

Correlation-Based Reduction of Environmental Input Variables for Efficient GRU-Based Prediction in IoT Systems

Peter Kolok, Michal Hodoň, Marek Tebelák, Tomáš Bača

University of Žilina, Faculty of Management Science and Informatics

Department of Technical Cybernetics

Univerzitná 8215/1, 010 26 Žilina, Slovakia

{peter.kolok, michal.hodon}@fri.uniza.sk,

tebelak2@stud.uniza.sk, tomas.baca@fri.uniza.sk

Abstract. This study investigates correlation-based feature reduction for GRU-based photovoltaic power prediction in IoT-oriented environmental monitoring systems. The aim was to determine whether a reduced subset of input variables can preserve or improve prediction accuracy compared with a full-input baseline model. Pearson correlation, Spearman correlation, and ENNEMI mutual-information-based dependency analysis were used to select relevant variables from the SOLETE Pombo hourly dataset. The best reduced-input model was obtained using Spearman threshold-based selection with a 48-hour input window. It used 11 instead of 16 input variables and achieved an RMSE of 0.4545, MAE of 0.2382, and R^2 of 0.9279. Compared with the baseline, RMSE decreased by 1.11% and MAE by 3.28%.

Keywords. GRU, photovoltaic power prediction, feature selection, correlation analysis, Spearman correlation, mutual information, IoT systems

1 Introduction

The rapid expansion of the Internet of Things (IoT) has increased the demand for intelligent methods capable of processing sensor data, modelling complex dynamic systems, and supporting automated monitoring applications. Neural networks have been applied to nonlinear system modelling and to the analysis of measurements collected by wireless sensor networks Formanek and Karpiš, 2020; Olešnaníková et al., 2016. Their ability to identify complex relationships within measured data makes them suitable for the development of intelligent sensor systems and data-driven monitoring solutions. In this context, the efficient processing and selection of sensor measurements represent important aspects of reducing unnecessary data transmission and improving the operation of resource-constrained IoT devices.

Environmental monitoring represents one of the most critical and widespread applications of IoT, impacting sectors from agriculture and hydrology to dis-

aster preparedness Haq et al., 2022, Pujara and Paudel, 2024. Accurate modeling and forecasting of environmental variables, such as temperature, rainfall, or pollution levels, are inherently complex due to the volatility, nonlinearity, and stochastic nature of meteorological time-series data Pujara and Paudel, 2024, Krammer et al., 2022. Traditional forecasting methods, including econometric models and simple trend analysis, often struggle to capture these intricate patterns, leading to substantial inaccuracies Tuyet et al., 2025, Wentz et al., 2022. Consequently, researchers have shifted toward Artificial Intelligence (AI) and Deep Learning (DL) architectures, which have proven superior in learning hidden patterns within large, complex datasets Nketiah et al., 2023, Haq et al., 2022. Among the various DL architectures, Recurrent Neural Networks (RNNs)—specifically Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU)—have emerged as the state-of-the-art for time-series forecasting Pujara and Paudel, 2024, Wang et al., 2023. These networks utilize gating mechanisms to handle long-term temporal dependencies and mitigate the vanishing gradient problem inherent in standard RNNs Tuyet et al., 2025, Nketiah et al., 2023. However, a significant issue in environmental forecasting is the management of high-dimensional input datasets. While there is a common belief that more input features improve accuracy, large feature sets often introduce redundancy and noise, which increases computational burden and the risk of overfitting Kibet et al., 2025, Krammer et al., 2022. This research addresses these challenges by focusing on the reduction of input environmental variables using correlation analysis. By identifying and selecting only the most influential predictors, the proposed approach aims to create a more efficient GRU-based prediction model. This minimalist approach aims to maintain high predictive accuracy while reducing the number of required input variables. Such reduction may simplify sensing, preprocessing, and communication requirements, making the approach relevant for resource-constrained IoT systems where processing power and energy availability are limited.

2 Related Work

The transition in environmental forecasting from traditional statistical methods to recurrent neural network architectures, such as LSTM and GRU, has been driven by the need to capture complex, non-linear, and time-dependent relationships in meteorological and environmental data Nketiah et al., 2023, Pujara and Paudel, 2024. Conventional forecasting approaches often struggle to represent delayed temporal effects and interactions among multiple environmental variables, particularly when the monitored processes are influenced by rapidly changing external conditions. Within this context, recurrent neural networks have become widely used for time-series forecasting tasks, while GRU models are often considered an efficient alternative to LSTM architectures because of their simpler gating structure, lower number of trainable parameters, and shorter training time Tuyet et al., 2025, Wang et al., 2023, Pujara and Paudel, 2024. These properties make GRU models particularly attractive for IoT-oriented forecasting systems, where computational resources, memory capacity, and energy availability are limited.

A critical challenge in deploying such models is the selection of appropriate input variables. Environmental monitoring systems commonly collect numerous meteorological, irradiance-related, and temporal variables, including temperature, humidity, atmospheric pressure, wind speed, solar irradiance, UV radiation, brightness, and time-derived features. However, not all measured variables contribute equally to the forecasting task. High-dimensional input datasets may introduce redundant, weakly informative, or noisy variables, increasing computational burden and the risk of overfitting Krammer et al., 2022, Kibet et al., 2025. Therefore, recent research has increasingly focused on feature selection and input dimensionality reduction as practical strategies for improving forecasting efficiency while preserving predictive accuracy.

In photovoltaic power forecasting, feature selection is particularly relevant because photovoltaic output is strongly influenced by a subset of dominant physical factors, especially irradiance-related and time-dependent variables. Mauladdawilah et al. investigated deep feature selection of meteorological variables for LSTM-based photovoltaic power forecasting in high-dimensional time-series data, emphasizing the need to identify the most informative meteorological inputs rather than relying on all available variables Mauladdawilah et al., 2025. Similarly, Kibet et al. proposed a minimalist deep-learning approach for solar power forecasting using key meteorological features, showing that reduced-input models can achieve competitive predictive performance compared with models using larger input feature sets Kibet et al., 2025. In addition, Ye et al. developed a short-term photovoltaic power forecasting model based on improved feature se-

lection combined with a TCN-LSTM-AM architecture and APO optimization, further demonstrating the importance of feature selection in photovoltaic forecasting pipelines Ye et al., 2025.

These studies indicate that careful input-variable selection can reduce unnecessary feature complexity while maintaining forecasting performance. Nevertheless, many existing approaches focus either on specific deep feature-selection mechanisms, transformer-based architectures, or hybrid forecasting models. In contrast, the present study focuses on a simple and interpretable comparison of correlation-based and mutual-information-based feature reduction methods for GRU-based photovoltaic power prediction. Specifically, Pearson correlation, Spearman correlation, and EN-NEMI mutual-information-based dependency analysis are used to construct reduced input subsets. The performance of these reduced-input GRU models is then compared with a full-input baseline model trained on all available environmental variables.

The main objective of this study is therefore to verify whether a smaller set of highly relevant environmental variables can provide sufficient information for accurate one-hour-ahead photovoltaic power forecasting. By focusing on input-variable reduction rather than architectural complexity, the proposed approach aims to support more efficient GRU-based prediction in IoT-oriented monitoring systems, where reducing sensing, preprocessing, and communication requirements is particularly important.

3 Materials and Methods

This section describes the dataset, preprocessing procedure, feature-selection strategy, GRU model configuration, and evaluation methodology used in this study. The main objective of the experimental design was to evaluate whether a reduced subset of environmental input variables, selected using correlation-based and mutual-information-based dependency analysis, can preserve or improve the predictive performance of a GRU model compared with a baseline model trained on all available input variables. The experimental workflow consisted of five main steps: dataset preparation, chronological data splitting, correlation and dependency analysis, construction of reduced feature subsets, and training and evaluation of GRU-based forecasting models. The target variable for the prediction task was the photovoltaic power output, while the input variables consisted of meteorological, irradiance-related, and time-derived features.

3.1 Dataset Description

The experiments were performed using the SOLETE Pombo dataset with an hourly resolution. After preprocessing, the dataset contained 10,969 samples. The target variable was photovoltaic power output, denoted as

P_Solar[kW]. The original input feature set consisted of 16 numerical variables, including environmental measurements, irradiance-related variables, and time-derived cyclic features. The measured environmental variables included temperature, humidity, wind speed, wind direction, atmospheric pressure, global horizontal irradiance, plane-of-array irradiance, azimuth, elevation, and the power output of the Gaia system. In addition, cyclic temporal features were generated to represent hourly, daily, and monthly periodicity. These features were included because photovoltaic power production is strongly influenced by time-dependent solar patterns Pombo et al., 2022. The full feature set was used as the baseline configuration. Reduced-input configurations were then created using Pearson correlation, Spearman correlation, and ENNEMI mutual-information-based dependency analysis. Each selected feature subset was evaluated using the same GRU architecture and training procedure in order to ensure a fair comparison between the full-input and reduced-input models.

3.2 Data Preprocessing

Prior to model training, the SOLETE Pombo dataset was preprocessed to ensure consistency and suitability for time-series forecasting. The dataset was first sorted chronologically in order to preserve the temporal structure of the measured data. Since the prediction task was formulated as a supervised time-series problem, maintaining the original order of observations was essential to avoid information leakage between past and future samples. Only numerical variables were used in the experiment, while missing or infinite values were removed from the dataset. After preprocessing, the final dataset contained 10,969 hourly samples. The target variable was photovoltaic power output, denoted as P_Solar [kW]. The input feature set consisted of environmental, irradiance-related, and temporal variables. In addition to the original measured variables, cyclic time features were generated to represent the periodic nature of photovoltaic power production. Specifically, the hour of the day, day of the year, and month were transformed using sine and cosine encoding. For a cyclic variable (t) with period (T), the transformation was defined as:

$$t_{sin} = \sin\left(2\pi \frac{t}{T}\right), \quad (1)$$

$$t_{cos} = \cos\left(2\pi \frac{t}{T}\right). \quad (2)$$

For example, the period was set to ($T = 24$) for the hour of the day, ($T = 365.25$) for the day of the year, and ($T = 12$) for the month. This transformation allows the model to capture cyclic temporal patterns, such as daily and seasonal changes, without introducing artificial discontinuities between consecutive time periods.

The dataset was divided chronologically into training, validation, and testing subsets using a 70:15:15 ratio. The training subset was used to fit the normalization parameters, while the validation and testing subsets were transformed using the same parameters. This ensured that no information from the validation or testing data influenced the preprocessing stage. All input variables and the target variable were standardized using z-score normalization:

$$x' = \frac{x - \mu}{\sigma}, \quad (3)$$

where x represents the original value, μ is the mean value calculated from the training subset, and σ is the corresponding standard deviation. After scaling, the data were converted into supervised learning sequences using a sliding-window approach. For a given window length W and forecasting horizon h , the input sequence was defined as:

$$X_t = [x_{t-W+1}, x_{t-W+2}, \dots, x_t], \quad (4)$$

while the corresponding target value was defined as:

$$y_t = P_{\text{Solar}}(t + h). \quad (5)$$

In this study, sliding windows with lengths of $W \in \{12, 24, 48\}$ time steps were evaluated, and the forecasting horizon was set to $h = 1$, corresponding to one-hour-ahead prediction. The experiments were implemented in Python using the PyTorch deep learning framework and executed on a CUDA-enabled workstation. Model training was performed on an NVIDIA GeForce RTX 3060 GPU using CUDA 12.1. GPU acceleration was used to reduce the training time of the GRU models across all tested feature subsets and sliding-window configurations.

3.3 Feature Selection Methods

The main objective of the feature selection stage was to reduce the dimensionality of the input space while preserving the most informative environmental variables for photovoltaic power prediction. The full feature set was first used as a baseline configuration. Subsequently, reduced feature subsets were created using Pearson correlation, Spearman correlation, and ENNEMI mutual-information-based dependency analysis. These methods were applied only to the training subset in order to avoid information leakage from the validation and testing data.

Pearson correlation was used to evaluate the strength of the linear relationship between each input variable and the target variable. For an input variable (X) and the target variable (Y), the Pearson correlation coefficient was calculated as:

$$r_{X,Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (6)$$

where x_i and y_i represent individual observations, $\bar{x}$ and $\bar{y}$ are the corresponding mean values, and n is the number of samples. Variables with higher absolute Pearson correlation values were considered more linearly related to the target variable.

Since environmental data may contain monotonic but not strictly linear relationships, Spearman correlation was also applied. Spearman correlation evaluates the relationship between ranked variables and is defined as:

$$\rho_{X,Y} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad (7)$$

where d_i is the difference between the ranks of the corresponding observations of variables X and Y . This method is useful when the dependency between variables is monotonic but does not necessarily follow a linear pattern.

In addition to Pearson and Spearman correlation, ENNEMI mutual-information-based dependency analysis was used to capture potentially nonlinear relationships between input variables and the target variable. Mutual information measures the amount of information shared between two variables and can be expressed as:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right), \quad (8)$$

where $p(x, y)$ is the joint probability distribution of X and Y , while $p(x)$ and $p(y)$ are their marginal probability distributions. Unlike linear correlation coefficients, mutual information can detect more general statistical dependencies, including nonlinear relationships.

For each feature selection method, input variables were ranked according to the absolute dependency score with respect to the target variable. Reduced feature subsets were then created using two strategies. The first strategy selected the top K variables, where $K \in \{3, 5, 7\}$. The second strategy selected all variables whose dependency score exceeded a predefined threshold, with thresholds set to 0.1, 0.2, and 0.3. These reduced feature subsets were then used as inputs to separate GRU models and compared against the full-input baseline model.

3.4 GRU Model Architecture

The prediction model used in this study was based on the GRU architecture. GRU networks are designed to process sequential data by maintaining a hidden state that is updated over time. Compared with LSTM networks, GRU models use a simpler gating mechanism, which reduces the number of trainable parameters while preserving the ability to model temporal dependencies.

At each time step t , the GRU receives an input vector x_t and the previous hidden state h_{t-1} . The reset gate r_t , update gate z_t , candidate hidden state $\tilde{h}_t$, and final hidden state h_t are computed as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z), \quad (9)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r), \quad (10)$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h), \quad (11)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t. \quad (12)$$

Here, σ denotes the sigmoid activation function, $\tanh$ is the hyperbolic tangent activation function, $\odot$ represents element-wise multiplication, and W , U , and b are trainable weight matrices and bias vectors.

The implemented GRU model consisted of one recurrent GRU layer with 64 hidden units. The output of the final time step was passed to a fully connected regression head consisting of a linear layer, ReLU activation, dropout, and a final linear output neuron. The dropout rate was set to 0.10 in order to reduce the risk of overfitting. The model was trained using the AdamW optimizer with a learning rate of 10^{-3} and weight decay of 10^{-5} . The Smooth L_1 loss function was used as the training objective because it combines the advantages of mean absolute error and mean squared error and is less sensitive to outliers than pure squared-error loss.

For a prediction $\hat{y}$ and target value y , the Smooth L_1 loss can be expressed as:

$$L(y, \hat{y}) = \begin{cases} 0.5(y - \hat{y})^2, & \text{if } |y - \hat{y}| < 1, \\ |y - \hat{y}| - 0.5, & \text{otherwise.} \end{cases} \quad (13)$$

This architecture was kept identical for all feature subsets in order to ensure that differences in model performance were caused by the selected input variables rather than by changes in network structure.

3.5 Training Procedure and Evaluation Metrics

All GRU models were trained using the same procedure to ensure a fair comparison between the full-input baseline model and the reduced-input models. For each experiment, the selected input variables were standardized using only the parameters calculated from the training subset. The same scaling parameters were then applied to the validation and testing subsets to prevent information leakage. The target variable was also standardized during training and transformed back to its original scale before the final evaluation.

The models were trained using mini-batch gradient descent with a batch size of 128. The maximum number of training epochs was set to 80. Early stopping was applied based on the validation loss with a patience of 12 epochs. A learning-rate reduction strategy was also used when the validation loss stopped improving. In addition, gradient clipping was applied to improve numerical stability during training.

The model architecture and training hyperparameters were kept identical across all experiments. Therefore, performance differences within the same sliding-window length can be attributed to the selected input feature subset, whereas comparisons across different window lengths reflect both feature selection and the amount of temporal context provided to the model. The evaluated window lengths were 12, 24, and 48 time steps. The forecasting horizon was fixed to one hour ahead.

The predictive performance of each model was evaluated using the root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). These metrics were calculated on the testing subset after transforming the predicted and measured values back to the original scale of the target variable.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (14)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (15)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (16)$$

In these equations, y_i represents the measured photovoltaic power output, $\hat{y}_i$ denotes the predicted value, $\bar{y}$ is the mean of the measured values, and n is the number of testing samples. Lower RMSE and MAE values indicate better predictive accuracy. A higher R^2 value indicates that the model explains a larger proportion of the variance in the target variable.

In addition to predictive accuracy, the reduced-input models were evaluated in terms of input dimensionality reduction. The feature reduction ratio was calculated as:

$$FR = \frac{N_{full} - N_{reduced}}{N_{full}} \times 100\%, \quad (17)$$

where N_{full} represents the number of input variables used in the baseline model and $N_{reduced}$ represents the number of input variables used in the reduced-input model. This ratio quantifies the reduction in the number of required input variables and reflects the potential decrease in sensing, preprocessing, and communication requirements in IoT-based monitoring systems.

The final comparison was performed between the best baseline model trained using all available input

variables and the best reduced-input model selected according to the lowest testing RMSE. A reduced-input model was considered effective if it achieved comparable or lower prediction error while using fewer input variables than the baseline configuration.

4 Results

This section presents the results of the proposed feature-reduction experiments for GRU-based photovoltaic power prediction. The performance of GRU models trained on reduced input feature subsets was compared with a baseline GRU model trained on all available input variables. The evaluation focused on prediction accuracy, input dimensionality reduction, and the comparison of different feature-selection strategies based on Pearson correlation, Spearman correlation, and ENNEMI mutual-information-based dependency analysis.

4.1 Correlation and Dependency Analysis

The correlation and dependency analysis was performed on the training subset in order to identify the input variables with the strongest relationship to the target photovoltaic power output. Since the feature-selection process was applied only to the training data, the validation and testing subsets did not influence the selection of input variables. This was important to avoid information leakage and to ensure a fair evaluation of the reduced-input GRU models.

Fig. 1 shows the absolute Spearman correlation values between the input variables and the target variable. Spearman correlation was included because environmental and photovoltaic data may contain monotonic relationships that are not strictly linear. The highest dependency values were observed mainly for irradiance-related variables and selected temporal features. This result is physically meaningful, since photovoltaic power output is directly influenced by the amount of solar irradiance reaching the photovoltaic panel and by the time-dependent position of the sun.

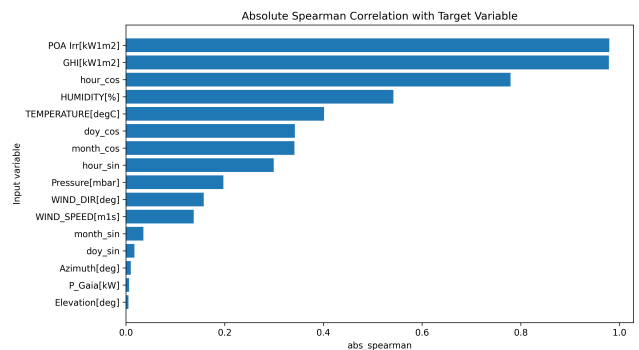


Figure 1. Absolute Spearman correlation between input variables and photovoltaic power output.

The correlation ranking indicates that not all measured environmental variables contributed equally to the prediction task. Variables with stronger monotonic relationships to the target variable were considered more relevant for GRU-based forecasting, while weakly dependent variables were treated as candidates for removal. This supports the main assumption of the experiment: using all available input variables does not necessarily improve prediction performance, because weakly informative or redundant variables may increase the dimensionality of the input space without adding useful predictive information.

The Spearman-based analysis was particularly important because the best-performing reduced-input model was obtained using the Spearman threshold-based feature subset with a threshold value of 0.1. This configuration selected 11 input variables from the original set of 16 variables. Specifically, the selected variables were plane-of-array irradiance, global horizontal irradiance, hour cosine, humidity, temperature, day-of-year cosine, month cosine, hour sine, atmospheric pressure, wind direction, and wind speed. Therefore, the correlation analysis provided the basis for reducing the input dimensionality before training the GRU models.

4.2 Performance Comparison of GRU Models

The predictive performance of the evaluated GRU models was compared using the testing RMSE, MAE, and R^2 metrics. Each model was trained using the same architecture and training procedure, while the input feature subset and sliding-window length were varied. The full-input model, denoted as ALL_FEATURES, was used as the baseline configuration and was compared with reduced-input models created using Pearson, Spearman, and ENNEMI-based feature selection.

Fig. 2 presents the testing RMSE values obtained for the evaluated GRU configurations. The results show that several reduced-input models achieved lower RMSE values than the baseline model trained on all available input variables. This indicates that using the complete set of input variables did not necessarily lead to the best predictive performance. In contrast, removing weakly dependent or redundant variables slightly improved the prediction accuracy in several configurations.

The best-performing configuration was obtained using the Spearman threshold-based feature subset with a threshold value of 0.1 and a sliding-window length of 48 time steps. This model used 11 input variables and achieved an RMSE of 0.4545, MAE of 0.2382, and R^2 of 0.9279. In comparison, the best full-input baseline model used all 16 input variables and achieved an RMSE of 0.4596, MAE of 0.2463, and R^2 of 0.9266. Therefore, the reduced-input model not only decreased the number of input variables but also slightly improved the predictive accuracy.

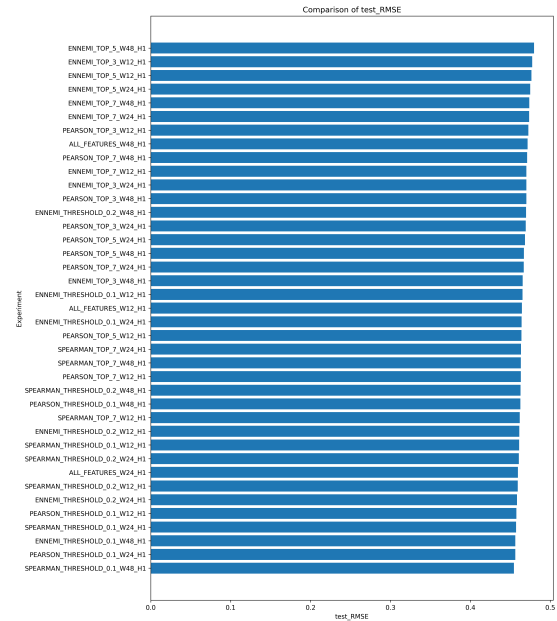


Figure 2. Comparison of testing RMSE values across the evaluated GRU configurations.

Table 1 summarizes the best-performing GRU configurations according to the testing RMSE. The results show that threshold-based feature selection generally produced better performance than selecting only a very small number of top-ranked variables. This suggests that a moderate reduction of the input space is more suitable for this forecasting task than an overly aggressive feature reduction.

Table 1. Best-performing GRU configurations according to testing RMSE.

Model	Window	Inputs	RMSE	MAE	R^2
SPEARMAN_THR_0.1	48	11	0.4545	0.2382	0.9279
PEARSON_THR_0.1	24	11	0.4562	0.2353	0.9276
ENNEMI_THR_0.1	48	16	0.4562	0.2373	0.9273
SPEARMAN_THR_0.1	24	11	0.4572	0.2394	0.9273
PEARSON_THR_0.1	12	11	0.4575	0.2393	0.9271
ALL_FEATURES	24	16	0.4596	0.2463	0.9266

The comparison confirms that the best reduced-input model outperformed the full-input baseline in all evaluated metrics. Although the improvement was relatively small, it is important because it was achieved with fewer input variables. This result supports the assumption that some measured environmental variables may be weakly informative or redundant for GRU-based photovoltaic power prediction.

4.3 Comparison of Feature Selection Methods

To further evaluate the effect of the applied feature-selection strategies, the best-performing configuration from each feature-selection family was compared. The comparison included the full-input baseline model, Pearson-based feature selection, Spearman-based fea-

ture selection, and ENNEMI mutual-information-based dependency analysis. This comparison makes it possible to determine not only which individual GRU configuration achieved the best prediction accuracy, but also which feature-selection strategy provided the most effective balance between prediction performance and input dimensionality reduction.

Table 2 summarizes the best result obtained for each feature-selection method.

Table 2. Best result achieved by each feature-selection method.

Method	Best configuration	Window	Inputs	RMSE	R^2
All features	ALL_FEATURES	24	16	0.4596	0.9266
Pearson	PEARSON_THR_0.1	24	11	0.4562	0.9276
Spearman	SPEARMAN_THR_0.1	48	11	0.4545	0.9279
ENNEMI	ENNEMI_THR_0.1	48	16	0.4562	0.9273

The Spearman-based method achieved the best overall performance, with an RMSE of 0.4545, MAE of 0.2382, and R^2 of 0.9279 while using 11 input variables. The Pearson-based method also improved the prediction performance compared with the full-input baseline and used the same number of input variables. The ENNEMI-based method achieved competitive prediction accuracy; however, its best-performing configuration retained all 16 input variables and therefore did not provide input dimensionality reduction in this experiment.

Although ENNEMI was included to capture nonlinear dependencies between input variables and the target variable, its best-performing threshold-based configuration retained all 16 input variables. This suggests that the ENNEMI dependency scores were less selective in this experiment than the Pearson- and Spearman-based criteria. Therefore, ENNEMI provided competitive predictive accuracy, but it did not contribute to input dimensionality reduction in the evaluated configuration.

However, this result is specific to the tested ENNEMI configuration and threshold values. Alternative mutual-information estimators, parameter settings, or threshold-selection strategies could produce more selective feature subsets. Therefore, the result should not be interpreted as a general limitation of mutual-information-based feature selection.

Among the evaluated methods, Spearman correlation provided the best balance between accuracy and dimensionality reduction. This suggests that monotonic relationships between environmental variables and photovoltaic power output are particularly relevant for this prediction task.

4.4 Baseline versus Best Reduced-Input Model

The final comparison was performed between the best-performing full-input baseline model and the best-performing reduced-input model. The baseline config-

uration, denoted as ALL_FEATURES, used all 16 available input variables with a sliding-window length of 24 time steps and achieved an RMSE of 0.4596, MAE of 0.2463, and R^2 of 0.9266. The best reduced-input model was obtained using the SPEARMAN_THR_0.1 configuration with a sliding-window length of 48 time steps. This model used 11 input variables and achieved an RMSE of 0.4545, MAE of 0.2382, and R^2 of 0.9279.

Across the best-performing configurations, the reduced-input model decreased RMSE by 1.11% and MAE by 3.28%. At the same time, the coefficient of determination increased slightly from 0.9266 to 0.9279. The number of input variables was reduced from 16 to 11, corresponding to a feature reduction of 31.25%. However, since the two best-performing configurations used different sliding-window lengths, this comparison reflects both the effect of feature selection and the amount of temporal context provided to the model.

To isolate the effect of feature selection from the effect of the sliding-window length, the models were also compared using the same 24-time-step window. Under this controlled setting, the SPEARMAN_THR_0.1 model achieved an RMSE of 0.4572, MAE of 0.2394, and R^2 of 0.9273, compared with an RMSE of 0.4596, MAE of 0.2463, and R^2 of 0.9266 for the full-input model. Therefore, the reduced-input model decreased RMSE by approximately 0.52% and MAE by 2.80% while using five fewer input variables. These results indicate that the removal of weakly dependent or redundant variables did not degrade predictive performance and provided a small improvement even when the same temporal window was used.

Fig. 3 compares the measured photovoltaic power output with the predictions of the full-input baseline GRU model and the best-performing reduced-input GRU model. Both models captured the main daily production patterns and followed the measured photovoltaic power output. Although local deviations can be observed in some parts of the test sequence, the reduced-input model achieved better overall error metrics while requiring fewer input variables. This supports the conclusion that the selected subset preserved the information required for accurate one-hour-ahead photovoltaic power forecasting under the evaluated experimental conditions.

4.5 Limitations and Generalizability

The presented results should be interpreted with respect to several limitations. First, the experiments were conducted using a single photovoltaic dataset with an hourly temporal resolution. Therefore, the identified feature subset and the selected Spearman threshold may depend on the measurement location, sensor configuration, climatic conditions, and temporal resolution of the SOLETE Pombo dataset.

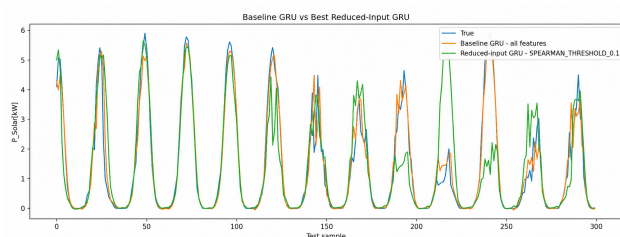


Figure 3. Comparison of measured photovoltaic power output, full-input GRU prediction, and best reduced-input GRU prediction.

Second, the evaluation was restricted to a single recurrent neural network architecture. Although the same GRU configuration was used in all experiments to isolate the effect of feature selection, other architectures, such as LSTM, temporal convolutional networks, or Transformer-based models, may respond differently to the removal of input variables.

Finally, the ENNEMI-based threshold configurations did not reduce the number of input variables under the tested conditions. This result does not imply that mutual-information-based selection is generally ineffective. Its selectivity may depend on the estimator configuration, sample size, normalization, and selected threshold values.

5 Conclusion

This study investigated the impact of correlation-based and mutual-information-based feature reduction on GRU-based photovoltaic power prediction. The main objective was to determine whether a reduced subset of environmental input variables could preserve or improve prediction accuracy compared with a baseline GRU model trained using all available input variables.

The experimental results showed that the best reduced-input model was obtained using the SPEARMAN_THR_0.1 feature-selection strategy with a sliding-window length of 48 time steps. This model used 11 input variables instead of the original 16 variables and achieved an RMSE of 0.4545, MAE of 0.2382, and R^2 of 0.9279. In comparison, the best full-input baseline model achieved an RMSE of 0.4596, MAE of 0.2463, and R^2 of 0.9266. Therefore, the reduced-input model decreased RMSE by 1.11% and MAE by 3.28%, while reducing the number of input variables by 31.25%.

These results indicate that using all available environmental variables does not necessarily lead to the best prediction performance. Weakly dependent or redundant variables may increase the dimensionality of the input space without providing additional predictive value. The Spearman-based feature-selection approach provided the best balance between prediction accuracy and input dimensionality reduction, suggesting that monotonic relationships between environmen-

tal variables and photovoltaic power output are particularly relevant for this forecasting task.

Overall, the results support the assumption that careful feature selection can improve the efficiency of GRU-based prediction models without sacrificing predictive performance. Although the observed improvement in accuracy was modest, it was achieved together with a substantial reduction in the number of input variables. These findings indicate the potential suitability of the proposed approach for IoT-oriented environmental monitoring systems. However, its general applicability should be verified using additional datasets, temporal resolutions, prediction horizons, and neural network architectures.

The main limitation of this study is that the experiments were performed using a single photovoltaic dataset with hourly resolution. Therefore, the obtained results may depend on the specific measurement location, sensor configuration, and temporal resolution of the SOLETE Pombo dataset. In addition, the study evaluated only GRU-based models, while other recurrent, convolutional, or attention-based architectures may respond differently to the same feature-reduction strategy. Furthermore, the direct impact of feature reduction on real sensor-node energy consumption was not measured in this study. Therefore, the reported reduction in the number of input variables should be interpreted as a potential reduction in sensing, preprocessing, and communication requirements rather than as a direct measurement of energy savings.

Future work will evaluate the proposed feature-selection procedure using additional photovoltaic and environmental datasets with different measurement locations and temporal resolutions, including finer sampling intervals. The experiments will also be extended to other forecasting architectures, particularly LSTM, temporal convolutional, and Transformer-based models.

Further research will investigate alternative mutual-information estimators and data-driven threshold-selection strategies. In addition to prediction accuracy, future experiments will evaluate model size, inference time, memory requirements, communication load, and real energy consumption on resource-constrained IoT hardware.

References

- Formanek, L., & Karpiš, O. (2020). Learning Lorenz Attractor Differential Equations Using Neural Network. *2020 5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, 1–6. <https://doi.org/10.1109/SEEDA-CECNSM49515.2020.9221785>

- Haq, M. A., Ahmed, A., Khan, I., Gyani, J., Mohamed, A., Attia, E.-A., Mangan, P., & Pandi, D. (2022). Analysis of environmental factors using ai and ml methods. *Scientific Reports*, 12(1), 13267. <https://doi.org/10.1038/s41598-022-16665-7>
- Kibet, D., So, M. S., & Shin, J.-H. (2025). Minimalist deep learning for solar power forecasting: Transformer-based prediction using key meteorological features. *Energies*, 18(24), 6395. <https://doi.org/10.3390/en18246395>
- Krammer, P., Kvassay, M., Forgáč, R., Očkay, M., Skovajsová, L., Hluchý, L., Skurčák, L., & Pavlov, L. (2022). Regression analysis and modeling of local environmental pollution levels for the electric power industry needs. *Computing and Informatics*, 41, 861–884. <https://doi.org/10.31577/cai20223861>
- Mauladdawilah, H., Balfaqih, M., Balfagih, Z., Pegalajar, M. d. C., & Jadraque Gago, E. (2025). Deep feature selection of meteorological variables for LSTM-based PV power forecasting in high-dimensional time-series data. *Algorithms*, 18(8), 496. <https://doi.org/10.3390/a18080496>
- Nketiah, E. A., Chenlong, L., Yingchuan, J., & Aram, S. A. (2023). Recurrent neural network modeling of multivariate time series and its application in temperature forecasting. *PLoS ONE*, 18(5), e0285713. <https://doi.org/10.1371/journal.pone.0285713>
- Olešnaníková, V., Karpiš, O., Chovanec, M., Šarafín, P., & Žalman, R. (2016). Water Level Monitoring Based on the Acoustic Signal Using the Neural Network. *2016 International Conference on Information and Digital Technologies (IDT)*, 203–206. <https://doi.org/10.1109/DT.2016.7557174>
- Pombo, D. V., Gehrke, O., & Bindner, H. W. (2022). Solete, a 15-month long holistic dataset including: Meteorology, co-located wind and solar pv power from denmark with various resolutions. *Data in Brief*, 42, 108046. <https://doi.org/10.1016/j.dib.2022.108046>
- Pujara, M., & Paudel, N. (2024). Rainfall prediction using long short-term memory and gated recurrent unit with various meteorological parameters. *Nepalese Journal of Statistics*, 8, 47–60. <https://doi.org/10.3126/njs.v8i1.73165>
- Tuyet, D. B. T., Minh, C. L. T., Mi, S. N. T., Tu, D. N., & Ngoc, H. P. T. (2025). Application of lstm and gru neural networks in forecasting the power output of wind power plant. *International Journal of Electrical and Electronics Research (IJEER)*, 13(2), 250–256. <https://doi.org/10.37391/ijeer.130208>
- Wang, Z., Xiao, S., Reuben, C., Wang, Q., & Wang, J. (2023). Soil nox emission prediction via recurrent neural networks. *Computers, Materials & Continua*, 77(1), 285–298. <https://doi.org/10.32604/cmc.2023.044366>
- Wentz, V. H., Maciel, J. N., Ledesma, J. J. G., & Junior, O. H. A. (2022). Solar irradiance forecasting to short-term pv power: Accuracy comparison of ann and lstm models. *Energies*, 15(7), 2457. <https://doi.org/10.3390/en15072457>
- Ye, N., Zhi, C., Yu, Y., Lin, S., & Liu, F. (2025). TCN-LSTM-AM short-term photovoltaic power forecasting model based on improved feature selection and APO. *Sensors*, 25(24), 7607. <https://doi.org/10.3390/s25247607>