A Dual-Network iTransformer Model for Robust and Efficient Time Series Forecasting

Ary Mazharuddin Shiddiqi [1,*)], Bagaskoro Kuncoro Ardi [2)], Bilqis Amaliah [3)], I Komang Ari Mogi [4)], Agung Mustika Rizki [5)], Bintang Nuralamsyah [6)], Ilham Gurat Adillion [7)], and Moch. Nafkhan Alzamzami [8)]

[1, 2, 3, 4, 5, 6, 7, 8)] Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

E-mail: ary.shiddiqi@its.ac.id[1)], 05111940000048@student.its.ac.id[2)], bilqis@its.ac.id[3)], 7025241014@student.its.ac.id[4)], 7025241006@student.its.ac.id[5)], bintang@its.ac.id [6)], ilhamgurata@its.ac.id[7)], and nafkhan@its.ac.id[8)]

ABSTRACT

This paper proposes a novel Dual-Network Time Series Forecasting Model that integrates a fast learner based on iTransformer with a slow learner for long-range context aggregation. The fast learner captures high-frequency local patterns efficiently, while the slow learner encodes broader temporal dependencies to improve robustness against noisy and non-stationary data. The model is evaluated across four benchmark datasets: Electricity Transformer Temperature (ETT), Exchange Rate, Weather, and Influenza-like Illness (ILI). Experiments are conducted on forecasting horizons ranging from 96 to 720 steps ahead. Performance is assessed using Mean Absolute Error (MAE) and Mean Squared Error (MSE), with the proposed model achieving state-of-the-art results, particularly on long-term forecasting tasks. Ablation studies and t-tests confirm the statistical significance and robustness of the dual-network design.

Keywords: Long-term time series forecasting, DualNet transformer, frequency-domain representation, inter-series dependencies

1. Introduction

Time-series forecasting is a crucial process across various domains, including economics, healthcare, and meteorology, where accurate and timely predictions are essential for informed decision-making. As data volume and complexity increase, the demand for efficient and effective forecasting methods becomes more pressing [1]. Researchers have developed numerous approaches to address this challenge, ranging from classical statistical methods, such as ARIMA, to advanced machine learning and deep learning models [2], [3]. These modern approaches are increasingly evaluated for their ability to outperform traditional techniques, particularly in capturing nonlinearities, long-term dependencies, and multi-scale patterns in real-world datasets.

Several algorithms have demonstrated effectiveness in time-series forecasting, including the Online Sequential Extreme Learning Machine (OS-ELM) [4], Online Recurrent Extreme Learning Machine (OR-ELM) [5], and the Inverted Transformer (iTransformer) [6]. OS-ELM is designed for high-speed online learning, enabling models to adapt to new data without retraining from scratch. Its advantage lies in its ability to process incoming data sequentially, making it highly suitable for real-time forecasting applications that require rapid responses to data changes.

Conversely, OR-ELM integrates recurrent components into the Extreme Learning Machine (ELM) framework, allowing it to capture temporal dependencies in time-series data. This capability is particularly beneficial for handling long-term dependencies commonly found in sequential data [5]. Building upon this, the Recurrent Extreme Learning Machine (Recurrent-ELM) introduces self-recurrent connections within the hidden layer, enhancing its ability to model temporal patterns in online process regression tasks [7]. Meanwhile, iTransformer adopts a novel approach by reversing the traditional Transformer architecture, treating the entire time series as a single token.

* Corresponding author.
Received: March 15th, 2025. Revised: April 13th, 2025. Accepted: May 14th, 2025.

Available online: July 8th, 2025.

DOI: https://doi.org/10.12962/j24068535.v23i1.a1264

This design improves model interpretability and facilitates the processing of complex multivariate information [6]. Further advancing Transformer architectures, the Time-Transformer AAE integrates Temporal Convolutional Networks and Transformers in a parallel design, effectively capturing both local and global features in time-series data and demonstrating superior performance in generating realistic synthetic sequences [8].

Despite the advancements in forecasting methods, significant challenges remain, particularly in handling non-stationary data, intricate seasonal patterns, and dynamic external factors. Research has shown that deep learning models, such as Transformers, have become a new standard in forecasting due to their superior ability to capture long-range dependencies compared to conventional methods [9], [10]. However, traditional Transformer architectures demand substantial computational resources and may be inefficient for high-frequency time-series data or datasets with multiple variables. Consequently, developing more efficient architectures, such as iTransformer and hybrid methods, is essential to enhance predictive accuracy while maintaining computational efficiency [6].

Recent approaches have focused on enhancing robustness to non-stationary data. For instance, Non-stationary Transformer (NST) integrates series stationarization and de-stationary attention mechanisms to better handle temporal shifts in data distributions [11]. Similarly, Spacetimeformer reformulates multivariate forecasting as a spatiotemporal sequence modeling task, enabling the Transformer to effectively capture both temporal and spatial dependencies [12]. Comprehensive reviews have also highlighted persistent challenges in deep learning-based time-series forecasting, particularly regarding seasonality, scale, and interpretability [13].

Studies have shown that time-series forecasting models are often vulnerable to noise introduced by sensor faults, outliers, or environmental variations, which can significantly degrade performance [14], [15]. To address this, recent work has proposed noise-resilient approaches, including curriculum learning with Transformers [14] and decomposition-based techniques for separating signal from noise components [15]. Despite these advances, integrating robust noise handling into lightweight and interpretable forecasting architectures remains an ongoing challenge. Thus, there is a need for approaches that not only improve forecast accuracy and efficiency but also ensure resilience against noise and data inconsistencies.

Moreover, as technology advances and the demand for more precise forecasts increases, research in time-series forecasting continues to evolve. Various deep learning-based approaches, such as long-short-term memory (LSTM) and temporal convolutional networks (TCN), have shown strong capabilities in capturing complex temporal patterns [16], [17]. Recent studies have explored architectural innovations, including graph neural networks and hybrid models, to address the limitations of traditional methods [13], [18]. However, challenges persist in terms of model interpretability and computational efficiency, particularly for real-time applications. Thus, there is a need for approaches that not only improve forecast accuracy but also enhance efficiency in both training and inference.

In this study, we propose the **Dual-Net iTransformer**, a model that combines the strengths of iTransformer with a dual-network approach to enhance accuracy and efficiency in time-series forecasting. Dual-Net iTransformer is expected to outperform existing methods, including OS-ELM, OR-ELM, and iTransformer, while addressing their limitations. Through this research, we aim to provide deeper insights into the effectiveness of these methods and demonstrate the contributions of Dual-Net iTransformer in improving time-series forecasting performance.

The remainder of this paper is organized as follows. Section 2 reviews the related works that form the foundation of our approach. Section 3 presents our methodology, detailing the proposed framework and technical implementation. Section 4 describes our experimental setup and presents the results of our evaluation. Section 5 provides a discussion of the findings, their implications, and limitations of the current work. Finally, Section 6 concludes the paper with a summary of contributions and outlines directions for future research.

2. Related Works

Time-series forecasting has been extensively studied using various machine learning and deep learning approaches. Traditional statistical models such as ARIMA [1] and exponential smoothing remain widely used but struggle to capture long-term dependencies and nonlinear patterns in complex datasets. With the rise of deep

learning, recurrent neural networks (RNNs)—particularly Long Short-Term Memory (LSTM) networks [16]—have gained popularity due to their ability to retain information across extended sequences. However, LSTMs often suffer from vanishing gradients and computational inefficiencies when handling large-scale time-series data [17].

To address these challenges, the **Extreme Learning Machine (ELM)** framework emerged as a fast and efficient alternative to conventional neural networks. ELM is a single hidden-layer feed forward neural network (SLFN) that has gained significant attention due to its rapid training speed, ease of implementation, and minimal human intervention [4], [5]. ELM-based techniques have been applied across multiple domains, including classification, regression, and feature learning, and have demonstrated promising results in fault detection and adaptive control tasks. Recent advancements include the integration of knowledge augmentation in Deep ELM for EEG seizure prediction, significantly improving accuracy and interpretability in biomedical signals [19]. Additionally, novel hardware implementations such as photonic ELM using microresonator arrays have been proposed to accelerate ELM processing for optical computing applications [20]. On the algorithmic front, ELM Ridge Regression Boosting has been introduced to enhance the robustness and predictive performance of ELM-based models in noisy or complex environments [21].

Building on this foundation, Online Sequential Extreme Learning Machine (OS-ELM) was introduced to enable sequential learning without requiring complete retraining, making it highly suitable for real-time forecasting applications [4]. Further, recurrent variants of ELM integrate feedback connections to enhance the modeling of long-term dependencies in streaming data [5]. These models offer significant computational advantages over traditional deep learning architectures, especially in scenarios that require continuous adaptation to dynamic environments. However, they struggle to capture complex multivariate interactions and long-range dependencies.

To overcome these limitations, researchers have explored Transformer-based architectures for time-series forecasting. [9] introduced efficient attention approximations, which build on the original Transformer concept and improve scalability for sequence modeling. Building on this, [22] and [23] proposed Informer and Autoformer, which optimize Transformers by reducing computational complexity and improving long-term forecast accuracy. Furthermore, [6] introduced iTransformer, which restructures the traditional Transformer by treating the entire time series as a single token, enhancing interpretability and computational efficiency. These studies underscore the increasing adoption of attention-based architectures for handling long-range dependencies in time-series forecasting.

Another emerging direction in time-series forecasting involves hybrid and adaptive models. [10] introduced Temporal Fusion Transformers (TFT), which integrate static and dynamic features to enhance interpretability in multivariate forecasting tasks. Similarly, SCINet [24] leverages sample convolution and interaction mechanisms to effectively capture temporal dependencies. Meanwhile, [25] developed N-BEATS, a deep neural network designed for interpretable time-series forecasting, demonstrating superior performance over conventional and RNN-based methods. Additionally, [26] introduced a meta-learning-based Transformer network for dynamic long-term forecasting, highlighting the potential of adaptive architectures in addressing time-series challenges.

Beyond architectural innovations, recent research has integrated probabilistic forecasting and meta-learning to improve forecasting reliability. [27] proposed DeepAR, an autoregressive recurrent network designed for probabilistic forecasting, allowing for more reliable uncertainty quantification. [28] developed a Transformer-based encoding approach for long-term forecasting, demonstrating improved generalization across multiple datasets. Additionally, [29] introduced DualNet, a continual learning framework that balances short-term adaptability with long-term memory retention, offering insights into lifelong learning in time-series models.

These advancements in deep learning, Transformer-based architectures, and extreme learning machines have significantly enhanced time-series forecasting accuracy and adaptability. However, challenges remain in balancing computational efficiency, interpretability, and scalability across different forecasting horizons. Future research must continue refining these methods to optimize performance while ensuring real-world applicability.

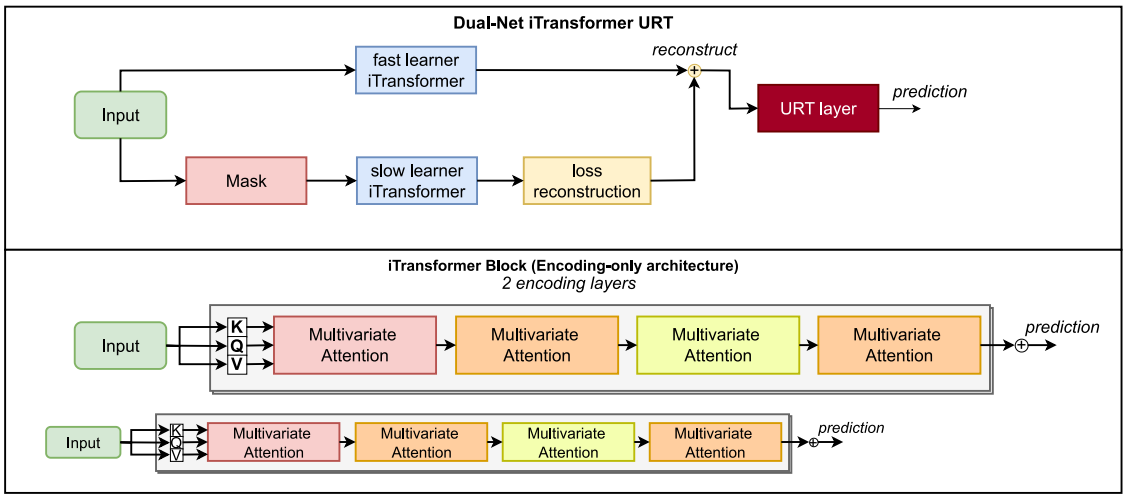

Fig. 1: Dual-Network iTransformer Universal Representation Transformer (URT) Architecture

3. Methodology

This research aims to develop Dual-Net iTransformer with Universal Representation Transformer (URT) [30], a time-series forecasting model that integrates two iTransformer architectures within a dual-network framework (Fig. 1). The proposed model comprises two primary components: Fast-learning iTransformer – Optimized for capturing short-term patterns in time-series data, Slow-learning iTransformer – Focused on long-term dependencies while effectively mitigating noise. To enhance forecasting accuracy and adaptability, these two components are combined using a URT, which generates more precise and robust final predictions by leveraging the complementary strengths of both networks. The notations and definitions of all variables used in the proposed Dual-Net iTransformer method are summarized in Table 1 for reference and clarity.

Each iTransformer model, both for the fast learners ($\frac{1}{M}\sum_{i=1}^{M} f_{\theta_i}(.)$) and slow learners ($g_{\{\theta\}}(.)$), operates through four main stages. The first stage is embedding (implemented by function f_{embed}), where each variable in the time series is transformed into tokens through an embedding layer, facilitating structured representation for subsequent processing as seen in (1).

$$X_t = \frac{1}{M}\sum_{i=1}^{M} f_{\theta_i}(X) \tag{1}$$

X_t represents the input data at time t. The second stage is the Multivariate Attention Layer, which captures relationships between variables using a self-attention mechanism as seen in (2).

$$Attention(Q, K, V) = softmax\left(Q\frac{K^T}{\sqrt{d_k}}\right)V \tag{2}$$

Q, K, V are the query, key, and value representations of the variables, and d_k is the key dimension. The third stage is the Feed-Forward Layer, which produces higher-level representations using a nonlinear activation function as seen in (3).

$$f_{ff}(x) = ReLU(W_1 x + b_1)W_2 + b_2 \tag{3}$$

W_1, W_2 are weights and b_1, b_2 are biases. The final stage is Layer Normalization, which normalizes feature scales and improves training stability as seen in (4).

$$\hat{x} = \frac{x - \mu}{\sigma} \tag{4}$$

μ and σ are the mean and standard deviation of the features.

Table 1: Variable Notations Used in Dual-Net iTransformer Method

Variable Notation	Explanation
X_t	Input time-series data at time t
$f_{\theta i}(.)$	Fast learner model function for the i-th learner
$g_\theta(.)$	Slow learner model function
f_{embed}	Embedding function that converts time-series values into token representations
$Attention(Q, K, V)$	Self-attention mechanism using query, key, and value representations
Q, K, V	Query, key, and value matrices in attention
d_k	Dimensionality of the key vectors
$f_{ff}(x)$	Feed-forward transformation with ReLU activation
W_1, W_2	Weight matrices in the feed-forward layer
b_1, b_2	Bias vectors in the feed-forward layer
μ, σ	Mean and standard deviation used for normalization
$\tilde{X}$	Masked input data
M	Binary mask tensor or number of fast learners (context-dependent)
L_m	Loss for masked data
L_{um}	Loss for unmasked data
L_{slow}	Total reconstruction loss for the slow learner
λ	Trade-off parameter between masked and unmasked loss
R	Aggregated representation from all fast learners
W_q, W_k	Weight matrices for query and key transformation
b_q, b_k	Bias vectors for query and key transformation
A	Attention weights computed via scaled dot-product attention
Y	Final prediction computed from attention-weighted fast learner outputs
MSE	Mean Squared Error, used as performance metric
MAE	Mean Absolute Error, used as performance metric
N	Total number of data points or prediction instances
$y_i, \hat{y}_i$	True and predicted values for the i-th instance

Dual-Net iTransformer URT differs from traditional Dual-Net models used in continual learning. In this research, the fast learner captures local patterns from different segments of the time series, while the slow learner captures global information and mitigates overfitting. To enhance the robustness of the model against noise, a Controlled Reconstruction Strategy is implemented, consisting of several steps. First, part of the data is randomly masked before being processed by the slow learner as seen in (5).

$$\tilde{X} = X \odot M \tag{5}$$

M is a binary mask tensor containing values of 0 or 1. Second, the loss is computed as a combination of masked and unmasked data as seen in (6) and (7).

$$L_m = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{X}_{m,i} - X_{m,i} \right)^2 \tag{6}$$

$$L_{um} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{X}_{um}, i - X_{um}, i \right)^2 \tag{7}$$

L_m represents the loss for masked data and L_{um} for unmasked data. The total loss for the slow learner is computed as (8).

$$L_{slow} = \lambda L_m + (1 - \lambda) L_{um} \tag{8}$$

λ is a trade-off parameter controlling the balance between reconstructing original and noisy data. λ has the range of $0 \leq \lambda \leq 1$ as per [26] as the reconstruction of the masked input is dependent on the reconstruction of the unmasked input.

After feature representations are generated, the URT is used to dynamically select the best learner. First, representations from all learners are aggregated (R) given in (9).

$$R = \frac{1}{M} \sum_{i=1}^{M} f_{\theta_i}(X) \tag{9}$$

M represents the number of learners. Next, a query-key attention mechanism determines the best learner based on data distribution given in (10) and (11).

$$Q = W_q R + b_q, \quad K = W_k R + b_k \tag{10}$$

$$A = softmax(QK^T) \tag{11}$$

Q and K represent the linear projections of the aggregated representation R into the query and key spaces, respectively. Here, W_q and W_k are learnable weight matrices, while b_q and b_k are corresponding bias vectors. A is the attention weights which performs scaled dot-product attention given in (12).

$$Y = \sum_{i=1}^{M} A_i f(\theta_i)(X) \tag{12}$$

To evaluate model performance, two primary metrics are used: Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE measures the average squared difference between predicted and actual values using (13).

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{13}$$

N refers to the total number of data points (or prediction instances). MSE is sensitive to large errors, making it useful for detecting outliers. Meanwhile, MAE calculates the absolute average error without considering the direction of the error as given in (14).

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{14}$$

The Dual-Net iTransformer method aims to enhance the accuracy of time-series forecasting by leveraging the complementary strengths of fast-learning and slow-learning mechanisms, while integrating the URT for optimal representation selection. By incorporating a Controlled Reconstruction Strategy, this model mitigates overfitting and enhances robustness against noise. The evaluation results will assess its effectiveness compared to existing state-of-the-art methods.

Algorithm 1 outlines the training procedure for Dual-Net iTransformer with URT, which consists of two main phases. The first phase trains the fast-learning and slow-learning iTransformer models, where the fast learner captures short-term patterns while the slow learner mitigates noise through reconstruction. The second phase involves training the URT, which aggregates predictions from multiple fast learners and assigns optimal weights to enhance forecast accuracy.

Algorithm 1: Dual-Net iTransformer with URT

Input: Time-series data, fast learner models $\{f_{\theta_i}(.)\}_{i=1}^{M}$, slow learner model $g_{\{\theta\}}(.)$, number of epochs E, number of batches B.
Output: Trained fast learner models $\{f_{\theta_i}(.)\}_{i=1}^{M}$
// Train fast learner and slow learner
for $e = 1$ to E **do**
  for $b = 1$ to B **do**
    Predict using fast learner models $\{f_{\theta_i}(.)\}_{i=1}^{M}$
    Compute fast learner loss L_{MSE}
    Update parameters of fast learner models $\{f_{\theta_i}(.)\}_{i=1}^{M}$ based on L_{MSE}
    Generate slow learner input mask m_t
    Predict using slow learner model to obtain $\hat{x}_t$
    Compute slow learner loss L_S
    Update slow learner model parameters $g_{\{\theta\}}(.)$ based on L_S
    Update fast learner models $\{f_{\theta_i}(.)\}_{i=1}^{M}$ based on $L_S + L_{MSE}$
  end
end
Save fast learner model parameters $\{f_{\theta_i}(.)\}_{i=1}^{M}$
// Train URT model
for $e = 1$ to E **do**
  Freeze fast learner models $\{f_{\theta_i}(.)\}_{i=1}^{M}$ trained with reconstruction loss from slow learner
  for $b = 1$ to B **do**
    for $m = 1$ to M **do**
      Obtain predictions from fast learner model m
    end
    Combine and average predictions of all fast learner models over batch range
    Train URT model parameters
    Compute weights for each variable in each learner
    Perform tensor multiplication for each variable in each learner $\{f_{\theta_i}(.)\}_{i=1}^{M}$
    Compute URT loss $L_{\{u_{MSE}\}}$
    Update URT parameters $\varphi_{\{w\theta\}}(.)$
  end
end

4. Experiments

4.1. Dataset

The datasets used to evaluate the models consist of four real-world datasets, which are commonly employed in long-term time series forecasting (LSTF) research. These datasets are as follows:

1. Electricity Transformer Temperature (ETT) (Zhou et al., 2021): This dataset includes electricity consumption data from two counties in China. It is widely used as a benchmark for evaluating LSTF models. The dataset is divided into four subsets: ETTh1 and ETTh2, which contain hourly measurements, and ETTm1 and ETTm2, which contain 15-minute measurements.

2. Weather (Wu, Xu, Wang, & Long, 2022): This dataset consists of road usage data collected by multiple sensors on freeways in the San Francisco Bay Area. The data is recorded hourly by the California Department of Transportation.
3. Influenza-like Illness (ILI) (Wu, Xu, Wang, & Long, 2022): The ILI dataset contains weekly records of influenza-like illness (ILI) cases reported by the Centers for Disease Control and Prevention (CDC) in the United States from 2002 to 2021. It specifically represents the ratio of patients diagnosed with ILI to the total number of patients.
4. Exchange (Wu, Xu, Wang, & Long, 2022): This dataset includes daily exchange rates from eight different countries, covering the period from 1990 to 2016.

For model training, the datasets are divided into separate subsets using predefined proportions. The **Weather, ILI, and Exchange** datasets are split using a **7:1:2** ratio, meaning **70%** of the data is used for training, **10%** for validation, and **20%** for testing. Meanwhile, the **ETT** dataset follows a **6:2:2** ratio, with **60%** allocated for training, **20%** for validation, and **20%** for testing. These partitioning strategies align with the standards commonly used in long-term time series forecasting models.

4.2. Results

The experimental results evaluating the performance of Dual-Net iTransformer are presented in Table 2. Mean Squared Error (MSE) and Mean Absolute Error (MAE) are used to assess and compare the proposed model against existing models, including OS-ELM, MANTRA, iTransformer, Reformer, and Informer. According to Table 2, Dual-Net iTransformer consistently outperforms all baseline models in long-term forecasting tasks. Specifically, as shown in Table 3, the model achieves lower MSE and MAE across multiple datasets, with an average improvement of 17.01% over MANTRA, and 6.48% over iTransformer.

In contrast to the other models, OS-ELM demonstrates strong performance on the ILI and Weather datasets, outperforming all other models on these datasets. Compared to Dual-Net iTransformer URT, OS-ELM achieves an average performance improvement of 19.64% on the Weather dataset and 152.54% on the ILI dataset. However, for the ETT and Exchange datasets, Dual-Net iTransformer URT remains superior, with average performance improvements of 126.74% over OS-ELM on the ETT dataset and 20.68% over OS-ELM on the Exchange dataset. The OR-ELM model also demonstrates promising results on the ILI dataset. While OS-ELM performs slightly better, OR-ELM still surpasses most of the other baseline models.

Despite OSELM's impressive performance on specific datasets, we encountered some technical limitations during experimentation. When using OSELM with the ETT dataset, we observed extremely high and volatile prediction values, occasionally resulting in overflow issues that produced NaN values. Similarly, ORELM experienced convergence problems with the ETT dataset, generating "LinAlgError: SVD did not converge" errors during training, particularly in the pseudo-inverse matrix calculations used in the algorithm.

Fig. 2 and Fig. 3 provide illustrative examples of the effectiveness of the proposed Dual-Net iTransformer in forecasting real-world time series data. In Fig. 2, the model accurately captures the complex temporal dynamics of the Weather dataset across 200 time steps, including periodic fluctuations and localized irregularities. This demonstrates its capability to adapt to diverse patterns within a relatively short prediction window. When the forecast horizon is extended to 300 data points in Fig. 3, the model continues to exhibit strong alignment with the ground truth, maintaining stable predictions even as the complexity and uncertainty of the sequence increase. These results highlight the model's robustness, generalization ability, and effectiveness in long-term forecasting tasks involving non-stationary and fluctuating data.

5. Discussions

The performance variations observed across different datasets can be attributed to their inherent characteristics. The Weather dataset, which exhibits strong seasonal patterns, presents challenges for the URT layer in the Dual-Net iTransformer architecture. This limitation is evident in the model's relatively weaker performance compared to OS-

Table 2: Performance of Dual-Net iTransformer on Multivariate Data Prediction

Dataset	Forecasting horizon	Dual-Net+URT		iTransformer		OSELM		ORELM		MANTRA		Autoformer		Informer		Reformer	
		MSE	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
ETT	96	**0.121**	**0.235**	0.297	0.349	1.012	1.986	0.847	0.999	0.212	0.295	0.255	0.326	0.650	0.453	0.658	0.619
	192	**0.154**	**0.268**	0.380	0.400	-	-	-	-	0.275	0.335	0.299	0.350	0.533	0.563	1.078	0.827
	336	**0.195**	**0.299**	0.428	0.432	-	-	-	-	0.327	0.365	0.337	0.372	1.363	0.887	1.549	0.972
	720	**0.245**	**0.336**	0.427	0.445	-	-	-	-	0.440	0.435	0.442	0.432	3.379	1.388	2.631	1.242
Weather	96	0.173	0.212	0.174	0.214	**0.064**	**0.074**	0.337	0.841	0.248	0.321	0.269	0.338	0.300	0.384	0.689	0.596
	192	0.223	**0.227**	**0.221**	0.254	-	-	-	-	0.281	0.338	0.297	0.354	0.598	0.544	0.752	0.638
	336	0.280	0.297	0.278	**0.296**	-	-	-	-	**0.239**	0.369	0.358	0.392	0.578	0.523	0.639	0.596
	720	0.360	0.351	**0.358**	**0.349**	-	-	-	-	0.405	0.414	0.450	0.452	1.059	0.741	1.130	0.792
Exchange	96	**0.085**	**0.205**	0.090	0.211	0.702	0.664	0.843	1.001	0.155	0.285	0.153	0.285	0.847	0.752	1.065	0.829
	192	**0.177**	**0.299**	0.194	0.315	-	-	-	-	0.266	0.377	0.295	0.395	1.204	0.895	1.188	0.906
	336	**0.320**	**0.409**	0.333	0.418	-	-	-	-	0.421	0.480	0.446	0.496	1.672	1.036	1.357	0.976
	720	**0.884**	**0.711**	0.914	0.726	-	-	-	-	1.168	0.847	1.503	0.919	2.478	1.310	1.510	1.106
ILI	24	2.424	0.979	2.415	0.984	**0.021**	**0.108**	0.249	0.121	3.238	1.224	3.680	1.346	5.764	1.677	4.400	1.382
	36	**2.078**	**0.938**	2.354	0.999	-	-	-	-	2.396	1.176	3.629	1.260	4.755	1.467	4.783	1.448
	48	**2.207**	**0.951**	2.268	0.976	-	-	-	-	2.941	1.144	3.376	1.258	4.763	1.469	4.832	1.465
	60	**2.183**	**0.959**	2.360	1.000	-	-	-	-	2.705	1.106	2.917	1.590	5.278	1.560	4.882	1.483

Table 3: Dual-Net iTransformer Performance Improvement over Other Models

Dataset	Forecasting horizon	iTransformer		OSELM		ORELM		MANTRA		Autoformer		Informer		Reformer	
		MSE	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
ETT	96	0.18	0.11	0.89	1.75	0.73	0.76	0.09	0.06	0.13	0.09	0.53	0.22	0.54	0.38
	192	0.23	0.13	-	-	-	-	0.12	0.07	0.14	0.08	0.38	0.29	0.92	0.56
	336	0.23	0.13	-	-	-	-	0.13	0.07	0.14	0.07	1.17	0.59	1.35	0.67
	720	0.18	0.11	-	-	-	-	0.20	0.10	0.20	0.10	3.13	1.05	2.39	0.91
Weather	96	0.001	0.002	**−0.11**	**−0.14**	0.16	0.63	0.08	0.11	0.10	0.13	0.13	0.17	0.52	0.38
	192	**−0.002**	0.03	-	-	-	-	0.06	0.11	0.07	0.13	0.38	0.32	0.53	0.41
	336	**−0.002**	**−0.001**	-	-	-	-	**−0.04**	0.07	0.08	0.10	0.30	0.23	0.36	0.30
	720	**−0.002**	**−0.002**	-	-	-	-	0.05	0.06	0.09	0.10	0.70	0.39	0.77	0.44
Exchange	96	0.005	0.01	0.62	0.46	0.76	0.80	0.07	0.08	0.07	0.08	0.76	0.55	0.98	0.62
	192	0.02	0.02	-	-	-	-	0.09	0.08	0.12	0.10	1.03	0.60	1.01	0.61
	336	0.01	0.01	-	-	-	-	0.10	0.07	0.13	0.09	1.35	0.63	1.04	0.57
	720	0.03	0.02	-	-	-	-	0.28	0.14	0.62	0.21	1.59	0.60	0.63	0.40
ILI	24	**−0.01**	0.01	**−2.40**	**−0.87**	**−2.17**	**−0.86**	0.81	0.24	1.26	0.37	3.34	0.70	1.98	0.40
	36	0.28	0.06	-	-	-	-	0.32	0.24	1.55	0.32	2.68	0.53	2.71	0.51
	48	0.06	0.03	-	-	-	-	0.73	0.19	1.17	0.31	2.56	0.52	2.63	0.51
	60	0.18	0.04	-	-	-	-	0.52	0.15	0.73	0.63	3.09	0.60	2.70	0.52

ELM on this dataset. The recurring nature of weather data at specific time intervals appears to be better captured by OS-ELM's approach, highlighting the influence of seasonality in forecasting performance.

The fast learner is designed to capture short-term, high-frequency patterns, while the slow learner is responsible for modeling long-term dependencies and mitigating noise. The URT layer dynamically integrates their outputs, ideally selecting the most relevant representation based on input characteristics. However, in highly seasonal data such as Weather, the URT may not properly measure the priority of the two learners. Visual analysis in Fig. 2 supports this finding, where the Dual-Net iTransformer demonstrates moderate accuracy in tracking weather patterns but

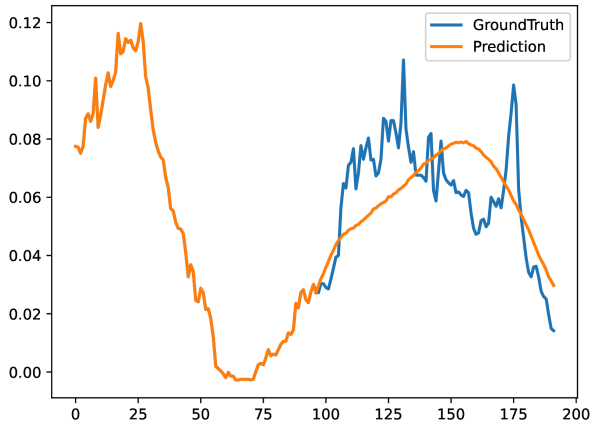

Fig. 2: Prediction from Weather Dataset for 200 data.

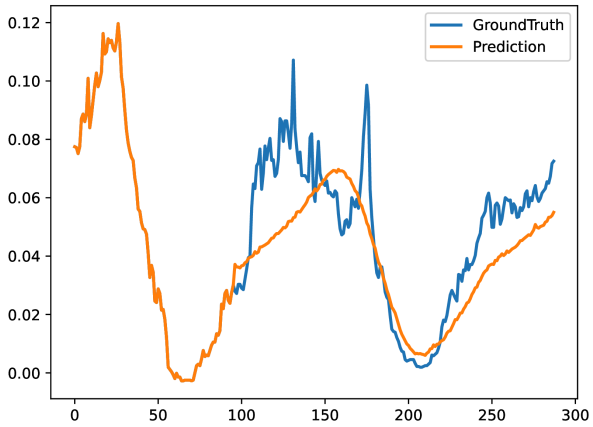

Fig. 3: Prediction from Weather Dataset for 300 data.

exhibits slight lag during sharp seasonal transitions. This suggests the need for future refinement of the URT layer when applied to datasets with strong periodic behavior.

Our experiments revealed that Dual-Net iTransformer URT's performance improves with increasing prediction sequence length, particularly on the ETT dataset with lengths of 96, 192, 336, and 720. This trend suggests that ensemble learning becomes more effective as the forecasting horizon extends in larger datasets. However, this pattern was not observed in datasets with shorter prediction sequences, such as ILI (24, 36, 48, and 60) and Weather, indicating that sequence length influences model performance differently depending on dataset characteristics. Fig. 3 illustrates the model's ability to remain stable over extended forecast windows. In this long-range scenario, the predictions closely follow the ground truth trend with minimal drift, underscoring the synergy between the fast and slow learners in mitigating cumulative error across time steps.

When compared to other Transformer-based architectures, such as Informer, LogTrans, and Reformer, our model demonstrates significant superiority in both MSE and MAE metrics. The average performance improvements are: 97.11% over Informer, 95.14% over LogTrans, and 91.36% over Reformer. These improvements stem from fundamental differences in how each architecture processes time-series data:

- **Informer** employs *sparse self-attention* to enhance computational efficiency but struggles with complex multivariate relationships. While its mechanism improves efficiency, it fails to fully capture intricate temporal dependencies in multivariate time series data.
- **LogTrans** uses *log-sparse attention* to efficiently model long-range dependencies but may miss critical information in highly variable time series due to its selective attention mechanism.
- **Reformer**, with its *locality-sensitive hashing and reversible layers*, prioritizes memory efficiency over prediction accuracy, leading to lower performance in complex forecasting tasks.

The enhanced performance of our Dual-Net iTransformer URT can be attributed to the Inverted Transformer's innovative embedding approach, which groups variables into unified embedding spaces. This method enables a more effective capture of temporal relationships between variables compared to conventional approaches. Additionally, our model's simplified architecture—utilizing only one fast learner and one slow learner instead of multiple learners, as employed by models like MANTRA—demonstrates that strategic architectural simplification can enhance performance while reducing computational requirements.

6. Conclusions and Future Works

We developed the Dual-Net iTransformer for long term time series forecasting which uses one fast learner and one slow learner with the URT layer as its ensemble layer. Our model demonstrated superior performance compared to OS-ELM and OR-ELM models across most datasets. When compared to the individual iTransformer model, Dual-Net iTransformer URT showed advantages on ETT, Exchange, and ILI datasets. However, this model was less

effective for the Weather dataset which has seasonal trends, where the individual iTransformer model performed better. This indicates that the ensemble approach in Dual-Net iTransformer is more effective for multivariate datasets with complex patterns but has limitations with seasonal datasets. As a future work, an aggregation techniques can help optimize our model's ability to capture complex patterns in time series data.

CRediT Authorship Contribution Statement

Ary M. Shiddiqi: Conceptualization, Methodology, Supervision, Writing – Original Draft. **Bagaskoro K. Ardi:** Data Curation, Methodology. **Bilqis Amaliah:** Methodology, Supervision. **I K. A. Mogi:** Data Curation, Methodology. **Agung M. Rizki:** Writing – Review & Editing, Funding Acquisition. **Bintang Nuralamsyah:** Validation, Writing – Review & Editing. **Ilham G. Adillion:** Project Administration, Funding Acquisition. **Moch. N. Alzamzami:** Project Administration, Writing – Review & Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data sharing is not applicable as the data are secondary data drawn from already published literature.

Declaration of Generative AI and AI-assisted Technologies in The Writing Process

The authors used generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and take full responsibility for the final publication.

References

[1] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. OTexts, 2018. [Online]. Available: https://otexts.com/fpp2/

[2] E. Spiliotis, "Time Series Forecasting with Statistical, Machine Learning, and Deep Learning Methods: Past, Present, and Future," *Advances in Forecasting with Artificial Intelligence*. Springer, pp. 45–67, 2023. doi: 10.1007/978-3-031-35879-1_3.

[3] V. Assimakopoulos, A.-A. Semenoglou, G. Mulder, and K. Nikolopoulos, "Statistical, Machine Learning, and Deep Learning Forecasting Methods: Comparisons and Ways Forward," *The Journal of the Operational Research Society*, vol. 74, no. 3, pp. 840–859, 2023, doi: 10.1080/01605682.2022.2118629.

[4] J. Wang, S. Lu, S.-H. Wang, and Y.-D. Zhang, "A review on extreme learning machine," *Multimedia Tools and Applications*, vol. 81, no. 29, pp. 41611–41660, May 2021, doi: 10.1007/s11042-021-11007-7.

[5] B. Deng, X. Zhang, W. Gong, and D. Shang, "An Overview of Extreme Learning Machine," in *2019 4th International Conference on Control, Robotics and Cybernetics (CRC)*, 2019, pp. 189–195. doi: 10.1109/crc.2019.00046.

[6] Y. Liu, T. Hu, H. Zhang, and others, "iTransformer: Inverted Transformers Are Effective for Time Series Forecasting," *arXiv preprint arXiv:2310.06625*, 2023, [Online]. Available: https://arxiv.org/abs/2310.06625

[7] L. Zha, K. Ma, G. Li, and Q. Fang, "An improved extreme learning machine with self-recurrent hidden layer for online process regression prediction," *Advanced Engineering Informatics*, vol. 54, p. 101736, 2022, doi: 10.1016/j.aei.2022.101736.

[8] Y. Liu, S. Wijewickrema, A. Li, C. Bester, S. O'Leary, and J. Bailey, "Time-Transformer: Integrating Local and Global Features for Better Time Series Generation," *arXiv preprint arXiv:2312.11714*, 2023, [Online]. Available: https://arxiv.org/abs/2312.11714

[9] K. Choromanski, V. Likhosherstov, D. Dohan, and others, "Rethinking Attention with Performers," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://arxiv.org/abs/2009.14794

[10] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2020.06.001.

[11] Y. Liu, H. Wu, J. Wang, and M. Long, "Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting," *Advances in Neural Information Processing Systems*, 2022, [Online]. Available: https://arxiv.org/abs/2205.14415

[12] J. Grigsby, Z. Wang, N. H. Nguyen, and Y. Qi, "Spacetimeformer: Attention-based Time-series Forecasting with Spatiotemporal Dependency," *arXiv preprint arXiv:2109.12218*, 2021, [Online]. Available: https://arxiv.org/abs/2109.12218

[13] A. Casolaro, V. Capone, G. Iannuzzo, and F. Camastra, "Deep Learning for Time Series Forecasting: Advances and Open Problems," *Information*, vol. 14, no. 11, p. 598, 2023, doi: 10.3390/info14110598.

[14] E. Eldele, M. Ragab, Z. Chen, M. Wu, and X. Li, "TSLANet: Rethinking Transformers for Time Series Representation Learning," *Advances in Neural Information Processing Systems*, vol. 235, pp. 12409–12428, 2024, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85203840463

[15] H. Xu, X. Zhang, D. Liang, and G. Wang, "Robust-STP: A Robust Seasonal-trend Decomposition Method for Partial Periodic Time Series," in *2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, 2022, pp. 691–695. doi: 10.1109/CCIS57298.2022.10016327.

[16] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017, doi: 10.1109/TNNLS.2016.2582924.

[17] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[18] A. Cini, I. Marisca, D. Zambon, and C. Alippi, "Graph Deep Learning for Time Series Forecasting," *ACM Comput. Surv.*, vol. 57, no. 12, Jul. 2025, doi: 10.1145/3742784.

[19] Y. Zhang, J. Wang, and H. Liu, "Deep Extreme Learning Machine with Knowledge Augmentation for EEG Seizure Prediction," *Frontiers in Neuroinformatics*, vol. 17, p. 1205529, 2023, doi: 10.3389/fninf.2023.1205529.

[20] S. Biasi, R. Franchi, L. Cerini, and L. Pavesi, "An array of microresonators as a photonic extreme learning machine," *APL Photonics*, vol. 8, no. 9, p. 96105, 2023, doi: 10.1063/5.0156189.

[21] C. Peralez-González, J. Pérez-Rodr\iguez, and A. M. Durán-Rosal, "Boosting ridge for the extreme learning machine globally optimised for classification and regression problems," *Scientific Reports*, vol. 13, no. 1, p. 11809, 2023.

[22] S. Li *et al.*, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," *Advances in Neural Information Processing Systems*, vol. 32, 2019, doi: https://doi.org/10.48550/arXiv.1907.00235.

[23] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting," in *Neural Information Processing Systems*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2106.13008

[24] M. Liu *et al.*, "SCINet: Time Series Modeling and Forecasting with Sample Convolution and Interaction," *Advances in Neural Information Processing Systems*, vol. 35, 2022, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/266983d0949aed78a16fa4782237dea7-Paper-Conference.pdf

[25] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: NEURAL BASIS EXPANSION ANALYSIS FOR INTERPRETABLE TIME SERIES FORECASTING," *8th International Conference on Learning Representations, ICLR 2020*, 2020, doi: 10.48550/arXiv.1905.10437.

[26] M. A. Masum *et al.*, "Dynamic Long-Term Time-Series Forecasting via Meta Transformer Networks," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 8, pp. 4258–4268, 2024, doi: https://doi.org/10.1109/TAI.2024.3365775.

[27] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020, doi: 10.1016/j.ijforecast.2019.07.001.

[28] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers." [Online]. Available: https://arxiv.org/abs/2211.14730

[29] Q. Pham, C. Liu, and S. Hoi, "DualNet: Continual Learning, Fast and Slow," in *Advances in Neural Information Processing Systems*, 2021. doi: https://doi.org/10.48550/arXiv.2110.00175.

[30] L. Liu, W. L. Hamilton, G. Long, J. Jiang, and H. Larochelle, "A Universal Representation Transformer Layer for Few-Shot Image Classification," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://arxiv.org/abs/2006.11702