Advances in Artificial Intelligence and Machine Learning; Research (Ahead of Print) Received 01-08-2026; Accepted 10-11-2026; Published 17-11-2027

# dCor-Transformer: Feature-Independent Attention for Load Forecasting

**Sabra ElFerchichi** seelferchichi@uj.edu.sa
*Department of Computer Science and Engineering*
*University of Jeddah*
*Jeddah, Saudi Arabia*

**Corresponding Author:** Sabra ElFerchichi

**Copyright** © 2026 Sabra ElFerchichi. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

Sustainable resource planning and modern power grids depend on accurate electricity demand forecasts. The underlying data present a modeling challenge due to: weather conditions, time-of-day patterns, consumption history, and market prices. They are simultaneously correlated and non-linear in their interactions. This study introduces the dCor-Transformer, a multivariate-to-univariate forecasting framework that combines a statistical feature preprocessing stage with a standard Transformer encoder. The preprocessing step draws on distance correlation theory to structure the input space prior to attention, with the goal of improving both computational efficiency and predictive focus. The framework is evaluated on two complementary benchmarks — the AEMO Victoria dataset for short-term load forecasting and the UCI Electricity dataset for long-term horizons — to assess its behavior across different forecasting regimes. Component-level analysis and multi-seed validation further reinforce the reliability and interpretability of the results. Findings show that dCor preprocessing yields the largest prediction gains on high-dimensional unstructured input like UCI Dataset, while in the case of well engineered feature sets like AEMO it achieves slightly lower than the Vanilla transformer accuracy while reducing training by 41%.

**Keywords:** Electricity demand forecasting, Distance correlation; Transformer, Feature preprocessing, Short/long time series forecasting.

## 1. INTRODUCTION

To maintain stability, smart grids operations are based on accurate electricity demand forecasting while handling generation scheduling and energy trading. These predictions enable grid operators to effectively manipulate the fluctuations and maintain the essential equilibrium between the transmission and distribution of networks[1], [2]. In fact accurate estimation of the variation in future electric load is of very important for both electric utility companies and consumers due to its application in the decision making and operation of the power grid. Consequently, forecasting electricity load still a challenging task. The relation between load variation and various influencing features such as ambient temperature, time-of-day, seasonality, holidays, and electricity prices is



**Citation:** Sabra ElFerchichi. dCor-Transformer: Feature-Independent Attention for Load Forecasting. **Advances in Artificial Intelligence and Machine Learning**. 2026. [Ahead of Print]. https://dx.doi.org/10.54364/AAIML.2026.66359

extremely complicated due to the stochastic and non-linear behavior of consumers and prices[3]. Many approaches have been explored to address this issue. Moving from classical statistical models, including ARIMA, which is effective only for stationary time series and struggles with nonlinear relationships [4], [5], to machine learning approaches like support vector regression (SVR), gradient boosting, and XGBoost [6], capable of capturing nonlinear patterns. However, their performance is mainly degraded in highly non-linear and non-stationary electricity load datasets [7]. Deep learning approaches, basically recurrent networks such as LSTM and GRU, demonstrate strong performance for short and medium time forecasting horizons [8]. Recently, models based on Transformer have been applied for time series forecasting such as Autoformer, Informer, PatchTST, and DLinear. They are using different mechanisms and techniques to either improve sequence modeling efficiency, forecasting accuracy or capture long-range dependencies [9], [10]. Although these models have succeeded, the preparation of the input features remains a significant bottleneck. Most current pipelines apply linear tools such as Pearson's correlation or PCA to filter and rank features [11], [12]. However, in power grids, the interaction between weather, prices and demand is complex and non-linear, which means these tools can not identify the most important predictors [13]. Consequently, Transformers may miss the deep and non-linear features that actually dictate energy patterns [14]. Distance Correlation (dCor) [15] can handle both limitations: it can capture linear and nonlinear statistical dependencies and it's equal to zero if and only if two variables are statistically independent. However, unlike mutual information estimators that are based on kernel density estimation, dCor does not require bandwidth selection and is easy to apply in practice [16]. Despite its theoretical advantages, dCor has not been explored as a feature pre-processing tool in deep learning-based load forecasting. Transformer-based methods such as Informer [9], Autoformer [10], PatchTST [11], and iTransformer [12] operate on raw or linearly normalized features, which means that the nonlinear dependency embedded in the input structure is unexplored. The main objective of this work is to introduce the dCor-Transformer in Multivariate to single context prediction. We apply dCor eigen-decomposition to project input features into a dCor-decorrelated space before training, enabling the attention mechanism to operate on well decorrelated, informational distinct signals. The rest of this paper is organized as follows: background and related work are reviewed in Section 2. In Section 3, we present the proposed methodology including data preparation and dCor decomposition. Section 4 illustrates the obtained experimental results and discussion, and finally the conclusion and perspectives are presented in Section 5.

## 2. BACKGROUND AND RELATED WORK

### 2.1 Distance Correlation Theory

Distance correlation (dCor) [15] is a multivariate measure of dependence between two random vectors X and Y of arbitrary (and potentially different) dimensions.

$$\mathrm{dCor}(X, Y) \in [0, 1] \quad \text{and} \quad \mathrm{dCor}(X, Y) = 0$$

if and only if X and Y are statistically independent [15] , [16]. This zero-equivalence property makes dCor a principled nonparametric alternative to mutual information without requiring density estimation or kernel bandwidth selection. The computation relies on pairwise Euclidean distance

matrices. For n observations of X and Y, the sample dCor is computed through doubly-centered distance matrices A and B, with computational complexity O($n^2$):

$$\mathrm{dCov}^2(X,Y) = \frac{1}{n^2}\sum_{i,j} A_{ij}B_{ij}$$

$$\mathrm{dCor}^2 = \frac{\mathrm{dCov}^2(X,Y)}{\sqrt{\mathrm{dVar}^2(X)\cdot \mathrm{dVar}^2(Y)}}$$

The spectral decomposition of the p×p dCor matrix

$$M = V\Lambda V^T$$

provides a rotation in the space of nonlinear statistical dependencies. This rotation ensures that the Transformer's attention mechanism receives clean signals that are already organized by their true predictive value. FIGURE 1 illustrates this distinction on the AEMO VIC1 feature set. The `rrp_ratio` (electricity price normalized by its 168h rolling mean) shows a Pearson correlation with demand of 0.07, indicating no linear dependence; while dCor measures 0.51, revealing substantial nonlinear dependency.

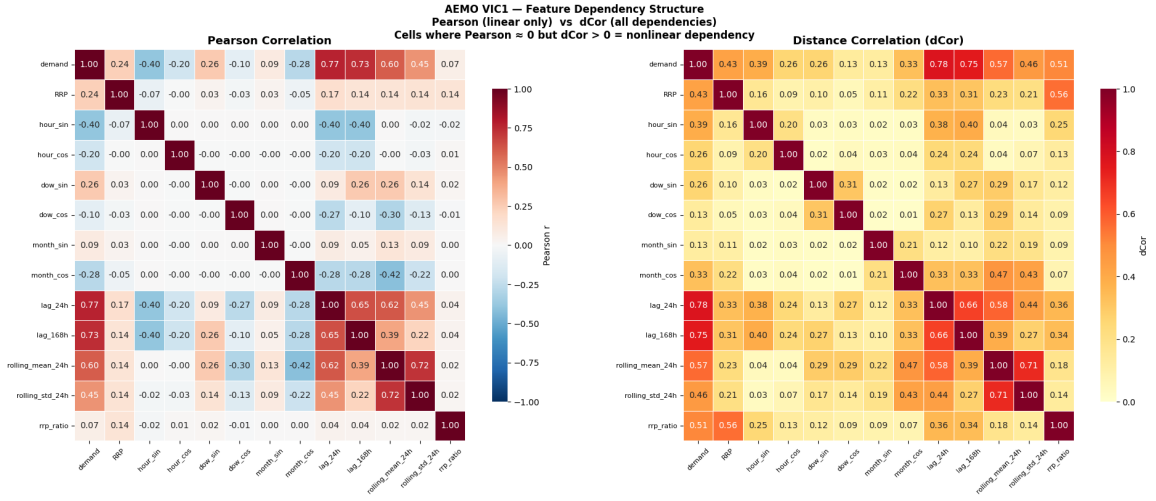



Figure 1: Feature dependency structure for AEMO VIC1.

## 2.2 Transformer Architectures for Time Series

The original Transformer architecture [17] was proposed for NLP, where self-attention is used instead of recurrence to perform parallel processing of sequences. Due to the different challenges of time series forecasting, several influential architectures have been proposed. While Informer [9] proposed ProbSparse attention for long sequences, Autoformer [10] incorporated decomposition-based architecture and auto-correlation mechanisms. Later, PatchTST [11] proposed to patch input sequences, allowing capture of local semantics and improved long-term forecasting. iTransformer [12] inverted the attention mechanism to operate across variables rather than time steps, achieving strong results on high-dimensional multivariate benchmarks, through architectural redesign rather

than additional complexity. Despite these improvements, Zeng et al [18] showed that a simple linear model (DLinear) can outperform complex Transformer variants on standard LTSF benchmarks. This suggests that architectural complexity alone is insufficient — what matters is whether the model receives informative, well-structured input. Our work addresses this gap: rather than modifying the Transformer architecture, we apply dCor-based feature preprocessing to ensure that attention operates on statistically independent, nonlinearly-selected inputs. While most recent architectures are evaluated under a Multivariate-to-Multivariate ($M \rightarrow M$) protocol predicting all variables simultaneously, our work adopts the Multivariate-to-Single ($M \rightarrow S$) protocol, predicting a single target variable that is electricity demand.

### 2.3 Short and Long Term Load Forecasting

First deep learning approaches to electricity load forecasting relied on recurrent architectures such as RNN and LSTM. Unfortunately, they suffer from vanishing gradients, limited parallelization, and poor scalability on long-range temporal relations. Transformer architectures solved these limitations using self-attention mechanisms that enable parallel sequence processing and capture long-range temporal dependencies [19]. For short-term forecasting, standalone models such as LSTM achieve MAPE around 2 to 4%, while hybrid frameworks can reduce this further [20] , [21]. For long-term forecasting, the UCI Electricity dataset has been widely used as a standard benchmark with Transformer-based models such as Informer [9] and Autoformer [10] commonly evaluated on it. To ensure fair comparison with these baselines as well as DLinear [18], we follow the standard $70/10/20$ temporal split under the Multivariate-to-Single protocol — details are provided in Section 3.

### 2.4 Feature Pre-processing for Deep Learning Forecasters

Many forecasting frameworks do not invest in feature pre-processing and simply apply raw features, MinMax normalization, or simple PCA without accounting for nonlinear dependencies. Although feature engineering significantly influences deep learning model accuracy in energy forecasting [7], the use of distance correlation as a pre−processing step remains an unexplored in the load forecasting literature. Our approach is most closely related to the N-Space Transformation of L'Heureux et al[22], which applies a learned linear projection before the Transformer encoder. However, their projection is learned by backpropagation with no statistical interpretation, and the projection dimension must be tuned as a hyperparameter. In contrast, in our proposed dCor-based approach, the projection matrix $V_k$ consists of the top-k eigenvectors of the distance correlation matrix M, ranked by nonlinear dependency strength. The projection is computed analytically from the training data before any gradient descent, adds zero trainable parameters to the model. The number of components k is determined by a variance-explained threshold rather than hyperparameter search. In fact, dependency-aware selection operates at two levels: the input feature configuration is determined via ablation study, selecting which variables to include based on their dCor scores with the target; and within each configuration, the projection retains the top-k eigenvectors of M ranked by nonlinear dependency strength. This two-level selection ensures that attention operates on a maximally informative, non-redundant input space.

# 3. METHODOLOGY

We illustrate the proposed dCor-Transformer pipeline in FIGURE 2. It consists of four stages:

1. **Stage1: Feature engineering and normalization:** data preparation.
2. **Stage2: dCor matrix computation and eigen-decomposition:** computing the dependence matrix $M = V\Lambda V^T$.
3. **Stage3: Transformer encoder training on projected features:** learning from the decomposed feature space.
4. **Stage4: Direct multi-step forecasting:** generating the final predictions.

All pre-processing parameters (scalers, dCor matrices, and eigenvectors) are calculated using only the training data and then applied to the validation and test sets.

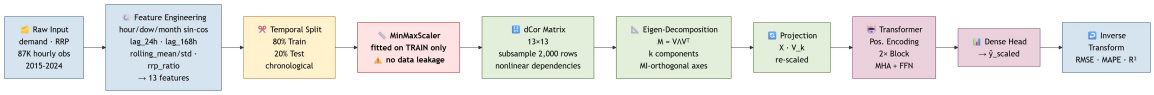



Figure 2: Proposed dCor-Transformer pipeline

## 3.1 Data Preparation and Feature Engineering

We evaluated the model on two benchmarks with distinct characteristics:

- AEMO VIC1: A 13 feature set including five grid variables (demand, price, weather), six cyclical time encodings, an autoregressive lag ($168h$), and a derived price-ratio.
- UCI Electricity: We utilize seven features: client demand and six cyclical time encoding. We found that adding lag features or rolling statistics actually degraded MSE by $34.3\%$. Since the model looks back $96$ hours, it already possesses sufficient historical context; adding redundant lags simply introduced noise that interfered with the dCor projection. We follow the standard Autoformer $M \rightarrow S$ benchmark protocol with a $70/10/20$ temporal split

## 3.2 dCor Eigen-Decomposition

Our approach focuses on replacing standard linear rotations with a Distance Correlation (dCor) projection. We compute a p×p distance correlation matrix M where each entry $M_{ij}$ represents the nonlinear dependency between features. We then perform an eigen-decomposition $M = V\Lambda V^T$ and project the original features onto the top-k eigenvectors. This creates a dCor-decorrelated rotation; the resulting components are ordered by their non linear dependency contribution, providing an interpretable inductive bias.
For the configuration ($k = p = 7$), the rotation is complete ($100\%$ variance explained in the

dCor sense), and the projection acts as a dependency-aware feature reordering. For efficiency, M is computed on a $2,000$-row training subsample. Crucially, we apply a second StandardScaler after projection to stabilize the unequal variances produced by the decomposition, ensuring smooth gradient descent during training. For the UCI ablation study (6 configurations A to F), the dCor matrix is computed once on the full feature set and sub-matrices are extracted for each configuration.

### 3.3 Transformer Architecture

In this work, we use an encoder-only Transformer designed for direct multi-step forecasting. The architecture processes the dCor-projected sequences through:

- A linear embedding layer and learned positional encodings.
- Two Transformer blocks utilizing multi-head self-attention and GELU activation.
- A Global Average Pooling (GAP) layer to condense temporal information.
- A dense regression head that outputs the entire forecasting horizon H simultaneously.

The vanilla Transformer is modified by adding L2 regularization, a learning rate warmup, and an early-stopping mechanism to prevent overfitting. Detailed hyperparameters for both datasets are presented in TABLE 1.

Activation functions were selected independently for each dataset through preliminary tuning. To compare directly with the baselines DLinear and Informer, we adopt a dual-scaler approach. We note that published Informer/DLinear benchmarks on this dataset are typically reported under the

Table 1: **Hyperparameter configurations for AEMO and UCI electricity experiments**

| Hyperparameter | AEMO | UCI (M to S) |
|---|---|---|
| Transformer blocks | 2 | 2 |
| Attention heads | 4 | 4 |
| Feed-forward dim | 192 | 128 |
| Dropout rate | 0.3 | 0.2 |
| Activation | ReLU | GELU |
| Pooling | Last timestep | GAP |
| Loss function | MSE | Huber ($\delta = 1.0$) |
| Optimizer / LR | Adam, 1e-3 | Adam, 1e-4 |
| LR schedule | ReduceLROnP. | Warmup 10 ep + RLRnP. |
| Early stopping | patience = 30 | patience = 30 |
| Look-back | 24h | 96h |
| Batch size | 64 | 64 |
| Max epochs | 100 | 300 |
| Input features ($k$) | 13 | 7 |
| Output | H=1 | H$\in \{96, 192, 336, 720\}$ |

Multivariate-to-Multivariate (M→M) protocol, predicting all 321 client series jointly; to our knowledge, no directly comparable M→S (single-target) benchmark figures exist in the literature for these models. Our re-implemented baselines were therefore evaluated under the identical M→S pipeline and dual-scaler protocol used for the proposed model (Section 3-3), making the UCI Electricity comparison presented in Section 4.2 internally consistent, even though it cannot be cross-checked against a published M→S number. For the same reason, more recent architectures such as PatchTST and iTransformer are not included in our comparison: both are designed and typically evaluated under the Multivariate-to-Multivariate (M→M) protocol, predicting all 321 series jointly, and are therefore outside the M→S scope of this study. We train our model on standard-scaled data, but calculate MSE and MAE in a MinMax [0, 1] space to match benchmarks. $R^2$ and sMAPE are computed on the original physical scale to provide a clear picture of real-world performance.

# 4. RESULTS AND DISCUSSION

## 4.1 Short-Term Forecasting: AEMO VIC1

The AEMO Victoria State electricity dataset (from Kaggle) covers $2015$–$2024$ with approximately $87,000$ hourly observations, providing a high-resolution benchmark with strong seasonal volatility. We conducted a comparative analysis using four models: LSTM, GRU, Vanilla Transformer, and our proposed dCor-Transformer. LSTM and GRU are configured with their empirically optimal parameters: a look-back window of 168 h (one full week) to capture weekly demand cycles, using $256/128$ units, no recurrent dropout, and early stopping patience of 30 epochs. In contrast, Transformer-based models are evaluated with a look-back window of 24 h. For all models, we use an identical set of features (13), a fixed $seed = 42$, and a uniform $80/20$ temporal split.

We can conclude from the results in TABLE 2 that: First, Transformer models achieved a $3.1$ to $4.3\times$ improvement in MAPE and $0.26$ to $0.32$ in $R^2$ over traditional RNNs (LSTM/GRU), confirming that attention mechanisms are significantly more capable of capturing the $168h$ weekly demand cycles than sequential state propagation. Second, the Vanilla Transformer and dCor-Transformer achieved both strong predictive performance ($R^2 > 0.96$), although the dCor-Transformer showed slightly lower accuracy. On a dataset with 13 engineered domain features, the self-attention mechanism implicitly learns the nonlinear dependency structure that dCor captures explicitly. Finally, the primary advantage of dCor here was computational: on identical hardware (Colab CPU), dCor−Transformer trained in 1,444 s versus 2,477 s for Vanilla Transformer ($-41.7\%$). By pre-organizing features into a dependency-ordered space, the dCor−Transformer trained 40% faster on CPU, converging in significantly fewer epochs than the Vanilla version (FIGURE 3).

Table 2: **Benchmark results on AEMO VIC1**

| Model | RMSE (MW) | MAPE (%) | $R^2$ | $MSE_{norm}$ | $MAE_{norm}$ |
|---|---|---|---|---|---|
| LSTM | 616.82 | 11.154 | 0.6539 | 0.007015 | 0.070182 |
| GRU | 589.28 | 10.189 | 0.6841 | 0.006402 | 0.063390 |
| Vanilla Transformer | 158.22 | 2.637 | 0.9772 | 0.000462 | 0.015377 |
| dCor-Transformer | 196.84 | 3.431 | 0.9648 | 0.000714 | 0.019914 |

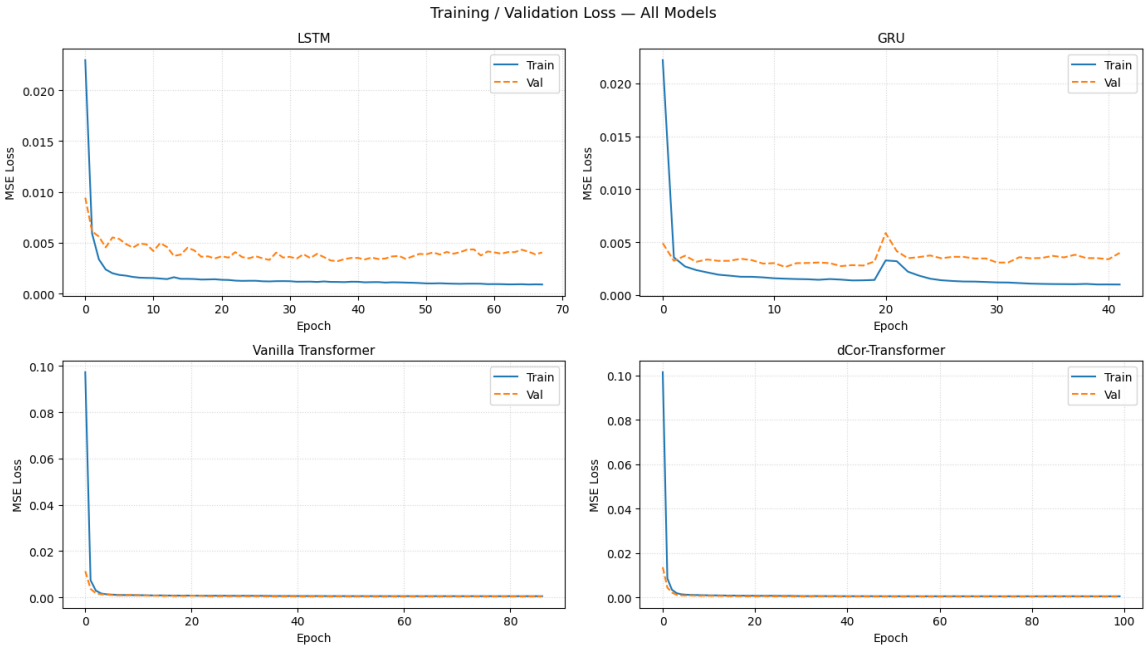



Figure 3: Training and validation loss curves for all four models.

### 4.2 Long-Term Forecasting: UCI Electricity (M→ S)

The UCI Electricity dataset contains hourly electricity consumption for 321 Portuguese clients: 26,304 hourly observations per client. We adopt the standard Autoformer benchmark protocol: 70/10/20 temporal split, fixed seed = 42, and the M→S (Multivariate-to-Single) task, where 321 input variates predict one target client (OT column). Note that this 321-variate scope describes the standard Autoformer M→S benchmark protocol under which the baselines (Informer, DLinear) are re-implemented and evaluated for comparability with the literature; the proposed dCor-Transformer itself receives only the 7 dCor-selected Config D features (Section 3.1) as input, not the full 321-variate set.

To eliminate protocol inconsistency, all baselines were re-implemented by the authors using an identical pipeline (Standard Scaler for training, MinMax Scaler for metrics). Results for horizon H = 96 are presented in TABLE 3.

Table 3: **Forecasting results on UCI electricity, H = 96**. *dCor-Transformer row reflects a single representative seed (seed=42); 5-seed mean ± SD ($MSE_{norm}$=0.003276±0.000055, $R^2$=0.676±0.005)

| Model | $MSE_{norm}$ | $MAE_{norm}$ | $R^2$ | sMAPE (%) | ΔMSE vs dCor |
|---|---|---|---|---|---|
| Informer | 0.009298 | 0.078813 | 0.1238 | 12.88 | +186.9% |
| DLinear | 0.003995 | 0.045828 | 0.6235 | 7.45 | +23.3% |
| Vanilla Transformer (1feat) | 0.009020 | 0.079648 | 0.1064 | 13.11 | +64.4% |
| dCor-Transformer* | 0.003241 | 0.043391 | 0.6794 | 7.16 | reference |

Table 4: **Ablation study results across 6 configurations**. *Config D row reflects a single representative seed (seed=42); 5-seed mean ± SD ($MSE_{norm}$=0.003276±0.000055, $R^2$=0.676±0.005). Relative rankings across configurations (B–F) are expected to hold, but absolute values for Config D should be read alongside the multi-seed estimate.

| Rk | Configuration | N features | N components | Var% | $MSE_{norm}$ | ΔMSE | $R^2$ | sMAPE |
|---|---|---|---|---|---|---|---|---|
| 1 | D – No lags* | 7 | 7 | 100% | 0.003241 | −34.3% | 0.6794 | 7.16% |
| 2 | C – No rolling | 9 | 8 | 98% | 0.003561 | −27.8% | 0.6477 | 7.52% |
| 3 | B – No lags | 9 | 8 | 98% | 0.003699 | −25.0% | 0.6341 | 7.49% |
| 4 | A – Full baseline | 11 | 8 | 94% | 0.004932 | baseline | 0.5121 | 8.73% |
| 5 | E – No temporal | 5 | 5 | 100% | 0.007459 | +51.2% | 0.2622 | 11.01% |
| 6 | F – Target only | 1 | 1 | 100% | 0.008955 | +81.6% | 0.1142 | 13.05% |

The dCor-Transformer outperforms specialized long-term baselines: it achieved a 23.3% improvement in MSE over DLinear and surpassed Informer by a significant margin (+186.9%). This confirms that the dCor projection effectively captures complex nonlinear inter-dependencies inherent in the 321 client series, which linear models and standard attention mechanisms often fail to isolate. As a lower bound, a Vanilla Transformer trained on the raw target series alone achieves $MSEnorm = 0.009020 (R^2 = 0.106)$, confirming that the $64.4\%$ MSE reduction of Config D is attributable to dCor-guided feature selection rather than architectural capacity. To quantify the contribution of each feature group, a systematic ablation was conducted across six configurations ($A \rightarrow F$) with identical model architecture, split, and seed. The dCor matrix is computed once on the full feature set (Config A) and sub-matrices are extracted for each configuration, ensuring performance differences reflect feature content rather than projection variability. Results are shown in TABLE 4.

The ablation results reveal two important results. First, temporal cyclical features are computationally essential: removing cyclical encodings (Configurations E and F) caused a $51$ to $81\%$ increase in MSE, proving that daily and weekly seasonal signals are the primary features of forecasting stability. Second, there is a counter-intuitive redundancy between lags and look-back windows: Config D (no lags/rolling stats) outperformed the full feature set (Config A) by 34.3%. With $lookback = 96h$, the Transformer already observes four days of raw history; adding explicit $24h/168h$ lags introduced dimensionality noise into the dCor projection. This finding is consistent with [23], confirming that structured temporal encoding dominates over raw lag accumulation.

### 4.3 Multi Seed Multi Horizons Validation

To assess statistical robustness beyond $H = 96$, we extended validation to the full horizon scope of TABLE 1 ($H \in \{96, 192, 336, 720\}$), with each horizon averaged over the same five independent seeds $(42, 123, 456, 789, 2024)$, using the identical data, feature-projection, and training pipeline. Results are summarized in TABLE 5.

Table 5: **Multi-seed validation summary on UCI electricity (M→S), Config D, across all four forecasting horizons.** Results are averaged over five independent seeds (42, 123, 456, 789, 2024) using the identical data, feature-projection, and training pipeline. Values are reported as mean ± standard deviation. *Note:* $H = 192/336/720$ use a 500-epoch budget, increased after trials with 300 epochs showed incomplete convergence (see the note in the Limitations). All values satisfy the stability criterion of CV(MSE)< $10\%$.

| Horizon | Epoch cap | $MSE_{norm}$ | $MAE_{norm}$ | $R^2$ | sMAPE (%) | CV(MSE) (%) |
|---|---|---|---|---|---|---|
| 96 | 300 | 0.003276±0.000055 | 0.04328±0.00029 | 0.676±0.005 | 7.13±0.06 | 1.7 |
| 192 | 500 | 0.003459±0.000135 | 0.04432±0.00076 | 0.659±0.013 | 7.26±0.12 | 3.9 |
| 336 | 500 | 0.004293±0.000402 | 0.05031±0.00280 | 0.578±0.040 | 8.15±0.42 | 9.4 |
| 720 | 500 | 0.003671±0.000085 | 0.04599±0.00072 | 0.635±0.009 | 7.48±0.11 | 2.3 |

### 4.4 Discussion and Limitations

From the results presented in the previous sections, a clear pattern is observed. Indeed, we found from the three experiments on the two datasets reveal that for datasets with complex, unstructured nonlinear dependencies (as in the UCI dataset), the dCor eigen-decomposition provides measurable forecasting gains. In contrast, for datasets with engineered features (as in the AEMO dataset), the self-attention mechanism essentially discovers those same structures alone. This aligns with the broader observation in the feature selection literature. Hence, pre-processing methods are most beneficial when raw features contain a latent structure that is not directly accessible to the downstream model [7]. Regarding the speed, the dCor-Transformer outperformed the Vanilla Transformer on large datasets, training in 1,444 s vs. 2,477s. The gain is primarily computational (-40% training time) with marginal accuracy differences. This accuracy gap is small in absolute term ($R^2 - 0.013$, $MAPE + 0.79pp$) consistent with the two models performing nearly identically overall. However, on a T4 GPU with a smaller dataset (about 2,106 observations), the dCor was 19% slower than Vanilla due to the initial math overhead. Consequently, the efficiency gain is most visible for CPU-based deployments on large-scale hourly data.

Some limitations should be considered for future refinement. the first one is about the divergent impact of lag features. it proved to be detrimental on the UCI dataset but beneficial on AEMO. Thus highlighting a significant domain-specific sensitivity and the utility of lagged variables in electricity forecasting may depend on the underlying regularity or noise of the specific grid data.

The second concerns the convergence at longer horizons. In fact, horizons $H = 192$, 336, and 720 required an increased epoch budget (500, versus 300 for $H = 96$) to reach convergence; even so, 3 of 5 seeds at $H = 720$ reached the epoch cap without triggering early stopping. That would suggest that results may still improve marginally with further training.
We also note that performance does not degrade monotonically with horizon. Indeed, when looking to horizon $H = 336$ ($R^2 = 0.578$), it underperforms the longer horizon $H = 720$ ($R^2 = 0.635$), suggesting that this is not simply a matter of incomplete training but instead points to an intrinsic interaction between the 336-hour horizon and the periodicities encoded in the temporal features.
Concerning generalizability, the AEMO VIC1 results (TABLE 2) are reported for a single seed, while the UCI Electricity results are validated across 5 seeds (TABLE 5). Indeed, our findings are demonstrated based on our experiments on two datasets only for energy load forecasting. Evaluat-

ing on additional load forecasting benchmarks would strengthen the generality of the conclusions. Finally, the dCor matrix $M$ is estimated on a fixed 2,000-row training subsample for computational efficiency; we did not formally test the sensitivity of the resulting eigen-decomposition to this subsample size, and we flag this as a practical assumption rather than a validated design choice.

## 5. CONCLUSION

In this paper, we have proposed a dCor eigen-decomposition pipeline for feature pre-processing in Transformer-based electricity demand forecasting, providing dCor-decorrelated rotation of the input feature space based on nonlinear statistical dependencies. The experiments across three settings demonstrated four primary contributions. First, on the AEMO VIC1 dataset, the dCor-Transformer achieved a modest accuracy loss camparing to Vanilla Transformer while reducing CPU training costs by 40% and outperforming tuned recurrent baselines by $3to4\times$. Second, under a strictly verified M→ S protocol ($H = 96$) on the UCI Electricity dataset, the dCor-Transformer surpassed DLinear by 23.3% and Informer by 186.9% within a pipeline where all baselines were re-implemented under identical conditions. Third, a systematic ablation study established that temporal cyclical encodings are the dominant predictors for electricity demand, and that explicit lag features become redundant—and even detrimental—when the Transformer's look-back window is sufficiently long.

We conclude that dCor pre−processing should be applied when input features are raw and unstructured. Future work will investigate dCor-guided attention mechanisms, extending the dCor prior into the attention score computation, evaluate the framework under the Multivariate-to-Multivariate (M→M) protocol on high-dimensional benchmarks.

## 6. ACKNOWLEDGMENT

The author thanks the Deputyship for Research & Innovation University of Jeddah, Ministry of Education in Saudi Arabia for funding this work under project number UJ-20-126-DR.

## References

[1] Dudek G, Piotrowski P, Baczyński D. Forecasting in Modern Power Systems: Challenges, Techniques, and Emerging Trends. Energies. 2025;18:3589.

[2] Biswal B, Deb S, Datta S, Ustun TS, Cali U. Review on Smart Grid Load Forecasting for Smart Energy Management Using Machine Learning and Deep Learning Techniques. Energy Rep. 2024;12:3654-3670.

[3] Pinheiro MG, Madeira SC, Francisco AP. Short-Term Electricity Load Forecasting—A Systematic Approach From System Level to Secondary Substations. Appl Energy. 2023;332:120493.

[4] Lim B, Zohren S. Time-Series Forecasting With Deep Learning: A Survey. Phil Trans R Soc A. 2021;379:20200209.

[5] Hyndman RJ, Athanasopoulos G. Forecasting: Principles and Practice. 3rd ed. Melbourne Australia: OTexts. 2021.

[6] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016:785-794.

[7] Nabavi SA, Mohammadi S, Motlagh NH, Tarkoma S, Geyer P. Deep Learning Modeling in Electricity Load Forecasting: Improved Accuracy by Combining Dwt and Lstm. Energy Rep. 2024;12:2873-2900.

[8] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9:1735-1780.

[9] Zhou H, Zhang S, Peng J, Zhang S, Li J, et al. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In Proceedings of the AAAI conference on artificial intelligence. 2021;35:11106-11115.

[10] Wu H, Xu J, Wang J, Long M. Autoformer: Decomposition Transformers With Auto-Correlation for Long-Term Series Forecasting. In Proceedings of the 35th International Conference on Neural Information Processing Systems. 2021;34:22419–22430.

[11] Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A Time Series Is Worth 64 Words: Long-Term Forecasting With Transformers. InProceedings of the international conference on learning representations Kigali Rwanda. 2023.

[12] Liu Y, Hu T, Zhang H, Wu H, Wang S, et al. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. InInternational conference on learning representations. 2024; 2024:11116-11140.

[13] Ragupathi C, Dhanasekaran S, Vijayalakshmi N, Salau AO. Prediction of Electricity Consumption Using an Innovative Deep Energy Predictor Model for Enhanced Accuracy and Efficiency. Energy Rep. 2024;12:5320-5337.

[14] Oliveira HS, Oliveira HP. Transformers for Energy Forecast. Sensors. 2023;23:6840.

[15] Székely GJ, Rizzo ML, Bakirov NK. Measuring and Testing Dependence by Correlation of Distances. Ann Statist. 2007;35:2769-2794.

[16] Székely GJ, Rizzo ML. Brownian Distance Covariance. Ann Appl Stat. 2009;3:1236-1265.

[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need. In Proceedings of the advances in neural information processing systems. Cambridge, MA: MIT Press. 2017;30.

[18] Zeng A, Chen M, Zhang L, Xu Q. Are Transformers Effective for Time Series Forecasting? In Proceedings of the AAAI conference on artificial intelligence. 2023;37:11121-11128.

[19] Hussan U, Wang H, Peng J, Jiang H, Rasheed H. Transformer-Based Renewable Energy Forecasting: A Comprehensive Review. Renew Sustain Energy Rev. 2026;226:116356.

[20] Wang Y, Hao Y, Zhang B, Zhang N. Short-term power load forecasting using SSA-CNN-LSTM method. Syst Sci Control Eng. 2024;12:2343297.

[21] Ullah K, Ahsan M, Hasanat SM, Haris M, Yousaf H, et al. Short-Term Load Forecasting: A Comprehensive Review and Simulation Study With CNN-LSTM HybridS Approach. IEEE Access. 2024;12:111858-111881.

[22] L'Heureux A, Grolinger K, Capretz MA. Transformer-Based Model for Electrical Load Forecasting. Energies. 2022;15:4993.

[23] Magalhães B, Bento P, Pombo J, Calado MR, Mariano S. Short-Term Load Forecasting Based on Optimized Random Forest and Optimal Feature Selection. Energies. 2024;17:1926.