; in this study, the wavelet basis functions are applied to degradation-ordered diagnostic feature sequences rather than time-domain signals. The scaling function $\varphi(t)$ corresponds to the low-frequency, or approximation, component of the signal and meets the two-scale relation given in Equation (3).

$$\varphi(t) = \sum_{n=0}^{N-1} h_n\, \varphi(2t-n) \tag{3}$$

The wavelet function $\psi(t)$ represents the high-frequency (detail) part of the signal (Equation (4)) and is defined in terms of $\varphi(t)$ (Equation (5)) as:

$$\psi(t) = \sum_{n=0}^{N-1} g_n\, \varphi(2t-n) \tag{3}$$

$$g_n = (-1)^n h_{N-1-n} \quad (4)$$

In multi-resolution analysis, scaled and shifted versions of $\varphi(t)$ and $\psi(t)$ form the basis functions showed in Equation (6) and (7):

$$\varphi_{j,k}(t) = 2^{-j/2}\varphi(2^{-j}t - k) \quad (5)$$

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k) \quad (6)$$

For orthogonal wavelets, the filters $h_n$ and $g_n$ satisfy the quadrature mirror relationship in Equation (5), ensuring that approximation and detail coefficients contain distinct frequency content [27].

This study tests several wavelet families, taking advantage of their unique mathematical characteristics and their capability to represent signals in both time-frequency domains, including Daubechies (dbN) [27] Symlets (symN) [28], Coiflets (coifN) [29], Biorthogonal and Reverse Biorthogonal [30], and Haar [31]. Wavelet denoising is applied independently to each diagnostic feature, rather than jointly across multiple variables, ensuring that no cross-feature dependencies are introduced during preprocessing. Selecting the mother wavelet whose characteristics best match the spectral and statistical properties of transformer signals ensures diagnostically important features are preserved.

### *2.3 Feature selection*

Feature selection was carried out using a combined Pearson correlation and Kneedle algorithm approach to identify the most relevant variables for THI prediction. The Pearson correlation coefficient measures the linear relationship between THI and condition parameters such as DGA and oil quality factors, where $X$ represents the selected condition parameter and $Y$ represents the THI, and is defined as:

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y} \quad (8)$$

Equation (8) defines the Pearson correlation, $r_{xy}$ a metric for assessing the magnitude and direction of the linear link between $X$ and $Y$. In this formulation, $cov(x,y)$ is the covariance, $\sigma_x$ and $\sigma_y$ correspond to the standard deviations of $X$ and $Y$. The coefficient veries between –1 to +1, with values near ±1, with extreme values indicating strong correlation and values around zero implying minimal or no linear relationship [32].

In this study, correlation coefficient values were calculated for all features, and their absolute magnitudes $|r_{xy}|$ were ranked in descending order to evaluate their association strength with THI. To identify the most suitable number of features, the Kneedle algorithm was applied to the sorted correlation values [33]. This algorithm identifies the elbow or knee point on the curve, representing the stage beyond which additional features contribute marginally to feature relevance and model stability rather than improving peak predictive accuracy. The method works by normalizing the curve, computing deviations obtained from the line joining the starting and ending points and selecting the index where this deviation is maximal.

By combining coefficients with the Kneedle algorithm, only the most informative variables were retained for the RF prediction model. This reduces feature dimensionality and improves model interpretability, establish findings in feature selection research [34]. The primary objective of the Pearson + Kneedle filtering step is to establish a more reliable and stable relationship between the input health indicators and the THI, rather than pursuing the highest possible accuracy on a single experimental scenario.

The threshold determined by the Kneedle algorithm is a data-driven quantity derived from the curvature of the ranked correlation distribution. Since this threshold is defined independently of model performance, it is intentionally not recalibrated based on prediction results in order to preserve methodological objectivity and reproducibility.

## 3 Machine learning model

As a tree-based ensemble technique, RF advances beyond individual Decision Trees (DT) by generating numerous DT models and integrating their outputs. Due to its ability to model nonlinear relationships, and low preprocessing requirements, RF has been widely used for THI prediction [13]. The ability to derive a proximity matrix from pattern similarity, while avoiding complex normalization or transformation, allows the model to be effectively applied to heterogeneous transformer diagnostic datasets [35].

Many studies show that RF often achieves the highest accuracy among machine learning models for THI prediction. For example, Wang et. al. [34] showed that RF provided better diagnostic accuracy than Support Vector Machines (SVM) in real transformer fault diagnosis. Senoussaoui et. al. in [36] also showed that RF performed better than J48 DT in classifying THI from 91 oil transformer samples. Other studies have also confirmed RF's predictive advantage. Rediansyah et. al. [37] tested seven AI methods, including kNN, SVM, AdaBoost, Naïve Bayes (NB), ANN, DT, and RF, on 504 units of 150 kV power transformers, and concluded that RF was the best-performing method with 97.3% accuracy in THI category classification. In another study, in THI prediction with different missing parameter replacement methods, RF reached the highest overall accuracy of 92% across multiple scenarios [38].

In this study, the RF model was selected as the main prediction method for THI estimation because of its proven success in transformer diagnostics as shown in earlier works as mentioned before. The model was trained on the preprocessed and feature selected dataset obtained after multi-family wavelet denoising and filter-based feature selection. Its ability to work with different types of features, and no strict need for normalization make it a strong choice for this task.. A fixed hyperparameter configuration was employed to balance predictive accuracy and generalization capability. Specifically, the number of trees was set to 100 and the minimum leaf size was fixed at 5, while the

remaining parameters were kept at their default MATLAB settings. Detailed performance results, including statistical comparisons for different wavelet families, are provided in the next section.

## 4 Results and discussion

As shown in Table 1, applying the RF model to the preprocessed datasets resulted in the symlet (sym2) wavelet family delivering the highest prediction accuracy, with an $R^2$ value of 0.879, RMSE of 6.129, and MAE of 3523. This strong performance can be explained by the sym2 wavelet's balanced multi-scale representation, which effectively removes high frequency noise while preserving key patterns in the data. Such capability is particularly important for THI prediction, where maintaining sharp transitions and essential signal features supports the identification of faults or degradation trends. In contrast, wavelet families like bior1.3 and rbio1.3 achieved noticeably lower accuracy ($R^2$ = 0.161 and 0.537, respectively). The lower performance is likely due to excessive smoothing, which removes useful information alongside noise, or minor distortions to the signal shape both of which reduce the quality of the input data and limit the RF model's predictive capability.

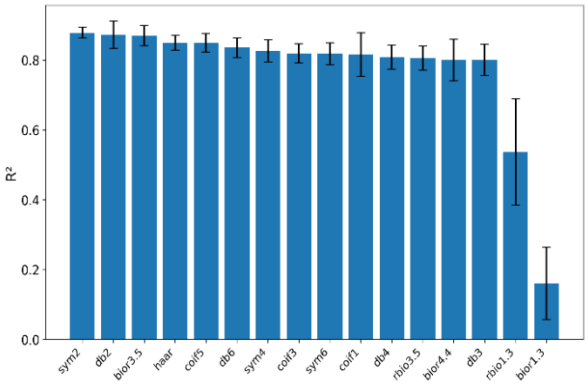


**Figure 3.** $R^2$ across wavelet families (Mean ± Std)

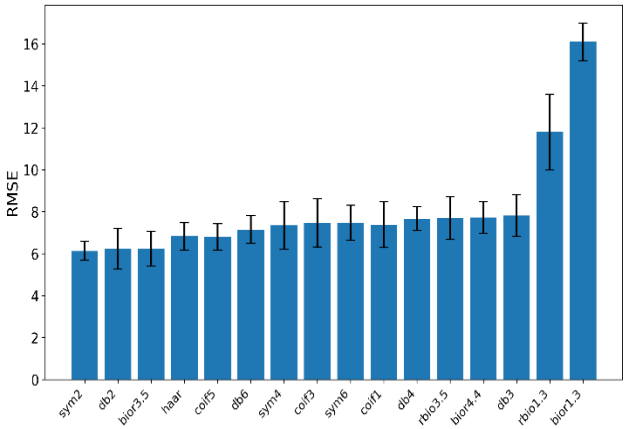


**Figure 4.** RMSE across wavelet families (Mean ± Std)

Figure 3 and 4 present the mean ± standard deviation of $R^2$ and RMSE values obtained from 5-fold cross-validation for different wavelet families. The error bars indicate the variability across folds, highlighting not only predictive accuracy but also model stability.

**Table 1.** Performance comparison of the random forest model for all wavelet families

| Wavelet | RMSE (Mean ± Std) | MAE (Mean ± Std) | MAPE (%) (Mean ± Std) | $R^2$ (Mean ± Std) |
|---|---|---|---|---|
| sym2 | 6.129 ± 0.453 | 3.523 ± 0.350 | 14.18 ± 2.72 | 0.879 ± 0.015 |
| db2 | 6.232 ± 0.962 | 3.631 ± 0.586 | 14.61 ± 2.91 | 0.873 ± 0.039 |
| bior3.5 | 6.246 ± 0.836 | 3.833 ± 0.619 | 16.05 ± 3.55 | 0.871 ± 0.029 |
| haar | 6.832 ± 0.647 | 4.296 ± 0.275 | 18.67 ± 2.98 | 0.850 ± 0.021 |
| coif5 | 6.809 ± 0.622 | 3.985 ± 0.241 | 16.53 ± 2.14 | 0.850 ± 0.026 |
| db6 | 7.149 ± 0.663 | 4.380 ± 0.459 | 18.44 ± 3.74 | 0.836 ± 0.028 |
| sym4 | 7.352 ± 1.126 | 4.555 ± 0.935 | 19.65 ± 3.90 | 0.827 ± 0.032 |
| coif3 | 7.477 ± 1.144 | 4.490 ± 0.504 | 18.74 ± 2.27 | 0.820 ± 0.027 |
| sym6 | 7.482 ± 0.830 | 4.571 ± 0.309 | 19.86 ± 2.19 | 0.819 ± 0.031 |
| coif1 | 7.390 ± 1.098 | 4.549 ± 0.735 | 19.96 ± 3.51 | 0.816 ± 0.062 |
| db4 | 7.670 ± 0.572 | 4.602 ± 0.433 | 18.85 ± 3.61 | 0.809 ± 0.034 |
| rbio3.5 | 7.695 ± 1.022 | 4.923 ± 0.719 | 21.12 ± 4.26 | 0.806 ± 0.035 |
| bior4.4 | 7.725 ± 0.762 | 4.693 ± 0.547 | 20.26 ± 3.53 | 0.801 ± 0.059 |
| db3 | 7.821 ± 0.997 | 4.911 ± 0.612 | 21.39 ± 1.79 | 0.801 ± 0.044 |
| rbio1.3 | 11.80 ± 1.81 | 9.70 ± 1.76 | 60.13 ± 14.75 | 0.537 ± 0.152 |
| bior1.3 | 16.11 ± 0.90 | 13.70 ± 0.78 | 88.92 ± 2.96 | 0.161 ± 0.103 |

Figure 5 shows that the Pearson correlation analysis applied to wavelet preprocessed datasets highlights a consistent set of variables with strong relevance to THI prediction. InterfacialV, DBDS, and Hydrogen consistently ranked as the top three features.

In Table 2, most wavelet families produced elbow absolute correlation thresholds within a narrow range (0.2715–0.2717) and kept six features, namely InterfacialV, DBDS, Hydrogen, Methane, Water Content and Ethylene, that reflects a high degree of overlap in feature selection. In contrast, the Coiflet-3 and Coiflet-5 families generated distinctly higher thresholds (0.3774–0.3775), reducing the feature set to only the top three variables, thus prioritizing the strongest predictors. This variation suggests that while higher thresholds can reduce dimensionality and produce more compact models, lower thresholds preserve a broader set of features that may enhance generalization. All these results indicate that certain wavelet families increase the correlations of the most influential variables, enabling stronger dimensionality reduction, while others maintain additional variables that support model robustness. Overall, the findings underline that the choice of wavelet family directly shapes the balance between model simplicity and predictive performance.

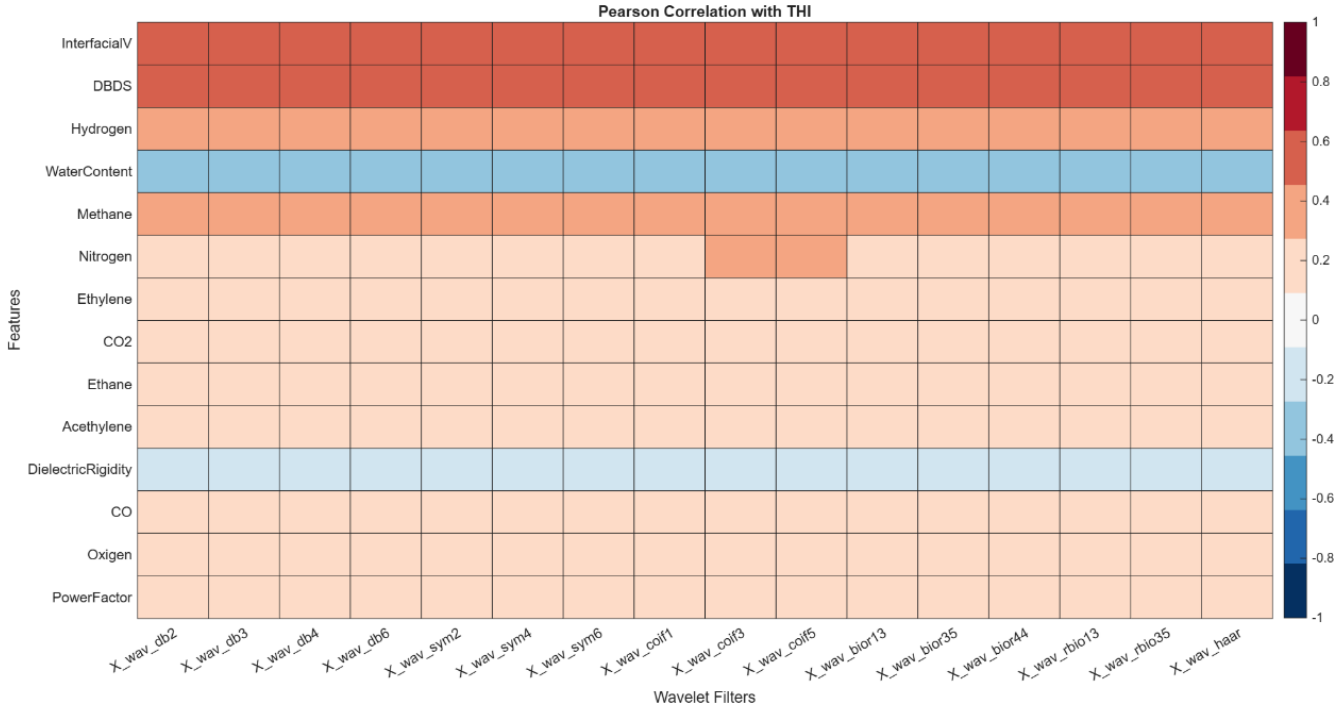


**Figure 5.** Pearson correlation heatmap of THI with different wavelet preprocessed features

**Table 2**. Feature Selection Results for Multi Wavelet Families

| Filter | Elbow Abs. Corr. | Number of Selected Features | Selected Features |
|---|---|---|---|
| db2 | 0.27171 | 6 | InterfacialV, DBDS, Hydrogen, Methane, WaterContent, Ethylene |
| db3 | 0.27171 | 6 | InterfacialV, DBDS, Hydrogen, Methane, WaterContent, Ethylene |
| db4 | 0.27159 | 6 | InterfacialV, DBDS, Hydrogen, WaterContent, Methane, Ethylene |
| db6 | 0.27175 | 6 | InterfacialV, DBDS, Hydrogen, WaterContent, Methane, Ethylene |
| sym2 | 0.27171 | 6 | InterfacialV, DBDS, Hydrogen, Methane, WaterContent, Ethylene |
| sym4 | 0.27175 | 6 | InterfacialV, DBDS, Hydrogen, WaterContent, Methane, Ethylene |
| sym6 | 0.27171 | 6 | InterfacialV, DBDS, Hydrogen, Methane, WaterContent, Ethylene |
| coif1 | 0.27174 | 6 | InterfacialV, DBDS, Hydrogen, WaterContent, Methane, Ethylene |
| coif3 | 0.37746 | 3 | InterfacialV, DBDS, Hydrogen |
| coif5 | 0.37749 | 3 | InterfacialV, DBDS, Hydrogen |
| bior13 | 0.27157 | 6 | InterfacialV, DBDS, Hydrogen, Methane, WaterContent, Ethylene |
| bior35 | 0.27146 | 6 | InterfacialV, DBDS, Hydrogen, Methane, WaterContent, Ethylene |
| bior44 | 0.27160 | 6 | InterfacialV, DBDS, Hydrogen, Methane, WaterContent, Ethylene |
| rbio13 | 0.27158 | 6 | InterfacialV, DBDS, Hydrogen, Methane, WaterContent, Ethylene |
| rbio35 | 0.27158 | 6 | InterfacialV, DBDS, Hydrogen, Methane, WaterContent, Ethylene |
| haar | 0.27160 | 6 | InterfacialV, DBDS, Hydrogen, WaterContent, Methane, Ethylene |

Table 3 shows that the best overall performance was achieved with the sym2 configuration, reaching an of $R^2$ 0.7720 alongside the lowest RMSE value of 8.391, an MAE of 5.103 , and a MAPE of 22.222. The sym4 (RMSE 8.412, $R^2$ 0.7705) and db6 (RMSE 8.516, $R^2$ 0.7661) families followed closely, indicating that these six-feature configurations provided consistent and reliable predictions. In contrast, coif3 and coif5, which retained only three features due to higher elbow thresholds, recorded $R^2$ values of 0.7327 and 0.7357, with RMSE values of 9.089 and 9.025, respectively. This suggests that excessive dimensionality reduction can lead to a notable drop in predictive accuracy. Interestingly, family such as bior1.3 ($R^2$ = 0.7452) preserved six features yet still produced the weakest result, with RMSE value of 8.855, highlighting that the relevance of the selected variables matters more than their number. Overall, Table 3 emphasizes that wavelet family selection directly influences model accuracy, and a balanced approach between feature count and quality yields the best outcomes.

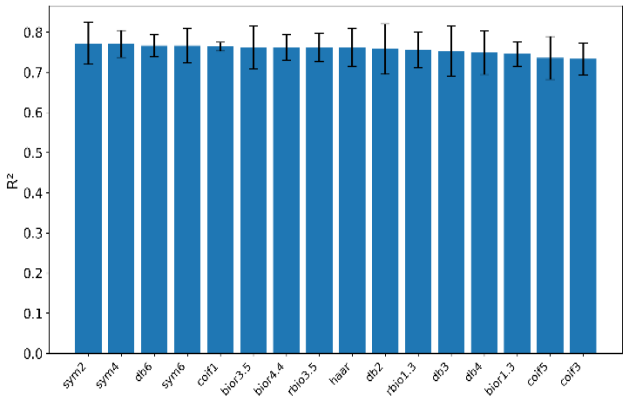


**Figure 6.** R² after Pearson+Kneedle feature selection (Mean ± Std)

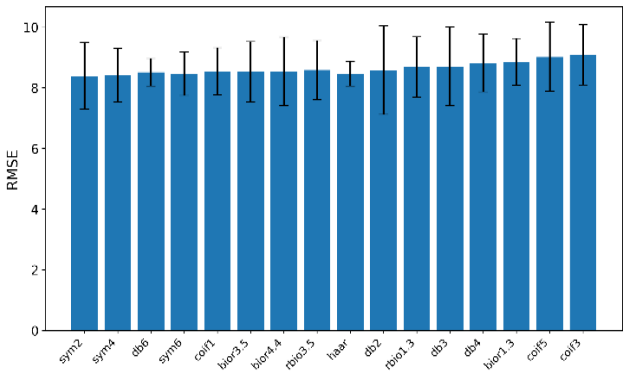


**Figure 7.** RMSE after Pearson+Kneedle feature selection (Mean ± Std)

Figure 6 and 7 illustrate the comparative R² and RMSE performance of different wavelet families before and after Pearson–Kneedle feature selection.

The combined evidence from Table 1 and 3 demonstrates that Symlet-2 consistently provided the highest predictive performance, both before and after Pearson correlation-based filtering. Its features retained strong relevance to THI prediction even after dimensionality reduction, achieving $R^2$ values of 0.879 (RMSE = 6.129) in the non-filtered set and 0.7720 (RMSE = 8.391) in the filter-based set. From a theoretical perspective, this behavior can be attributed to the near-symmetric structure of the Symlet family, which minimizes phase distortion during signal decomposition. In the context of transformer diagnostic measurements such as DGA ratios and oil quality indicators preserving the relative magnitude and ordering of degradation-related variations is critical. sym2 provides sufficient localization to suppress minor fluctuations while maintaining these physically meaningful relationships.

**Table 3.** Performance comparison of the random forest model for all wavelet families using filtered data

| Wavelet Family | Features | RMSE (Mean ± Std) | MAE (Mean ± Std) | MAPE (%) (Mean ± Std) | R² (Mean ± Std) |
|---|---|---|---|---|---|
| sym2 | 6 | 8.391 ± 1.092 | 5.103 ± 0.645 | 22.222 ± 1.136 | 0.7720 ± 0.0523 |
| sym4 | 6 | 8.412 ± 0.876 | 5.043 ± 0.413 | 21.545 ± 1.260 | 0.7705 ± 0.0337 |
| db6 | 6 | 8.516 ± 0.454 | 5.111 ± 0.151 | 22.183 ± 2.148 | 0.7661 ± 0.0271 |
| sym6 | 6 | 8.463 ± 0.714 | 5.155 ± 0.467 | 22.453 ± 2.341 | 0.7660 ± 0.0421 |
| coif1 | 6 | 8.549 ± 0.773 | 5.283 ± 0.556 | 23.157 ± 1.748 | 0.7637 ± 0.0108 |
| bior3.5 | 6 | 8.531 ± 0.992 | 5.050 ± 0.668 | 22.215 ± 4.941 | 0.7616 ± 0.0531 |
| bior4.4 | 6 | 8.542 ± 1.131 | 5.313 ± 0.802 | 23.150 ± 3.051 | 0.7616 ± 0.0328 |
| rbio3.5 | 6 | 8.589 ± 0.970 | 5.253 ± 0.532 | 22.223 ± 3.430 | 0.7615 ± 0.0346 |
| Haar | 6 | 8.471 ± 0.411 | 5.161 ± 0.276 | 22.562 ± 2.725 | 0.7614 ± 0.0475 |
| db2 | 6 | 8.583 ± 1.462 | 5.179 ± 0.845 | 22.022 ± 4.607 | 0.7573 ± 0.0625 |
| rbio1.3 | 6 | 8.685 ± 0.992 | 5.360 ± 0.524 | 23.407 ± 3.420 | 0.7563 ± 0.0444 |
| db3 | 6 | 8.702 ± 1.295 | 5.498 ± 0.892 | 23.658 ± 4.305 | 0.7517 ± 0.0628 |
| db4 | 6 | 8.823 ± 0.955 | 5.380 ± 0.372 | 22.746 ± 4.478 | 0.7485 ± 0.0541 |
| bior1.3 | 6 | 8.855 ± 0.774 | 5.354 ± 0.416 | 23.596 ± 4.551 | 0.7452 ± 0.0307 |
| coif5 | 3 | 9.025 ± 1.140 | 5.450 ± 0.699 | 23.069 ± 4.439 | 0.7357 ± 0.0533 |
| coif3 | 3 | 9.089 ± 0.997 | 5.673 ± 0.622 | 24.920 ± 2.811 | 0.7327 ± 0.0403 |

In contrast, families such as Bior1.3 and Reverse Bior1.3 in the non-filtered case, Coiflet-3 and Coiflet-5 in the filter-based case, recorded the lowest accuracies, underscoring the sensitivity of some configurations to feature reduction. Biorthogonal and Coiflet wavelets, while offering stronger smoothness or vanishing moment properties, tend to introduce excessive smoothing or feature attenuation in this application, which may suppress diagnostically relevant variations alongside noise. These findings highlight that while higher correlation thresholds and aggressive dimensionality reduction, as seen in Coiflet-3 and Coiflet-5,

can improve efficiency, they may also reduce predictive accuracy. Conversely, approaches like Symlet-2, which retain a slightly broader yet still relevant set of features, strike a more effective balance between model simplicity and accuracy, capturing greater variance and delivering more reliable THI predictions

## 5 Limitations of the study

The dataset used in this study contains 470 transformer samples, which is relatively limited for data-driven modelling. While 5-fold cross-validation was employed to mitigate overfitting and assess model stability, the limited sample size may still restrict the model's ability to fully capture rare degradation patterns or extreme operating conditions.

Although wavelet-based preprocessing is commonly applied in signal analysis, its use on tabular diagnostic data introduces potential risks. In this study, this limitation is addressed by ordering the samples along a degradation-based axis and applying the wavelet transform independently to each diagnostic feature. Nevertheless, the applicability of wavelet denoising to other tabular datasets without such an inherent ordering may require additional justification or alternative preprocessing strategies.

The results are derived from a single publicly available transformer dataset. Differences in asset characteristics, measurement practices, environmental conditions, and grid operating regimes across different power systems may affect model performance. Therefore, the generalizability of the proposed approach to other power grids or real-world utility datasets should be validated using larger and more diverse transformer populations in future studies.

Despite these limitations, the study provides a structured and reproducible framework for analysing the influence of wavelet preprocessing and feature selection on THI prediction, serving as a foundation for further investigation with expanded datasets.

## 6 Conclusion

This research presents the first systematic investigation in the THI prediction literature examining how different wavelet families influence model performance. A comprehensive comparison was conducted using the Random Forest algorithm on both non-filtered and Pearson correlation-based filter configurations, supported by an extensive review of prior studies to position this work within existing research gaps. The results clearly show that the preprocessing wavelet family has a measurable impact on both feature selection and predictive accuracy. Symlet-2 achieved the highest performance, with $R^2$ values of 0.879 (RMSE = 6.129) for the non-filtered dataset and 0.7720 (RMSE = 8.391) for the filtered dataset, while Coiflet-3 and Coiflet-5 offered computational efficiency but recorded an $R^2$ decrease of approximately 0.04–0.50 compared with Symlet-2. Other families, such as Bior1.3, exhibited a measurable decrease in accuracy, particularly in the filtered case, whereas the remaining configurations achieved $R^2$ values between 0.161 and 0.7452. From a mathematical perspective, Symlet wavelets are modified versions of Daubechies wavelets designed to improve symmetry while preserving orthogonality and compact support. For sym2, the filter length remains short, limiting excessive smoothing while still enabling effective noise suppression. The near linear phase response of sym2 reduces distortion in localized feature variations, ensuring that abrupt changes in diagnostic parameters such as sudden increases in dissolved gas concentrations or oil degradation indicators are preserved after preprocessing. In contrast, higher-order wavelets with longer support (e.g., Coiflets and Biorthogonal families) distribute local variations over wider neighborhoods, which can attenuate physically meaningful transitions along the degradation axis. Therefore, sym2 provides an appropriate compromise between noise attenuation, making it particularly well-suited for transformer health index prediction based on degradation-ordered diagnostic data.

These outcomes highlight that wavelet family selection should match operational requirements, as reducing the number of features can improve efficiency but may lower accuracy; while keeping a slightly larger yet relevant set captures more variance and enhances reliability. Beyond accuracy metrics, this distinction has practical implications for transformer asset management, where balancing computational cost with predictive performance can directly influence maintenance planning and investment decisions. Future work should validate these findings using real-world transformer datasets from different operating environments, transformer age profiles, and diagnostic test combinations to ensure their applicability in practical grid asset management scenarios.

## Conflict of interest

The authors declare that they have no conflicts of interest related to this work.

**Similarity rate (iThenticate):** %7

## References

[1] F. E. Bezerra, F. A. Z. Garcia, S. I. Nabeta, G. F. M. de Souza, I. E. Chabu, J. C. Santos, S. Nagao Junior, and F. H. Pereira, Wavelet-like transform to optimize the order of an autoregressive neural network model to predict the dissolved gas concentration in power transformer oil from sensor data. Sensors, 20(9), 2730, 2020. https://doi.org/10.3390/s20092730.

[2] S. Li, X. Li, Y. Cui and H. Li, Review of transformer health index from the perspective of survivability and condition assessment. Electronics, 12(11), 2407, 2023. https://doi.org/10.3390/electronics12112407.

[3] H. Guo and L. Guo, Health index for power transformer condition assessment based on operation history and test data. Energy Reports, 8, 9038–9045, 2022. https://doi.org/10.1016/j.egyr.2022.07.041.

[4] N. El-Rashidy, Y. A. Sultan and Z. H. Ali, Predicting power transformer health index and life expectation based on digital twins and multitask LSTM-GRU model. Scientific Reports, 15, 1359, 2025. https://doi.org/10.1038/s41598-024-83220-x.

[5] A. Alqudsi and A. El-Hag, Application of machine learning in transformer health index prediction.

Energies, 12(14), 2694, 2019. https://doi.org/10.3390/en12142694.

[6] M. A. A. Putra, Suwarno and R. A. Prasojo, Improving transformer health index prediction performance using machine learning algorithms with a synthetic minority oversampling technique. Energies, 18(9), 2364, 2025. https://doi.org/10.3390/en18092364.

[7] M. M. Islam, G. Lee and S. N. Hettiwatte, Application of a general regression neural network for health index calculation of power transformers. International Journal of Electrical Power and Energy Systems, 93, 308–315, 2017. https://doi.org/10.1016/j.ijepes.2017.06.008.

[8] I. B. Taha, Power transformers health index enhancement based on convolutional neural network after applying imbalanced-data oversampling. Electronics, 12(11), 2405, 2023. https://doi.org/10.3390/electronics12112405.

[9] R. Zemouri, Power transformer prognostics and health management using machine learning: A review and future directions. Machines, 13(2), 125, 2025. https://doi.org/10.3390/machines13020125.

[10] S. S. M. Ghoneim and I. B. M. Taha, Comparative Study of Full and Reduced Feature Scenarios for Health Index Computation of Power Transformers. IEEE Access, 8, 181326-181339, 2020. https://doi.org/10.1109/ACCESS.2020.3028689

[11] A. S. Mogos, X. Liang and C. Y. Chung, Enhancing transformer health index prediction using dissolved gas analysis data through integration of LightGBM and robust EM algorithms. IEEE Access, 12, 108472–108483, 2024. https://doi.org/10.1109/ACCESS.2024.3439248.

[12] N. Islam, R. Khan, S. K. Das, S. K. Sarker, M. M. Islam, M. Akter and S. M. Muyeen, Power transformer health condition evaluation: A deep generative model aided intelligent framework. Electric Power Systems Research, 218, 109201, 2023. https://doi.org/10.1016/j.epsr.2023.109201.

[13] S. T. Zahra, S. K. Imdad, S. Khan, S. Khalid and N. A. Baig, Power transformer health index and life span assessment: A comprehensive review of conventional and machine learning based approaches. Engineering Applications of Artificial Intelligence, 139, Part A, 109474, 2025. https://doi.org/10.1016/j.engappai.2024.109474.

[14] D. Rediansyah, R. A. Prasojo, Suwarno and A. Abu-Siada, Artificial intelligence-based power transformer health index for handling data uncertainty. IEEE Access, 9, 150637–150648, 2021. https://doi.org/10.1109/ACCESS.2021.3125379.

[15] S. H. Syed and V. Muralidharan, Feature extraction using discrete wavelet transform for fault classification of planetary gearbox: A comparative study. Applied Acoustics, 188, 108572, 2022. https://doi.org/10.1016/j.apacoust.2021.108572.

[16] N. Dhobale, S. S. Mulik and S. P. Deshmukh, Naïve Bayes and Bayes net classifier for fault diagnosis of end mill tool using wavelet analysis: A comparative study. Journal of Vibration Engineering & Technologies, 10, 1721–1735, 2022. https://doi.org/10.1007/s42417-022-00478-z.

[17] Z. Abda, M. Chettih and B. Zerouali, Assessment of neuro-fuzzy approach based different wavelet families for daily flow rates forecasting. Modeling Earth Systems and Environment, 7, 1523–1538, 2021. https://doi.org/10.1007/s40808-020-00855-1 .

[18] J. A. Domínguez-Navarro, T. B. Lopez-Garcia and S. M. Valdivia-Bautista, Applying wavelet filters in wind forecasting methods. Energies, 14(11), 3181, 2021. https://doi.org/10.3390/en14113181.

[19] T. Guo, T. Zhang, E. Lim, M. López-Benítez, F. Ma and L. Yu, A review of wavelet analysis and its applications: Challenges and opportunities. IEEE Access, 10, 58869-58903, 2022. https://doi.org/10.1109/ACCESS.2022.3179517.

[20] A. M. Dahman, A. A. Abou El-Ela, M. I. Zaki and R. A. El Sehiemy, A proposed wavelet analysis based fault diagnosis scheme of power transformers using fault signatures and CT saturation. Results in Engineering, 105820, 2025.

[21] Failure Analysis in Power Transformers, https://www.kaggle.com/code/sradha92/failure-analysis-in-power-transformers, Accessed 15 August 2025.

[22] İ. Öz, Comparative analysis of wavelet families in image compression, featuring the proposed new wavelet. Turkish Journal of Science and Technology, 19(1), 279–294, 2024. https://doi.org/10.55525/tjst.1428424.

[23] A. K. Karaev, O. S. Gorlova, V. V. Ponkratov, M. L. Sedova, N. S. Shmigol and M. L. Vasyunina, A comparative analysis of the choice of mother wavelet functions affecting the accuracy of forecasts of daily balances in the treasury single account. Economies, 10(9), 213, 2022.

[24] N. Hussain, M. Hasanzade, D. W. Breiby and M. N. Akram, Performance comparison of wavelet families for noise reduction and intensity thresholding in fourier ptychographic microscopy. Optics Communications, 519, 128400, 2022.

[25] F. Vatansever., F. Uysal and A. Uzun, Ayrık Dalgacık Dönüşümü ile Gürültü Süzme. 338-342, 2002. https://www.emo.org.tr/ekler/7841cc9e552bd5c_ek.pdf

[26] Ç. K. Çavaş, Protein verilerinin ayrık dalgacık dönüşümü ile analizi. Bayburt Üniversitesi Fen Bilimleri Dergisi, 6(1), 20–29, 2023. https://doi.org/10.55117/bufbd.1192229

[27] C. Vonesch, T. Blu and M. Unser, Generalized Daubechies Wavelet Families. IEEE Transactions on Signal Processing, 55(9), 4415-4429, 2007, https://doi.org/10.1109/TSP.2007.896255

[28] X. Wang, G. Gong and N. Li, Automated recognition of epileptic EEG states using a combination of symlet wavelet processing, gradient boosting machine, and grid search optimizer. Sensors, 19(2), 219, 2019.

[29] A. Antoniadis, Smoothing noisy data with coiflets, Statistica Sinica, 651–678, 1994.

[30] R. Szewczyk, K. Grabowski, M. Napieralska, W. Sankowski, M. Zubert and A. Napieralski, A reliable iris recognition algorithm based on reverse biorthogonal wavelet transform. Pattern Recognition Letters, 33(8), 1019–1026, 2012.
[31] S. Mallat, A Wavelet Tour of Signal Processing, Elsevier, 1999.
[32] D. Domingo, A. B. Kareem, C. N. Okwuosa, P. M. Custodio and J. W. Hur, Transformer core fault diagnosis via current signal analysis with Pearson correlation feature selection. Electronics, 13(5), 926, 2024. https://doi.org/10.3390/electronics13050926.
[33] V. Satopaa, J. Albrecht, D. Irwin and B. Raghavan, Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. 31st International Conference on Distributed Computing Systems Workshops, pp. 166-171, Minneapolis, MN, USA, 2011. https://doi.org/10.1109/ICDCSW.2011.20 .
[34] J. Wang, G. Li and W. Zhang, Combine-net: an improved filter pruning algorithm. Information, 12(7), 264, 2021. https://doi.org/10.3390/info12070264.
[35] X. Chen, H. Cui and L. Luo, Fault Diagnosis of Transformer Based on Random Forest, 2011 Fourth International Conference on Intelligent Computation Technology and Automation, pp. 132-134, Shenzhen, China, 2011, https://doi.org/10.1109/ICICTA.2011.40
[36] M. E. A. Senoussaoui, M. Brahami and I. Fofana, Transformer oil quality assessment using random forest with feature engineering. Energies, 14(7), 1809, 2021, https://doi.org/10.3390/en14071809.
[37] D. Rediansyah, R. A. Prasojo and Suwarno, Study on Artificial Intelligence Approaches for Power Transformer Health Index Assessment. 2021 International Conference on Electrical Engineering and Informatics (ICEEI), pp. 1-4, Kuala Terengganu, Malaysia, 2021. https://doi.org/10.1109/ICEEI52609.2021.9611109
[38] G. Chintia, R. A. Prasojo and Suwarno, Power Transformer Insulation System Health Index with Missing Data Prediction using Random Forest. 2023 IEEE 3rd International Conference in Power Engineering Applications (ICPEA), pp. 5-8, Putrajaya, Malaysia, 2023, https://doi.org/10.1109/ICPEA56918.2023.10093216 .