IRSTI 28.23.37 https://doi.org/10.26577/jpcsit4220266

Tin Trung Chau , Miras Shaltayev , Kulash Talapiden ,
Muhammad Auwal Shehu , Ahmad Bala Alhassan*

Nazarbayev University, Astana, Kazakhstan
*e-mail: ahmad.alhassan@nu.edu.kz

ENHANCING WIND SPEED FORECASTING WITH LARGE LANGUAGE MODELS: A CASE STUDY OF KAZAKHSTAN

Abstract. Accurate wind speed forecasting plays an important role in the planning, operation, and control optimization of modern wind energy systems. Kazakhstan, the country with the largest wind energy potential in Central Asia, is facing significant challenges due to the highly variable, nonlinear, and non-stationary nature of wind. This study proposes a wind-speed forecasting framework based on frozen pre-trained Transformer backbones (BERT/BART) with lightweight projection-based adaptation for time-series inputs. The study employs frozen pre-trained Transformer backbones with self-attention and lightweight projection-based adaptation of Bidirectional Encoder Representations from Transformers (BERT) and Bidirectional and Auto-Regressive Transformers (BART) as the core of the model. In this retrospective analysis, hourly wind speed data from five representative cities across Kazakhstan are used to evaluate simulated operational forecast performance at seven time steps: 1, 3, 6, 36, 72, 144, and 432 hours. The LLMs were compared with AutoRegressive Integrated Moving Average (ARIMA), Segmented Recurrent Neural Network (SegRNN), and Patch Time Series Transformer (PatchTST) models, and the results showed superior accuracy and higher stability in long-term forecasting. These results support the potential of pre-trained Transformer backbones as effective sequence models for wind-speed forecasting under diverse climatic conditions.

Keywords: Wind energy forecasting, Large language models, Bidirectional Encoder Representations from Transformers, Bidirectional and Auto-Regressive Transformers, Time series forecasting.

1. Introduction

Kazakhstan is one of the countries with the largest wind energy potential in Central Asia, with a vast territory characterized by steppe and desert terrain. Approximately 50% of Kazakhstan's territory is suitable for wind energy development, including Aktobe, Aktau, Mangystau, and Astana, with average wind speeds of 7.5 m/s at altitudes of 50–100 meters [1]. According to international reports, Kazakhstan's total technical wind energy potential is estimated to reach 920 terawatt-hours (TWh) per year, significantly exceeding the country's current electricity demand. This indicates that wind power can be a key pillar of the energy transition, enabling carbon neutrality by 2060 [2], [3].

However, the exploitation of wind energy depends primarily on wind speed. In theory, wind power scales with the cube of wind speed, which means that even a small change in wind speed can have a significant impact on energy production [4], [5]. Meanwhile, wind speed in Kazakhstan is nonlinear, strongly seasonally fluctuates, and is influenced by topography, particularly in the coastal areas of the Caspian Sea and the eastern mountainous region [6]. This variability makes it challenging to plan, operate, and dispatch wind power systems, reducing the reliability of power generation and the efficiency of integration into the national grid.

In this context, the development of advanced wind speed forecasting models with high accuracy is a necessary step to improve forecasting accuracy and reliability. These forecasting models play a crucial role in the entire value chain of wind power systems, from planning and operation to control optimization [7], [8], [9]. Furthermore, accurate wind speed forecasting also helps to improve the stability and reliability of the power system, especially as the share of renewable energy in the grid increases. Modern forecasting models enable the optimization of turbine operations, reduce dispatch costs, and facilitate the integration of wind power with other energy sources, such as solar power or storage systems [10]. As a result, Kazakhstan can enhance the efficiency of wind resource exploitation, ensure national energy security, and promote a sustainable transition to a low-carbon economy.

Given the importance of forecasting, this paper proposes an advanced wind speed forecasting framework that uses frozen pre-trained Transformer backbones to model temporal dependencies in wind-speed sequences. Unlike traditional statistical methods or models, pre-trained Transformer architectures provide strong sequence modeling capacity to model complex, nonlinear relationships and long-term temporal dependencies, which are crucial for capturing the stochastic, highly variable nature of wind speed [11].

The model is applied directly to wind speed data from several cities in Kazakhstan to evaluate its performance under different climatic and topographic conditions. The results are expected to assess whether frozen pre-trained Transformer backbones can effectively capture temporal variation patterns and improve both short- and long-term wind-speed forecasting accuracy. Furthermore, this approach contributes to improving the quality of wind energy forecasting in Kazakhstan. It highlights the potential application of pre-trained Transformer backbones in renewable energy forecasting and management.

Given the importance of accurate wind speed forecasting, numerous studies have been conducted to enhance the reliability and quality of forecasts. Traditional forecasting methods, such as ARIMA [12], Seasonal-ARIMA [13], and ARIMA-ANN-Kalman [14], are relatively easy to implement and straightforward models. These models are primarily based on strict linear assumptions, and the data must be stable and consistent [15]. Therefore, these models often struggle to model the strong nonlinear and random features inherent in wind data, especially in areas with complex topography. Additionally, traditional deep learning models, such as CNNs [16], RNNs [17], LSTMs [18], and GRUs [19], have demonstrated promising performance in time series forecasting. However, these deep learning models still have limitations in maintaining long-term dependencies and are also susceptible to vanishing gradients when processing long data series.

The advent of the Transformer architecture marked a significant step forward, thanks to its self-attention mechanism, which enables the modeling of global relationships in time series. Typical variants such as Informer [20], Autoformer [21], and FEDformer [22] have demonstrated superior performance in handling non-stationarity and long-term dependencies in wind data.

Based on this foundation, pre-trained Transformer backbones such as BERT [23], BART [24], GPT [25], and LLaMA [26] have been extended from natural language processing to time series forecasting, opening up new approaches for reusing pre-trained sequence models in non-text domains. These models are capable of capturing long-term dependencies and modeling complex nonlinear relationships in sequential data. Two main approaches explored in recent research include (1) Intramodal Transfer Learning (ITL), which is capable of directly fine-tuning pre-trained Transformer models on sequence data to learn deep temporal features, and (2) Cross-modal Knowledge Transfer (CKT), which allows converting sequence data into text format to leverage pre-trained model priors [27].

These qualities of LLMs are beneficial for wind speed forecasting. Wind data are non-stationary, subject to seasonal cycles, varied terrain, and random noise, making modeling complex. Traditional statistical and deep learning models often struggle to handle nonlinear variations and long-term dependencies in time series data.

In contrast, pre-trained Transformer backbones with global self-attention and contextual sequence representations can flexibly adapt to data fluctuations, learning both short-term and long-term trends in wind behavior. Thanks to parameter reuse and lightweight adaptation, these models can generalize across multiple regions or turbines, reducing training costs and increasing adaptability to local conditions.

Therefore, integrating frozen pre-trained Transformer backbones into wind speed forecasting represents a significant step forward, moving from purely data-driven regression models to pre-trained backbone-based sequence forecasting systems, thereby enhancing accuracy, reliability, and support for intelligent energy management in modern wind power systems.

Objectives and contributions

Based on the identified research gaps and challenges, this study proposes developing an pre-trained Transformer backbone-based wind-speed forecasting framework for Kazakhstan meteorological data. The main contributions of the paper can be summarized as follows:

- Proposing a wind speed forecasting framework based on frozen pre-trained Transformer backbones with a self-attention mechanism and fixed pre-trained sequence modeling capability to model nonlinearity, non-stationarity, and long-term dependence in wind data.
- Applying and validating the model on real-world wind datasets in many cities of Kazakhstan to reflect diverse climatic and topographic conditions, thereby evaluating the model's generalization ability and stability.
- Comparing the performance of the proposed frozen pre-trained Transformer backbones with current advanced models to evaluate the empirical accuracy and reliability of the proposed method.
- Providing a reference framework for time series and renewable energy forecasting applications, contributing to grid operation optimization, and with potential relevance to Kazakhstan's carbon neutrality strategy.

2. Materials and methods

2.1 Dataset and preprocessing

The study uses hourly wind speed data collected from NASA's Prediction of Worldwide Energy Resources (POWER) project. Wind speed data were collected from January 2010 to July 2025 at hourly intervals over cities including Aktobe, Aktau, Kokshetau, Taldykorgan, and Uralsk. These regions were selected to reflect the diversity of climate and topography, from coastal to inland and mountainous areas. Each dataset consists of time-labeled wind speed time series. Data were collected at a height of 80 m, corresponding to the typical hub height of wind turbines. Figure 1 presents the location of five cities.

The data were split into training, test, and validation sets with proportions of 70%, 20%, and 10%, respectively. Subsequently, the data were cleaned and interpolated to handle missing or outlier values using cubic spline interpolation, and short-term noise was mitigated using Savitzky–Golay filtering. Crucially, the parameters for these filtering operations were fitted exclusively on the training set and then applied to the validation and test sets. Next, the data were standardized using Z-score normalization, with the mean and standard deviation computed from the training data only. Finally, the normalized time series was divided into sliding windows. This chronological processing converts NASA POWER data from raw meteorological data into a structured time series suitable for Transformer and LLM models, ensuring consistency, stability, and a rigorous out-of-sample evaluation.

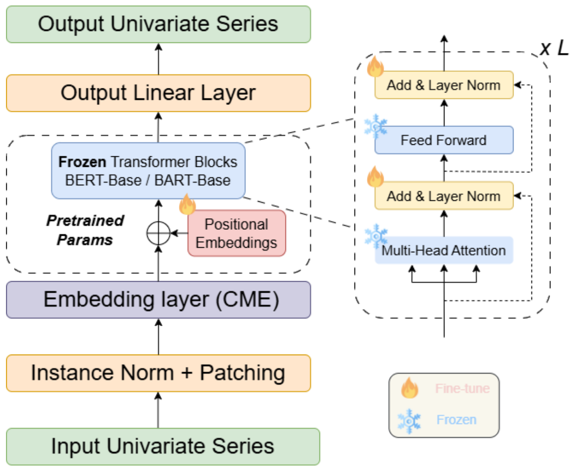
Figure 1
Geographical locations of the five study sites in Kazakhstan

2.2 Model structure

The diagram in Figure 2 illustrates the model structure, including global Z-score standardization of the input sequence, patching, linear projection, and positional encoding. To improve training stability and ensure consistent scaling across datasets, the input sequences are first standardized using the global Z-score normalization described in Section 2.1, where the mean and standard deviation are estimated from the training set only. Then, it is passed through the pre-trained Transformer blocks (frozen) and combined with fine-tuning layers (Add & Layer Norm). The output is passed through a linear layer to create the predicted sequence. This structure leverages the long-range sequence modeling capacity of the frozen pre-trained Transformer blocks and reduces training costs by fixing most of the original model's parameters.

Figure 2

Architecture of LLM Reprogramming for Univariate Time Series Forecasting



Problem Definition. The problem of forecast-ting wind speed time series is formulated as follows:

Given a historical observation data series of length L, denoted by

$$\mathbf{V} = (v_1, v_2, \dots, v_L), \quad (1)$$

Where $v_t \in \mathbb{R}^M$ – a multivariate vector consisting of M features recorded at time t. In this study, only univariate wind speed is forecasted, so $M = 1$.

The goal of the model is to learn the time-dependent relationship in the input data series $\mathbf{V}$ to forecast a future value series of length T, denoted by

$$\hat{\mathbf{Y}} = (v_{L+1}, v_{L+2}, \dots, v_{L+T}). \quad (2)$$

In other words, the problem is to construct a prediction function $f(\cdot)$ such that

$$\hat{\mathbf{Y}} = f(\mathbf{V}),$$

to minimize the error between the forecast value $\hat{\mathbf{Y}}$ and the actual value $\mathbf{Y}$. In the context of this study, v_t represents the wind speed at time t, and the model must accurately predict the wind speed in the following T hours based on the observation data from the previous L hours.

Data Normalization. The observed wind speed series $\mathbf{V} = (v_1, v_2, \dots, v_L)$ is normalized to reduce the influence of amplitude fluctuations and stabilize the training process. The mean and standard deviation are calculated as follows:

$$\mu = \frac{1}{L}\sum_{t=1}^{L} v_t \,, \sigma = \sqrt{\frac{1}{L}\sum_{t=1}^{L}(v_t - \mu)^2 + \varepsilon}, \quad (3)$$

where

(ε) – a very small constant to avoid division by 0.

The series is normalized according to the formula:

$$\tilde{v}_t = \frac{v_t - \mu}{\sigma}. \quad (4)$$

The resulting series $\widetilde{\mathbf{V}} = (\tilde{v}_1, \tilde{v}_2, \dots, \tilde{v}_L)$has a mean of 0 and a variance of 1.

It should be noted that this study employs dataset-level standardization fitted on the training set and then applied consistently to the validation and test sets, rather than instance-wise normalization performed separately for each input window.

Patching. The normalized sequence$\widetilde{\mathbf{V}}$ is divided into patches to reduce the input length and increase the ability to capture the local structure in the time domain. With each patch length $\boldsymbol{P}$ and stride $\boldsymbol{S}$, the number of patches obtained is calculated as:

Each patch is denoted as:

$$N = \left\lfloor \frac{L-P}{S} \right\rfloor + 2, n = 1,2, \dots, N \quad (5)$$

The set of all patches forms a matrix$\mathbf{P} = [p_1, p_2, \dots, p_N]^\top \in \mathbb{R}^{N \times P}$. Patching helps the model to process in parallel and learn more effectively the local features of the sequence.

Embedding. Each patch$p_n$is projected into the hidden space of dimension d_{model} via a linear embedding layer:

$$\mathbf{E}_n = \mathbf{W}_e \mathbf{p}_n + \mathbf{b}_e,$$

where

($\mathbf{W}_e \in \mathbb{R}^{d_{\text{model}} \times P}$) – the learnable weight matrix

($\mathbf{b}_e \in \mathbb{R}^{d_{\text{model}}}$)- translation vector.

To preserve the temporal order information between patches, a positional encoding component is added to each embedding vector:

$$\mathbf{Z}_n = \mathbf{E}_n + PE_n,$$

where PE_n is defined as:

$$PE_{(n,2i)} = \sin\left(\frac{n}{10000^{\frac{2i}{d_{\text{model}}}}}\right),$$

$$PE_{(n,2i+1)} = \cos\left(\frac{n}{10000^{2i/d_{\text{model}}}}\right). \quad (6)$$

The result is the embedded signature string:

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_N] \in \mathbb{R}^{N \times d_{\text{model}}}. \quad (7)$$

LLM Backbone. The embedding sequence $\mathbf{Z}$ is fed into a Transformer encoder consisting of L_t layers, each layer including a multi-head self-attention mechanism and a feed-forward network (FFN). The calculation formula for an attention head is defined as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \quad (8)$$

with $\mathbf{Q} = \mathbf{ZW}_Q, \mathbf{K} = \mathbf{ZW}_K, \quad \mathbf{V} = \mathbf{ZW}_V,$ and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are the learned weight matrices.

The results from all the heads are concatenated and projected linearly:

$$\mathbf{O} = \text{Concat}(\text{Att}_1, \dots, \text{Att}_H)\mathbf{W}_O.$$

Then, the residual connection block and FFN network are applied:

$$\mathbf{H}^{(l)} = \text{LayerNorm}\left(\mathbf{Z}^{(l-1)} + \text{MHA}\left(\mathbf{Z}^{(l-1)}\right)\right),$$
$$\mathbf{Z}^{(l)} = \text{LayerNorm}\left(\mathbf{H}^{(l)} + \text{FFN}(\mathbf{H}^{(l)})\right).$$

After the L_t layers, the final output of the encoder is:

$$\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_N] \in \mathbb{R}^{N \times d_{\text{model}}}.$$

This $\mathbf{H}$ representation contains contextual features and long-term dependencies in the wind speed series.

Forecasting Head. The output of the encoder $\mathbf{H}$ is passed through a linear regression layer to predict the future series. The projection is performed as follows:

$$\widehat{\mathbf{Y}} = \mathbf{HW}_o + \mathbf{b}_o,$$

where $\mathbf{W}_o \in \mathbb{R}^{d_{\text{model}} \times T}$and $\mathbf{b}_o \in \mathbb{R}^T$.

The result is the forecast wind speed series in the following T time steps:

$$\widehat{\mathbf{Y}} = (\hat{v}_{L+1}, \hat{v}_{L+2}, \dots, \hat{v}_{L+T}).$$

In general, the entire forward propagation process is compactly represented as:

$$\widehat{\mathbf{Y}} = f_{\text{Transformer}}(\text{Embed}(\widetilde{\mathbf{V}})),$$

where

$(f_{\text{Transformer}}(\cdot))$ – the nonlinear mapping of the large language model (LLM), including the embedding, self-attention, and linear regression layers.

Loss Function. The model is trained by minimizing the error between the predicted and actual wind speeds. The loss function used is **MSE**, which is defined as follows:

$$MSE = \frac{1}{NT}\sum_{i=1}^{N}\sum_{\tau=1}^{T}(v_{i,\tau} - \hat{v}_{i,\tau})^2, \quad (9)$$

where i indexes the training sample (or sliding window), and τ denotes the forecast step within the prediction horizon.

The optimization objective of the model is expressed as:

$$\theta^* = \text{argmin}_\theta MSE(\theta) \quad (10)$$

where

(θ) – the set of parameters of the model, including the weights of the embedding, attention, and regression layers.

The Adam algorithm updates weights with an adaptive learning rate, helping the training process converge quickly and stably. To avoid overfitting, dropout and early stopping techniques are used during training.

The linear embedding layer projects continuous time-series patches into the input space of the frozen Transformer backbone, enabling the reuse of pre-trained attention mechanisms for temporal dependency modeling. This allows the frozen Transformer blocks to serve as generalized pattern recognizers, leveraging their robust pre-trained attention mechanisms to capture complex temporal dependencies. For the BART model, specifically, only its encoder module was retained to match the unified encoder-only mathematical formulation described above, while its decoder was discarded.

3. Experiments

3.1 Datasets

In this study, five wind speed datasets from cities representing different regions were used to evaluate the performance of the LLMs. The experiments were designed to verify the model's forecast accuracy, its generalization across different climate domains, and its stability over the forecast horizon. The datasets covered both coastal and inland areas, reflecting the diversity of wind characteristics in space and time. The main information of the data is presented in Table 1.

Table 1
Datasets properties

Datasets	Location	Acquisition time	Sample rate	Samples
Aktobe	50.28 N, 57.17 E	Jan. 2010 – Jul. 2025	1 hour	136584
Aktau	47.12 N, 51.88 E	Jan. 2010 – Jul. 2025	1 hour	136584
Kokshetau	53.28 N, 69.38 E	Jan. 2010 – Jul. 2025	1 hour	136584
Taldykorgan	45.02 N, 78.38 E	Jan. 2010 – Jul. 2025	1 hour	136584
Uralsk	51.23 N, 51.34 E	Jan. 2010 – Jul. 2025	1 hour	136584

3.2. Evaluation and Benchmarking Setup

To evaluate the performance and generalization of the proposed wind speed forecasting model, the dataset is split into three non-overlapping parts: 70% for training, 20% for testing, and 10% for validation. All data are normalized using the mean and standard deviation calculated from the training set to ensure consistent scaling across experiments. The evaluation will focus on two main criteria: model accuracy and stability. Accuracy measures the deviation between the forecasted and actual values, while stability reflects the model's ability to adapt across different climate zones and forecast horizons. Standard evaluation metrics used include Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which are calculated by the formula:

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T} | v_t - \hat{v}_t |, \mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} (v_t - \hat{v}_t)^2}$$

To validate the effectiveness of the proposed LLM-based forecasting framework, the model performance is compared with a group of representtative baseline models. The baseline models include:

(1) ARIMA: a traditional autoregressive statistical model;

(2) SegRNN: a segmented recurrent network model for time series data;

(3) PatchTST: a Transformer model with a patch division mechanism specifically for time series forecasting.

In addition, two large language models (LLMs), including (4) BART and (5) BERT, are put into the experiment to evaluate the adaptability and feasibility of LLMs in wind speed forecasting.

For reproducibility and fair comparison, the main hyperparameter settings of the proposed models and all baseline methods are summarized in Table 2.

All models are trained and evaluated under the same experimental conditions to ensure fairness in comparison. Each experiment was repeated three times with different random seeds, and the mean and standard deviation of the evaluation indexes were reported to minimize the influence of random factors and ensure reproducibility.

Table 2
Main hyperparameter settings of the proposed models and baseline methods

Model	Input length	Patch / Segment length	Stride / No. segments	Hidden size	Main setting	Training
ARIMA	336	—	—	—	Dataset-specific *(p,d,q)* via minimum AIC on training data	—
SegRNN	336	48	7 segments	512	1-layer GRU	30 epochs
PatchTST	336	16	stride = 8	128	3 encoder layers, 4 heads	100 epochs
BERT/BART (proposed)	336	16	stride = 8	768	3 frozen layers, 12 heads	Adam, LR (=10^{-4}), batch size = 32

4. Results

To analyze and compare the overall performance of LLMs, experiments were conducted across five datasets with forecasting horizons of 1, 3, 6, 36, 72, 144, and 432 hours. The forecast errors of LLMs, along with comparisons with other models, are presented in Tables 3 and 4 across the five datasets.

Table 3
MAE of different models on five datasets

Data set	Models	Forecasting horizon (step)						
		1	**3**	**6**	**36**	**72**	**144**	**432**
Aktobe	ARIMA	0.3692 ±0.012	0.6565 ±0.015	1.1874 ±0.021	2.1598 ±0.035	2.6912 ±0.041	3.1485 ±0.045	3.4411 ±0.051
	SegRNN	0.0917 ±0.005	0.2008 ±0.007	0.3514 ±0.009	0.6635 ±0.012	0.7951 ±0.015	0.7892 ±0.014	0.8341 ±0.018
	PatchTST	0.1161 ±0.006	0.1975 ±0.008	0.3218 ±0.011	0.6622 ±0.014	0.7431 ±0.012	0.7825 ±0.015	0.8182 ±0.017
	BART	0.0931 ±0.004	0.1985 ±0.006	0.3223 ±0.008	0.6605 ±0.011	0.7361 ±0.010	0.7725 ±0.011	0.7963 ±0.012
	BERT	0.0712 ±0.003	0.1528 ±0.005	0.3142 ±0.007	0.6614 ±0.009	0.7295 ±0.008	0.7704 ±0.010	0.7955 ±0.011

Continuation of the table

Data set	Models	Forecasting horizon (step)						
		1	3	6	36	72	144	432
Aktau	ARIMA	0.3533 ±0.011	0.6701 ±0.016	1.1805 ±0.020	2.1612 ±0.032	2.6645 ±0.038	3.1388 ±0.042	3.4315 ±0.048
	SegRNN	0.0911 ±0.005	0.2025 ±0.008	0.3541 ±0.010	0.6455 ±0.011	0.6885 ±0.014	0.7824 ±0.016	0.8361 ±0.019
	PatchTST	0.0998 ±0.004	0.1981 ±0.007	0.3238 ±0.009	0.6295 ±0.012	0.7152 ±0.013	0.7845 ±0.015	0.8126 ±0.018
	BART	0.0931 ±0.003	0.1985 ±0.005	0.3218 ±0.007	0.6292 ±0.010	0.6815 ±0.011	0.7802 ±0.012	0.8118 ±0.014
	BERT	0.0901 ±0.003	0.1648 ±0.005	0.3095 ±0.006	0.6195 ±0.009	0.6766 ±0.010	0.7775 ±0.011	0.7996 ±0.013
Kokshetau	ARIMA	0.3942 ±0.013	0.6631 ±0.017	1.1965 ±0.022	2.2155 ±0.036	2.6892 ±0.040	3.2175 ±0.047	3.4355 ±0.052
	SegRNN	0.1292 ±0.006	0.1818 ±0.007	0.3625 ±0.011	0.6432 ±0.013	0.6908 ±0.015	0.7925 ±0.017	0.8105 ±0.018
	PatchTST	0.1155 ±0.005	0.1738 ±0.006	0.3218 ±0.009	0.6295 ±0.011	0.6912 ±0.013	0.7442 ±0.015	0.7785 ±0.016
	BART	0.1135 ±0.004	0.1848 ±0.006	0.3202 ±0.008	0.6265 ±0.010	0.6915 ±0.011	0.7365 ±0.012	0.7792 ±0.014
	BERT	0.1042 ±0.004	0.1752 ±0.005	0.2965 ±0.007	0.6192 ±0.009	0.6885 ±0.010	0.7342 ±0.011	0.7735 ±0.012
Taldykorgan	ARIMA	0.3521 ±0.011	0.6138 ±0.015	1.1685 ±0.021	1.9855 ±0.031	2.4582 ±0.037	3.0265 ±0.042	3.2562 ±0.048
	SegRNN	0.0812 ±0.004	0.1992 ±0.007	0.3508 ±0.009	0.6612 ±0.012	0.6822 ±0.014	0.7855 ±0.016	0.8235 ±0.019
	PatchTST	0.0858 ±0.004	0.1768 ±0.006	0.3355 ±0.009	0.6262 ±0.011	0.6805 ±0.013	0.7872 ±0.015	0.8208 ±0.018
	BART	0.0818 ±0.003	0.1778 ±0.005	0.3222 ±0.008	0.6355 ±0.010	0.6812 ±0.011	0.7865 ±0.013	0.8122 ±0.014
	BERT	0.0782 ±0.003	0.1695 ±0.005	0.3205 ±0.007	0.6265 ±0.009	0.6775 ±0.010	0.7725 ±0.012	0.8118 ±0.013
Uralsk	ARIMA	0.3592 ±0.012	0.6542 ±0.016	1.1722 ±0.021	2.1485 ±0.035	2.6765 ±0.040	3.1254 ±0.045	3.4585 ±0.051
	SegRNN	0.0905 ±0.005	0.1988 ±0.007	0.3452 ±0.010	0.6565 ±0.012	0.6822 ±0.014	0.7765 ±0.016	0.8175 ±0.018
	PatchTST	0.1188 ±0.005	0.2085 ±0.008	0.3225 ±0.009	0.6555 ±0.012	0.6775 ±0.013	0.7812 ±0.015	0.8021 ±0.017
	BART	0.0894 ±0.004	0.1978 ±0.006	0.3215 ±0.008	0.6512 ±0.010	0.6755 ±0.011	0.7715 ±0.012	0.7978 ±0.014
	BERT	0.0895 ±0.003	0.1922 ±0.005	0.3195 ±0.007	0.6562 ±0.009	0.6795 ±0.010	0.7445 ±0.011	0.8005 ±0.013

Table 4
RMSE of different models on five datasets

Data set	Models	Forecasting horizon (step)						
		1	3	6	36	72	144	432
Aktobe	ARIMA	0.4576 ±0.014	0.8318 ±0.020	1.4725 ±0.028	2.7391 ±0.042	3.3362 ±0.049	3.9125 ±0.055	4.2685 ±0.062
	SegRNN	0.1162 ±0.006	0.2541 ±0.009	0.4362 ±0.012	0.8255 ±0.016	1.0092 ±0.019	0.9785 ±0.018	1.0592 ±0.021
	PatchTST	0.1468 ±0.007	0.2448 ±0.010	0.4091 ±0.013	0.8382 ±0.018	0.9245 ±0.017	0.9935 ±0.020	1.0178 ±0.022
	BART	0.1154 ±0.005	0.2525 ±0.008	0.4002 ±0.010	0.8385 ±0.015	0.9128 ±0.014	0.9778 ±0.016	1.0108 ±0.018
	BERT	0.0892 ±0.004	0.1948 ±0.006	0.3915 ±0.009	0.8371 ±0.013	0.9252 ±0.012	0.9775 ±0.014	1.0092 ±0.016
Aktau	ARIMA	0.4488 ±0.013	0.8285 ±0.019	1.4635 ±0.026	2.7395 ±0.041	3.3698 ±0.048	3.8925 ±0.054	4.3452 ±0.060
	SegRNN	0.1152	0.2512	0.4505	0.8188	0.8535	0.9898	1.0575

Continuation of the table

Data set	Models	Forecasting horizon (step)						
		1	3	6	36	72	144	432
		±0.006	±0.009	±0.013	±0.015	±0.017	±0.020	±0.022
	PatchTST	0.1248 ±0.005	0.2472 ±0.008	0.4115 ±0.011	0.7805 ±0.014	0.9075 ±0.016	0.9728 ±0.018	1.0282 ±0.020
	BART	0.1176 ±0.004	0.2528 ±0.007	0.4012 ±0.009	0.7962 ±0.012	0.8458 ±0.014	0.9902 ±0.016	1.0065 ±0.018
	BERT	0.1118 ±0.004	0.2061 ±0.006	0.3855 ±0.008	0.7842 ±0.011	0.8562 ±0.013	0.9672 ±0.015	0.9941 ±0.017
Kokshetau	ARIMA	0.4912 ±0.015	0.8222 ±0.020	1.4885 ±0.028	2.8085 ±0.043	3.3358 ±0.050	4.0712 ±0.058	4.2625 ±0.063
	SegRNN	0.1635 ±0.007	0.2272 ±0.009	0.4498 ±0.013	0.8005 ±0.016	0.8765 ±0.018	1.0022 ±0.021	1.0275 ±0.022
	PatchTST	0.1432 ±0.006	0.2172 ±0.008	0.3992 ±0.011	0.7835 ±0.014	0.8575 ±0.016	0.9255 ±0.018	0.9855 ±0.020
	BART	0.1451 ±0.005	0.2338 ±0.007	0.3972 ±0.010	0.7772 ±0.012	0.8582 ±0.014	0.9342 ±0.016	0.9662 ±0.018
	BERT	0.1332 ±0.004	0.2221 ±0.006	0.3751 ±0.009	0.7708 ±0.011	0.8561 ±0.013	0.9288 ±0.015	0.9812 ±0.017
Taldykorgan	ARIMA	0.4452 ±0.013	0.7785 ±0.018	1.4495 ±0.026	2.4668 ±0.038	3.0495 ±0.045	3.7592 ±0.052	4.0412 ±0.058
	SegRNN	0.1021 ±0.005	0.2522 ±0.009	0.4455 ±0.012	0.8225 ±0.016	0.8632 ±0.018	0.9965 ±0.021	1.0425 ±0.023
	PatchTST	0.1075 ±0.005	0.2252 ±0.008	0.4245 ±0.011	0.7922 ±0.014	0.8445 ±0.016	0.9785 ±0.019	1.0205 ±0.021
	BART	0.1038 ±0.004	0.2222 ±0.007	0.3998 ±0.010	0.8062 ±0.013	0.8622 ±0.015	0.9951 ±0.018	1.0282 ±0.020
	BERT	0.0972 ±0.003	0.2162 ±0.006	0.3991 ±0.009	0.7925 ±0.011	0.8592 ±0.014	0.9782 ±0.016	1.0295 ±0.018
Uralsk	ARIMA	0.4455 ±0.013	0.8135 ±0.019	1.4835 ±0.027	2.7242 ±0.041	3.3875 ±0.049	4.0265 ±0.056	4.2915 ±0.061
	SegRNN	0.1148 ±0.005	0.2488 ±0.009	0.4281 ±0.012	0.8175 ±0.016	0.8635 ±0.018	0.9855 ±0.020	1.0172 ±0.022
	PatchTST	0.1488 ±0.006	0.2642 ±0.010	0.4005 ±0.011	0.8298 ±0.015	0.8435 ±0.016	0.9688 ±0.019	0.9955 ±0.021
	BART	0.1112 ±0.004	0.2475 ±0.008	0.4092 ±0.010	0.8082 ±0.013	0.8551 ±0.014	0.9785 ±0.017	1.0122 ±0.019
	BERT	0.1145 ±0.004	0.2435 ±0.007	0.4045 ±0.010	0.8335 ±0.014	0.8605 ±0.015	0.9238 ±0.016	1.0155 ±0.018

Experimental results on five wind speed datasets show a large difference in performance between the models. Transformer-based baselines and the proposed frozen pre-trained Transformer backbone models, such as PatchTST, BART, and BERT, significantly outperform the classical ARIMA model. This difference is measured using two metrics, MAE and RMSE, over forecasting horizons ranging from 1 to 432 steps. In addition, the resulting errors also increase as the forecasting horizon increases. This is a common feature of long-term forecasting, reflecting increased uncertainty and the accumulation of errors over time.

The ARIMA model has the highest error in most cases. With long forecasting steps (144-432), the RMSE error increases sharply, reaching from 3.9 to 4.4. This shows the limitation of the linear model in describing the nonlinear characteristics and fluctuations of wind.

Meanwhile, the SegRNN and PatchTST models significantly improve the forecasting results at short time steps (1-72) with MAE ranging from 0.08 to 0.354 and RMSE ranging from 0.1 to 1. This shows that the latter models better capture the temporal relationship than ARIMA.

For BART and BERT, these models achieve the best results on most datasets and forecasting steps. In particular, with a long forecasting step of 432, the MAE and RMSE of BERT are 50-70% lower than those of ARIMA. These results indicate that frozen pre-trained Transformer backbones can achieve better accuracy in long-term forecasting under the present experimental setting.

Geographically, regions exposed to strong continental steppe winds (such as Aktobe) and coastal areas (such as Aktau) exhibit higher errors at long forecasting steps due to stronger, less predictable wind fluctuations.

Figures 3 and 4 show the MAE and RMSE plots of the models on the five datasets, respectively. MAE and RMSE increase with the forecasting step. However, the increase in the proposed BERT/BART-based models is slower and more stable than that of other models, suggesting that the frozen pre-trained Transformer blocks provide effective long-range sequence modeling in this forecasting task.

To assess whether the performance differences were statistically significant, the Diebold-Mariano (DM) test was conducted to compare the proposed BERT and BART models against the strongest baseline, PatchTST. The complete DM statistics and corresponding p-values for all forecasting horizons are provided in the Appendix A2. Overall, the DM test results support the superiority of the proposed models across most forecasting horizons, with stronger statistical evidence observed particularly at the 144-step and 432-step horizons.

While models achieve high accuracy using standard metrics (MSE, MAE), these metrics treat all errors equally. In reality, wind power generation is proportional to the cube of wind speed $P \propto v^3$. Therefore, forecasting errors at high wind speeds impact grid stability much more severely than those at low speeds. Future research should integrate a power-weighted loss function during LLM fine-tuning to better capture extreme wind dynamics and align with real-world operational requirements.

Figure 3
Comparison of the MAE of different models across five cities in Kazakhstan

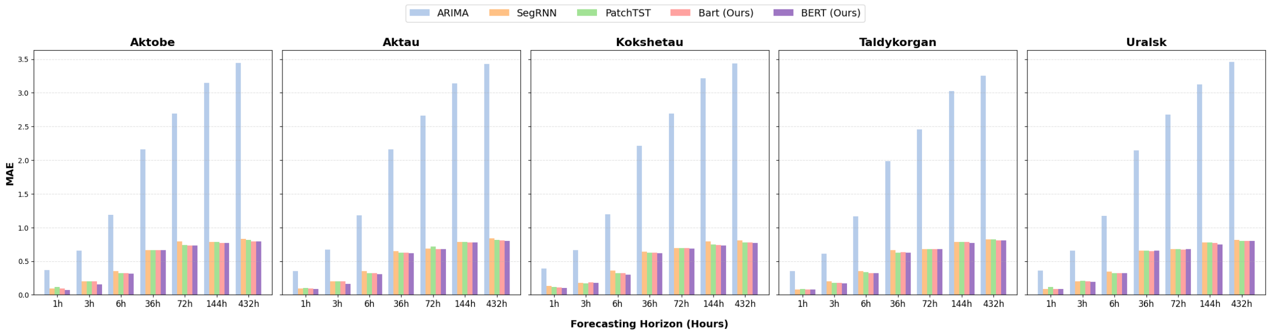
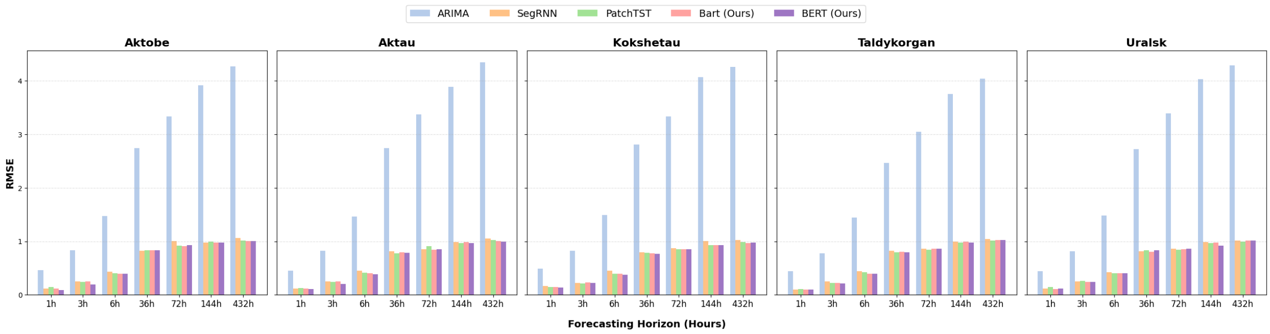


Figure 4
Comparison of the RMSE of different models across five cities in Kazakhstan



5. Conclusion

This study proposed a wind speed forecasting framework based on frozen pre-trained Transformer backbones (BERT and BART) to improve forecasting accuracy across different climatic and topographic regions of Kazakhstan. Five different datasets and seven forecasting horizons were used to evaluate the performance of the proposed framework. Under the present experimental setup, the proposed BERT- and BART-based models achieved lower MAE and RMSE than the selected statistical and deep learning baselines, while maintaining relatively stable performance at long forecasting horizons. The results suggest that frozen pre-trained Transformer backbones, when combined with lightweight adaptation, can effectively model nonlinear, non-stationary, and long-range temporal dependencies in wind-speed data. Overall, this study supports the potential of pre-trained Transformer backbones as promising tools for wind-speed forecasting and provides an application-oriented reference for planning and operation in wind power systems.

Funding

The research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP26198931).

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

Table A1

Summary aggregated across the five datasets of Diebold-Mariano test results for BERT/BART versus PatchTST across forecasting horizons

Horizon	BERT vs PatchTST	BART vs PatchTST
1	DM = 1.72, (p=0.089)	DM = 1.81, (p=0.074)
3	DM = 1.96, (p=0.054)	DM = 2.04, (p=0.046)
6	DM = 2.11, (p=0.039)	DM = 2.18, (p=0.033)
36	DM = 2.28, (p=0.027)	DM = 2.35, (p=0.022)
72	DM = 2.41, (p=0.021)	DM = 2.53, (p=0.016)
144	DM = 2.86, (p=0.008)	DM = 2.97, (p=0.006)
432	DM = 3.12, (p=0.004)	DM = 3.26, (p=0.002)

Author Contributions

Conceptualization, T.T.C. and A.B.A.; Methodology, T.T.C.; Software, T.T.C.; Validation, T.T.C., M.S. and K.T.; Formal Analysis, T.T.C.; Investigation, T.T.C., M.S. and M.A.S.; Resources, A.B.A.; Data Curation, T.T.C.; Writing – Original Draft Preparation, T.T.C.; Writing – Review & Editing, T.T.C., M.A.S. and A.B.A.; Visualization, T.T.C.; Supervision, A.B.A.; Project Administ-ration, A.B.A.; Funding Acquisition, A.B.A.

References

1. T. T. Chau, A. B. Alhassan, M. Shaltayev, K. Talapiden, M. A. Shehu, and T. D. Do, “Assessment of wind energy potential using hybrid adaptive bandwidth kernel density technique: A case study of major cities in Kazakhstan,” *Energy Conversion and Management: X*, vol. 29, p. 101439, Jan. 2026, doi: 10.1016/j.ecmx.2025.101439.
2. U. Alparslan, “Green energy corridors for Central Asia and the Caucasus”. Accessed: Feb. 22nd, 2026. [Online]. Available: https://ember-energy.org/app/uploads/2024/11/Report-Green-energy-corridors-for-Central-Asia-and-the-Caucasus.pdf
3. IEA, “Kazakhstan 2022”, Accessed: Oct. 09, 2025. [Online]. Available: https://iea.blob.core.windows.net/assets/fc84229e-6014-4400-a963-bccea29e0387/Kazakhstan 2022.pdf
4. T. T. Chau, T. N. Nguyen, and T. D. Do, “Assessing wind energy exploitation potential in several regions of Viet Nam using Kernel density estimation model,” *CTU Journal of Innovation and Sustainable Development*, vol. 16, pp. 25–34, 2024, doi: 10.22144/ctujoisd.2024.319.
5. T. T. Chau, T. T. H. Nguyen, L. Nguyen, and T. D. Do, “Wind Speed Probability Distribution Based on Adaptive Bandwidth Kernel Density Estimation Model for Wind Farm Application,” *Wind Energy*, vol. 28, no. 2, p. e2970, Feb. 2025, doi: 10.1002/WE.2970.
6. A. M. Khan, A. B. Alhassan, A. Kolumbetov, A. Haruna, V. Gali, N. G. M. Thao, and T. D. Do, “Optimizing demand-side energy management for stand-alone wind-solar microgrids in rural settlements: A case study for nomadic Yurt in Kazakhstan,” *Energy Conversion and Management: X*, vol. 30, p. 101587, May 2026, doi: 10.1016/j.ecmx.2026.101587.
7. T. T. Chau, H. T. Nguyen, and T. D. Do, “Online Nonlinear Quadratic Tracking Control Accompanied by Selective Current Harmonic Compensator for MPPT of PMSG Drives,” *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2026, doi: 10.1109/JESTPE.2026.3656176.
8. T. D. Do, H. T. Nguyen, A. S. Al-Sumaiti, K. Al Hosani, and D. D. Nguyen, “Nonlinear Quadratic Regulator Design With a Hybrid Switching Scheme for Wind Turbine Generators,” *IEEE Trans Syst Man Cybern Syst*, vol. 53, no. 9, pp. 5525–5535, Sep. 2023, doi: 10.1109/TSMC.2023.3272222.

9. T. T. Chau, T. N. Nguyen, L. Nguyen, A. B. Alhassan, and T. D. Do, “A maximum power point tracking control for wind energy conversion systems using regularized data-enabled predictive control,” *Engineering Applications of Artificial Intelligence*, vol. 170, p. 114005, Apr. 2026, doi: 10.1016/j.engappai.2026.114005.

10. M. A. Shehu, K. Talapiden, T. T. Chau, A. Haruna, M. Aly, V. Gali, T. D. Do, and A. B. Alhassan, “Control and energy management of standalone microgrids in remote areas: A review of recent advances, challenges, and opportunities for future research,” *Engineering Science and Technology, an International Journal*, vol. 75, p. 102288, Mar. 2026, doi: 10.1016/j.jestch.2026.102288.

11. G. Zhu, W. Jia, Z. Xing, L. Xiang, A. Hu, and R. Hao, “CMLLM: A novel cross-modal large language model for wind power forecasting,” *Energy Convers Manag*, vol. 330, p. 119673, Apr. 2025, doi: 10.1016/J.ENCONMAN.2025.119673.

12. H. R. Alsamamra, S. Salah, and J. H. Shoqeir, “Performance analysis of ARIMA Model for wind speed forecasting in Jerusalem, Palestine,” *Energy Exploration & Exploitation*, vol. 42, no. 5, pp. 1727–1746, Sep. 2024, doi: 10.1177/01445987241248201.

13. L. Xiaolei, L. Zi, and F. Ziming, “Short-term offshore wind speed forecast by seasonal ARIMA – A comparison against GRU and LSTM,” *Energy (Oxf.)*, vol. 227, no. 120492 Jul 2021, doi: 10.1016/j.energy.2021.120492.

14. H. Liu, H. qi Tian, and Y. fei Li, “Comparison of two new ARIMA-ANN and ARIMA-Kalman hybrid methods for wind speed prediction,” *Appl Energy*, vol. 98, pp. 415–424, Oct. 2012, doi: 10.1016/J.APENERGY.2012.04.001.

15. G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, “JOURNAL OF TIME SERIES ANALYSIS BOOK REVIEW TIME SERIES ANALYSIS: FORECASTING AND CONTROL, 5TH EDITION, by,” *J. Time. Ser. Anal*, vol. 37, pp. 709–711, 2016, Accessed: Oct. 09, 2025. [Online]. Available: https://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021

16. J. Duan *et al.*, “A combined short-term wind speed forecasting model based on CNN–RNN and linear regression optimization considering error,” *Renew Energy*, vol. 200, pp. 788–808, Nov. 2022, doi: 10.1016/J.RENENE.2022.09.114.

17. J. Zheng and J. Wang, “Short-term wind speed forecasting based on recurrent neural networks and Levy crystal structure algorithm,” *Energy*, vol. 293, p. 130580, Apr. 2024, doi: 10.1016/J.ENERGY.2024.130580.

18. X. Ai, S. Li, and H. Xu, “Short-term wind speed forecasting based on two-stage preprocessing method, sparrow search algorithm and long short-term memory neural network,” *Energy Reports*, vol. 8, pp. 14997–15010, Nov. 2022, doi: 10.1016/J.EGYR.2022.11.051.

19. K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

20. Z. Fu *et al.*, “An Informer-BiGRU-temporal attention multi-step wind speed prediction model based on spatial-temporal dimension denoising and combined VMD decomposition,” *Energy*, vol. 326, p. 136265, Jul. 2025, doi: 10.1016/J.ENERGY.2025.136265.

21. G. Ban, Y. Chen, Z. Xiong, Y. Zhuo, and K. Huang, “The univariate model for long-term wind speed forecasting based on wavelet soft threshold denoising and improved Autoformer,” *Energy*, vol. 290, p. 130225, Mar. 2024, doi: 10.1016/J.ENERGY.2023.130225.

22. T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, “FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting,” *Proc Mach Learn Res*, vol. 162, pp. 27268–27286, Jan. 2022, Accessed: Oct. 09, 2025. [Online]. Available: https://arxiv.org/pdf/2201.12740

23. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies – Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Oct. 09, 2025. [Online]. Available: https://arxiv.org/pdf/1810.04805

24. M. Lewis *et al.*, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Oct. 2019, doi: 10.18653/v1/2020.acl-main.703.

25. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving Language Understanding by Generative Pre-Training”, Accessed: Oct. 09, 2025. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

26. H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

27. M. Jin *et al.*, “Time-LLM: Time Series Forecasting by Reprogramming Large Language Models,” *12th International Conference on Learning Representations, ICLR 2024*, Oct. 2023, Accessed: Oct. 09, 2025. [Online]. Available: https://arxiv.org/pdf/2310.01728

Information about the authors:

Tin Trung Chau received the B.S. degree in Automation and Control Engineering Technology and the M.S. degree in Electrical Engineering from the Vinh Long University of Technology Education, Vietnam, in 2020 and 2022, respectively. He is a Ph.D. student in Robotics at the School of Engineering and Digital Sciences, Nazarbayev University, Kazakhstan. His research interests include robotics, renewable energy conversion systems, and machine learning (ORCID iD: 0009-0009-7831-8864).

Miras Shaltayev is a BSc student in Computer Science at the School of Engineering and Digital Sciences at Nazarbayev University. Kazakhstan. His research interests include machine learning, deep learning and renewable energy (ORCID iD: 0009-0008-4457-3763).

Kulash Talapiden received her Bachelor's degree (Hons.) in Technological Machinery and Equipment from Kazakh Agrotechnical University in 2020, and her M.S. degree in Robotics from Nazarbayev University. Her research interests include

control systems, optimization techniques for control system design and analysis, feedback control in motor drives, and the control of autonomous systems (ORCID iD: 0000-0002-1757-7311).

Muhammad Auwal Shehu received the B.Eng. degree in electrical engineering from Bayero University, Kano, Nigeria, in 2012 and the M.Eng. degree in mechatronics and automatic control from the University of Technology Malaysia in 2016. His primary research interests include the design of nonlinear control systems and state estimators for underactuated mechatronic systems (ORCID iD: 0000-0002-2158-7619).

Ahmad Bala Alhassan received a Ph.D. degree in Mechanical Engineering from Xi'an Jiaotong University, China, in 2022. He is currently a Postdoctoral Scholar in the Department of Robotics and Mechatronics at Nazarbayev University, Kazakhstan. His current research interests include modeling, simulation, and control of mechatronic systems. He has also served as a reviewer for many refereed journals, including IEEE Access, ISA Transactions, and Robotics and Autonomous Systems. He has been a Member of IEEE, including IEEE Young Professionals, IEEE Control Systems Society, and IEEE Industrial Electronics Society (ORCID iD: 0000-0002-0179-7686).

Submission received: 19 January, 2026
Revised: 16 March, 2026
Accepted: 16 March, 2026