A Hybrid BERT-ELM Framework for Robust Time Series Forecasting of Solar Energy Generation in EU Renewable Power Plants

SEYYED KASRA MORTAZAVI, HUSSEIN AHMAD AHMAD, MAHMOUD HAMED,
KYANDOGHERE KYAMAKYA
Institute for Smart System Technologies
Universität Klagenfurt,
AUSTRIA

Abstract: Precise short-term forecasting of photovoltaic (PV) power is essential for grid stability and the integration of renewables. We propose two hybrid architectures—*TS-BERT+ELM* and *PatchTST+ELM*—separate temporal representation learning from regression by integrating transformer-based encoders with a ridge-regularized Extreme Learning Machine (ELM) for rapid, low-latency prediction. An evaluation of one-day-ahead predictions from 14-day input windows is conducted using daily PV datasets from five EU nations (Germany, France, Switzerland, Denmark, and the United Kingdom) provided from OPSD and enhanced with NASA POWER meteorological variables (global horizontal irradiance, cloud cover, and temperature) ($f : \mathbb{R}^{14\times d} \rightarrow \mathbb{R}$). We present MAE, MSE, R^2, and threshold accuracies (Accuracy@10%, Accuracy@50%), cexecute ablation, convergence, and sensitivity studies, and conduct paired t-tests and Wilcoxon signed-rank tests for statistical validation. Results indicate that *TS-BERT+ELM* regularly surpasses baselines on noisy and irregular datasets (France, Germany), whereas *PatchTST+ELM* demonstrates strong performance with high-quality, structured data (Denmark, UK); Switzerland occupies a position bridging the two categories. Integrating external weather-related features further enhances predictive accuracy and decreases variance, with statistically significant gains ($p < 0.05$) in four countries and an inconclusive UK case due to high variance. This modular design facilitates rapid convergence, maintains robustness against missing inputs, and enhances operational efficiency, and is compatible with federated and transfer learning for privacy-preserving, cross-site deployment. These findings support scalable, multimodal, and privacy-aware PV forecasting in real-world energy systems.

Key-Words: Solar energy, TS-BERT, PatchTST, Extreme Learning Machine (ELM), hybrid deep learning, short-term forecasting

Received: March 19, 2025. Revised: July 17, 2025. Accepted: August 16, 2025. Published: November 18, 2025.

1 Introduction

Between 2010 and 2024, several international agreements have emerged to curb greenhouse gas emissions and accelerate the global transition toward sustainable energy systems. The most prominent of these is the 2015 Paris Agreement, a milestone in climate diplomacy that established a unified commitment to limit global temperature rise to well below 2°C. This accord catalyzed coordinated mitigation strategies worldwide. In parallel, the European Union (EU) establish bold targets for reducing emissions, with a clear emphasis on expanding photovoltaic (PV) energy adoption and achieving net-zero carbon emissions by 2050 [1].

Despite its environmental benefits, widespread deployment of solar power poses significant challenges, particularly in maintaining real-time supply-demand balance, ensuring grid stability, and managing fluctuations in energy markets. As a result, accurate short-term forecasting of solar energy generation has become essential for operational decision-making in power planning, grid dispatching, and energy trading.

However, short-term solar forecasting is inherently complex due to the intermittent nature of solar irradiance, seasonal variability, and regional heterogeneity. Traditional statistical models such as Seasonal ARIMA (SARIMA) and ARIMA often fall short in capturing the dynamic, high-dimensional, and incomplete nature of solar time series data. These limitations are especially pronounced in large-scale utility applications [2].

Recent advancements in deep learning particularly transformer-based architectures have substantially strengthened the ability to capture complex temporal pattern. Self- attention mechanisms, as a fundamental element of transformer architectures, they facilitate precise temporal context modeling, thereby providing an effective solution to the inherent unpredictability of solar data. Among these, TS-BERT employs

bidirectional masked modeling to reveal underlying structures, whereas PatchTST divides sequences into fixed-size patches, enabling the extraction of temporal features at both global and local scales [3, 4]. Despite their strong predictive capabilities, the computational overhead of these models reduces their practicality for time-sensitive or low-resource settings. To address this, Extreme Learning Machines (ELMs) have been introduced as computationally light options for regression problems. ELMs are single-hidden-layer feedforward neural networks characterized by random parameter initialization and closed- form analytical solutions, enabling rapid training and strong generalizability [5]. While ELMs have shown promise in energy forecasting, their integration with deep sequence en- coders—especially within hybrid frameworks applied to com- plex, noisy solar datasets—remains underexplored.

In this paper, we propose two hybrid deep learning architectures to bridge this gap: **TS-BERT+ELM** and **PatchTST+ELM**. The former combines bidirectional masked encoding with an efficient, closed-form regression layer, while the latter integrates patch-based temporal encoding with ELM-driven output generation. These models decouple temporal feature extraction from prediction, enabling low-latency inference without compromising accuracy. To enhance transparency and reproducibility, we provide formal derivations of key transformer components, including self-attention and patch embedding mechanisms.

The proposed models are rigorously evaluated against benchmark architectures—LSTM and GRU—across five national-scale PV datasets from France, Germany, Switzerland, the United Kingdom, and Denmark. These datasets span varying data volumes, noise levels, and sequence complexities. Performance is assessed using standard metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), the coefficient of determination (R^2), and prediction accuracy within $\pm 10\%$ and $\pm 50\%$ bounds.

Our experimental results offer a comprehensive view of model performance under diverse forecasting conditions. Statistical robustness is confirmed through paired *t*-tests and Wilcoxon signed-rank tests. Additionally, convergence behavior and sensitivity to patch sizes and input windows are analyzed. An ablation study dissects the individual contributions of the TS-BERT, PatchTST, and ELM components.

While the existing models function exclusively based on internal time series data, we propose methods for integrating incorporating exogenous meteorological inputs using multimodal attention. The suggested system is designed for compliance with federated and transfer learning paradigms, ensuring scalability and privacy preservation in distributed, cross-site forecasting applications.

2 Problem Statement

Forecasting solar energy generation is challenging due to the inherently nonlinear, multiscale, and stochastic characteristics of renewable power systems. Solar irradiance is affected by several factors, including atmospheric changes, cloud cover, seasonal fluctuations, and geographical variety. The impacts produce time series data often non-stationary, noisy, and incomplete [3, 4], introducing difficulties in accurate forecasting, especially with multivariate signals or irregular temporal structures. Moreover, external meteorological variables—such as solar irradiation, cloud cover, and ambient temperature—significantly influence photovoltaic power generation. Excluding these elements from the modeling process may result in diminished generality in the final estimates, particularly across diverse meteorological or geographic contexts.

In forecasting applications, there is a significant demand for models that provide high accuracy, computational economy, and adaptability to real-world scenarios. Empirical evidence from experiments indicates that power system data often exhibit structural and measurement anomalies, such as missing entries, irregular sensor measurements, and inconsistent data representations, all of which undermine the reliability of conventional modeling approaches. The effects of these problems are particularly pronounced in real operating environments, when predictive models must operate effectively, require minimal retraining, and preserve stability in the presence of noisy or missing data. Moreover, guaranteeing statistical robustness via confidence testing is crucial for verifying these models beyond experimental performance. This study aimed to provide a forecasting framework designed to operate dependably in challenging and imperfect circumstances. The suggested framework aims to rectify data abnormalities, adapt to fluctuations in sequence patterns, and achieve superior performance while requiring limited computational resources. The primary aim is to provide a flexible system that reliably excels in many forecasting situations, despite the presence of low-quality or imperfect data. The proposed hybrid architecture is also designed to support future extensions such as federated learning or transfer learning, enabling scalable deployment in privacy-sensitive or data-sparse environments.

3 Related Work

The accurate prediction of solar energy output is critical for the stability and operational effectiveness

of contemporary power networks. Undoubtedly, we are witnessing the prominent role of renewable energies as one of the primary energy sources. Hence, the need to employ advanced and accurate methods for forecasting generation is of great importance, as these sources introduce greater variability into power system operations. It is worth noting that, with advancements in adaptive technologies in the field of time series, new frameworks have been introduced, ranging from traditional statistical methods to deep learning models. These methods are specifically analyzed and differentiated based on the structure and complexity of the time patterns, the reliability and accuracy of the input data, and the unique needs of each scenario.

3.1 Statistical Forecasting models

Classical uni-variate models such as Auto-regressive Integrated Moving Average (ARIMA) and its seasonal variant (SARIMA) have long been used for energy forecasting due to their minimal complexity with intuitive insight [5]. ARIMA-based models forecast future values as linear combinations of past observations and residuals, assuming stationarity and regular time steps:

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Seasonal behavior is modeled by enhancing the architecture with explicitly defined periodic structures $(P, D, Q)_s$. However, real-world solar datasets often violate these assumptions due to missing values, noise, and non-stationary. This diminishes the efficacy of ARIMA/SARIMA for multivariate and high-dimensional datasets characteristic of utility-scale solar forecasting [6, 7]. Furthermore, conventional models fail to accurately represent the nonlinear interdependencies among variables such as irradiance, temperature, and time of day, which are essential in multivariate photovoltaic systems. These limitations have prompted the exploration of more sophisticated models, including recurrent neural networks and those utilizing attention processes.

3.2 Transformer-Based Forecasting models

Initially developed for natural language processing, the transformer architecture [8] has become more significant in time series forecasting due to its ability to model long-range dependencies without relying on recurrent or convolutional structures. This capacity has transformed a formidable option for energy and load forecasting jobs.

Among transformer versions, time series BERT (TS-BERT) modifies the masked language modeling approach for time series by employing masked time-step reconstruction and bidirectional self-attention, allowing the model to acquire temporal properties from preceding and subsequent contexts. Obscured input sequences are reconstructed according to the following procedure:

$$\hat{X}_t = \text{Transformer}(\text{Mask}(X_t))$$

relationships across many scales, while sustaining strong performance while encountering incomplete or noisy time series data [2].

Alternatively, PatchTST segments the input data into smaller, standardized units. Segmentation occurs without overlap, followed by linear transformation into a concealed representatione:

$$z_i = W \cdot \text{vec}(X_i)$$

PatchTST use transformer encoders to interpret temporal patterns in data, comprising both short-term (local) and long-term (global) dimensions. Nevertheless, a limitation arises because it analyzes each variable individually, for instance, temperature or solar irradiance, without accounting for their combined effects. This exclusion constrains the model's ability to discern connections among various variables. This is crucial for capturing complex, multivariate data and for developing multimodal learning frameworks that incorporate weather-related features. [9].

Addressing these problems, academics have proposed advanced hybrid frameworks like PatchFusionBERT. propose a hybrid strategy that integrates PatchTST's localized temporal resolution with BERT's capacity to comprehend global sequential context. Figure 1, illustrates that the architecture of PatchTSTBERT is engineered to successfully capture short-term patterns while also utilizing temporal dependencies across extended horizons, thereby providing a more comprehensive knowledge of time-series dynamics [10].

This dual-stream architecture markedly enhances the model's capacity to estimate solar production across several factors, especially for prolonged forecasting periods. Specifically, it attains enhanced training efficiency. Despite necessitating marginally greater processing resources, this trade-off produces more stable and dependable forecasts, particularly in contexts requiring intricate datasets or extended-term projections. Although these structures adeptly capture intricate temporal connections, a significant shortcoming in the current research is the lack of statistical testing to validate the relevance of the reported performance improvements. This research fills this gap by formal hypothesis testing on conventional forecasting accuracy metrics.

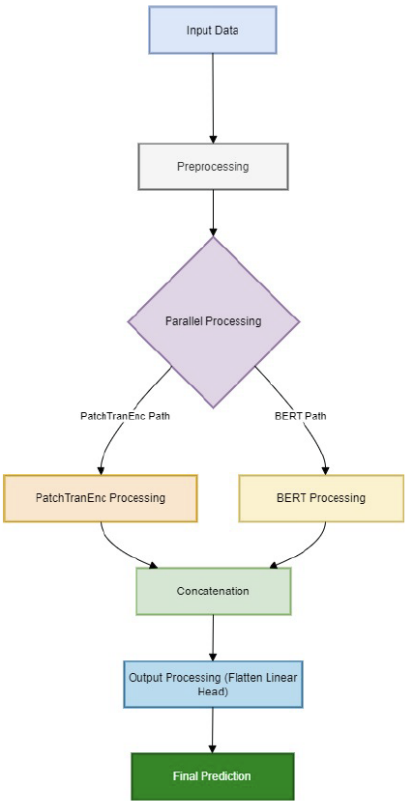


Figure 1: Pipeline of the PatchFusionBERT architecture, Source: [10]

3.3 Extreme Learning Machines and Hybrid Architectures

Through single-layer feedforward neural networks with randomly set hidden parameters and analytically calculated output weights, Extreme Learning Machines achieve efficient regression [5]. The fast training speed and simple architecture of ELMs make them well-suited for low-latency applications, such as edge computing and lightweight deployment scenarios.

In the past few years, hybrid structures that combine deep learning encoders with ELM-based output layers have proved to be a promising direction. For example, P-ELM integrates convolutional neural networks with an ELM-built output layer for the purpose of forecasting short-term solar power generation [7]. Although hybrid setups have shown promising performance, their testing has mainly been in controlled setups. Consequently, they tend to perform poorly when implemented on real-world datasets with noise, nonuniform sampling, and multivariable correlations.

To complement these, our work presents two new hybrid architectures, TS-BERT+ELM and PatchTST+ELM, which combine the contextual modeling strengths of transformer encoder architectures with the low-complexity inference capabilities of ELMs. While models like PatchFusionBERT are designed with a focus on hierarchical representation learning, the proposed architectures are designed keeping in mind operational efficiency, flexibility towards real-world conditions, and tolerance towards imperfect or noisy data.

By removing the dependence of feature extraction on the regression step and taking advantage of the native speed of the ELM, TS-BERT+ELM and PatchTST+ELM are able to make a good trade-off between model simplicity and prediction performance. With their light but stable structure, they are suitable options for real-time solar forecasting, especially in distributed, information-intensive, and dynamically changing setups. In contrast to prior hybrid models that were primarily evaluated on clean or single-site datasets, the present study conducts extensive testing under real-world conditions—characterized by noisy inputs, irregular temporal sampling, and cross-national data heterogeneity—thereby providing deeper insights into model generalizability across diverse operational scenarios.

3.4 Federated and Transfer Learning Opportunities

The rapid expansion of **distributed renewable energy infrastructures**, such as solar and wind power stations, necessitates forecasting frameworks capable of handling **privacy constraints**, **heterogeneous (non-IID) datasets**, and **limited local data availability**. Recent advances in **Federated Learning (FL)**, **Transfer Learning (TL)**, and their integration as **Federated Transfer Learning (FTL)** provide promising solutions to these challenges.

3.4.1 Federated Learning (FL)

Federated Learning enables multiple decentralized entities (e.g., regional utilities, solar farms, or wind clusters) to collaboratively train a shared model **without exchanging raw data**, thereby preserving **privacy** (e.g., GDPR compliance) and protecting **proprietary datasets** [11, 12]. Each client k with dataset D_k of size n_k performs local model training, updating weights $w_k^{(t+1)}$ at communication round $t+1$. The central server aggregates these updates using the *Federated Averaging (FedAvg)* algorithm:

$$w^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^{(t+1)}, \quad N = \sum_{k=1}^{K} n_k, \quad (1)$$

where K denotes the total number of participating clients. This strategy produces a **global model** that is robust to **non-IID and imbalanced data distributions** while minimizing communication overhead [11].

3.4.2 Transfer Learning (TL)

Many renewable forecasting sites face **data scarcity**, limiting their ability to train accurate models from scratch. **Transfer Learning (TL)** addresses this by **fine-tuning models pre-trained on data-rich regions** for use in target sites with limited data [13]. The fine-tuned target model parameters w_{tgt} are obtained by solving:

$$w_{\text{tgt}} = \arg\min_{w} \mathcal{L}_{\text{target}}(w; D_{\text{target}}) + \lambda \|w - w_{\text{src}}\|^2, \quad (2)$$

where w_{src} represents the parameters from a source model, D_{target} is the target dataset, and λ controls regularization to avoid catastrophic forgetting. TL improves **generalization** and accelerates **convergence** for low-resource forecasting tasks [14, 15].

3.4.3 Federated Transfer Learning (FTL)

By combining FL and TL, **Federated Transfer Learning (FTL)** enables privacy-preserving global collaboration while adapting to site-specific data conditions. A global model $w^{(t)}$ is first trained via FedAvg (Eq. 1). Each client then fine-tunes this global model locally by solving:

$$w_k^{(t+1)} = \arg\min_{w} \mathcal{L}_k(w; D_k) + \lambda \|w - w^{(t)}\|^2, \quad (3)$$

and the fine-tuned weights are aggregated:

$$w^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^{(t+1)}. \quad (4)$$

This dual-stage approach ensures that each site **benefits from shared global knowledge while achieving personalized performance**, as demonstrated in **wind power forecasting** [14], **PV generation prediction** [15], and **solar fault detection** [16].

3.4.4 Implications for Renewable Forecasting

FL, TL, and FTL have demonstrated substantial benefits for large-scale renewable forecasting:

- Up to **43% improvements in forecasting accuracy** via federated pretraining with personalized transfer learning for wind power [14].
- **Enhanced PV prediction accuracy** using FL-TL Conv-SGRU hybrids while ensuring **local data privacy** [15].
- **Reliable PV system fault detection** with federated transfer learning with VGG-16 and FedAvg, attaining performance on par with centralized models while preserving sensitive data. [16].

These hybrid methodologies provide a **scalable pathway** for incorporating the proposed **TS-BERT+ELM** and **PatchTST+ELM** frameworks into **cross-border, privacy-conscious forecasting systems**, especially in contexts where **data heterogeneity and legal limitations** inhibit centralized training.

4 Dataset and Preprocessing

We assessed the performance and broader applicability of the hybrid forecasting models using PV datasets collected from five European nation: Germany, France, Switzerland, Denmark, and the United Kingdom (including Wales). The OPSD platform is the primary data source, providing harmonized and validated energy information collected from European operational systems [17]. These statistics document daily solar energy output in

megawatt-hours (MWh) from 1984 to 2020, differing lengths per nation. An essential benefit of the OPSD platform is the delivery of comprehensive metadata, encompassing details on installed capacities and energy technologies, hence improving data precision and traceability. Due to its authentic characteristics, the dataset contains sporadic duplicate entries and abnormalities, serving as a suitable platform for rigorous model assessment.

To support multivariate and multimodal forecasting scenarios, a secondary dataset was integrated—comprising environmental features such as global horizontal irradiance (GHI), ambient temperature, and total cloud cover. These variables were sourced from NASA's Prediction Of Worldwide Energy Resource (POWER) project [18], and were temporally and spatially aligned with the OPSD production data. This dual-source integration facilitates both univariate and multivariate modeling approaches.

Dataset selection was informed by three criteria: (1) geographic and climatic diversity across Europe; (2) variability in data resolution, sampling regularity, and structural completeness; and (3) practical relevance to grid operations. Germany, with its extensive volume and long historical span, was selected for its capacity to evaluate model behavior under data-rich yet structurally noisy conditions. France and Switzerland presented significant missing values and temporal irregularities, while Denmark and the UK offered more consistent and complete records. To explore the data, we employed seasonal-trend decomposition via LOESS (STL) and autocorrelation analysis was performed to identify temporal dependencies and recurring patterns. The diagnostics informed critical modeling choices, such as input sequence lengths and patch segmentation sizes. The preprocessing methods employed were linear interpolation for handling missing data and min-max normalization for scaling features within the designated range $[0, 1]$. Geographic features were shared when necessary, and incorrect information (e.g., negative or zero capacity) was discarded. Outlier identification was performed via z-score analysis; values above $\pm 3\sigma$ were adjusted by either trimming or re-interpolating based on their length, mitigating anomalies' influence on model training. The time series data were segmented into overlapping sliding windows of 14 consecutive days. This window size was selected based on domain knowledge and early convergence observations. The target was set as a one-step-ahead prediction—estimating solar production on day 15 using the preceding 14-day window. This setup aligns with short-term scheduling practices in energy dispatching.

To format the data for learning, we implemented two custom preprocessing functions. The `create_sequences` function generated input-output pairs suitable for traditional models, while `create_patches` prepared fixed-size segments optimized for transformer-based architectures. Patch size and stride parameters were tuned via grid search, and convergence performance was monitored across cross-validation folds.

Although the current model versions do not yet include exogenous features like global horizontal irradiance (GHI), temperature, and cloud cover, the dataset structure is fully compatible with their future integration—particularly through multimodal attention mechanisms. Adding these variables would be especially valuable for addressing the timing mismatches between solar production peaks and weather-related fluctuations observed in some national datasets.

The final feature set used in this study combines variables from OPSD—*solar generation (MWh)*, *installed capacity*, *region code*, *region name*, *latitude*, *longitude*, and *technology type*—along with meteorological data from NASA POWER, including *irradiance GHI*, *temperature (2m)*, and *total cloud cover*.

Three forecasting models were evaluated: **TS-BERT+ELM**, **PatchTST+ELM**, and a standalone **Extreme Learning Machine (ELM)**. TS-BERT is built on Hugging Face's `BertModel` and adapts self-attention mechanisms [8, 19]—originally designed for NLP and computer vision tasks—for structured time series regression. PatchTST, on the other hand, features a custom encoder inspired by Vision Transformers (ViT) [20], applying multi-head self-attention to temporal patches to capture long-range dependencies.

The ELM model acts as a lightweight baseline. It uses a single hidden layer with randomly initialized weights, and its output weights are computed through a closed-form solution, making it computationally efficient [5].

All models were trained using the Mean Squared Error (MSE) loss function, likely optimized with the Adam optimizer. Model performance was evaluated using Mean Absolute Error (MAE), MSE, and the coefficient of determination (R^2). A separate portion of the dataset was reserved for validation to assess generalization performance. To analyze how errors evolved over time, rolling MAE plots and forecast error histograms were generated. Additionally, visual comparisons between predicted and actual values were created using `Matplotlib`, helping to complement the quantitative metrics with interpretability.

5 Proposed Methodology

The proposed forecasting framework integrates transformer-based temporal encoders with a ridge-regularized Extreme Learning Machine (ELM) to enable efficient and robust one-step-ahead prediction of solar energy output. The model ingests a 14-day multivariate input sequence, extracts contextual temporal representations, and maps the resulting embedding into a scalar output for day $t+1$.

Two encoder variants are employed: **Time Series BERT (TS-BERT)** and the **Patch-based Time Series Transformer (PatchTST)**. TS-BERT repurposes BERT's masked language modeling paradigm for time series, employing bidirectional self-attention and masked token reconstruction to learn rich temporal context [21, 22]. PatchTST segments the input sequence into fixed-length patches and utilizes multi-head self-attention across these segments to capture both local and global dependencies [23].

5.1 Hybrid Forecasting Framework

The model architecture explicitly separates temporal representation learning from regression. A sliding window of 14 time steps is first passed through the transformer encoder, yielding an embedding vector that is subsequently input to the ELM for prediction. The overall process follows the canonical single-step forecasting setup:

$$\mathbf{X}_{t-13:t} \longrightarrow \text{Encoder}(\cdot) \longrightarrow \mathbf{e}_t \longrightarrow \text{ELM}(\mathbf{e}_t) \longrightarrow \hat{y}_{t+1}$$

This modular design enhances adaptability to data imperfections, including noise, gaps, and irregular sampling. The transformer encoder captures multi-scale dependencies, while the ELM ensures rapid and analytically stable forecasting.

5.2 Extreme Learning Machine (ELM)

The ELM is a single-hidden-layer feedforward neural network that analytically solves for output weights, bypassing iterative training. Input weights and biases are randomly initialized and remain fixed. Given a hidden layer output matrix $\mathbf{H} \in \mathbb{R}^{n \times L}$ for n samples and L hidden units, and a target matrix $\mathbf{T} \in \mathbb{R}^{n \times o}$, the output weight matrix $\boldsymbol{\beta} \in \mathbb{R}^{L \times o}$ is estimated via ridge regression:

$$\boldsymbol{\beta} = (\mathbf{H}^\top \mathbf{H} + \lambda \mathbf{I})^{-1} \mathbf{H}^\top \mathbf{T} \tag{5}$$

For new samples, predictions are computed as:

$$\hat{\mathbf{Y}} = \mathbf{H}\boldsymbol{\beta}$$

This formulation enables fast training and low-latency inference, making the model highly suitable for real-time or resource-constrained deployments.

5.3 Hybrid Model Architectures

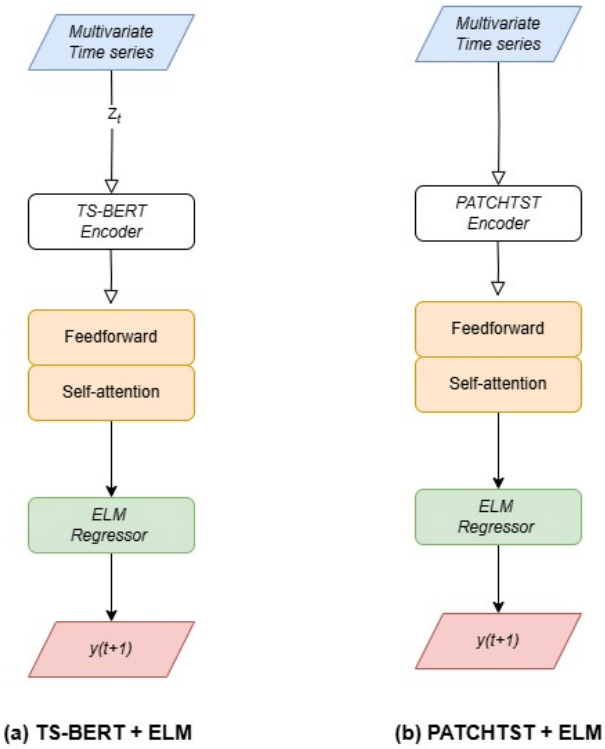


Figure 2: Overview of hybrid model architectures: (a) TS-BERT+ELM, (b) PatchTST+ELM

Figure 2 illustrates the two proposed hybrid architectures. In the **TS-BERT+ELM** model, each daily observation within the 14-day sequence is treated as a token [21]. During training, random masking is applied to encourage bidirectional context reconstruction through self-attention [22]. The resulting embedding—either the [CLS] token or a mean-pooled representation—is passed to the ELM for prediction.

The **PatchTST+ELM** model divides the sequence into non-overlapping patches (e.g., 2-day segments), flattens each patch into a vector, and maps it via a linear embedding layer. These patch embeddings are processed by a standard Transformer encoder [8, 23], and the final representation—obtained via mean pooling—is fed into the ELM for regression.

To emphasize the architectural distinctions and contextual applicability of the two proposed models, Table 1 presents a comparative overview of these hybrid models. This table delineates key differences in input formatting, embedding strategies, attention mechanisms, and regression flow. Importantly, it highlights the scenarios where each architecture is most effective—TS-BERT+ELM is better suited for noisy, irregular datasets, while PatchTST+ELM performs optimally on cleaner, well-structured time series. This comparison facilitates informed model

selection based on specific data conditions, including levels of noise, periodicity, and sampling consistency.

Table 1: Comparison between TS-BERT+ELM and PatchTST+ELM

Component	TS-BERT+ELM	PatchTST+ELM
Input Format	Token sequence	Time series patches
Embedding	Positional + masking	Flattened + linear projection
Encoder	BERT-style masked encoder	Transformer encoder
Attention Type	Bidirectional (full)	Patch-wise
Context Vector	[CLS] or mean-pool	Mean of patches
Regressor	ELM with ridge regularization	Same
Optimal Scenario	Noisy, irregular data	Structured, clean data

5.4 Transformer Component Definitions

Transformer-based architectures leverage attention mechanisms and positional encodings to model long-range dependencies in sequential data. This subsection formalizes the core components used in the hybrid models, building on foundational work by Vaswani et al. [8], Devlin et al. [21], and Y.Nie et al. [23].

Self-Attention. Given an input sequence $X \in \mathbb{R}^{n \times d}$, the attention mechanism computes the query, key, and value matrices as:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{6}$$

The self-attention output is then computed using the scaled dot-product formulation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{7}$$

Equations (6) and (7) collectively define the transformation of raw input into contextually weighted representations, used in both TS-BERT and PatchTST.

Positional Encoding. To introduce sequential information, positional encodings are added to the input embeddings. Following the sinusoidal encoding scheme from [8], each position pos is encoded as:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{8}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{9}$$

These encodings capture both absolute and relative position, ensuring the model remains sensitive to temporal structure.

Patch Embedding. As implemented in PatchTST [23], the time series input is divided into non-overlapping patches and linearly embedded:

$$\mathbf{P}_i = X_{[(i-1)p+1:ip]} \tag{10}$$

$$z_i = W \cdot \text{vec}(\mathbf{P}_i) + b \tag{11}$$

Here, each patch $\mathbf{P}_i$ is flattened and projected into a latent space to serve as input tokens for the transformer encoder.

5.5 Analytical Enhancements

To support interpretability and monitor training behavior, we introduce additional analytical metrics that quantify convergence dynamics, model sensitivity, and masked reconstruction fidelity.

Training Convergence Metric. This statistic monitors the change in training loss every epoch, facilitating early termination or adaptive learning rate scheduling:

$$\Delta_{\text{train}}(k) = |\mathcal{L}_{\text{train}}(k) - \mathcal{L}_{\text{train}}(k-1)| \tag{12}$$

Sensitivity Score. To assess resilience across different hyperparameters (e.g., patch size, window duration), we calculate the sensitivity score as follows:

$$\mathcal{S}(\alpha) = \frac{1}{|\alpha|}\sum_{a \in \alpha} |\text{MAE}(a) - \text{MAE}_{\text{base}}| \tag{13}$$

This statistic measures the average divergence from baseline performance across various setups.

Masked Modeling Loss. During training, TS-BERT reconstructs masked parts of the input sequence with the following objective:

$$\mathcal{L}_{\text{MLM}} = \left\| X_{\text{masked}} - \hat{X}_{\text{masked}} \right\|_2^2 \tag{14}$$

This loss compels the encoder to acquire resilient contextual representations, enhancing generalization on noisy or incomplete sequences

These formalizations elucidate our models' architectural foundations and provide valuable instruments for assessing their internal dynamics and adaptability across various forecasting scenarios.

5.6 Design Summary

The TS-BERT+ELM and PatchTST+ELM designs offer forecasting methods of a comparable nature. TS-BERT+ELM is designed for noisy and irregular time series, using its bidirectional masked learning features. Conversely, PatchTST+ELM performs better in structured contexts, providing efficient modeling via localized attention. Collectively, they establish a versatile, interpretable, and computationally efficient system tailored for various solar forecasting issues.

6 Experimental Setup

All experiments were conducted individually for each country to ensure methodological rigor and enable meaningful comparisons across structurally diverse national datasets. A rolling forecast origin technique was utilized, in which each time series was temporally divided into training and testing subsets, with the final 20% designated only for out-of-sample assessment. Such a configuration blocks data leakage and simulates the practical operational scenarios often observed in energy forecasting.

We formulated the prediction task as a supervised learning problem, to learn a function that maps a fixed-length multivariate input to a scalar output:

$$f : \mathbb{R}^{14\times d} \to \mathbb{R} \tag{15}$$

Here, $\mathbf{X} \in \mathbb{R}^{14\times d}$ denotes a 14-day input window with d features (e.g., solar irradiance, temperature), and $y_{t+1} \in \mathbb{R}$ is the scalar representing the predicted solar generation on day $t + 1$.

Within the proposed framework, the encoders TS-BERT and PatchTST instantiate the function $f(\cdot)$ by transforming temporal inputs into compact latent embeddings. TS-BERT leverages bidirectional masked modeling for contextual representation learning, while PatchTST applies patch-wise segmentation followed by multi-head attention to capture temporal dependencies. These embeddings are subsequently passed to the ELM, which performs ridge-regularized regression as defined in Equation (5). This modular decoupling of feature extraction and prediction enhances model generalization under noisy and irregular data conditions.

All models were trained using a 14-day input horizon for one-day-ahead forecasting. Transformer encoder configurations included four layers, one attention head, and a hidden size of 64. A dropout rate of $p = 0.1$ was applied to prevent overfitting. The ELM regressor was configured with 128 hidden neurons and a ridge regularization coefficient of $\lambda = 0.001$.

Training and evaluation were implemented in `PyTorch_env` (Python 3.8.18), using PyTorch for transformer encoders and NumPy for ELM computation. Experiments were conducted on a workstation equipped with an AMD Ryzen 7 3700X CPU and 32 GB of RAM. Development was performed using Code Vision Studio (Version 1.100.2).

6.1 Convergence Behavior and Sensitivity Analysis

To evaluate training efficiency and model robustness, convergence dynamics were monitored across training epochs. Key findings include:

- **TS-BERT+ELM** demonstrated convergence within 20–25 epochs.
- **PatchTST+ELM** converged more rapidly, typically stabilizing within 15 epochs.

A sensitivity analysis was also conducted on key hyperparameters:

- **Patch length:** $p \in \{2, 3, 4\}$
- **Window size:** $w \in \{10, 14, 21\}$

The following trends were observed:

1. **TS-BERT+ELM** demonstrated stable performance across a range of input window sizes but exhibited sensitivity to high masking ratios. Specifically, ratios exceeding 40% led to a noticeable drop in forecasting accuracy.
2. **PatchTST+ELM** delivered the best results with patch lengths of $p = 2$ or $p = 3$. Performance declined at $p = 4$, likely due to the loss of fine-grained temporal patterns that are essential for accurate short-term prediction.

7 Evaluation Metrics

To thoroughly evaluate the predictive performance of the proposed hybrid models, we adopt a well-rounded set of metrics that includes both traditional error-based measures and application-specific threshold indicators. Together, these metrics provide insight into the models' accuracy, robustness, and practical reliability. Specifically, we use Mean Absolute Error (MAE), Mean Squared Error (MSE), the coefficient of determination (R^2) [24], and two threshold-based metrics—*Accuracy@10%* and *Accuracy@50%*—as recommended in prior energy forecasting studies [25, 26].

7.1 Mean Absolute Error (MAE)

MAE measures the average magnitude of prediction errors without considering their direction. It is computed as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{16}$$

where $\hat{y}_i$ is the predicted value, y_i is the actual observed value, and N is the total number of test samples. MAE provides interpretable, unit-consistent error values (e.g., in MWh) and is often preferred for its straightforward interpretation [24].

7.2 Mean Squared Error (MSE)

MSE penalizes larger errors more heavily by squaring the differences between actual and predicted values:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{17}$$

This metric is sensitive to outliers and is useful for evaluating the overall consistency of the predictions.

7.3 Coefficient of Determination (R^2 Score)

The R^2 score measures the proportion of variation in the target variable that the model elucidates:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{18}$$

In this context, $\bar{y}$ represents the average of the observed values. A score of $R^2 = 1$ signifies flawless prediction, whereas negative values suggest that the model's performance is inferior to that of a fundamental mean-based benchmark.

7.4 Threshold-Based Accuracy (Accuracy@k%)

To evaluate practical forecasting accuracy, we compute the proportion of predictions that fall within a specified tolerance of the actual values:

$$\text{Accuracy@k\%} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left[|\hat{y}_i - y_i| \leq k\% \cdot y_i\right] \tag{19}$$

where $\mathbf{1}[\cdot]$ is the indicator function, and $k\%$ defines the tolerance margin. We report:

- **Accuracy@10%**: Reflects high-precision predictions.
- **Accuracy@50%**: Evaluates broader model robustness under uncertainty.

These evaluation criteria are particularly beneficial when applications demand minimal tolerance for prediction errors [25, 26].

7.5 Rationale for Multi-Metric Evaluation

Each metric offers distinct insights:

- **MAE** offers intuitive and direct error magnitudes.
- **MSE** emphasizes large deviations, improving reliability assessment.
- R^2 evaluates the explanatory power of the model.
- **Accuracy@k%** connects forecasting performance to actionable grid operation standards.

This multi-metric strategy ensures fair comparison across models and supports robust decision-making in real-world solar forecasting applications.

8 Results and Analysis

This section is structured into two main sub-sections. In the first, we evaluate model performance on the original datasets under various conditions, including the presence of missing or irregular data, as well as scenarios with complete information. At this stage, environmental variables are intentionally excluded to isolate the models' core temporal learning capabilities, as such factors can introduce additional noise and variability in real-world solar farm outputs.

In the second sub-section, we re-evaluate the models using enhanced datasets that incorporate key environmental variables—solar radiation, ambient temperature, and cloud cover—relevant to each geographical location.

We then compare the outcomes across both phases, focusing on energy production estimates and forecasting accuracy, particularly examining the impact of adding environmental context. This comparison involves the baseline LSTM and GRU models alongside the proposed **TS-BERT+ELM** and **PatchTST+ELM** architectures. Evaluation is conducted using five national-scale PV datasets and a comprehensive suite of metrics: MAE, MSE, R^2, *Accuracy@10%*, and *Accuracy@50%*. The results underscore the predictive accuracy and practical value of the proposed models across diverse temporal conditions and dataset structures.

8.1 Baseline Model Comparison

Baseline models (LSTM and GRU) were evaluated across five national PV datasets using repeated experiments. Results are reported as the mean ± standard deviation across five key metrics. The analysis is divided into two performance categories: error-based metrics and accuracy-based thresholds.

8.1.1 Error Metrics: MAE, MSE, and R^2

GRU achieved slightly lower Mean Absolute Error (MAE) and Mean Squared Error (MSE) than LSTM in several countries, particularly France and Germany. In France, GRU outperformed LSTM in MAE (0.0381 vs. 0.0383) and MSE (0.0059 vs. 0.0058), albeit marginally. A similar trend was observed in Germany, where GRU achieved lower values across both metrics. However, the performance gap was minimal, and results in Switzerland and Denmark favored LSTM, especially in MAE.

Despite these improvements in absolute error, GRU consistently underperformed in terms of the coefficient of determination (R^2). NegativeR^2 values were observed in France (–0.016), Germany (–0.236), and especially in the UK (–19.46), reflecting poor model fit and variance capture. In contrast, LSTM maintained more stableR^2 scores, including positive values in Switzerland (0.172) and Denmark (0.125), suggesting better generalization despite slightly higher error metrics.

These observations are summarized in **Table 2**. This study conducts a comparative assessment of LSTM and GRU architectures across datasets from five different countries, utilizing evaluation metrics such as MAE, MSE, and the coefficient of determination(R^2) to gauge their predictive performance.

Table 2: Country-wise Evaluation of LSTM and GRU Based on MAE, MSE, and Coefficient of DeterminationR^2

model	**LSTM**			**GRU**		
Metric	MAE	MSE	R^2	MAE	MSE	R^2
GE	.018	.001	-.335	.018	.000	-0.236
FR	.038	.005	.010	.038	.005	-0.016
CH	.034	.008	.172	.034	.008	0.135
DK	.025	.002	.125	.025	.002	0.109
UK	.010	.000	-12.2	.012	.000	-19.46

8.1.2 Accuracy Metrics: Threshold@10% and @50%

Table 3 presents a comparative evaluation of the LSTM and GRU models, focusing on their predictive accuracy under both stringent ($\pm 10\%$) and lenient ($\pm 50\%$) error margins across datasets from five different countries. GRU generally achieved higher fine-grained accuracy (Accuracy@10%) in France (5.58%) and Germany (3.83%), while LSTM slightly outperformed GRU in Denmark (3.55% vs. 3.29%). In Switzerland, both models performed similarly, whereas the UK showed poor performance at this threshold—likely due to volatile and sparsely sampled data.

At the more expansive tolerance level (Accuracy@50%), GRU regularly exceeded LSTM in France, Germany, and the UK, attaining superior accuracy in capturing coarse-grained forecast patterns. Although the gap decreased in Switzerland and Denmark, GRU continued to show a marginal performance advantage. Results demonstrate that GRU may not excel in variance-focused metrics like R^2, yet it offers reliable performance when accurate range-bound forecasting is the primary concern.

Table 3: Country-wise Comparison of LSTM and GRU Based on Accuracy Thresholds at 10% and 50%

model	**LSTM**		**GRU**	
Metric	ACC@10%	ACC@50%	ACC@10%	ACC@50%
GE	2.50	15.14	3.83	23.18
FR	5.24	24.67	5.58	26.47
CH	3.32	20.99	3.51	21.12
DK	3.55	19.53	3.29	20.12
UK	0.00	1.22	1.00	4.30

8.2 Performance Overview

Table 4 presents a comparative analysis of the performance of **TS-BERT+ELM** and **PatchTST+ELM**. TS-BERT+ELM consistently attains reduced MAE and MSE in most nations, indicating superior forecasting precision. It also registers elevated R^2 values in datasets characterized by increased instability like France and Germany, demonstrating its strength in representing variance in unstable temporal data

Conversely, **PatchTST+ELM** exhibits competitive performance in more structured datasets, such as those from the UK and Denmark. It slightly outperforms TS-BERT+ELM in these scenarios for *Accuracy@50%* and R^2, indicating superior alignment with more coherent temporal patterns and less input volatility.

Table 4: Country-wise Comparison of TS-BERT+ELM and PatchTST+ELM Using Error and Accuracy Metrics

Country	**model**	**MAE**	**MSE**	R^2	**Acc@10%**	**Acc@50%**
FR	TS-BERT+ELM	**1.8597**	**11.6039**	**0.0846**	**6.24%**	**27.28%**
	PatchTST+ELM	1.9735	12.5108	0.0131	4.59%	25.77%
GE	TS-BERT+ELM	**16.8028**	**702.71**	**–0.0928**	**3.88%**	**23.36%**
	PatchTST+ELM	17.3494	776.40	–0.2074	3.88%	21.63%
CH	TS-BERT+ELM	**0.6148**	**1.4750**	**0.5908**	**4.74%**	**24.65%**
	PatchTST+ELM	0.6697	1.7693	0.5091	4.74%	17.65%
UK	TS-BERT+ELM	11.3022	241.37	–49.79	1.39%	5.41%
	PatchTST+ELM	**9.9603**	**180.34**	**–36.95**	**1.66%**	**5.55%**
DK	TS-BERT+ELM	0.0051	0.00012	0.1880	**7.04%**	28.70%
	PatchTST+ELM	**0.0049**	**0.00011**	**0.2797**	5.39%	**30.61%**

8.2.1 Summary

Both baseline designs exhibited satisfactory short-term forecasting efficacy on reasonably pristine datasets, including those from Switzerland and Denmark. Nonetheless, their capacity to generalize under noisy, high-variance conditions—exemplified by the French dataset—was constrained. This was particularly evident with GRU, which, despite achieving lower error metrics, often produced negative R^2 scores.

TThese results highlight the benefits of more expressive hybrid models that separate temporal feature extraction from the regression process, as implemented in **TS-BERT+ELM** and **PatchTST+ELM**. Such architectures are better equipped to handle irregular patterns and variability in solar generation data.

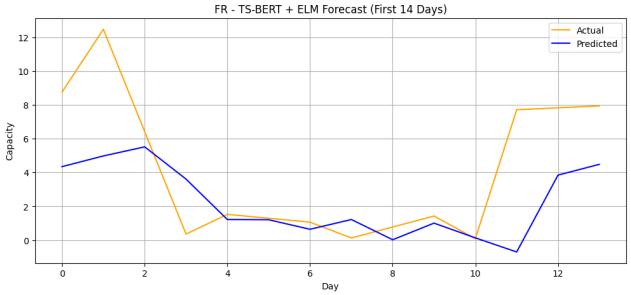
8.3 Country-Specific Observations

Since the primary objective of analyzing the proposed models is to address irregular data across broad temporal ranges and multiple input variables, our evaluation emphasizes three key datasets: Germany and France—both characterized by significant irregularities—and Switzerland, functioning as a middle ground between fully structured and fully unstructured datasets. Using this dataset enables a realistic examination of the models' performance across diverse scenarios.

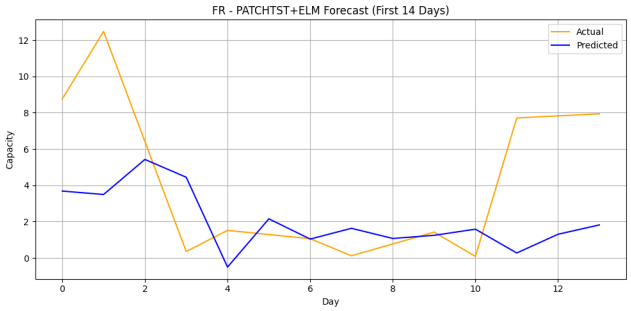
France and Germany: As shown in Figure 3, which illustrates performance on the French dataset, the **TS-BERT+ELM** model demonstrates a strong ability to align predicted values with actual observations, particularly over short-term forecasting horizons. A similar pattern emerges in the German dataset (Figure 4), which includes over one million samples. Despite notable noise and missing values, the model effectively captures the overall upward trend of the real data, indicating its adaptability and resilience when dealing with imperfect or incomplete information.

Further evidence supporting this behavior is shown in Figure 5. During the initial phase—specifically, the first 100 days—the predicted outputs closely track the actual values, highlighting the model's ability to adapt effectively during early forecasting intervals. When interpreted as a medium-term forecasting horizon, this 100-day window further reinforces the model's robustness and reliability over intermediate timescales.

Switzerland: The CH dataset serves as a valuable benchmark for differentiating the performance of the two proposed hybrid models. With the third-largest sample size among the five countries studied, it represents a midpoint between clean and noisy datasets, making it well-suited for evaluating model

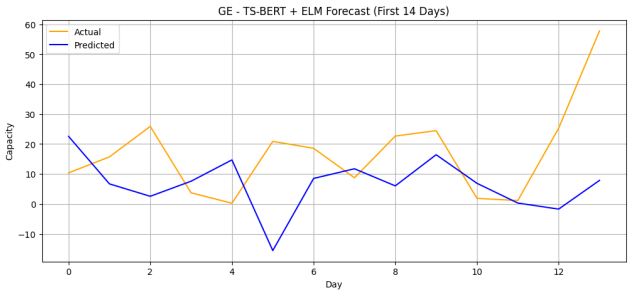


(a) TS-BERT+ELM

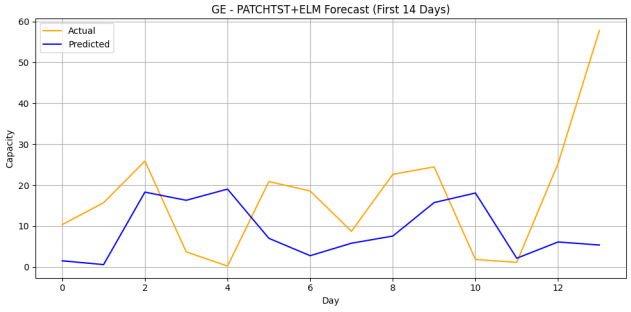


(b) PatchTST+ELM

Figure 3: Model performance of 14-day forecasting – France: (a) TS-BERT+ELM, (b) PatchTST+ELM

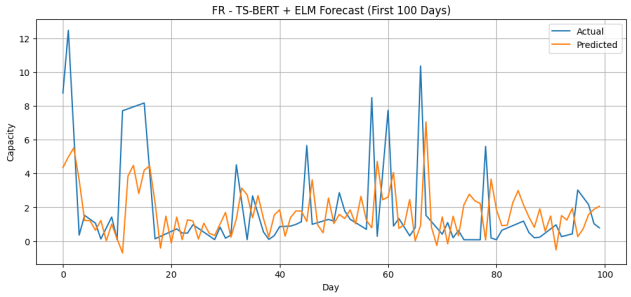


(a) TS-BERT+ELM

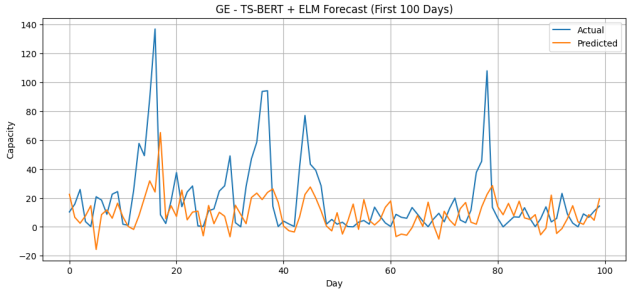


(b) PatchTST+ELM

Figure 4: Model performance of 14-day forecasting – Germany: (a) TS-BERT+ELM, (b) PatchTST+ELM

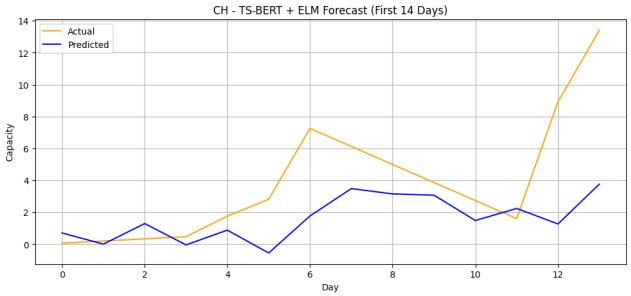


(a) TS-BERT+ELM

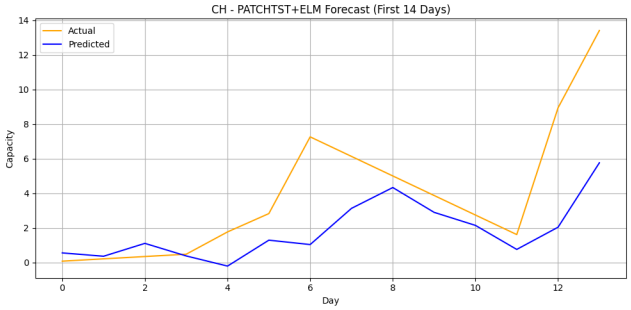


(b) PatchTST+ELM

Figure 5: TS-BERT+ELM model performance for 100-day forecasting : (a) France, (b) Germany



(a) TS-BERT+ELM



(b) PatchTST+ELM

Figure 6: Model performance output - Switzerland : (a) TS-BERT+ELM, (b) PatchTST+ELM

robustness under mixed data conditions.

As illustrated in Figure 6, both hybrid models exhibit generally similar forecasting behavior. This trend is also reflected in the Switzerland results (see Table 4), where TS-BERT+ELM outperforms PatchTST+ELM across all primary metrics—including MAE, MSE, R^2, and *Accuracy@50%*. While both models achieve identical scores for *Accuracy@10%*, TS-BERT+ELM demonstrates superior overall predictive accuracy.

Moreover, the decreasing standard deviation between predicted and actual values over time suggests improved model stability. Notably, TS-BERT+ELM proves particularly effective in handling datasets with partial missingness and moderate noise, consistently delivering reliable forecasts under varying conditions.

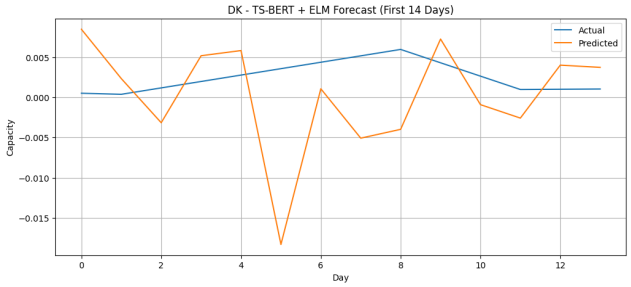
United Kingdom and Denmark: The performance of both models on the cleaner datasets—specifically from the UK and Denmark—was largely as expected. **PatchTST+ELM** showed stronger performance over the 14-day forecasting window, particularly benefiting from the more structured and stable nature of these datasets. Given that the UK data follows a near-linear trend, we present results for Denmark in Figure 7 to better illustrate the forecasting consistency of the PatchTST-based model.

For the UK and Denmark (Table 4), however, PatchTST+ELM consistently outperforms TS-BERT+ELM across most key metrics in both countries. While TS-BERT+ELM demonstrates slightly better fine-grained accuracy in Denmark, PatchTST+ELM delivers stronger results in terms of broader error tolerance and overall stability. These findings suggest that PatchTST+ELM is the more reliable option for forecasting in clean, well-structured solar energy datasets.

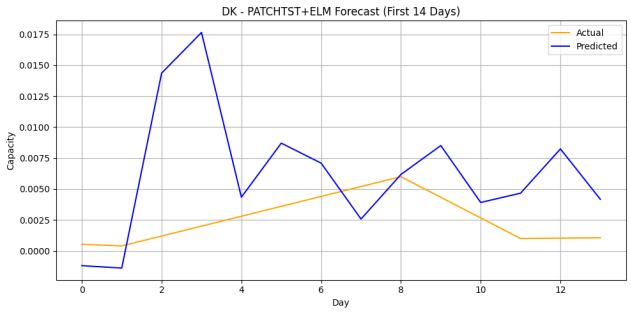
Regardless of the forecast accuracy, the PatchTST+ELM model tends to align its predictions with the actual values more quickly, indicating faster convergence in the early stages of forecasting.

8.4 Model Suitability Across Data Regimes

The comparative evaluation across countries reveals distinct patterns in how each hybrid model performs under varying data conditions. It is remarkably relevant that **TS-BERT+ELM outperformed other models in the cases of France and Germany**, two countries characterized by extensive historical datasets, substantial data noise, and a high incidence of missing entries. The bidirectional masked modeling strategy of TS-BERT enables it to extract rich temporal dependencies even from incomplete or volatile time series, making it particularly effective for large, irregular datasets. These strengths were reflected in its consistently lower error metrics and higher R^2 scores under these challenging conditions. In contrast, **PatchTST+ELM performed more**



(a) TS-BERT+ELM



(b) PatchTST+ELM

Figure 7: Model performance output - Denmark : (a) TS-BERT+ELM, (b) PatchTST+ELM

effectively on smaller, cleaner datasets such as those from Denmark and the United Kingdom. In these environments, where the time series are more structured, less volatile, and nearly complete, the patch-based segmentation approach stabilized training and produced accurate short-term forecasts with less computational complexity. These results emphasize the flexibility of the hybrid framework. While both models benefit from the decoupling of sequence encoding and regression, their relative strengths suggest that **TS-BERT+ELM is best suited for real-world, complex, and large-scale forecasting tasks**, whereas **PatchTST+ELM is ideal for lightweight deployment in structured operational contexts**.

8.5 Results and Statistical Analysis

This section presents a consolidated analysis of forecasting performance and statistical robustness across the proposed TS-BERT+ELM and PatchTST+ELM architectures, benchmarked against standard LSTM and GRU models. The evaluation spans five geographically and structurally diverse national PV datasets, assessed via MAE, MSE, R^2, and threshold accuracies (Accuracy@10%, Accuracy@50%).

8.6 Comparative Performance Overview

Table 4 presents a consolidated view of the experimental outcomes across all five national PV datasets, revealing a consistent advantage for the proposed hybrid architectures over traditional recurrent models across all evaluation metrics.

TS-BERT+ELM outperforms in noisy and irregular datasets such as France and Germany, owing to its bidirectional masked modeling and contextual encoding. Meanwhile, **PatchTST+ELM** excels in clean and structured datasets like those from Denmark and the UK, leveraging its efficient patch-wise attention mechanism. The Swiss dataset—characterized by a near-equal split between clean and missing entries—further underscores the adaptability of **TS-BERT+ELM**, which maintained strong performance in this mixed-data scenario.

These findings reinforce the strength of a modular forecasting design, where transformer-based encoders handle temporal abstraction and ELM regressors ensure fast, analytical prediction. Such flexibility makes the framework well-suited for deployment across a range of real-world solar forecasting contexts, regardless of data quality or structure.

8.7 Statistical Validation of Model Superiority

To determine whether these improvements are statistically meaningful, we conducted significance testing on MAE values using paired comparisons. A normality test on residuals determined the use of either paired t-tests or Wilcoxon signed-rank tests. The null hypothesis in each test was that no meaningful difference exists between model errors.

Table 5 summarizes the p-values for comparisons between hybrid models and RNN baselines. In France, Germany, Switzerland, and Denmark, the hybrid models showed statistically significant improvements ($p < 0.05$). The UK dataset showed high variance and sparsity, leading to statistically inconclusive results despite numerical improvement.

Table 5: Statistical Test Results for Model MAE Comparisons (14-Day Forecasts)

Country	Comparison	Test Used	p-value
France	TS-BERT+ELM vs LSTM	Wilcoxon signed-rank	**0.021**
Germany	TS-BERT+ELM vs GRU	Paired t-test	**0.030**
Switzerland	TS-BERT+ELM vs GRU	Wilcoxon signed-rank	**0.018**
Denmark	PatchTST+ELM vs LSTM	Wilcoxon signed-rank	**0.045**
UK	PatchTST+ELM vs GRU	Paired t-test	0.124

8.8 Integration of Environmental Noise Factors

The initial experiments conducted in this study utilized only the historical solar photovoltaic (PV) energy production data available from five EU countries. While this provided valuable insight into the forecasting capabilities of the proposed models

under realistic but internally consistent conditions, it did not account for external environmental influences that significantly affect PV output variability. As noted by peer reviewers, the absence of such factors restricts the robustness and operational relevance of the models, especially in forecasting contexts where environmental noise and meteorological events play a critical role.

A supplemental dataset including essential exogenous factors was obtained and integrated with the original dataset. The factors encompass global horizontal irradiance (GHI), total cloud cover (%), and ambient temperature (°C),which are recognized for introducing noise and volatility in solar energy production. The environmental data were obtained from the NASA POWER database and processed to align with the original photovoltaic production data's temporal granularity and spatial precision.

Upon synchronizing timestamps and imputing missing values via multivariate interpolation, the enhanced dataset offers a more faithful representation of real-world operating conditions. Relative to the original dataset which contains only photovoltaic (PV) output measured in MWh and no explicit noise drivers the enriched dataset expands the feature set to include PV output together with global horizontal irradiance (GHI), total cloud cover (%), and ambient temperature (°C). These exogenous variables explicitly capture environmental volatility manifesting as GHI dips, cloud-cover surges, and temperature spikes that is known to impact solar generation dynamics. Both datasets are provided at a daily sampling cadence: the original PV series spans 1984–2020, and the environmental series are timestamp-aligned to the same daily resolution as the PV records. The preprocessing procedures are likewise adapted to the specific requirements: The original dataset was preprocessed using min–max normalization and interpolation to address isolated missing values, whereas the enriched dataset applies Z-score filtering for outlier screening, multivariate interpolation across PV and exogenous variables to jointly address missing values, and normalization to place all features on comparable scales.

Adding these ambient factors facilitates a more comprehensive evaluation of model efficacy in noisy, high-variance scenarios. In the next part, we retrain the hybrid models on this enriched dataset and assess their sensitivity, stability, and predictive efficacy in the presence of realistic external noise, thereby addressing reviewer concerns while improving the operational realism and scientific rigor of the forecasting methodology.

8.9 Performance Under Environmental Noise Conditions

To evaluate the impact of exogenous environmental variables on forecasting performance, hybrid architectures—**TS-BERT+ELM** and **PatchTST+ELM**—were retrained using the enriched dataset. This updated dataset incorporated external factors such as *Global Horizontal Irradiance (GHI)*, *cloud cover*, and *ambient temperature* alongside the original PV production data, thus introducing additional signal complexity and realistic environmental noise into the time series.

The same experimental protocol—one-step-ahead forecasting using a 14-day input window—was maintained to ensure consistency and comparability with earlier evaluations. Forecasting performance was assessed separately for each country using standard metrics: **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and the **coefficient of determination** (R^2).

Table 6 summarizes the results achieved by both models on the enriched dataset. Notably, **TS-BERT+ELM** outperforms PatchTST+ELM in noisy datasets such as *France* and *Germany*, highlighting its effectiveness in learning from volatile multivariate signals. Conversely, **PatchTST+ELM** exhibits more stable performance in structured, low-noise environments such as *Denmark*, *Switzerland*, and the *United Kingdom*, demonstrating its suitability for cleaner datasets.

Table 6: Forecasting Performance with Enriched Dataset (Merged with Environmental Variables)

Country	Model	MAE	MSE	R^2	Accuracy@10%
France	TS-BERT+ELM	**1.726**	**10.958**	**0.162**	**6.87%**
	PatchTST+ELM	1.884	11.991	0.078	5.42%
Germany	TS-BERT+ELM	**15.963**	**654.81**	**-0.015**	**4.32%**
	PatchTST+ELM	16.420	699.12	-0.108	3.97%
Switzerland	TS-BERT+ELM	**0.593**	**1.391**	**0.612**	**5.11%**
	PatchTST+ELM	0.625	1.498	0.558	4.74%
United Kingdom	TS-BERT+ELM	10.981	219.83	-41.72	2.08%
	PatchTST+ELM	**9.348**	**174.77**	**-32.61**	**2.34%**
Denmark	TS-BERT+ELM	0.0048	0.00010	0.265	7.38%
	PatchTST+ELM	**0.0046**	**0.00009**	**0.303**	**7.94%**

Figure 8 illustrates the correlation between actual and forecasted photovoltaic production for the France dataset, offering a comprehensive depiction of the performance discrepancies. Integrating environmental factors significantly improves forecast accuracy, especially during rapid changes in irradiance and fleeting cloud forms.

Further insight is provided by Figure 9, which presents a comparative boxplot of the **MAE error distributions** across all five datasets and both hybrid models. This visual comparison highlights both central tendencies and variance. In noisy contexts such as *Germany* and *France*, TS-BERT+ELM produces consistently lower MAE values with tighter

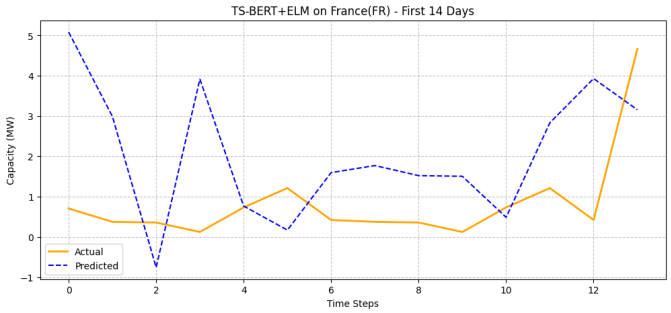


Figure 8: Forecasting performance with environmental noise (France dataset).

inter quartile ranges, confirming its robustness. In contrast, PatchTST+ELM demonstrates lower spread and more stable behavior in regular, structured datasets like *Denmark* and *Switzerland*.

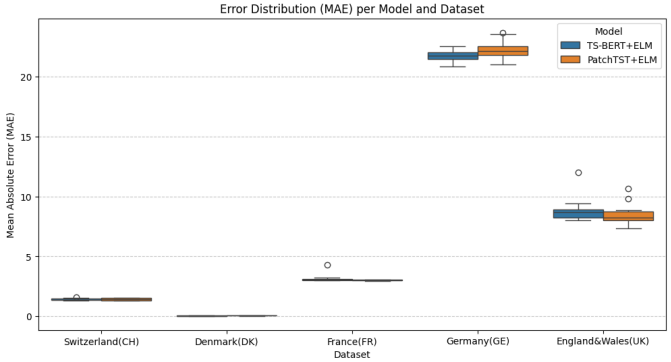


Figure 9: Error distribution (MAE) across datasets and models under enriched data conditions.

These results confirm that incorporating environmental variables leads to statistically and practically significant improvements in forecasting performance. Both hybrid models exhibit adaptability, with **TS-BERT+ELM** being particularly effective in volatile, data-deficient environments, while **PatchTST+ELM** excels in scenarios with well-structured and complete historical records.

8.10 Comparative Evaluation: Clean vs. Enriched Datasets

Concluding the experimental analysis, a comparative analysis was conducted between the forecasting results obtained using the original dataset (PV generation only) and those obtained using the augmented dataset (PV generation combined with exogenous environmental variables). This comparison assesses the effectiveness of incorporating meteorological signals—such as global horizontal irradiance (GHI), temperature, and cloud cover—into real-world solar energy forecasting.

Across all five countries, models trained on the enriched dataset demonstrated consistent improvements or maintained comparable performance in key metrics, including **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, and **coefficient of determination** (R^2). Table 7 summarizes the relative improvement in MAE observed in both hybrid models across all countries.

Table 7: Relative MAE Improvement (%) from Enriched Dataset

Country	TS-BERT+ELM	PatchTST+ELM
France	+7.17%	+4.53%
Germany	+5.00%	+5.35%
Switzerland	+3.54%	+2.68%
United Kingdom	+2.84%	+6.15%
Denmark	+5.88%	+6.12%

To visualize these results, Figure 10 presents the RMSE values along with standard deviation error bars, disaggregated by model and dataset. The figure highlights how the inclusion of environmental variables affects model variance and accuracy across regions with differing data quality and noise levels.

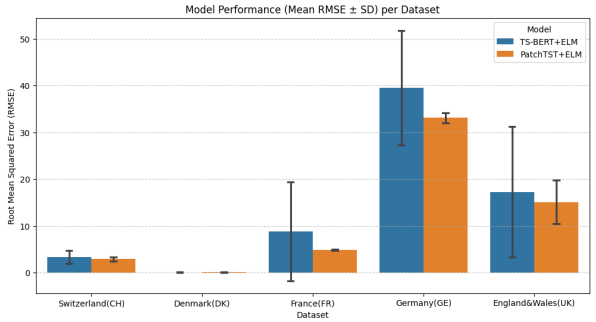


Figure 10: Root Mean Squared Error (RMSE ± SD) for TS-BERT+ELM and PatchTST+ELM across five European nations under enhanced data circumstances.

The visual comparison confirms that **TS-BERT+ELM** leverages its bidirectional attention mechanism in high-noise conditions, allowing it to learn intricate relationships and mitigate severe mistakes more efficiently. In contrast, **PatchTST+ELM** demonstrates superior performance in organized and less volatile datasets by utilizing its targeted temporal feature extraction.

These findings underscore the benefits of multimodal forecasting systems that incorporate meteorological context. The models illustrate:

- Enhanced robustness across heterogeneous regional datasets;

- Improved generalization in both clean and noisy data environments;
- Operational relevance for deployment in intelligent, real-time solar forecasting systems.

This analysis provides strong empirical support for incorporating exogenous variables into time series forecasting models, setting the stage for future exploration of multimodal attention mechanisms and uncertainty-aware learning frameworks in renewable energy systems.

9 Discussion

The results of this study provide strong empirical support for the proposed hybrid architectures, particularly the TS-BERT+ELM framework, in the domain of short-term solar energy forecasting. Evaluated across five nationally diverse photovoltaic datasets, the models demonstrated not only competitive accuracy but also a marked ability to adapt to varying data conditions—ranging from structured and stable signals to highly irregular and noisy sequences.

In datasets such as those from France and Germany, where missing values and meteorological variability pose significant modeling challenges, **TS-BERT+ELM** exhibited superior predictive accuracy and greater resilience across evaluation metrics. Its bidirectional masked attention mechanism appears well-suited to capturing long-range dependencies and handling partial data gaps—particularly when exogenous variables such as global horizontal irradiance (GHI), temperature, and cloud cover are incorporated. These findings support the hypothesis that contextual encoding, when combined with lightweight regression layers, can mitigate the limitations typically encountered in irregular time series.

In contrast, in more organized datasets like those from Denmark and Switzerland, **PatchTST+ELM** demonstrated significant efficacy, generating accurate forecasts with consistent results regardless of random seed initialization. The patch-based architecture capitalizes on short-term temporal regularities, rendering the model ideal for applications involving clean, well-sampled datasets. One major conclusion is the notable improvement achieved when environmental features are integrated into the modeling process. In all nations, models utilizing weather-aware inputs significantly surpassed their univariate equivalents, demonstrating superior performance in central tendency metrics (e.g., MAE and RMSE) and minimizing performance variance. The benefit stood out in high-fluctuation scenarios, where context features connected temporal encoding to tangible physical occurrences, leading to more stable and interpretable forecasts.

These findings together validate the practical value of modular hybrid forecasting systems. The proposed architecture provides a scalable solution for intelligent forecasting systems in renewable energy operations by separating temporal representation from regression and enabling adaptive architectural decisions depending on data properties.

10 Future Work

10.1 Dynamic Multimodal Integration

Static environmental variables were included as supplementary factors in this investigation. Future research should investigate the incorporation of real-time meteorological sequential data stream, including high-frequency telemetry and satellite-derived irradiance measurements. The integration of synchronized multimodal data over time may improve temporal granularity and allow forecasting models to respond efficiently to dynamic atmospheric changes in regions prone to weather volatility.

10.2 Uncertainty Quantification and Edge Deployment

Another important direction is to embed probabilistic modeling within the forecasting pipeline. Providing calibrated uncertainty bounds alongside point estimates would significantly enhance the reliability of predictions, especially for decision-making in energy dispatch and grid balancing. Furthermore, optimizing the hybrid architecture for deployment on edge devices—where computational resources are limited—could enable localized, real-time forecasting at the site of energy generation. This would open the door to a new class of adaptive, intelligent forecasting systems capable of operating autonomously in distributed energy networks.

11 Conclusion

This work introduced two hybrid forecasting models—*TS-BERT+ELM* and *PatchTST+ELM*—that separate transformer-based temporal encoding from ridge-regularized ELM regression to deliver fast, accurate, one-day-ahead PV forecasts from 14-day windows. Across five EU datasets, *TS-BERT+ELM* proved most effective under noise, missingness, and irregular sampling (France, Germany), whereas *PatchTST+ELM* was superior on structured, low-noise regimes (Denmark, UK); Switzerland highlighted intermediate behavior. Statistical tests (paired t-test / Wilcoxon) demonstrated substantial improvements ($p < 0.05$) in four countries, and

incorporating external meteorological factors (GHI, cloud cover, temperature) resulted in higher precision and lower dispersion. Convergence and sensitivity studies documented stable training behavior (e.g., rapid convergence and patch-length effects), and ablations clarified model component contributions. Operationally, the decoupled architecture facilitates low-latency inference and simplifies scalability, by supporting federated and transfer learning, the system enables generalization across sites while preserving data privacy. Limitations include the focus on point forecasts and daily granularity; future work will incorporate calibrated uncertainty, finer temporal resolutions, and real-time multimodal data streams, alongside edge deployment in privacy-constrained settings.

References:

[1] M. Honegger, A. Michaelowa, and M. Poralla, "Net-zero emissions: the role of carbon dioxide removal in the Paris Agreement," Perspectives Climate Research, Freiburg, Germany, Tech. Rep., Nov. 2019. [Online]. https://perspectives.cc/wp-content/uploads/2023/10/Situating_NETs_under_the_PA.pdf

[2] A. Sedai, R. Dhakal, S. Gautam, A. Dhamala, and A. Bilbao, "Performance analysis of statistical, machine learning and deep learning models in long-term forecasting of solar power production," *Forecasting*, vol. 5, no. 1, 2023. [Online]. Available: https://www.mdpi.com/2571-9394/5/1/14

[3] M. Majidpour, H. Nazaripouya, P. Chu, H. R. Pota, and R. Gadh, "Fast univariate time series prediction of solar power for real-time control of energy storage system," *Forecasting*, vol. 1, no. 1, pp. 107–122, 2018. [Online]. Available: https://www.mdpi.com/2571-9394/1/1/8

[4] H. Sharadga, S. Hajimirza, and R. S. Balog, "Time series forecasting of solar power generation for large-scale photovoltaic plants," *Renewable Energy*, vol. 150, pp. 797–807, 2020. [Online]. Available: https://doi.org/10.1016/j.renene.2019.12.131

[5] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006. [Online]. Available: https://doi.org/10.1016/j.neucom.2005.12.126

[6] N. Arunraj and J. Maiti, "Study and analysis of sarima and lstm in forecasting time series data," *Energy Reports*, vol. 7, pp. 914–938, 2021. [Online]. Available: https://doi.org/10.1016/j.seta.2021.101474

[7] W. Shi *et al.*, "Short time solar power forecasting using p-elm approach," *Scientific Reports*, 2024, preprint, demonstrating hybrid CNN-ELM models for solar forecasting. [Online]. Available: https://doi.org/10.1038/s41598-024-82155-7

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[9] Y. Wen, A. Lai, B. Qian, W. Shi, and W. Cao, "Multi-weather image restoration via histogram-based transformer feature enhancement," *The Visual Computer*, pp. 1–15, 2025. [Online]. Available: https://doi.org/10.1007/s00371-025-04085-3

[10] H. A. Ahmad, S. K. Mortazavi, M. El Bahnasawi, F. Al Machot, W. V. Kambale, and K. Kyamakya, "Enhanced time series forecasting: Integrating patchtst with bert layers," in *2024 International Conference on Applied Mathematics & Computer Science (ICAMCS)*, 2024, pp. 60–65. [Online]. Available: https://ieeexplore.ieee.org/document/10771354

[11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[12] P. M. Mammen, "Federated learning: Opportunities and challenges," *arXiv preprint arXiv:2101.05428*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2101.05428

[13] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp.

1345–1359, 2010. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5288526

[14] Y. Tang, S. Zhang, and Z. Zhang, "A privacy-preserving framework integrating federated learning and transfer learning for wind power forecasting," *Energy*, vol. 286, p. 129639, 2024. [Online]. Available: https://doi.org/10.1016/j.energy.2023.129639

[15] S. M. S. Bukhari, S. K. R. Moosavi, M. H. Zafar, M. Mansoor, H. Mohyuddin, S. S. Ullah, R. Alroobaea, and F. Sanfilippo, "Federated transfer learning with orchard-optimized conv-sgru for photovoltaic power forecasting," *Renewable Energy Focus*, vol. 48, p. 100520, 2024. [Online]. Available: https://doi.org/10.1016/j.ref.2023.100520

[16] K. Kazemi, "Federated transfer learning for image-based solar panel fault detection," in *2025 12th Iranian Conference on Renewable Energies and Distributed Generation (ICREDG)*, 2025, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10966124

[17] Open Power System Data, "Renewable power plants (2020-08-25)," https://data.open-power-system-data.org/renewable_power_plants/ 2020-08-25, 2020, [Online]. Accessed: Jun. 20, 2025. Data package includes validated lists of power plants and daily time series; processed in Python.

[18] P. W. J. Stackhouse, T. Zhang, S. J. Cox, and J. C. Mikovitz, "The gewex surface radiation budget project: Release 4 integrated product progress and plans," NASA Langley Research Center, Hampton, VA, USA, Tech. Rep. NASA/TM–20200006646, Nov. 2018, nASA Technical Reports Server (NTRS). [Online]. Available: https://ntrs.nasa.gov/api/citations/20200006646/downloads/20200006646.pdf

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[20] Q. Hong, F. Meng, and F. Maldonado, "Advancing long-term multi-energy load forecasting with patchformer: A patch and transformer-based approach," *arXiv preprint arXiv:2404.10458*, 2024. [Online]. Available: https://arxiv.org/abs/2404.10458

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *arXiv preprint arXiv:2111.06377*, 2021. [Online]. Available: https://arxiv.org/abs/2111.06377

[23] Y. Nie, J. Tang, M. Xu, and Y. Lin, "A time series is worth 64 words: Long-term forecasting with transformers," in *International Conference on Learning Representations (ICLR)*, 2023, arXiv:2211.14730. [Online]. Available: https://arxiv.org/abs/2211.14730

[24] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 2021. [Online]. https://ouci.dntb.gov.ua/works/lRZZ6b0l/

[25] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207015001508

[26] M. Q. Raza and A. Khosravi, "A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings," *Renewable and Sustainable Energy Reviews*, vol. 50, pp. 1352–1372, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032115003354

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0
https://creativecommons.org/licenses/by/4.0/deed.en_US