This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2026.3682814, IEEE Transactions on Instrumentation and Measurement

Multi-Scale Decoupling of Industrial Dynamics Via Trend-Fluctuation Interaction-Aware Transformer for Quality Prediction

Lin Xiao, Pingping Wang, Yijing Fang, Zidong Wang, *Fellow, IEEE*

***Abstract*—Accurate prediction of key quality variables is crucial for monitoring and optimizing modern industrial processes. However, most existing methods remain constrained by single-scale modeling, making it difficult to capture long-term global trends and short-term local fluctuations simultaneously. In addition, the dynamic couplings between these multi-scale components are often overlooked, leading to insufficient feature extraction. To address these limitations, a multi-scale trend-fluctuation interaction-aware transformer (MTI-Former) is proposed in this paper. First, a decoupling layer based on discrete wavelet transform (DWT) is designed to decompose industrial data into low-frequency trend and high-frequency fluctuation signals. Then, an adaptive high-pass enhancement filter is introduced to amplify critical high-frequency details and improve the perception of local disturbances. Cross-scale coupling is modeled through a trend-fluctuation interaction-aware attention module, which captures dynamic interactions between trends and fluctuations. Subsequently, a trend-fluctuation decoupling attention module applies a dual-path cross-attention mechanism to separately extract global dependencies and local variations. Finally, a gating mechanism fuses these representations to generate comprehensive multi-scale temporal predictions. The effectiveness of MTI-Former is verified on two real industrial datasets, and extensive results show that it outperforms several state-of-the-art methods in industrial quality prediction.**

***Index Terms*—Industrial quality prediction, industrial dynamics decoupling, trend-fluctuation interaction, multi-scale modeling, transformer.**

I. Introduction

Real-time monitoring and optimization of modern industrial processes are critical tasks for improving production efficiency [1]. The accurate prediction of key quality indices provides an effective way to reflect the operational state of industrial manufacturing [2, 3]. However, due to harsh industrial environments and high analytical costs, most quality indices cannot be measured online [4, 5]. As a result, soft sensor technology has been developed, in which easily measurable process variables correlated with quality indices are used as inputs to construct mathematical models for estimating unmeasurable quality variables [6, 7]. Data-driven soft sensor models have been widely applied for real-time monitoring of critical process variables [8–10].

Conventional soft sensor modeling has primarily relied on methods such as principal component analysis (PCA) [11], partial least squares (PLS) [12], and artificial neural networks (ANNs) [13–15]. These approaches are limited by their linear or shallow structures, which restrict their ability to represent the nonlinear characteristics of complex industrial processes [16]. In recent years, deep learning has been increasingly adopted and has demonstrated strong capabilities in soft sensing [17, 18]. Various deep architectures have been proposed, including stacked quality-driven autoencoders for quality-relevant feature representation [19], an interpretable disentangled transfer learning (IDTL) model for production prediction [20], bimodal manifold autoencoders for address modality gaps in flotation processes [21], and uncertainty-aware stacked autoencoders for adapting to dynamic operating conditions [22].

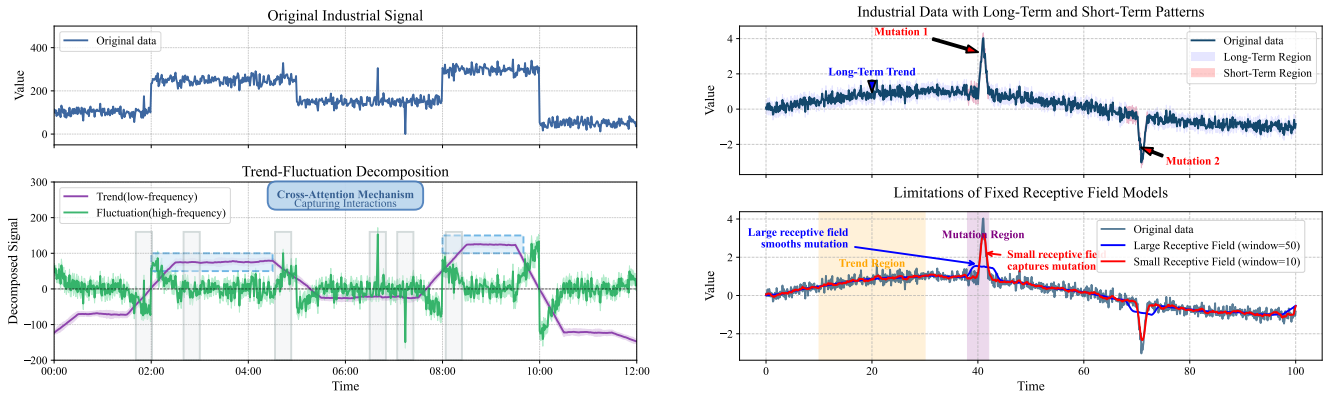
In recent years, transformer-based methods have received increasing attention in industrial process modeling, as their self-attention mechanisms enable the capture of long-range dependencies and complex interactions [23]. A Gaussian-based interval-aware transformer (GIA-Trans) was proposed in [24, 25] to handle irregularly sampled industrial time series data. A data-mode related interpretable transformer (DMRI-Former) [26] was developed for industrial process prediction. The Informer model [27] was further introduced to mitigate the long-tail problem of attention mechanisms and achieved strong performance in power forecasting tasks.

Despite the advances in deep learning and transformer-based methods [28–30], the inherent multiscale properties of industrial data and their dynamic informational couplings remain challenging for quality prediction. Industrial signals typically consist of two highly coupled components: a global trend and local fluctuations. The global trend reflects long-term, slowly varying process characteristics and represents the primary structure of the data, as illustrated by the low-frequency curve in Fig. 1(a). In contrast, local fluctuations describe rapid short-term variations arising from dynamic process disturbances, as shown by the high-frequency curve in Fig. 1(a). These two components differ markedly in statistical behavior and temporal evolution, and strong dynamic interactions exist between them. Local fluctuations may induce cumulative deviations in the global trend, and modeling them jointly at a single scale

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62403195 and the NSFC General Program under Grant 62573109. (Corresponding author: Yijing Fang)

L. Xiao, P. Wang, Y. Fang are with College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China and with Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha 410081, China. (E-mail: xiaolin860728@163.com; pingpingwang200211@163.com; yijingfang0401@163.com)

Zidong Wang is with the Department of Computer Science, Brunel UniversityLondon, Uxbridge, Middlesex, UB8 3PH, United Kingdom. (Email:Zidong.Wang@brunel.ac.uk)



(a) (b)

Fig. 1: The challenge of modeling complex multi-scale dynamics in industrial data. (a) A typical industrial signal decomposed into its two highly coupled components: the stable global trend (low-frequency) and the rapid local fluctuations (high-frequency). (b) An illustrative example of how conventional single-scale modeling fails to effectively capture both components, leading to insufficient feature extraction.

often results in insufficient feature extraction. Therefore, it is essential to decouple these components and model them according to their individual dynamic properties.

Signal decomposition methods have recently attracted increasing attention due to their capability to reveal multi-scale structures in industrial time-series data [31]. FEDFormer [32] was introduced to perform decomposition using the fast Fourier transform and frequency enhancement blocks, although cross-frequency coupling was not considered. A graph-based dual-stream architecture was developed in [33] using the short-time Fourier transform and multi-graph attention; however, its fixed window design limits frequency adaptability. CTFNet [34] employed convolution and decomposition to capture multiscale patterns, but the reliance on predefined kernels restricts adaptive modeling. WCED-Trans was developed in [35] using DWT and deformable attention to estimate the state of health of lithium-ion batteries. DwtFormer [36] was introduced to use wavelets to transform data into a two-dimensional space and learn features into and between cycles. Similarly, W-Transformers [37] was designed to incorporate wavelets within Transformers to capture irregularities by separating the signal from noise. However, by processing decomposed components independently, these approaches often neglect the physical meaning of each wavelet decomposition level and overlook the intrinsic interactions in which high-frequency disturbances influence the evolution of the global trend, thereby posing a significant challenge in achieving a comprehensive representation of complex industrial dynamics. As illustrated in Fig. 1(b), single time-scale models with fixed receptive fields typically fail to capture long-term trends and short-term fluctuations simultaneously.

Motivated by the limitations of existing single-scale and transformer-based methods in modeling cross-scale industrial dynamics, this paper proposes a multi-scale trend-fluctuation interaction-aware transformer (MTI-Former). A decoupling layer based on the DWT is constructed to decompose the input into low-frequency trend and high-frequency fluctuation sequences. To enhance the representation of transient disturbances, an adaptive high-pass enhancement filter (AHEF) is introduced to selectively amplify critical high-frequency details. To model cross-scale dynamic couplings, a trend-fluctuation interaction-aware attention (TFIA) mechanism is designed to capture dependencies between trends and fluctuations. A trend-fluctuation decoupling attention (TFDA) module is further developed to extract global long-term patterns and local dynamic variations from the multi-scale interactive features obtained by TFIA. Finally, a gating mechanism is adopted to fuse trend, fluctuation, and interactive representations, yielding a comprehensive multi-scale temporal description of complex industrial processes. The main contributions of this paper are summarized as follows:

1) A multi-scale framework named MTI-Former is proposed to improve the prediction accuracy of key quality variables. Industrial data are first decoupled into trend and fluctuation components, after which their dynamic cross-scale couplings are modeled to capture distinct global dependencies and local variations.
2) An AHEF module is designed to selectively amplify critical high-frequency information, enabling more effective characterization of transient process dynamics.
3) A TFIA mechanism is introduced to model dynamic cross-scale couplings between trends and fluctuations, thereby providing coordinated awareness of multi-scale temporal behaviors.
4) A TFDA module is presented to further extract global long-term dependencies and local dynamic variations. This module employs a dual-path cross-attention architecture to disentangle these patterns from the multi-scale interactive representations produced by the TFIA module.
5) Extensive experiments conducted on two industrial process datasets validate the effectiveness and superiority of the proposed MTI-Former compared with several state-of-the-art soft sensing approaches.

II. PRELIMINARIES

The wavelet transform (WT) is a time-frequency analysis tool that employs multi-resolution decomposition to capture both local and global characteristics of signals across different scales. Owing to this capability, WT is well suited for modeling cross-frequency relationships in complex industrial processes.

A discrete time series is denoted by $\mathbf{X} = \{x_k\}_{k=0}^{T-1}$, where T represents the number of signal observation points. To handle discrete data, the discrete wavelet transform (DWT) framework based on multiresolution analysis (MRA) is adopted. In this decomposition, a scaling function $\phi(t)$ and a mother wavelet $\psi(t)$ are utilized, which satisfy the following two-scale relations:

$$\phi(t) = \sqrt{2}\sum_n h_n \phi(2t-n), \quad \psi(t) = \sqrt{2}\sum_n g_n \phi(2t-n), \tag{1}$$

where $\{h_n\}$ and $\{g_n\}$ are the low-pass and high-pass filter coefficients respectively. These coefficients are determined by the chosen wavelet family such as Daubechies, and the orthogonality condition $\sum_n h_n h_{n-2k} = \delta_{k,0}$ is maintained, where δ is the Kronecker delta.

Based on Eq. (1), the signal is decomposed recursively by the DWT. Let $a_{l,k}$ and $d_{l,k}$ be defined as the approximation and detail coefficients at scale l and location k. Starting from the original signal as the finest-scale approximation, where $a_{0,k} = x_k$, the coefficients at the next coarser scale $l+1$ are computed via convolution and downsampling:

$$a_{l+1,k} = \sum_m h_{m-2k} a_{l,m}, \quad d_{l+1,k} = \sum_m g_{m-2k} a_{l,m}. \tag{2}$$

Through this decomposition, the signal is transformed into multi-scale coefficients, in which long-term trends are represented by the low-frequency component $a_{l,k}$, while short-term local fluctuations are captured by the high-frequency component $d_{l,k}$ [38].

To reconstruct the original sequence x_k from the wavelet coefficients, the inverse discrete wavelet transform (iDWT) is employed. The reconstruction is performed from the coarsest scale back to the finest scale by upsampling and filtering. The approximation coefficients at scale l are recovered as:

$$a_{l,m} = \sum_k \tilde{h}_{m-2k} a_{l+1,k} + \sum_k \tilde{g}_{m-2k} d_{l+1,k}, \tag{3}$$

where $\tilde{h}$ and $\tilde{g}$ are the synthesis filters, which are defined as $\tilde{h}_n = h_{-n}$ and $\tilde{g}_n = g_{-n}$ for orthogonal wavelets. Since this reconstruction process is mathematically exact, $\mathbf{X}$ can be recovered by the iDWT without information loss.

By embedding this wavelet decomposition and reconstruction mechanism within the transformer architecture, the capability of the model to characterize dynamic coupling relationships across multiple temporal scales is enhanced.

III. MTI-FORMER

To address the multi-scale characteristics inherent in industrial process data, an encoder-based architecture named MTI-Former is proposed, as illustrated in Fig. 2. In the decoupling layer, the input industrial data are first decomposed into multiple low-frequency and high-frequency components using the DWT. Each high-frequency component is then processed by an AHEF module to adaptively enhance critical fluctuation details.

To characterize the complex cross-scale dependencies present in industrial data, a TFIA module is designed to interactively model the relationships between the low-frequency components and the enhanced high-frequency components. After interaction, the original low-frequency and enhanced high-frequency components are reconstructed into trend and fluctuation signals using the iDWT. A TFDA module is subsequently employed to extract long-term trend dependencies and local fluctuation dynamics from these reconstructed multi-scale signals. Finally, a gating mechanism is utilized to adaptively fuse the decoupled trend signals, fluctuation signals from the TFDA module, and the multi-scale interactive features generated by the TFIA module. The fused representation is then fed into a regression layer to produce the final prediction of the key quality variable.

A. Adaptive High-Pass Enhancement Filter

Assume that the input industrial data are denoted by $\mathbf{S} \in \mathbb{R}^{T\times d}$, where T is the sequence length and d is the feature dimension. The DWT is applied to $\mathbf{S}$ to perform a hierarchical, tree-structured decomposition over L levels. To accommodate odd-length sequences during decomposition, a replication padding strategy is implemented by replicating the last value of the sequence for odd-length inputs. At each decomposition level $l \in [1, L]$, the current low-frequency approximation coefficient $\mathbf{x}_{\text{lf}}^{l-1}$ is decomposed into a new approximation coefficient $\mathbf{x}_{\text{lf}}^{l} \in \mathbb{R}^{\frac{T}{2^l}\times d}$ and a high-frequency coefficient $\mathbf{x}_{\text{hf}}^{l} \in \mathbb{R}^{\frac{T}{2^l}\times d}$. The process is initialized as $\mathbf{x}_{\text{lf}}^{0} = \mathbf{S}$. Formally, the decomposition is expressed as

$$\left\{\mathbf{x}_{\text{hf}}^{l}, \mathbf{x}_{\text{lf}}^{l}\right\} = \text{DWT}(\mathbf{x}_{\text{lf}}^{l-1}). \tag{4}$$

To generate adaptive high-pass filters, a joint feature representation $\mathbf{Z}^l$ is constructed by concatenating the low-frequency features $\mathbf{x}_{\text{lf}}^{l}$ with the high-frequency features $\mathbf{x}_{\text{hf}}^{l}$ for each decomposition level:

$$\mathbf{Z}^l = \mathbf{x}_{\text{lf}}^{l} \oplus \mathbf{x}_{\text{hf}}^{l}, \tag{5}$$

where $\oplus$ denotes the concatenation operation along the feature dimension, resulting in $\mathbf{Z}^l \in \mathbb{R}^{\frac{T}{2^l}\times 2d}$. This joint representation captures both global trends and fine-grained fluctuations.

To capture dependencies from local temporal neighborhoods, $\mathbf{Z}^l$ is processed by a lightweight 1×3 convolutional layer. This layer performs adaptive weighted averaging over neighboring values, smoothing high-frequency spikes while preserving the underlying low-frequency trend. Subsequently, a softmax activation is applied to normalize the convolution output, thereby directly obtaining the dynamic low-pass filter weights $\mathbf{W}_{\text{lp}}^{l}$:

$$\mathbf{W}_{\text{lp}}^{l} = \text{softmax}\left(\text{Conv}_{1\times 3}(\mathbf{Z}^l)\right). \tag{6}$$

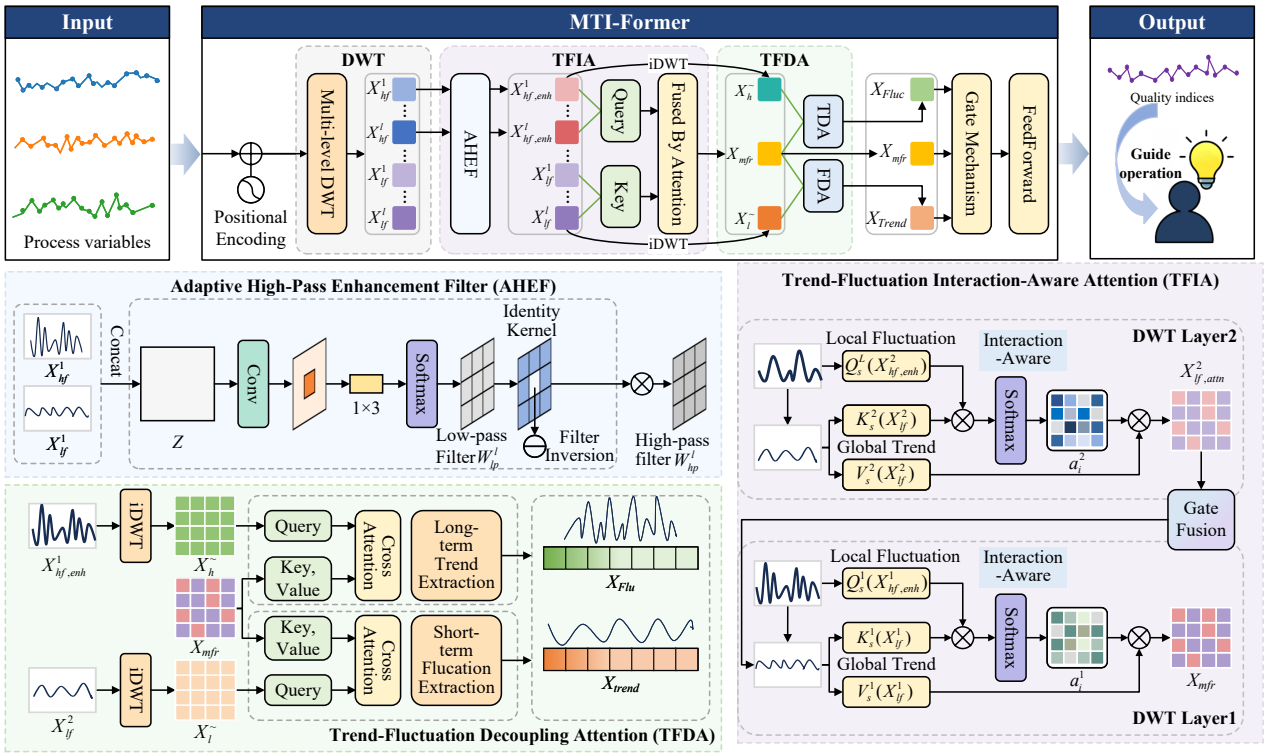


Fig. 2: Architecture of MTI-Former.

Inspired by [39], the corresponding high-pass filter weights $\mathbf{W}_{\text{hp}}^{l}$ are derived by subtracting the generated low-pass filter from the identity matrix $\mathbf{E}$ which acts as an all-pass filter:

$$\mathbf{W}_{\text{hp}}^{l} = \mathbf{E} - \mathbf{W}_{\text{lp}}^{l}. \tag{7}$$

After applying the high-pass filter and adding a residual connection, the enhanced high-frequency features $\mathbf{x}_{\text{hf,enh}}^{l}$ are obtained as:

$$\mathbf{x}_{\text{hf,enh}}^{l} = \mathbf{x}_{\text{hf}}^{l} + \mathbf{W}_{\text{hp}}^{l} \odot \mathbf{x}_{\text{hf}}^{l}, \tag{8}$$

where $\odot$ represents the hadamard product. Unlike conventional static filtering approaches, the proposed enhancement mechanism generates filter weights dynamically based on local data patterns through a lightweight convolution and softmax operation. This adaptive design enables the filter to adapt to time-varying process conditions and disturbances, thereby significantly improving the model's ability to capture transient dynamics in industrial processes. The pseudo-code description of the proposed AHEF is outlined in Algorithm 1.

Algorithm 1 AHEF Algorithm

Input: Industrial data $\mathbf{S} \in \mathbb{R}^{T \times d}$, Decomposition levels L, Identity matrix $\mathbf{E}$.
Output: Enhanced high-frequency features $\{\mathbf{x}_{\text{hf,enh}}^{l}\}_{l=1}^{L}$.
Apply replication padding to $\mathbf{S}$ to handle variable sequence lengths.
Initialize low-frequency approximation: $\mathbf{x}_{\text{lf}}^{0} = \mathbf{S}$.
for $t = 1, \ldots, L$ **do**
    Perform hierarchical decomposition via DWT based on Eq. (4):
        $\{\mathbf{x}_{\text{hf}}^{l}, \mathbf{x}_{\text{lf}}^{l}\} = \text{DWT}(\mathbf{x}_{\text{lf}}^{l-1})$
    Construct joint feature representation $\mathbf{Z}^{l}$ based on Eq. (5):
        $\mathbf{Z}^{l} = \mathbf{x}_{\text{lf}}^{l} \oplus \mathbf{x}_{\text{hf}}^{l}$
    Calculate dynamic low-pass filter weights $\mathbf{W}_{\text{lp}}^{l}$ via Conv and Softmax based on Eq. (6).
    Derive adaptive high-pass filter weights $\mathbf{W}_{\text{hp}}^{l}$ based on Eq. (7):
        $\mathbf{W}_{\text{hp}}^{l} = \mathbf{E} - \mathbf{W}_{\text{lp}}^{l}$
    Obtain enhanced high-frequency features $\mathbf{x}_{\text{hf,enh}}^{l}$ via residual connection (Eq. (8)):
        $\mathbf{x}_{\text{hf,enh}}^{l} = \mathbf{x}_{\text{hf}}^{l} + \mathbf{W}_{\text{hp}}^{l} \odot \mathbf{x}_{\text{hf}}^{l}$
end for
Return the improved features $\{\mathbf{x}_{\text{hf,enh}}^{l}\}_{l=1}^{L}$.

B. Trend-Fluctuation Interaction-Aware Attention

A central challenge in modeling complex industrial processes lies in capturing the coupling between their multi-scale dynamics. Transient fluctuations may accumulate over time and eventually influence long-term system trends. Consequently, short-term operational disturbances and long-horizon behaviors become inherently intertwined, necessitating the explicit modeling of their interactions. To address this, the TFIA module is designed to characterize these cross-scale dependencies between trend and fluctuation components.

At each wavelet decomposition level $l \in \{1, 2, \ldots, L\}$, the enhanced high-frequency components $\mathbf{x}_{\text{hf,enh}}^{l} \in \mathbb{R}^{\frac{T}{2^l} \times d}$ are utilized to explore the correlations embedded within the long-

term patterns of the low-frequency components $\mathbf{x}_{\text{lf}}^{l} \in \mathbb{R}^{\frac{T}{2^l} \times d}$. The query vector $\mathbf{Q}$ for level l is defined as:

$$\mathbf{Q} = \mathbf{x}_{\text{hf,enh}}^{l} \mathbf{W}_{Q}^{l}. \tag{9}$$

For each level $l \in \{1, \ldots, L\}$, both the key and value matrices are projected from the low-frequency trend component $\mathbf{x}_{\text{lf}}^{l}$. Specifically, the key and value vectors for level l are computed as:

$$\mathbf{K} = \mathbf{x}_{\text{lf}}^{l} \mathbf{W}_{K}^{l}, \quad \mathbf{V} = \mathbf{x}_{\text{lf}}^{l} \mathbf{W}_{V}^{l}, \tag{10}$$

where $\mathbf{W}_{Q}^{l}, \mathbf{W}_{K}^{l}, \mathbf{W}_{V}^{l} \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices.

Subsequently, a fluctuation-aware trend representation $\mathbf{x}_{\text{lf,attn}}^{l}$ is computed through the standard scaled dot-product attention mechanism:

$$\begin{aligned} \mathbf{x}_{\text{lf,attn}}^{l} &= \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\ &= \text{softmax}\left(\frac{\mathbf{Q}(\mathbf{K})^{\top}}{\sqrt{d_k}}\right)\mathbf{V}, \end{aligned} \tag{11}$$

where d_k denotes the key dimension and the scaling factor $\sqrt{d_k}$ prevents excessively large dot-product magnitudes and stabilizes optimization.

To balance the influence of short-term fluctuations with the preserved long-term structure, the enriched trend representation $\mathbf{x}_{\text{lf,attn}}^{l}$ is fused with the original low-frequency component $\mathbf{x}_{\text{lf}}^{l}$ at each level $l \in \{L, L-1, \ldots, 1\}$:

$$\mathbf{x}_{\text{lf,fuse}}^{l} = \beta \cdot \mathbf{x}_{\text{lf,attn}}^{l} + (1 - \beta) \cdot \mathbf{x}_{\text{lf}}^{l}, \tag{12}$$

where $\beta \in [0, 1]$ is a learnable fusion parameter. The fused trend component $\mathbf{x}_{\text{lf,fuse}}^{l}$ is then combined with the enhanced high-frequency component $\mathbf{x}_{\text{hf,enh}}^{l}$ and reconstructed via iDWT to form the intermediate representation for the next level:

$$\mathbf{X}^{l-1} = \text{iDWT}\left(\mathbf{x}_{\text{lf,fuse}}^{l},\ \mathbf{x}_{\text{hf,enh}}^{l}\right). \tag{13}$$

This iterative process continues until $l = 1$, and the final reconstructed output $\mathbf{X}^0$ is taken as the multi-scale fused representation $\mathbf{X}_{\text{mfr}}$. This representation integrates the full spectrum of cross-scale dependencies, providing a refined multi-scale description of the underlying industrial process dynamics.

C. Trend-Fluctuation Decoupling Attention

Unlike conventional methods that process time-varying features on a single temporal scale, industrial data typically contain trend and fluctuation components that exhibit different temporal characteristics. The low-frequency trend reflects long-term structural behavior, whereas the high-frequency fluctuation captures short-term variations induced by operational disturbances. To model these differences effectively, the TFDA module is introduced to separately extract the unique temporal information embedded within each component.

To decouple the trend component, a global trend reference signal $\widetilde{\mathbf{X}}_l$ is constructed using only the low-frequency coefficients from the first decomposition level:

$$\widetilde{\mathbf{X}}_l = \text{iDWT}_{\text{low-only}}(\mathbf{x}_{\text{lf}}^{1}) \tag{14}$$

Similarly, the dynamic fluctuation reference signal $\widetilde{\mathbf{X}}_h$ is obtained by reconstructing the signal using only the high-frequency coefficients from the last decomposition level:

$$\widetilde{\mathbf{X}}_h = \text{iDWT}_{\text{high-only}}(\mathbf{x}_{\text{hf}}^{L}). \tag{15}$$

Following the acquisition of the decoupled signals, two parallel cross-attention mechanisms are deployed to extract features specific to the trend and fluctuation components. The multi-scale fused representation $\mathbf{X}_{\text{mfr}}$, generated by the TFIA module, serves as the shared feature pool from which long-term and short-term dependencies are selectively extracted. The decoupled signals $\tilde{\mathbf{X}}_l$ and $\tilde{\mathbf{X}}_h$ are used as queries to identify the most relevant patterns from $\mathbf{X}_{\text{mfr}}$.

a) Long-term Trend Feature Extraction: The trend decoupling attention (TDA) mechanism extracts global trend features by taking the trend reference $\tilde{\mathbf{X}}_l$ as the query and the multi-scale feature $\mathbf{X}_{\text{mfr}}$ as both key and value:

$$\begin{aligned} \mathbf{X}_{\text{trend}} &= \text{Attention}(\mathbf{Q}_{\text{trend}}, \mathbf{K}_{\text{mfr}}, \mathbf{V}_{\text{mfr}}) \\ &= \text{softmax}\left(\frac{\mathbf{Q}_{\text{trend}}\mathbf{K}_{\text{mfr}}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}_{\text{mfr}}, \end{aligned} \tag{16}$$

$$\mathbf{Q}_{\text{trend}} = \tilde{\mathbf{X}}_l \mathbf{W}_{Q}^{\text{trend}}, \quad \mathbf{K}_{\text{mfr}} = \mathbf{X}_{\text{mfr}} \mathbf{W}_{K}^{\text{trend}}, \quad \mathbf{V}_{\text{mfr}} = \mathbf{X}_{\text{mfr}} \mathbf{W}_{V}^{\text{trend}}. \tag{17}$$

The scaling factor $\sqrt{d_k}$ stabilizes the variance of the dot-product attention scores.

b) Dynamic Fluctuation Feature Extraction: The fluctuation decoupling attention (FDA) mechanism captures short-term dynamic variations by using the fluctuation reference $\tilde{\mathbf{X}}_h$ as the query:

$$\begin{aligned} \mathbf{X}_{\text{fluc}} &= \text{Attention}(\mathbf{Q}_{\text{fluc}}, \mathbf{K}_{\text{mfr}}, \mathbf{V}_{\text{mfr}}) \\ &= \text{softmax}\left(\frac{\mathbf{Q}_{\text{fluc}}\mathbf{K}_{\text{mfr}}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}_{\text{mfr}}, \end{aligned} \tag{18}$$

$$\mathbf{Q}_{\text{fluc}} = \tilde{\mathbf{X}}_h \mathbf{W}_{Q}^{\text{fluc}}, \quad \mathbf{K}_{\text{mfr}} = \mathbf{X}_{\text{mfr}} \mathbf{W}_{K}^{\text{fluc}}, \quad \mathbf{V}_{\text{mfr}} = \mathbf{X}_{\text{mfr}} \mathbf{W}_{V}^{\text{fluc}}. \tag{19}$$

Here, $\mathbf{W}$ denotes the learnable projection matrices. The resulting representations $\mathbf{X}_{\text{trend}}$ and $\mathbf{X}_{\text{fluc}}$ encode long-term structural patterns and transient dynamics, respectively, ensuring that the distinct temporal compositions of industrial data are effectively captured.

To construct a unified representation, an adaptive gating mechanism is employed to fuse the extracted trend features, fluctuation features, and multi-scale representation. Specifically, the features of each component are first aggregated via global average pooling (GAP) and are subsequently concatenated to form a comprehensive feature $\mathbf{g}$:

$$\mathbf{g} = \text{Concat}\left(\text{GAP}(\mathbf{X}_{\text{trend}}), \text{GAP}(\mathbf{X}_{\text{fluc}}), \text{GAP}(\mathbf{X}_{\text{mfr}})\right). \tag{20}$$

Subsequently, the coefficients $[\lambda_t, \lambda_f, \lambda_m] \in [0, 1]$, which balance the contribution of each component, are obtained through a linear transformation followed by a softmax activation:

$$[\lambda_t, \lambda_f, \lambda_m] = \text{Softmax}(\mathbf{W}\mathbf{g} + \mathbf{b}), \tag{21}$$

where $\mathbf{W}$ and $\mathbf{b}$ denote the weight matrix and bias term respectively. Finally, the fused representation is derived via the adaptive weighted aggregation:

$$\mathbf{X}_{\text{fused}} = \lambda_t \odot \mathbf{X}_{\text{trend}} + \lambda_f \odot \mathbf{X}_{\text{fluc}} + \lambda_m \odot \mathbf{X}_{\text{mfr}}. \tag{22}$$

Algorithm 2 MTI-Former Training Algorithm

Input: Industrial data series $\mathbf{S} \in \mathbb{R}^{T\times d}$, decomposition levels L.
Output: Predicted quality variable sequence $\hat{\mathbf{Y}}$.
Apply replication padding to $\mathbf{S}$ and decompose into $\{\mathbf{x}_{\text{lf}}^l, \mathbf{x}_{\text{hf}}^l\}_{l=1}^L$ via DWT.
for $l = 1$ **to** L **do**
    Compute joint feature $\mathbf{Z}^l$ and adaptive filters $\mathbf{W}_{\text{lp}}^l, \mathbf{W}_{\text{hp}}^l$ (Eqs. (5)-(7)).
    Enhance high-freq: $\mathbf{x}_{\text{hf,enh}}^l = \mathbf{x}_{\text{hf}}^l + \mathbf{W}_{\text{hp}}^l \odot \mathbf{x}_{\text{hf}}^l$ (Eq. (8)).
end for
for $l = 1$ **to** L **do**
    Compute trend interaction $\mathbf{x}_{\text{lf,attn}}^l$ via Attention$(\mathbf{Q}_i, \mathbf{K}_j, \mathbf{V}_j)$ (Eq. (11)).
    Fuse trend: $\mathbf{x}_{\text{lf,fuse}}^l = \beta\cdot\mathbf{x}_{\text{lf,attn}}^l + (1-\beta)\cdot\mathbf{x}_{\text{lf}}^l$ (Eq. (12)).
end for
Obtain multi-scale representation $\mathbf{X}_{\text{mfr}}$ via iterative iDWT (Eq. (13)).
Construct references $\widetilde{\mathbf{X}}_l, \widetilde{\mathbf{X}}_h$ via single-branch iDWT (Eqs. (14)-(15)).
Extract features $\mathbf{X}_{\text{trend}}, \mathbf{X}_{\text{fluc}}$ via Cross-Attention with $\mathbf{X}_{\text{mfr}}$ (Eqs. (16)-(19)).
Compute gating weights $[\lambda_t, \lambda_f, \lambda_m]$ via GAP and Softmax (Eqs. (20)-(21)).
Adaptive Fusion: $\mathbf{X}_{\text{fused}} = \lambda_t \odot \mathbf{X}_{\text{trend}} + \lambda_f \odot \mathbf{X}_{\text{fluc}} + \lambda_m \odot \mathbf{X}_{\text{mfr}}$.
Prediction: $\hat{\mathbf{Y}} = \text{FFN}(\mathbf{X}_{\text{fused}})$.
Return $\hat{\mathbf{Y}}$.

The resulting representation is then fed into a feed-forward network (FFN) to generate the final prediction:

$$\hat{\mathbf{Y}} = \text{FFN}(\mathbf{X}_{\text{fused}}). \tag{23}$$

For a prediction horizon of H, the output sequence is

$$\hat{\mathbf{Y}} = [\hat{y}_{t+1}, \ldots, \hat{y}_{t+H}], \tag{24}$$

where $\hat{y}_{t+i}$ denotes the predicted quality variable at time $t+i$.

The objective of MTI-Former is to achieve high-precision soft sensing for quality indices. Accordingly, the loss function is formulated as the mean-squared error between the predicted and actual values, expressed as:

$$J(\theta) = \frac{1}{H}\sum_{i=1}^{H}(\hat{y}_{t+i} - y_{t+i})^2, \tag{25}$$

where y_{t+i} represents the ground-truth quality variable at time $t+i$. This loss function guides the model to minimize prediction errors, ensuring high-fidelity forecasting of quality variables. The complete training procedure for the MTI-Former framework is summarized in Algorithm 2.

IV. Industrial Applications

In this section, the proposed MTI-Former is experimentally evaluated on two representative industrial systems: the debutanizer column process and the ironmaking process. To ensure a fair and rigorous comparison, several advanced time-series prediction models, including CrossFormer [40], iTransformer [41], DLinear [42], Informer [27], PatchTST [43], TimesNet [44], and TSMixer [45], are implemented under the same experimental settings and assessed alongside the proposed method. All experiments are conducted using Python 3.8 with PyTorch 2.5.0.

A. Debutanizer Column

The debutanizer column is a key distillation unit in natural gas and refinery processing. It is employed to separate propane, butane, and heavier hydrocarbons from the naphtha stream, thereby facilitating desulfurization and the production of stabilized gasoline.

TABLE I: The best hyperparameter configuration of MTI-Former for C4 prediction in the debutanizer column.

Symbol	Description	Predict window length		
		1	3	5
len	Length of window size	15	15	15
d_{model}	Embedding dimension	128	256	256
h	Number of attention heads	8	8	8
N_e	Encoder layer number	2	2	3
d_{ff}	Nonlinear dimension	128	512	512
L	Number of Wavelet Decomposition Layers	2	2	2
W_f	Wavelet function	S	S	S

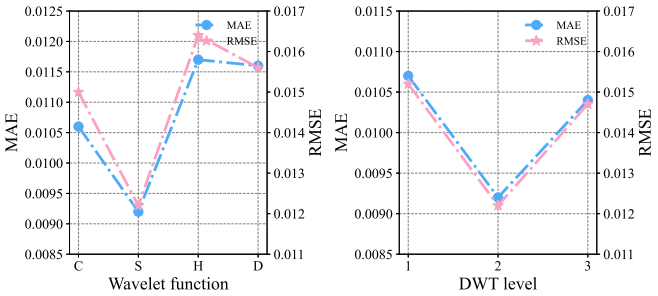


Fig. 3: Hyperparameter sensitivity analysis for predicting C4 in the debutanizer column.

The system consists of six components: a heat exchanger, a top condenser, a bottom reboiler, a reflux pump, a feed pump, and a reflux accumulator.Its primary objective is optimal column fractionation, maximizing top stabilized gasoline yield while minimizing butane (C4) in the bottoms. Eight variables were selected, including seven C4-correlated auxiliary variables based on process mechanisms and correlation analysis, alongside key quality variables from debutanizer simulations. A total of 2300 labeled samples were collected, of which 2000 were allocated for model training and the remaining 300 were reserved for testing.

To ensure optimal model training performance, all data underwent normalization. For a rigorous comparative analysis, a unified grid search strategy was adopted to optimize both the proposed MTI-Former and all baseline models within an identical search space. Specifically, the optimizer

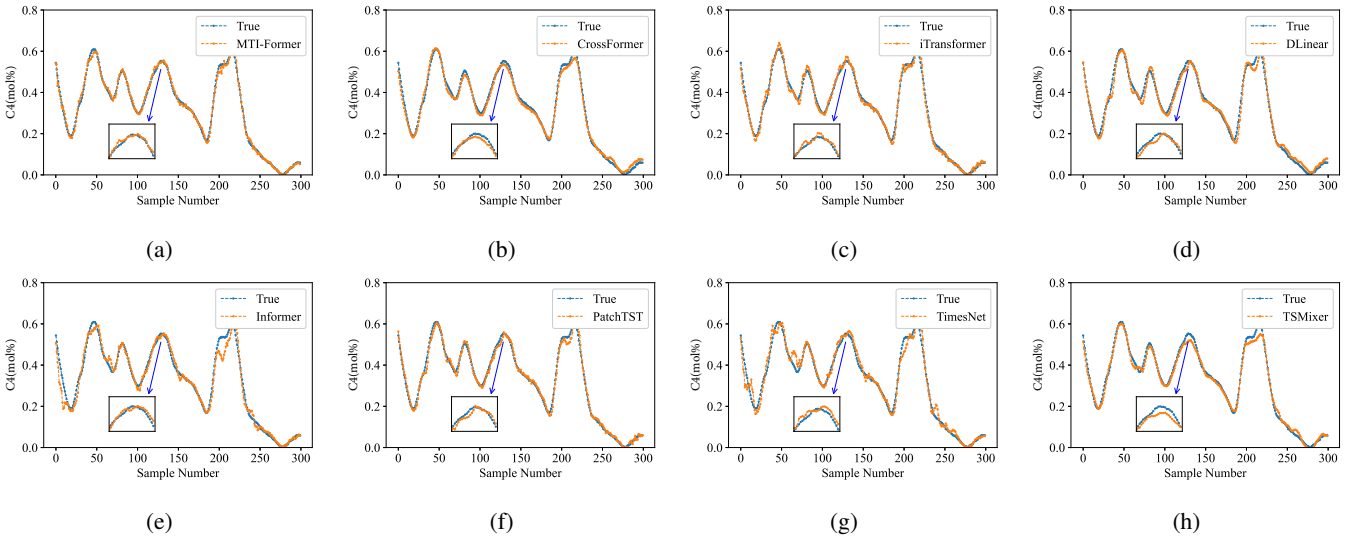


Fig. 4: Comparison of different methods for predicting C4 in the debutanizer column: (a) MTI-Former, (b) CrossFormer, (c) iTransformer, (d) DLinear, (e) Informer, (f) PatchTST, (g) TimesNet, (h) TSMixer.

TABLE II: A comparison of eight approaches to predicting C4 in the debutanizer column.

Method	Metrics	Predict window length		
		1	3	5
MTI-Former	MAE	**0.0092 ± 0.0002**	**0.0164 ± 0.0003**	**0.0290 ± 0.0005**
	RMSE	**0.0122 ± 0.0003**	**0.0200 ± 0.0005**	**0.0391 ± 0.0008**
CrossFormer	MAE	0.0130 ± 0.0006	0.0178 ± 0.0008	0.0390 ± 0.0015
	RMSE	0.0157 ± 0.0007	0.0216 ± 0.0011	0.0463 ± 0.0022
DLinear	MAE	0.0161 ± 0.0009	0.0226 ± 0.0012	0.0379 ± 0.0018
	RMSE	0.0208 ± 0.0011	0.0300 ± 0.0016	0.0438 ± 0.0025
PatchTST	MAE	0.0129 ± 0.0005	0.0248 ± 0.0011	0.0332 ± 0.0016
	RMSE	0.0155 ± 0.0006	0.0330 ± 0.0014	0.0440 ± 0.0021
TimesNet	MAE	0.0219 ± 0.0012	0.0287 ± 0.0019	0.0444 ± 0.0026
	RMSE	0.0300 ± 0.0018	0.0390 ± 0.0023	0.0725 ± 0.0038
TSMixer	MAE	0.0136 ± 0.0007	0.0252 ± 0.0013	0.0350 ± 0.0017
	RMSE	0.0175 ± 0.0009	0.0324 ± 0.0015	0.0451 ± 0.0023
iTransformer	MAE	0.0139 ± 0.0006	0.0185 ± 0.0009	0.0316 ± 0.0014
	RMSE	0.0165 ± 0.0008	0.0260 ± 0.0012	0.0461 ± 0.0020
Informer	MAE	0.0202 ± 0.0011	0.0238 ± 0.0016	0.0326 ± 0.0021
	RMSE	0.0303 ± 0.0017	0.0335 ± 0.0020	0.0478 ± 0.0029

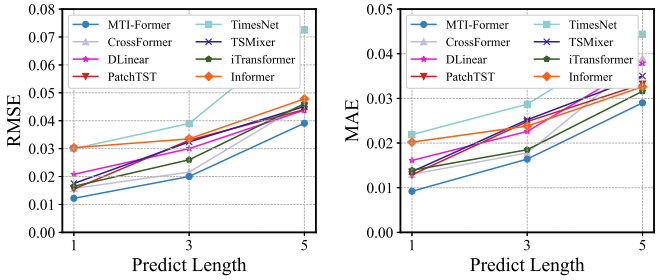


Fig. 5: Performance curve with the predicted length for predicting C4 in the debutanizer column: (a) RMSE, (b) MAE.

was set to Adam [46], the learning rate was selected from $\{10^{-3}, 5\times10^{-4}, 10^{-4}\}$, the number of layers from $\{1, 2, 3\}$, the series dimension d_{model} from $\{128, 256, 512\}$, and the number of attention heads h from $\{2, 4, 8\}$. For baseline models with architecture-specific hyperparameters, the optimal search spaces and recommended practices provided in their original papers were followed. The final hyperparameter configurations were determined based on the minimum mean squared error loss achieved on the validation set. To effectively extract multi-scale frequency-domain features from industrial data, a two-level WT was employed. Table I presents the detailed hyperparameter optimization combinations for MTI-Former. To investigate hyperparameter sensitivity, the impacts of the DWT basis function and the decomposition level were analyzed, with the corresponding results depicted in Fig. 3. Among the coiflets (C), symlets (S), haar (H), and daubechies (D) families, the symlets wavelet proved most effective at decoupling dynamic features of the signal. Furthermore, a two-level decomposition strikes the best balance, as a single level provides insufficient feature separation while three levels result in unnecessary complexity.

The quantitative results for C4 prediction are summarized in Table II. The results indicate that for the tested prediction horizons of 1, 3, and 5 steps, MTI-Former achieves the lowest MAE and RMSE values. In contrast, while CrossFormer shows comparable performance for short-term predictions, its global cross-variable attention incorporates irrelevant relationships. iTransformer, by treating variables as tokens, fails to adequately capture local temporal patterns. The probsparse attention mechanism of the Informer overlooks critical dynamic dependencies. DLinear is unsuitable for the nonlinear characteristics of industrial data due to its linear mapping. Although PatchTST shows promise in short-term forecasting, it disrupts the continuity of global trends, limiting its long-term efficacy. TimesNet, with its fixed-kernel 2D convolutions, lacks the adaptability for dynamically changing frequencies in industrial data, resulting in the poorest performance. TSMixer, despite its multi-scale mixing design, is constrained by its linear

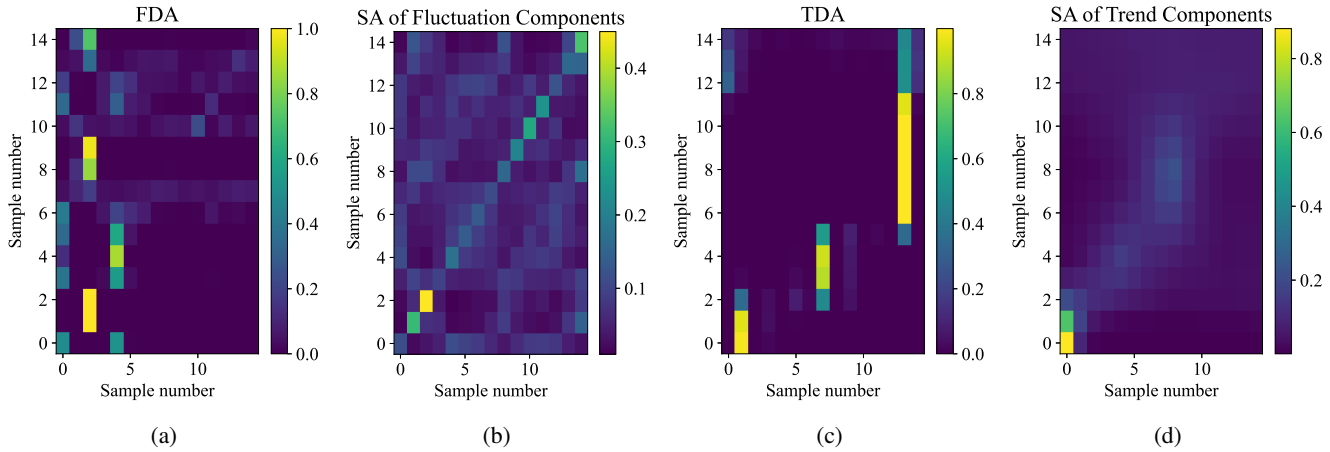


Fig. 6: Comparative visualization of attention heatmaps for fluctuation and trend components for 15 selected samples from the debutanizer column process. The heatmaps contrast our proposed TFDA module with standard SA applied to the decoupled components. (a)FDA: Heatmap of the cross-attention between the fluctuation component and multi-scale features obtained from the TFIA module;(b) Fluctuation SA: Heatmap of the SA mechanism within the fluctuation component;(c) TDA: Heatmap of the cross-attention between the trend component and the multi-scale features obtained from the TFIA module; (d)Trend SA: Heatmap of the self-attention within the trend component.

and low-order interactions, rendering it less effective than MTI-Former. Conversely, the proposed MTI-Former extracts comprehensive representations across distinct frequencies via multi-scale decomposition. By employing TFIA to capture complex interactions among these multi-scale features and utilizing TFDA to independently enhance local details and global trends, our model achieves robust predictive performance in both short and long-term forecasting scenarios.

TABLE III: Ablation study of MTI-Former components for C4 prediction.

Model Components				Metrics	
AHEF	TFIA	TDA	FDA	MAE	RMSE
✓	✓	✓	✓	0.0092	0.0122
✗	✓	✓	✓	0.0120	0.0169
✓	✗	✓	✓	0.0117	0.0157
✓	✓	✗	✓	0.0107	0.0149
✓	✓	✓	✗	0.0126	0.0178

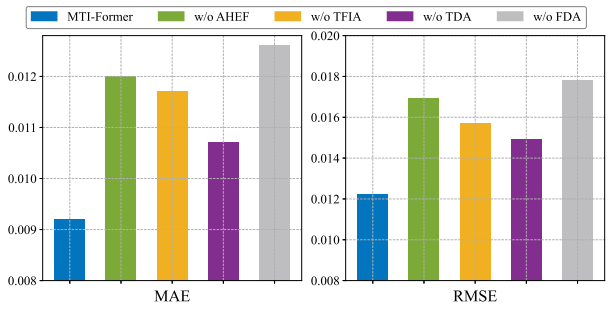


Fig. 7: Performance of MTI-Former variants for C4 prediction in the ablation study.

The prediction curves for a 1-step prediction horizon are shown in Fig. 4. The linear model DLinear struggles to track changes in the process variable. Informer, TimesNet, and TSMixer deviate from the ground truth, especially at peaks and troughs. Although recent transformer variants, including CrossFormer, iTransformer, and PatchTST, capture the overall trend effectively, they still show deviations at points of rapid variation. In contrast, the proposed MTI-Former demonstrates superior tracking capability, capturing the process dynamics more accurately compared to other methods.

Fig. 5 shows the effect of prediction length on model performance. The proposed MTI-Former method clearly maintains significantly higher prediction accuracy than all competing methods at all tested prediction lengths. Although the performance of all models deteriorates as the prediction horizon extends, the proposed MTI-Former exhibits the most gradual decline, thereby demonstrating its robustness and superiority in multi-step forecasting scenarios.

To systematically validate the effectiveness and individual contributions of the key components within the proposed MTI-Former architecture, a series of comprehensive ablation experiments were conducted on the debutanizer column C4 prediction task. Four MTI-Former variants are designed as follows:

- **w/o AHEF:** MTI-Former removed the AHEF module. The high-frequency and low-frequency components obtained from DWT layer are fed directly into the TFIA module, resulting in the loss of crucial transient details.
- **w/o TFIA:** MTI-Former removed the TFIA module. An inverse DWT is directly performed on the high-frequency and low-frequency components generated by DWT decomposition, and the resulting reconstructed sequence is subsequently processed by a standard self-attention mechanism, which ignores the interaction-aware fusion of trend and fluctuation information across scales.
- **w/o TDA:** MTI-Former removed the TDA branch from the TFDA module, ignoring the specific decoupling of long-term trend features.
- **w/o FDA:** MTI-Former removed the FDA branch from the TFDA module, ignoring the specific decoupling of short-term fluctuation features.

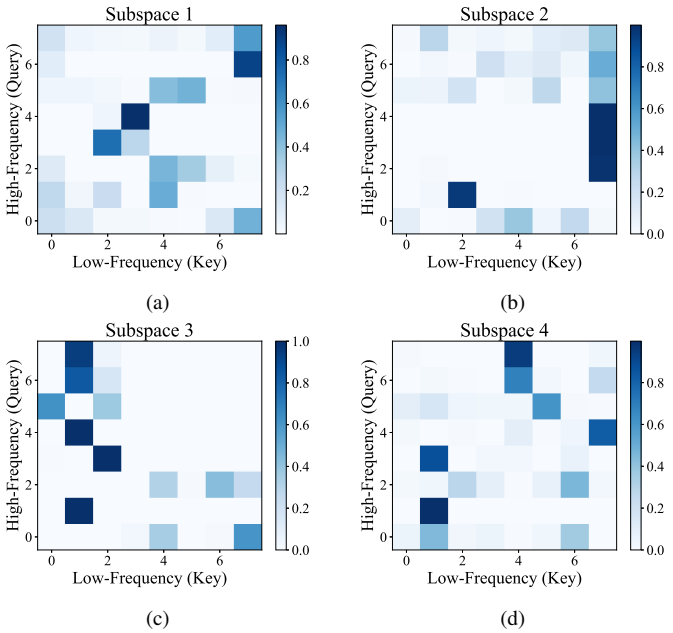


Fig. 8: Visualization of the TFIA module for C4 prediction. Detailed heatmaps of the cross-attention scores between the trend component and the fluctuation components.(a)Attention map of subspace 1;(b)Attention map of subspace 2;(c)Attention map of subspace 3;(d)Attention map of subspace 4.

As observed from Table III and Fig. 7, the complete MTI-Former achieves the best overall performance. The w/o AHEF variant performs worse because it does not amplify high-frequency components associated with important process disturbances. A pronounced performance degradation is also seen in the w/o TFIA variant, which cannot model the interaction between long-term operating trends and short-term fluctuations. Furthermore, removing either the TDA or FDA branch leads to clear performance losses, underscoring the importance of explicitly extracting both trend-specific and fluctuation-specific features.

Fig. 6 visualizes attention maps for 15 samples to compare the TFDA module with a standard self-attention (SA) mechanism. Standard SA is applied to the trend and fluctuation components reconstructed via iDWT. The proposed TFDA module yields substantially more structured attention patterns than SA. In particular, the FDA map in Fig. 6(a) exhibits a sparse, focused distribution that highlights specific transient variations, whereas the TDA map in Fig. 6(c) displays vertically aligned structures that reflect long-term dependencies. By contrast, the SA-based maps in Figs. 6(b) and 6(d) are diffuse and less selective. These visualizations demonstrate the efficacy of TFDA in decoupling and emphasizing distinct temporal properties, whereas the standard SA struggles to disentangle the unique temporal dynamics inherent in the industrial process.

To elucidate the internal mechanism of the TFIA module, the cross-attention maps of the trend and fluctuation components from the first decomposition level are visualized in Fig. 8, depicting the respective attention patterns from four subspaces. In this visualization, the fluctuation component acts as the query, while the trend component serves as the key. The resulting sparse and focused attention patterns clearly indicate the inherent correlation between transient fluctuations and long-term trends, validating the necessity of explicitly modeling their interactions.

B. Ironmaking Process

The blast furnace is the core unit in the ironmaking process, where iron-bearing materials, fuel, and fluxes are charged from the top in designated proportions. The burden descends through the furnace, undergoing continuous heat transfer, melting, and reduction reactions, and finally generates molten iron.

TABLE IV: Detailed description of the input variables in the blast furnace ironmaking process.

Input	Variable description	Unit
U1	Si(n-1)	wt%
U2	Belly gas index	$m^3/min{\cdot}m^2$
U3	Cold air volume	m^3/h
U4	Top pressure	kPa
U5	Differential pressure	kPa
U6	Theoretical combustion temperature	°C
U7	Oxygen-rich rate	%
U8	Standardized wind speed	m/s
U9	Hot blast temperature	°C
U10	Blast kinetic energy	kJ/s

TABLE V: The best hyperparameter configuration of MTI-Former for silicon content prediction in the blast furnace ironmaking process.

Symbol	Description	Predict window length		
		1	3	5
len	Length of window size	20	20	20
d_{model}	Embedding dimension	128	256	256
h	Number of attention heads	8	8	8
N_e	Encoder layer number	2	2	3
d_{ff}	Nonlinear dimension	256	512	512
L	Number of Wavelet Decomposition Layers	3	3	3
W_f	Wavelet function	D	D	D

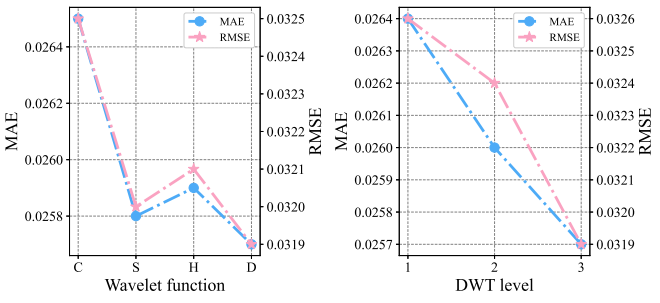


Fig. 9: Hyperparameter sensitivity analysis for blast furnace ironmaking process.

The data was collected from the real-world operation records in a steel plant, with the sampling interval set to 1 hour.

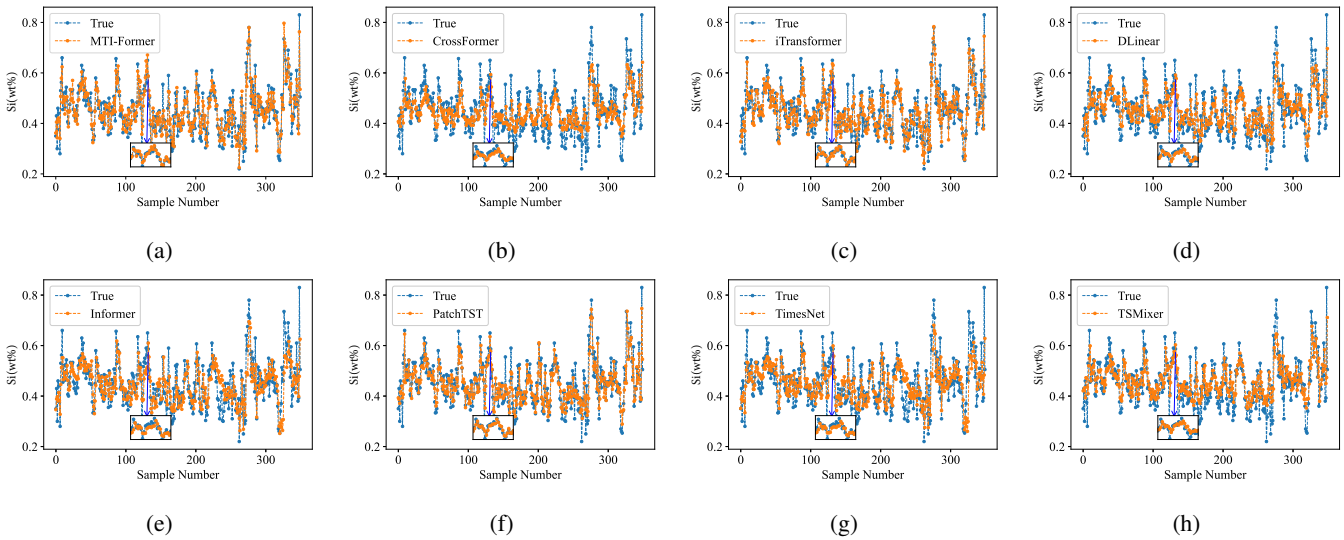


Fig. 10: Comparison of different methods for predicting silicon content in the blast furnace ironmaking process: (a) MTI-Former, (b) CrossFormer, (c) iTransformer, (d) DLinear, (e) Informer, (f) PatchTST, (g) TimesNet, (h) TSMixer.

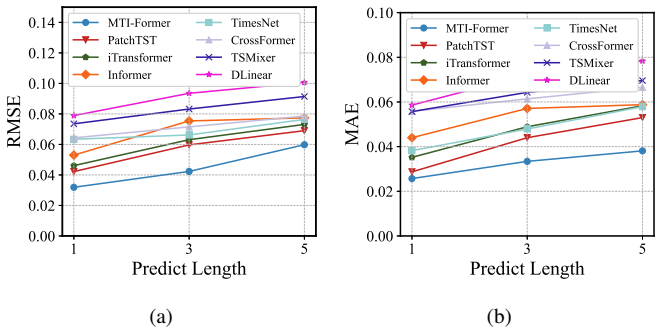


Fig. 11: Performance curve with the predicted length for silicon content in the blast furnace ironmaking process: (a) RMSE, (b) MAE.

The dataset contains approximately 3% missing entries due to sensor instability, which were filled using linear interpolation to maintain temporal continuity. The target silicon content exhibits substantial process variability, evidenced by a 22.71% coefficient of variation and a moving-window variance using a window size of 20 that ranges from 0.00071 to 0.038163. To eliminate the impact of sensor noise and abnormal operating conditions, the raw data underwent preprocessing to remove samples containing obvious outliers. The selection of model input variables followed a rigorous screening process combining ironmaking mechanisms, on-site operator experience, and data-driven correlation analysis. Based on smelting mechanisms, variables physically related to reduction reactions and the furnace thermal state were pre-selected. Moreover, the practical experience of on-site operators was incorporated to ensure the selected variables align with key monitoring indicators used in actual production. Finally, Pearson Correlation Coefficients (PCC) were calculated to quantify the relationship between candidate variables and silicon content, ensuring that only variables with significant statistical correlation were retained. Ultimately, 10 key process variables were selected as inputs. The candidate input variables of the model are listed in Table IV. The target variable is the silicon content ([Si]) in molten iron. A total of 2367 labeled samples were collected,

TABLE VI: A comparison of eight approaches to predicting silicon content in the blast furnace ironmaking process.

Method	Metrics	Predict window length		
		1	3	5
MTI-Former	MAE	**0.0257 ± 0.0006**	**0.0334 ± 0.0008**	**0.0381 ± 0.0011**
	RMSE	**0.0319 ± 0.0007**	**0.0423 ± 0.0010**	**0.0598 ± 0.0015**
CrossFormer	MAE	0.0558 ± 0.0018	0.0614 ± 0.0022	0.0666 ± 0.0025
	RMSE	0.0643 ± 0.0021	0.0715 ± 0.0026	0.0786 ± 0.0031
DLinear	MAE	0.0586 ± 0.0024	0.0720 ± 0.0029	0.0784 ± 0.0035
	RMSE	0.0789 ± 0.0032	0.0935 ± 0.0038	0.1006 ± 0.0042
PatchTST	MAE	0.0287 ± 0.0011	0.0440 ± 0.0016	0.0530 ± 0.0020
	RMSE	0.0422 ± 0.0015	0.0597 ± 0.0021	0.0690 ± 0.0028
TimesNet	MAE	0.0382 ± 0.0015	0.0479 ± 0.0019	0.0579 ± 0.0023
	RMSE	0.0634 ± 0.0024	0.0662 ± 0.0027	0.0764 ± 0.0032
TSMixer	MAE	0.0557 ± 0.0020	0.0643 ± 0.0025	0.0696 ± 0.0029
	RMSE	0.0735 ± 0.0028	0.0832 ± 0.0033	0.0914 ± 0.0039
iTransformer	MAE	0.0352 ± 0.0013	0.0489 ± 0.0017	0.0583 ± 0.0022
	RMSE	0.0461 ± 0.0016	0.0631 ± 0.0024	0.0731 ± 0.0030
Informer	MAE	0.0440 ± 0.0019	0.0571 ± 0.0023	0.0588 ± 0.0026
	RMSE	0.0530 ± 0.0022	0.0754 ± 0.0031	0.0774 ± 0.0034

with 2067 selected for model training and 300 reserved for testing.

Table V presents the detailed hyperparameter optimization combinations for MTI-Former. The optimal hyperparameter combinations for both the proposed method and all baseline models were determined through a unified grid search strategy. A hyperparameter sensitivity analysis was conducted to optimize the DWT component. As illustrated in Fig. 9, this analysis evaluated the impact of various wavelet basis functions and decomposition levels. The daubechies ('D') wavelet function produces the most accurate results. Additionally, a decomposition level of three was determined to be optimal, as it consistently minimized both MAE and RMSE.

Table VI compares eight methods for predicting the silicon

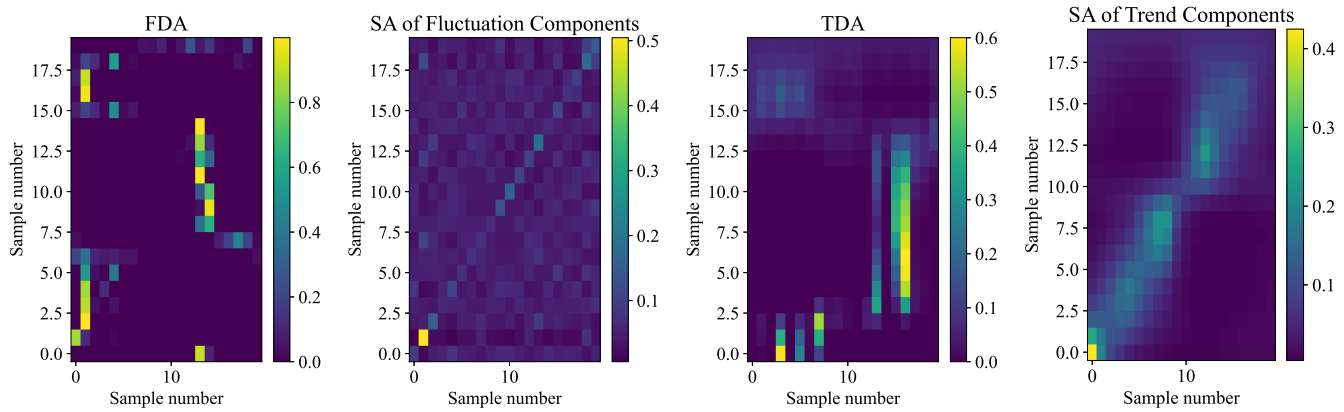


Fig. 12: Comparative visualization of attention heatmaps for fluctuation and trend components for 15 selected samples from the blast furnace ironmaking process. The heatmaps contrast our proposed TFDA with standard SA applied to the decoupled components: (a) FDA: Heatmap of the cross-attention between the fluctuation component and multi-scale features obtained from the TFIA module; (b) Fluctuation SA: Heatmap of the self-attention within the fluctuation component; (c) TDA: Heatmap of the cross-attention between the trend component and multi-scale feature obtained from the TFIA module; (d) Trend SA: Heatmap of the self-attention within the trend component.

content in blast furnace ironmaking process. The proposed MTI-Former demonstrates consistent superiority across all prediction horizons, achieving the best performance in both MAE and RMSE metrics. This can be attributed to its capability in handling the complex multiscale characteristics inherent in industrial data through DWT and novel attention mechanisms. PatchTST and iTransformer show relatively competitive performance in short-term predictions but are consistently outperformed by our approach, particularly in longer horizons. Other methods, including CrossFormer, DLinear, TimesNet, TSMixer, and Informer, exhibit substantially larger errors.

TABLE VII: Ablation study of MTI-Former components for blast furnace ironmaking.

Model Components				Metrics	
AHEF	TFIA	TDA	FDA	MAE	RMSE
✓	✓	✓	✓	0.0257	0.0319
✗	✓	✓	✓	0.0273	0.0343
✓	✗	✓	✓	0.0278	0.0349
✓	✓	✗	✓	0.0264	0.0326
✓	✓	✓	✗	0.0261	0.0321

To visually assess the prediction results, the predicted curves of the eight methods for silicon content in the blast furnace ironmaking process are compared against the ground truth curves, as shown in Fig. 10. It is evident that the predicted curve of the proposed MTI-Former closely aligns with the ground truth, further demonstrating that the proposed model better adapts to the multivariate, strongly coupled industrial process.

Fig. 11 shows the error trends of the models across prediction step lengths. The error growth curve of MTI-Former has the lowest slope, indicating better long-term prediction accuracy. Its accuracy degrades more slowly than other methods, demonstrating greater robustness.

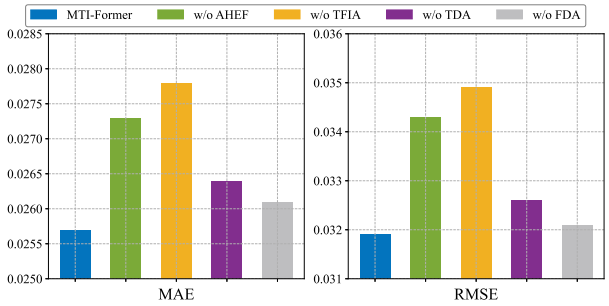


Fig. 13: Performance of MTI-Former variants for blast furnace ironmaking in the ablation study.

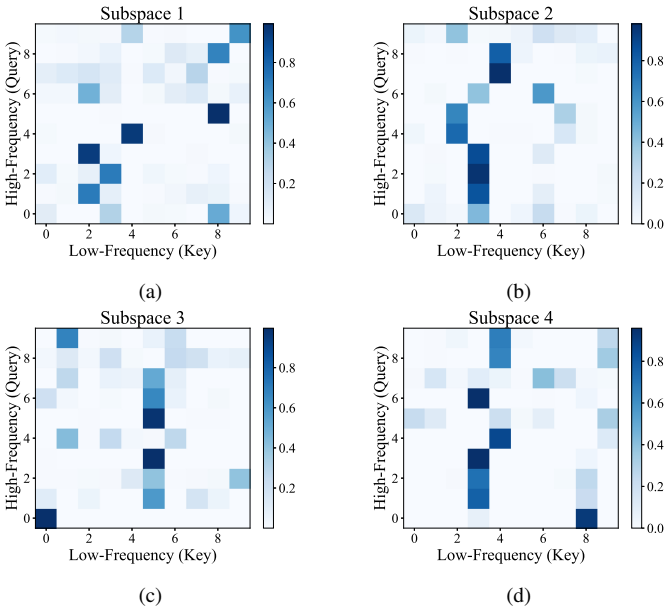


Fig. 14: Visualization of the TFIA for blast furnace ironmaking. Detailed heatmaps of the cross-attention scores between the trend component and the fluctuation components.(a)Attention map of subspace 1;(b)Attention map of subspace 2;(c)Attention map of subspace 3;(d)Attention map of subspace 4.

As shown in Table VII and Fig. 13, the ablation study on the blast furnace ironmaking process further validates the effectiveness of MTI-Former. The performance of the model is most sensitive to the removal of the TFIA module, confirming that capturing the interaction between long-term operating trends and short-term process disturbances is key to obtaining high accuracy. Concurrently, removing the AHEF, TDA, or FDA modules leads to an increase in prediction error, indicating that the feature amplification and decoupling strategies are indispensable for the final prediction.

To further investigate the internal mechanisms of the proposed model, Fig. 12 illustrates that the TFDA module generates highly structured attention patterns. Specifically, the FDA map in Fig. 12(a) leverages a sparse attention mechanism to blue precisely localize process disturbances, while the TDA map in Fig. 12(c) captures long-term dependencies, both of which stand in clear contrast to the diffuse and unfocused patterns exhibited by the standard SA maps in Figs. 12(b) and 12(d). This effective decoupling of temporal dynamics is enabled by the TFIA module, whose internal cross-attention mechanism is depicted in Fig. 14. The resulting sparse attention blocks indicate the strong intrinsic correlations between transient process fluctuations and the long-term quality indices.

C. Efficiency study

The comparative results are presented in Table VIII. All models were run with a batch size of 32. Training time is reported as the total seconds for the entire training run on the training set. Inference time is reported in milliseconds (ms) for the entire test set. As observed, models including DLinear, TSMixer, iTransformer, and PatchTST exhibit the lowest computational costs. However, these methods tend to overlook the dynamic interaction between long-term dependencies and short-term fluctuations in industrial data, which leads to suboptimal performance.

TABLE VIII: Computational cost comparison on the debutanizer column and blast furnace ironmaking datasets.

Dataset	Method	Computational Cost	
		Training (s)	Inference (ms)
Debutanizer Column	MTI-Former	204.92	30.56
	CrossFormer	496.00	17.84
	DLinear	23.46	0.86
	Informer	239.47	22.00
	iTransformer	60.84	2.91
	PatchTST	55.02	7.10
	TimesNet	242.00	24.00
	TSMixer	43.43	1.30
Blast Furnace Ironmaking	MTI-Former	165.73	56.08
	CrossFormer	183.49	18.37
	DLinear	13.50	0.36
	Informer	138.00	32.80
	iTransformer	37.45	16.68
	PatchTST	54.74	14.64
	TimesNet	169.00	28.05
	TSMixer	25.13	1.61

Conversely, complex transformer-based architectures demonstrate a significant computational burden. Notably, CrossFormer exhibits the highest training overhead on the debutanizer dataset (496.00s).

The proposed MTI-Former achieves a practical trade-off between efficiency and accuracy. Although the multi-scale preprocessing and cross-scale attention introduce extra operations, slightly increasing inference latency compared to some baselines, the training time of MTI-Former is significantly reduced. On the debutanizer dataset, MTI-Former requires 204.92 s for the full training run, representing approximately 58% reduction compared with CrossFormer. This training efficiency gain stems from the decoupled architecture, by separating trend and fluctuation components into lightweight, specialized attention paths, MTI-Former avoids redundant pairwise token interactions that dominate the computation of full mixed-attention designs. Moreover, the inference latencies of MTI-Former (30.56 ms on debutanizer and 56.08 ms on blast-furnace ironmaking) are well within the response-time requirements of real-world industrial monitoring systems, where typical sampling intervals range from one to several seconds. Therefore, the modest increase in inference cost is justified by substantially improved prediction accuracy and robustness.

TABLE IX: Model complexity analysis on the debutanizer column and blast furnace ironmaking datasets.

Model	Complexity	
	Params(M)	Time complexity
MTI-Former	2.55	$\mathcal{O}(T^2 D)$
Transformer	1.47	$\mathcal{O}(T^2 D)$
CrossFormer	6.62	$\mathcal{O}(D \cdot T_{\text{seg}}^2 + D \cdot T)$
DLinear	0.042	$\mathcal{O}(T \cdot D)$
Informer	1.6	$\mathcal{O}(T \log T \cdot D)$
iTransformer	0.8	$\mathcal{O}(N^2 \cdot D)$
PatchTST	0.8	$\mathcal{O}((\frac{T}{T_{\text{patch}}})^2 \cdot D)$
TimesNet	11.2	$\mathcal{O}(T \log T \cdot D)$
TSMixer	0.03	$\mathcal{O}(T \cdot D)$

To further evaluate the computational efficiency of MTI-Former for real-time industrial applications, Table IX compares its parameter count and theoretical complexity against several baseline models. Specifically, T denotes the sequence length, D represents the model dimension, T_{seg} is the segment length, N represents the number of variates, and T_{patch} is the patch size. The results demonstrate that MTI-Former, with 2.55M parameters, is significantly more lightweight than complex architectures such as TimesNet (11.2M) and CrossFormer (6.62M). Although linear baselines such as DLinear and TSMixer exhibit lower parameter counts, they are constrained in their ability to capture complex nonlinear trend-fluctuation interactions, which leads to inferior predictive accuracy. Furthermore, regarding theoretical time complexity, MTI-Former exhibits a complexity of $\mathcal{O}(T^2 D)$. Although this implies a quadratic dependency on the sequence length T, sliding window sizes of $T = 15$ and $T = 20$ are adopted for the debutanizer column and blast furnace ironmaking datasets, respectively. Consequently, the computational cost of the

quadratic term remains manageable, ensuring that MTI-Former maintains high operational efficiency without sacrificing its capability to capture complex temporal dependencies.

D. Industrial Deployment

To facilitate practical deployment, MTI-Former can be integrated into existing supervisory control architectures. Process variables can be collected from distributed control systems or programmable logic controllers through standard protocols such as OPC or Modbus. The acquired time-series data are transmitted to an edge computing unit or industrial server for preprocessing, including automated outlier filtering, missing value imputation, normalization and DWT. The model then performs real-time inference and writes predictions back to the control system, while periodic offline retraining allows the model to adapt to evolving operating conditions. When migrating to a new industrial process, the core architecture of the model remains unchanged. Engineers can adapt the model by first re-selecting relevant variables based on process mechanisms and statistical correlation analysis such as the pearson correlation coefficient. Subsequently, model hyperparameters can be optimized using a validation-based grid search strategy. Specifically, a practical guideline is to select the prediction window length such it can cover at least one typical operational control cycle of the target process, while the wavelet decomposition level L can be determined according to the dominant frequency of process noise identified through preliminary spectral analysis.

V. Conclusion

Accurate soft sensing remains critical for industrial process monitoring, but existing approaches are often been constrained by single-scale modeling, making it difficult to simultaneously capture global trends, local fluctuations, and their dynamic couplings. To address these limitations, an MTI-Former is proposed in this paper. The framework has employed DWT for multi-scale decomposition, an AHEF module for adaptive enhancement of high-frequency details, a TFIA mechanism for modeling interactions between trends and fluctuations, and a TFDA mechanism for extracting their distinct temporal patterns. The effectiveness of MTI-Former has been demonstrated on two real industrial datasets. Specifically, the model has achieved improvements exceeding 10% across three prediction horizons for both debutanizer C4 prediction and blast furnace silicon prediction when compared with several state-of-the-art baselines.

In the future, several research directions can be further explored. First, it is of practical significance to extend MTI-Former to handle multivariate quality indices simultaneously. Second, adaptive wavelet selection and learnable decomposition strategies can be integrated to improve robustness under varying operating conditions. Third, explore lightweight model variants so as to facilitate deployment on industrial edge devices. Finally, investigating MTI-Former within closed-loop optimization or predictive control frameworks may enable deeper integration of soft sensing and real-time decision-making in industrial systems.

References

[1] X. Yuan, N. Xu, L. Ye, K. Wang, F. Shen, Y. Wang, C. Yang, and W. Gui, "Attention-based interval aided networks for data modeling of heterogeneous sampling sequences with missing values in process industry," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 5253–5262, 2024.

[2] N. P. Lawrence, S. K. Damarla, J. W. Kim, A. Tulsyan, F. Amjad, K. Wang, B. Chachuat, J. M. Lee, B. Huang, and R. B. Gopaluni, "Machine learning for industrial sensing and control: A survey and practical perspective," *Control Engineering Practice*, vol. 145, p. 105841, 2024.

[3] D. Liu, Y. Wang, C. Liu, X. Yuan, and C. Yang, "Multirateformer: An efficient transformer-based hierarchical network for multistep prediction of multirate industrial processes," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–13, 2024.

[4] C. Ou, H. Zhu, Y. A. W. Shardt, L. Ye, X. Yuan, Y. Wang, C. Yang, and W. Gui, "Missing-data imputation with position-encoding denoising autoencoders for industrial processes," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–11, 2024.

[5] J. Gao, J. Luo, S. Yin, C. Gong, S. Wang, and G. Zhang, "Adaptive digital twin framework for pmsm thermal safety monitoring: Integrating bayesian self-calibration with hierarchical physics-aware network," *Machines*, vol. 14, no. 2, 2026.

[6] X. Yuan, Y. Wang, C. Wang, L. Ye, K. Wang, Y. Wang, C. Yang, W. Gui, and F. Shen, "Variable correlation analysis-based convolutional neural network for far topological feature extraction and industrial predictive modeling," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–10, 2024.

[7] X. Yin, X. Li, Y. Zhang, W. Zhou, and Z. Liu, "Multi-task network based health status assessment of cutting tools," *International Journal of Network Dynamics and Intelligence*, vol. 4, no. 2, p. 100008, 2025.

[8] J. Zhu, W. Gui, Z. Chen, and Z. Jiang, "Monitoring multiple operational statuses of blast furnace via multifeature fusion from burden surface video images," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–20, 2025.

[9] W. Wang, C. Yang, S. Lou, and Y. Yang, "Causalsyng: Multivariate collaborative causal inference with dynamic knowledge-data synergy for industrial soft sensors," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–23, 2025.

[10] Z. Sun and C. Han, "Linear state estimation for multi-rate ncss with multi-channel observation delays and unknown markov packet losses," *International Journal of Network Dynamics and Intelligence*, pp. 100 005–100 005, 2025.

[11] Z. Xu, N. Xu, K. Wang, X. Yuan, Y. Wang, C. Yang, W. Gui, S. Cheng, and L. Ye, "A sampling interval-adaptive transformer for industrial time sequence modeling with heterogeneou s sampling rates in quality prediction," *Engineering Applications of Artificial Intelligence*, vol. 162, p. 112374, 2025.

[12] P. Facco, F. Doplicher, F. Bezzo, and M. Barolo, "Moving average pls soft sensor for online product quality estimation in an industrial batch polymerization process," *Journal of Process Control*, vol. 19, no. 3, pp. 520–529, 2009.

[13] S. Lou, C. Yang, W. Wang, H. Zhang, Y. Zhao, and P. Wu, "Toward in-depth mastery of statistical properties: Novel stationary moment analysis with application to continuous industrial anomaly detection," *IEEE Transactions on Cybernetics*, vol. 55, no. 7, pp. 3417–3430, 2025.

[14] A. H. Khan, X. Cao, C. Luo, S. Zhang, W. Guo, V. N. Katsikis, and S. Li, "Spiking neural networks: A comprehensive survey of training methodologies, hardware implementations and applications," *Artificial Intelligence Science and Engineering*, vol. 1, no. 3, pp. 175–207, 2025.

[15] X. Shao, J. Zhang, M. Lyu, and Y. Lu, "Event-based nonfragile state estimation for memristive neural networks with multiple

time-delays and sensor saturations," *International Journal of Systems Science*, vol. 56, no. 3, pp. 618–637, 2025.

[16] J. G. Gallareta, C. González-Menorca, P. Muñoz, and M. V. Vasic, "Advancements in soft-sensor technologies for quality control in process manufacturing: A review," *IEEE Sensors Journal*, vol. 25, no. 9, pp. 14 575–14 588, 2025.

[17] Y. S. Perera, D. Ratnaweera, C. H. Dasanayaka, and C. Abeykoon, "The role of artificial intelligence-driven soft sensors in advanced sustainable process industries: A critical review," *Engineering Applications of Artificial Intelligence*, vol. 121, p. 105988, 2023.

[18] G. Tian, Y. Yang, and S. Wen, "Time-series stock price forecasting based on neural networks: A comprehensive survey," *Artificial Intelligence Science and Engineering*, vol. 1, no. 4, pp. 255–277, 2025.

[19] X. Yuan, J. Zhou, B. Huang, Y. Wang, C. Yang, and W. Gui, "Hierarchical quality-relevant feature representation for soft sensor modeling: A novel deep learning strategy," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 3721–3730, 2020.

[20] H. Ding, K. Hao, L. Chen, and X. Cai, "Variational information inference: An interpretable disentangled transfer learning quality prediction for multirate industrial processes," *IEEE Transactions on Cybernetics*, vol. 55, no. 10, pp. 4620–4633, 2025.

[21] C. Liu, Y. Wang, Y. Fang, C. Yang, and W. Gui, "Operating condition recognition of industrial flotation processes using visual and acoustic bimodal autoencoder with manifold learning," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 5, pp. 7428–7439, 2024.

[22] Y. Fang, Z. Jiang, W. Gui, and L. Shen, "Towards reliable control: Uncertainty-aware domain preserving stacked autoencoder for data-driven modeling in large-scale industrial systems," *Control Engineering Practice*, vol. 164, p. 106383, 2025.

[23] Z. Chen, C. Du, B. Zhang, C. Chen, and W. Gui, "Multi-step joint probabilistic forecasting of offshore wind power: A confidence-triggered clustering missing-data tolerant model," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 12, pp. 9802–9812, 2025.

[24] Y. Zhang, P. Zhou, and E. Ma, "Anomaly detection of industrial smelting furnace incorporated with accelerated sampling denoising diffusion probability model and conv-transformer," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–11, 2024.

[25] Z. Yang, K. Wang, X. Yuan, Y. Wang, C. Yang, W. Gui, L. Ye, and F. Shen, "Gaussian-based interval-aware transformer with interval embedding for data sequence modeling with irregular sampling frequency in industrial processes," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 8, pp. 5811–5821, 2025.

[26] D. Liu, Y. Wang, C. Liu, X. Yuan, C. Yang, and W. Gui, "Data mode related interpretable transformer network for predictive modeling and key sample analysis in industrial processes," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 9, pp. 9325–9336, 2023.

[27] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[28] H. SHAN, C. WEI, N. RAMOS, X. YANG, C. GUO, H. LI, and Y. CHEN, "Neuromorphic computing in the era of large models," *Artificial Intelligence Science and Engineering*, vol. 1, no. 1, pp. 17–30, 2025.

[29] Y. Wang, C. Wen, and X. Wu, "Fault detection and isolation of floating wind turbine pitch system based on kalman filter and multi-attention 1dcnn," *Systems Science & Control Engineering*, vol. 12, no. 1, p. 2362169, 2024.

[30] D. Wang, C. Wen, and X. Feng, "Deep variational luenberger-type observer with dynamic objects channel-attention for stochastic video prediction," *International Journal of Systems Science*, vol. 55, no. 4, pp. 728–740, 2024.

[31] Y. Li, Z. Yao, L. Mao, B. Shi, L. Yao, and J. Song, "Cnn-assisted adaptive signal decomposition method for correcting the distorted calibration signals of pressure sensors," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–10, 2024.

[32] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *International conference on machine learning*. PMLR, 2022, pp. 27 268–27 286.

[33] F. Yan, X. Zhang, and C. Yang, "A graph-based time-frequency two-stream network for multistep prediction of key performance indicators in industrial processes," *IEEE Transactions on Cybernetics*, vol. 54, no. 11, pp. 6867–6880, 2024.

[34] Z. Zhang, Y. Chen, D. Zhang, Y. Qian, and H. Wang, "Ctfnet: Long-sequence time-series forecasting based on convolution and time–frequency analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 16 368–16 382, 2024.

[35] T. Bai, D. Peng, and J. Tian, "Wavelet transformer-based multi-channel and multiresolution information perceptron for lithium-ion battery state of health estimation," *IEEE Transactions on Transportation Electrification*, vol. 11, no. 4, pp. 9470–9482, 2025.

[36] Y. Cao and X. Zhao, "Dwtformer: Wavelet decomposition transformer with 2d variation for long-term series forecasting," in *2023 IEEE 6th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)*, vol. 6, 2023, pp. 1548–1558.

[37] L. Sasal, T. Chakraborty, and A. Hadid, "W-transformers: A wavelet-based transformer framework for univariate time series forecasting," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2022, pp. 671–676.

[38] M. Danielak, O. Borova, P. Kiełbasa, J. Gwiżdż, and Ł. Gierz, "Detection of production discontinuities in the extrusion process based on current signature by means of a discrete wavelet transform," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2580053, 2025.

[39] S. A. Magid, Y. Zhang, D. Wei, W.-D. Jang, Z. Lin, Y. Fu, and H. Pfister, "Dynamic high-pass filtering and multi-spectral attention for image super-resolution," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4268–4277.

[40] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *The eleventh international conference on learning representations*, 2023.

[41] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023.

[42] S. T. Hussain Rizvi, N. Kanwal, and M. Naeem, "Bridging simplicity and sophistication using glinear: A novel architecture for enhanced time series prediction," *Digital Signal Processing*, p. 105702, 2025.

[43] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.

[44] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," *arXiv preprint arXiv:2210.02186*, 2022.

[45] V. Ekambaram, A. Jati, N. Nguyen, P. Sinthong, and J. Kalagnanam, "Tsmixer: Lightweight mlp-mixer model for multivariate time series forecasting," in *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, 2023, pp. 459–469.

[46] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.