LOWER SILESIAN UNIVERSITY DSW IN WROCŁAW

DATA SCIENCE – DATA ANALYSIS USING AI

THESIS

Paweł Pacybulenko

Thesis Title:

Prediction of Photovoltaic Energy Using a Neural Network

Lower Silesian University DSW in Wrocław 2025

STATEMENT

I hereby declare that this thesis, entitled
“Prediction of Photovoltaic Energy Using a Neural Network,”
was written independently by me.

I further declare that this work does not infringe upon any copyrights as defined in the Act of February 4, 1994, on Copyright and Related Rights (Journal of Laws 1994 No. 24, item 83, as amended), and that it does not contain any data or information obtained unlawfully.

This thesis has not been previously submitted for the purpose of obtaining a professional degree at any other institution of higher education.

Paweł Pacybulenko

Acknowledgements

I would like to express my sincere gratitude to **SOLCAST** for providing access to historical and forecasted solar irradiance data, which played a crucial role in the development of this thesis.

Table of Contents

Abstract

The aim of this thesis was to develop a predictive model for forecasting photovoltaic energy production using an artificial neural network. Real-world data on the operating parameters of a photovoltaic installation were retrieved from the inverter via the Modbus protocol and stored on an MQTT server. Historical data and solar irradiance forecasts provided by SOLCAST were also incorporated.

The dataset prepared for the model included, among others: global horizontal irradiance (GHI), timestamp, air temperature, instantaneous and cumulative energy production. Additionally, sine and cosine transformations of the hour (`sin_hour` and `cos_hour`) were introduced as input features to help the model capture daily seasonality and cyclicality.

The neural network was trained and validated, and its performance was evaluated using MAE, RMSE, and the coefficient of determination (R^2). The results indicate high effectiveness of the proposed approach in forecasting photovoltaic energy production based on weather data.

To further improve accuracy, additional input features such as DHI (Diffuse Horizontal Irradiance) could be introduced, helping the model distinguish between direct and diffuse solar radiation. Moreover, using weather data with a time interval shorter than 15 minutes could allow the model to capture rapid irradiance changes, such as sudden sunshine after cloud cover, enhancing short-term prediction accuracy.

Due to the wide variety of photovoltaic installations—differing in tilt angle, orientation, and shading—it is recommended to train a dedicated model for each system. However, a well-prepared model based on historical data from a specific inverter can effectively adapt to the individual characteristics of a given installation.

This study confirms the applicability of artificial intelligence methods in the analysis and management of renewable energy systems.

Introduction

Energy production from renewable sources, particularly from photovoltaic installations, is rapidly growing both in Poland and globally. As production increases, so does the need for efficient management and short- and medium-term generation forecasting. Accurate predictions of PV energy production enable better energy consumption planning, energy storage management, and optimization of energy system operations.

In recent years, artificial intelligence methods, including neural networks, have gained increasing importance and have proven effective in modeling complex nonlinear relationships. Applying these methods to forecast PV energy production can significantly improve prediction accuracy, especially under variable weather conditions.

Objective

The objective of this work is to develop a predictive model that, based on input data such as solar irradiance, temperature, and time of day, enables the forecasting of photovoltaic system output power using an artificial neural network.

Scope

The scope of the work includes analysis of data from a PV installation, feature engineering, design and training of the neural network model, and evaluation of prediction performance using model quality metrics (including MAE, RMSE, R^2).

Structure

The thesis consists of four main chapters. The first chapter presents a literature review and theoretical background related to photovoltaics and artificial neural networks. The second chapter describes the methodology used, including data sources and selected analytical tools. The third chapter discusses the model training process and analysis of the results. The fourth chapter contains interpretation of the results and final conclusions.

1. Literature Review and Theoretical Background

1.1 Introduction to Photovoltaics

Photovoltaics is a technology that converts solar radiation directly into electricity using photovoltaic cells. This phenomenon is based on the so-called photovoltaic effect, discovered in the 19th century by Alexandre Edmond Becquerel. It involves generating electromotive force in a semiconductor material under exposure to light.

Modern PV cells are most often made of silicon, which, due to its semiconductor properties, allows efficient conversion of solar energy into electricity. There are three main types of PV cells:

- Monocrystalline – with the highest efficiency (above 20%) and uniform black color, currently dominant on the market.

- Polycrystalline – cheaper to produce, with slightly lower efficiency (approx. 15–18%), now largely phased out due to lower performance.
- Thin-film – flexible and lightweight but less efficient; mainly used in specialized applications.

PV installation efficiency is affected by many factors, including:

- Irradiance – the main factor determining the amount of energy generated,
- Cell temperature – increased temperature usually lowers conversion efficiency,
- Panel tilt angle and orientation – optimal alignment maximizes energy yield,
- Shading – even partial shading can significantly reduce energy production.

Photovoltaics is environmentally friendly, emission-free, and increasingly economically competitive. Due to its growing role in the energy mix, accurate forecasting of PV energy production is essential for its effective use.

1.2 Characteristics and Variability of PV Energy Production

Electricity production from photovoltaic installations is highly variable over time. This variability results primarily from the strong dependence of the energy generation process on weather conditions, as well as geographic location, season, time of day, and technical characteristics of the installation itself.

The most important factor influencing PV production is the level of solar irradiance, especially GHI (Global Horizontal Irradiance) and DHI (Diffuse Horizontal Irradiance), which vary daily, seasonally, and randomly. During the day, irradiance peaks around noon and drops to zero after sunset. Seasonally, production is highest in summer and lowest in winter due to sun angle and day length.

PV module temperature also plays a significant role—higher temperatures reduce module efficiency, decreasing energy production under the same irradiance. Weather phenomena such as clouds, precipitation, fog, and air pollution also affect the amount of radiation reaching the module surface.

Short-term fluctuations in production can occur within minutes, e.g., due to passing clouds. These are difficult to predict using classical methods and present challenges for grid operators.

Production is also influenced by:

- Panel orientation and tilt relative to the sun,
- Shading (e.g., by trees, buildings, or other objects),
- Module degradation over time (efficiency loss),
- Panel surface soiling (dust, leaves, snow, etc.).

Due to these factors, accurate PV production modeling and forecasting requires many variables and the ability to handle complex, nonlinear dependencies. Traditional statistical methods often fall short, so artificial intelligence models—particularly neural networks—are increasingly used due to their ability to adapt to nonstationary and nonlinear data.

This thesis focuses on the most important factor influencing production—GHI. Although DHI was not included in the current model version, it should be added in future development to improve forecast accuracy.

1.3 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational structures inspired by the structure and function of the human brain. They consist of layers of neurons—data-processing units—interconnected by weights. These networks can model complex nonlinear relationships and are successfully used in many fields, including image recognition, natural language processing, and time-series forecasting, such as PV energy production.

In the context of PV energy prediction, ANNs are particularly useful due to their ability to detect complex patterns in meteorological and system measurement data. Traditional statistical methods often struggle with the high variability and seasonality of weather data, whereas neural networks can capture both linear and nonlinear correlations, making them highly effective for such tasks.

The most commonly used neural networks for predictive tasks include Multi-Layer Perceptrons (MLPs), Recurrent Neural Networks (RNNs), and their advanced variant, Long Short-Term Memory (LSTM). This thesis uses an MLP model, which, despite its simplicity, proved sufficient to achieve satisfactory PV energy forecast results.

1.4 Overview of AI Applications in Renewable Energy

Artificial intelligence (AI) is increasingly applied in the energy sector, particularly in renewable energy sources (RES) such as photovoltaics, wind, and hydro power. With the rising share of unstable sources in the energy mix, tools capable of effectively analyzing and forecasting energy-related phenomena are gaining importance.

One of the main applications of AI in RES is energy generation forecasting. Machine learning models such as neural networks, decision trees, support vector machines (SVM), and random forests enable accurate energy production forecasts based on weather data, historical data, and technical information about installations.

Numerous scientific studies document successful implementations of such models. MLP neural networks are used for short-term PV energy forecasting. RNNs and LSTMs are applied for time-series analysis in longer-term forecasts. Hybrid models combining classical statistical approaches with AI methods also perform well, especially with limited data.

Beyond forecasting, AI is also used in other aspects of renewable energy:

- Fault detection and prediction – identifying PV panel faults from inverter and sensor data,
- Energy storage optimization – intelligent charge/discharge management based on forecasted production and consumption,
- PV and hybrid system control – real-time operational decision-making based on environmental and energy data,
- Demand-supply management in microgrids – automatic energy balancing using AI.

- As technology advances and more data become available (e.g., from weather stations, smart meters, or SCADA systems), the role of AI in energy systems will continue to grow. In photovoltaics in particular, where production is highly environment-dependent, machine learning and deep learning methods can significantly improve forecast accuracy and system stability.

2. Methodology and Implementation

2.1 Data Collection and Description

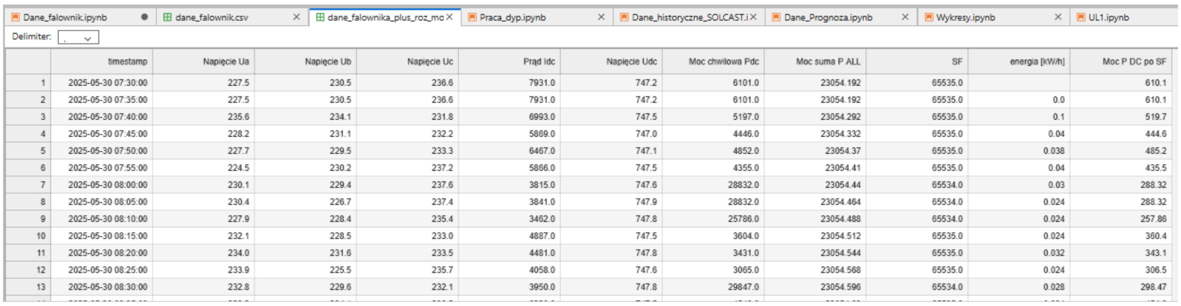
The data used in this work come from a real photovoltaic installation. They were collected in real time from the inverter using the Modbus RTU protocol. The recorded values were then transmitted via an intermediary device to an MQTT server, from which they were downloaded and saved locally in CSV format.

The following data were collected from the inverter:

- Date and time of measurement,
- Voltage and current generated by the installation,
- Instantaneous power (W),
- Total power (total energy produced in kWh).

Ambient and PV module temperatures were not available in the dataset.

An example of the data collected from the inverter:



	timestamp	Napięcie Ua	Napięcie Ub	Napięcie Uc	Prąd Idc	Napięcie Udc	Moc chwilowa Pdc	Moc suma P ALL	SF	energia [kW/h]	Moc P DC po SF
1	2025-05-30 07:30:00	227.5	230.5	236.6	7931.0	747.2	6101.0	23054.192	65535.0		610.1
2	2025-05-30 07:35:00	227.5	230.5	236.6	7931.0	747.2	6101.0	23054.192	65535.0	0.0	610.1
3	2025-05-30 07:40:00	235.6	234.1	231.8	6993.0	747.5	5197.0	23054.292	65535.0	0.1	519.7
4	2025-05-30 07:45:00	228.2	231.1	232.2	5869.0	747.0	4446.0	23054.332	65535.0	0.04	444.6
5	2025-05-30 07:50:00	227.7	229.5	233.3	6467.0	747.1	4852.0	23054.37	65535.0	0.038	485.2
6	2025-05-30 07:55:00	224.5	230.2	237.2	5866.0	747.5	4355.0	23054.41	65535.0	0.04	435.5
7	2025-05-30 08:00:00	230.1	229.4	237.6	3815.0	747.6	28832.0	23054.44	65534.0	0.03	288.32
8	2025-05-30 08:05:00	230.4	226.7	237.4	3841.0	747.9	28832.0	23054.464	65534.0	0.024	288.32
9	2025-05-30 08:10:00	227.9	228.4	235.4	3462.0	747.8	25786.0	23054.488	65534.0	0.024	257.86
10	2025-05-30 08:15:00	232.1	228.5	233.0	4887.0	747.5	3604.0	23054.512	65535.0	0.024	360.4
11	2025-05-30 08:20:00	234.0	231.6	233.5	4481.0	747.8	3431.0	23054.544	65535.0	0.032	343.1
12	2025-05-30 08:25:00	233.9	225.5	235.7	4058.0	747.6	3065.0	23054.568	65535.0	0.024	306.5
13	2025-05-30 08:30:00	232.8	229.6	232.1	3950.0	747.8	29847.0	23054.596	65534.0	0.028	298.47

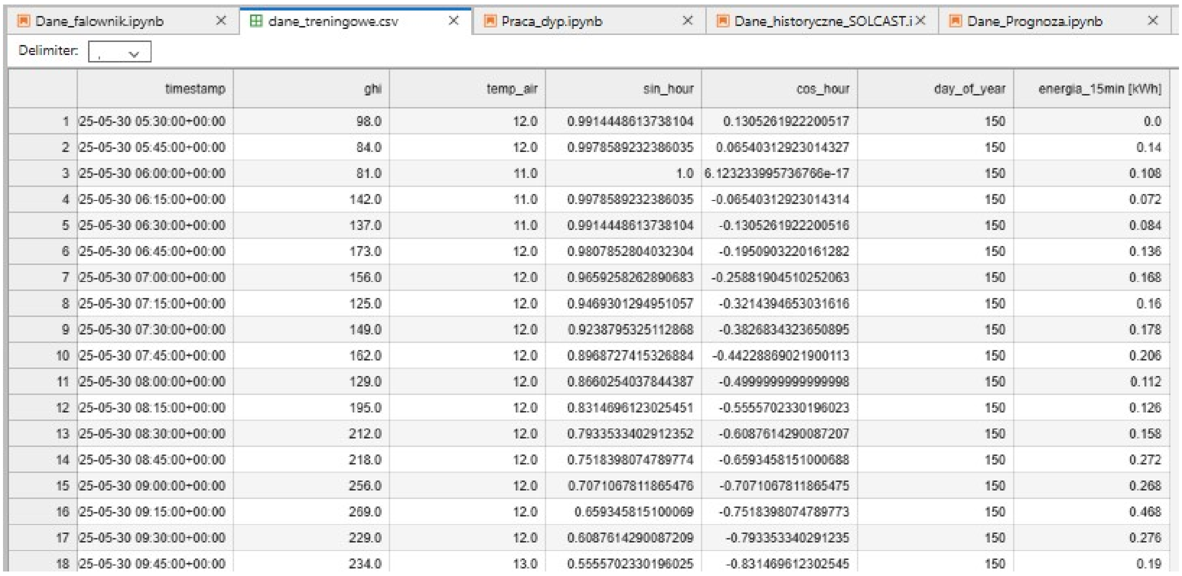
The following data were collected from the inverter:

- Date and time of measurement,
- Voltage and current generated by the installation,
- Instantaneous power (W),
- Total power (total energy produced in kWh).

Ambient and PV module temperatures were not available in the dataset.

Solar irradiance values (global irradiance GHI, expressed in W/m²) and air temperature—key variables for building the predictive model—were obtained from an external source: the historical meteorological database provided by SOLCAST. These data were time-synchronized with the measurements taken from the inverter.

An example of data retrieved from the inverter and from the SOLCAST database, prepared for neural network model training:



Delimiter: ,

	timestamp	ghi	temp_air	sin_hour	cos_hour	day_of_year	energia_15min [kWh]
1	25-05-30 05:30:00+00:00	98.0	12.0	0.9914448613738104	0.1305261922200517	150	0.0
2	25-05-30 05:45:00+00:00	84.0	12.0	0.9978589232386035	0.06540312923014327	150	0.14
3	25-05-30 06:00:00+00:00	81.0	11.0	1.0	6.123233995736766e-17	150	0.108
4	25-05-30 06:15:00+00:00	142.0	11.0	0.9978589232386035	-0.06540312923014314	150	0.072
5	25-05-30 06:30:00+00:00	137.0	11.0	0.9914448613738104	-0.1305261922200516	150	0.084
6	25-05-30 06:45:00+00:00	173.0	12.0	0.9807852804032304	-0.1950903220161282	150	0.136
7	25-05-30 07:00:00+00:00	156.0	12.0	0.9659258262890683	-0.25881904510252063	150	0.168
8	25-05-30 07:15:00+00:00	125.0	12.0	0.9469301294951057	-0.3214394653031616	150	0.16
9	25-05-30 07:30:00+00:00	149.0	12.0	0.9238795325112868	-0.3826834323650895	150	0.178
10	25-05-30 07:45:00+00:00	162.0	12.0	0.8968727415326884	-0.44228869021900113	150	0.206
11	25-05-30 08:00:00+00:00	129.0	12.0	0.8660254037844387	-0.4999999999999998	150	0.112
12	25-05-30 08:15:00+00:00	195.0	12.0	0.8314696123025451	-0.5555702330196023	150	0.126
13	25-05-30 08:30:00+00:00	212.0	12.0	0.7933533402912352	-0.6087614290087207	150	0.158
14	25-05-30 08:45:00+00:00	218.0	12.0	0.7518398074789774	-0.6593458151000688	150	0.272
15	25-05-30 09:00:00+00:00	256.0	12.0	0.7071067811865476	-0.7071067811865475	150	0.268
16	25-05-30 09:15:00+00:00	269.0	12.0	0.659345815100069	-0.7518398074789773	150	0.468
17	25-05-30 09:30:00+00:00	229.0	12.0	0.6087614290087209	-0.793353340291235	150	0.276
18	25-05-30 09:45:00+00:00	234.0	13.0	0.5555702330196025	-0.831469612302545	150	0.19

The collected data covered a 4-day period and were saved in CSV files. Before analysis, they were preliminarily verified for completeness and temporal consistency. The data were then split into training and test sets, prepared for further processing in the predictive model.

2.2 Data Preparation (Cleaning, Normalization, Feature Engineering)

After collecting data from the inverter and the SOLCAST platform, the dataset was preprocessed for analysis. The data were initially checked for completeness, temporal alignment, and the presence of outliers.

Missing values (e.g., brief data loss from MQTT) were either removed or imputed using linear interpolation. The data were then time-synchronized, allowing for precise matching of irradiance values with corresponding inverter measurement timestamps.

Initially, the predictive model considered the following input variables:

- Hour of measurement (numerical value from 0 to 23),
- Global horizontal irradiance (GHI) in W/m²,
- Instantaneous power in W,
- Cumulative energy in kWh,
- Air temperature.

However, due to observed time shifts caused by transmission delays from the inverter via MQTT and the high dynamics of solar irradiance changes, better modeling results were achieved using a derived variable—partial energy production, calculated as the difference in cumulative energy every 5 minutes.

This helped reduce the impact of momentary synchronization errors and yielded a more stable and representative output variable for machine learning.

Input variables were normalized using the min-max method, transforming the data into the [0, 1] range. This rescaling was necessary because neural networks perform optimally when all input features are on a similar scale.

As part of basic feature engineering, additional variables such as day of the year (to capture seasonality) and diffuse horizontal irradiance (DHI) were also considered. Although DHI was available in the SOLCAST dataset, it was mistakenly omitted during initial model development. Including such variables is recommended in future model iterations, especially for more advanced analyses.

In the final model version, the following input variables were used:

- Global irradiance (GHI),
- Air temperature,
- Temporal coordinates sin_hour and cos_hour, which help better reflect the daily cyclicality of energy production.

For the purpose of energy production forecasting, a separate dataset of forecast irradiance values from the SOLCAST platform was also prepared. These data allow testing the predictive model in real-world conditions—i.e., when only the forecast irradiance is known, and the actual PV production is not yet available.

2.3 Neural Network Model Construction

An artificial neural network (ANN) was used to predict photovoltaic energy production, with its architecture tailored to the data characteristics and short-term forecast horizon. The model aimed to forecast the amount of electricity generated over the next 15 minutes based on current weather parameters and time.

The model took the form of a classic feedforward MLP (Multi-Layer Perceptron) designed as a regression model. Its architecture consisted of:

- An input layer with four neurons corresponding to:
    - Global solar irradiance (GHI),
    - Air temperature (temp_air),
    - Encoded hour using sine (sin_hour) and cosine (cos_hour) functions,
- Two hidden layers:
    - First layer: 128 neurons with ReLU activation,
    - Second layer: 64 neurons with ReLU activation,
- An output layer with one neuron returning the predicted energy production value [kWh].

The network was trained using the backpropagation algorithm with the Adam optimizer, ensuring stable and fast convergence. Mean Squared Error (MSE) was used as the loss function, while Mean Absolute Error (MAE) was monitored as an additional metric.

Input data were standardized using StandardScaler, which allowed the network to properly handle features with different units and scales.

The dataset was split into:

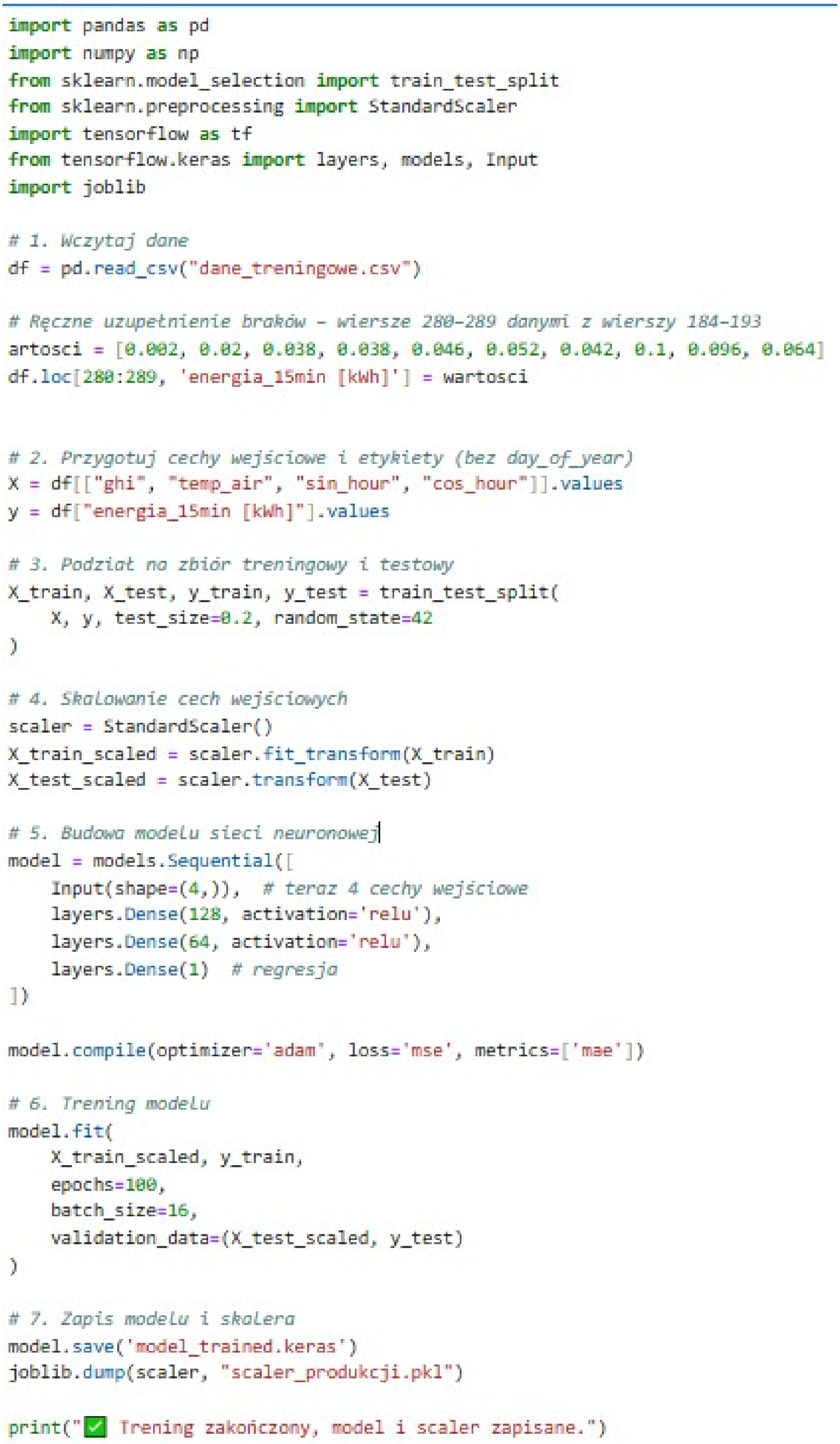
- Training set (80%) – for model learning,
- Test set (20%) – for evaluating overall prediction performance.

Model quality metrics on the test data:

- MAE: 0.056 kWh
- RMSE: 0.082 kWh
- R^2: 0.896

These results confirm that the model generalizes well and can be effectively used for short-term photovoltaic energy production forecasting based on meteorological data.

The training script for the model is shown below:



```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
import tensorflow as tf
from tensorflow.keras import layers, models, Input
import joblib

# 1. Wczytaj dane
df = pd.read_csv("dane_treningowe.csv")

# Ręczne uzupełnienie braków - wiersze 280-289 danymi z wierszy 184-193
artosci = [0.002, 0.02, 0.038, 0.038, 0.046, 0.052, 0.042, 0.1, 0.096, 0.064]
df.loc[280:289, 'energia_15min [kWh]'] = wartosci


# 2. Przygotuj cechy wejściowe i etykiety (bez day_of_year)
X = df[["ghi", "temp_air", "sin_hour", "cos_hour"]].values
y = df["energia_15min [kWh]"].values

# 3. Podział na zbiór treningowy i testowy
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

# 4. Skalowanie cech wejściowych
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# 5. Budowa modelu sieci neuronowej
model = models.Sequential([
    Input(shape=(4,)),  # teraz 4 cechy wejściowe
    layers.Dense(128, activation='relu'),
    layers.Dense(64, activation='relu'),
    layers.Dense(1)  # regresja
])

model.compile(optimizer='adam', loss='mse', metrics=['mae'])

# 6. Trening modelu
model.fit(
    X_train_scaled, y_train,
    epochs=100,
    batch_size=16,
    validation_data=(X_test_scaled, y_test)
)

# 7. Zapis modelu i skalera
model.save('model_trained.keras')
joblib.dump(scaler, "scaler_produkcji.pkl")

print("✅ Trening zakończony, model i scaler zapisane.")
```

2.4 Tools and Work Environment

All computations, data processing, and model training were performed in the Python 3.12.7 environment using popular data analysis and machine learning libraries:

- pandas – for processing and analyzing tabular data (CSV),
- numpy – for numerical operations,
- scikit-learn – for data splitting, metrics, and preliminary models,
- TensorFlow/Keras – for building and training the neural network,
- matplotlib – for visualizing prediction results,
- joblib – for serializing trained models and loading them efficiently.

The working environment was configured on a local computer running Windows 11, equipped with an Intel Core i9 processor, 32 GB RAM, and an NVIDIA RTX 3060 graphics card, which allowed for comfortable model training and rapid experimentation with neural network architectures.

The code was developed and executed in the Jupyter Notebook environment, which facilitated easy documentation and step-by-step testing of each phase.

To collect data from the photovoltaic installation, a LAN Controller 3.0 was used. It was responsible for reading inverter parameters and transmitting them to the MQTT server. Additionally, TinyControl SW 1.62 software was used for device configuration and monitoring, enabling continuous and stable real-time data measurement.

3. Experiments and Results Analysis

3.1 Model Training and Validation

The artificial neural network model was trained on real production data synchronized with irradiance data obtained from the SOLCAST platform. The data were split into training and validation sets in an 80:20 ratio. The learning process was carried out experimentally, testing various model configurations to achieve the best possible fit and to maximize the R^2 coefficient of determination. During the experiments, the following parameters were modified:

- Number of neurons in hidden layers,
- Number of layers,
- Number of training epochs,
- Batch size.

The Adam optimization algorithm was used, with Mean Squared Error (MSE) applied as the loss function.

Thanks to an iterative approach to hyperparameter tuning, it was possible to significantly improve prediction accuracy. During validation, a clear decrease in MSE was observed, indicating effective model fitting to the data. Preliminary results suggest that the model successfully learned to replicate the general dynamics of energy production variability depending on irradiance, temperature, and time of day.

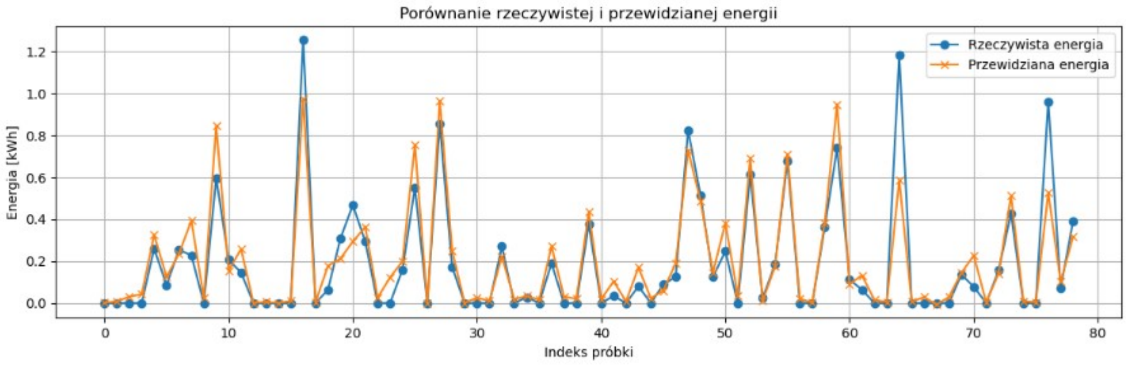


Figure 1. Comparison of Actual and Predicted Energy (Energy [kWh] vs. Sample Index)

In Figure 1, actual values (blue dotted line) and model predictions (orange line with crosses) are compared over sequential time samples.

Insights from the chart analysis:

1. **Trajectory convergence**: The general shape of both curves is very similar, indicating good model alignment with the actual data.
2. **Extreme values reproduction**: The model accurately captures energy production peaks, though it does not always match their exact magnitude.
3. **Low-production forecasts**: For values close to zero, the model maintains high precision and avoids generating falsely elevated predictions.
4. **Largest deviations**: Discrepancies between actual and predicted values are most visible during sudden changes, especially for higher values above 1 kWh. Possible reasons include:
   - Rapidly changing weather conditions (e.g., partial cloud cover followed by sudden clearing),
   - Limited number of features included in the model, preventing full reflection of complex atmospheric influences,
   - Low number of observations for high-irradiance days (most data were collected during periods with low GHI).

The model maintains a high level of agreement with actual values, confirming that the regressor effectively learned the key patterns linking input conditions to energy production.

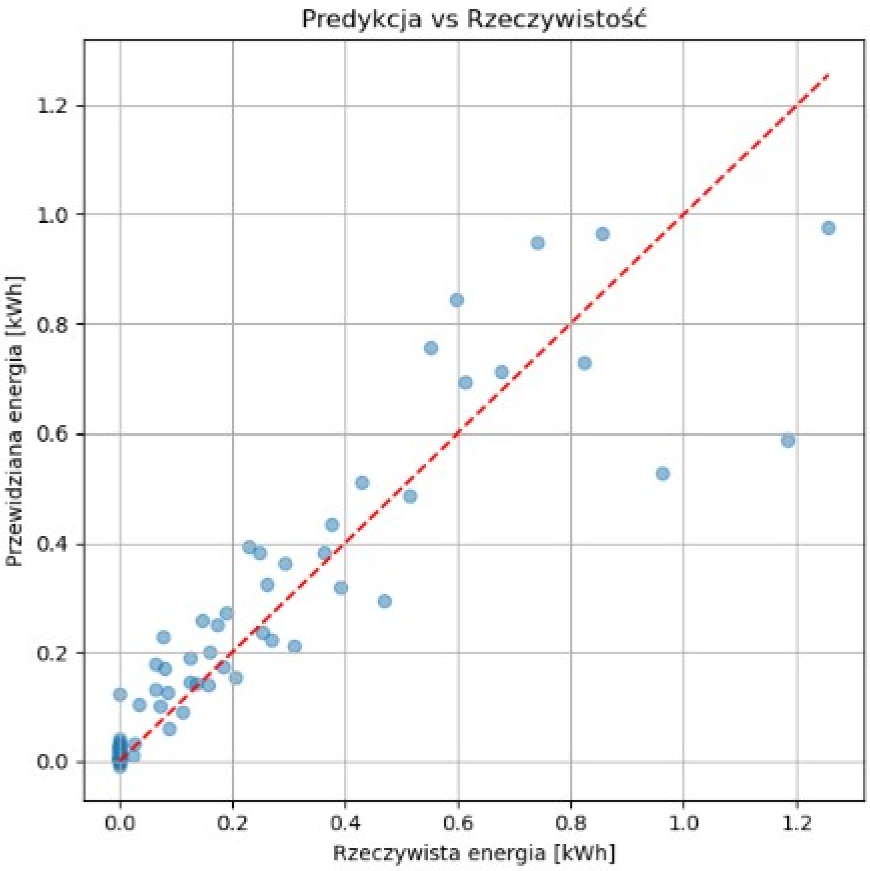


Figure 2. Scatter Plot: Actual vs. Predicted Energy Production [kWh]

Figure 2 shows a scatter plot comparing the actual and predicted values of energy production (in kWh) by the predictive model. The red dashed line represents perfect agreement (y = x).

Insights from the chart analysis:

1. **Closeness to the y = x line**: Most points lie close to the y = x line, indicating good model fit and effective mapping of input variables to energy output.
2. **Concentration in the lower range (0–0.2 kWh)**: A high density of points appears in the lower range, consistent with the training dataset which contained more samples from low-irradiance days.
3. **Greater dispersion for higher values (>0.6 kWh)**: A larger spread of points is observed for higher energy values, which may indicate:
   - Increased data variability under high solar irradiance conditions,
   - Insufficient number of samples from high-irradiance days (most data came from periods with lower GHI).

Despite these differences, the overall agreement between predictions and reality remains high, and the model successfully captured the key patterns in the data.

3.2 Model Quality Metrics (MAE, RMSE, R²)

The model's performance was evaluated using standard regression metrics:

- **MAE (Mean Absolute Error)** – the average magnitude of prediction errors,
- **RMSE (Root Mean Square Error)** – the square root of the average squared errors,
- **R² (Coefficient of Determination)** – a measure of how well the model fits the actual data.

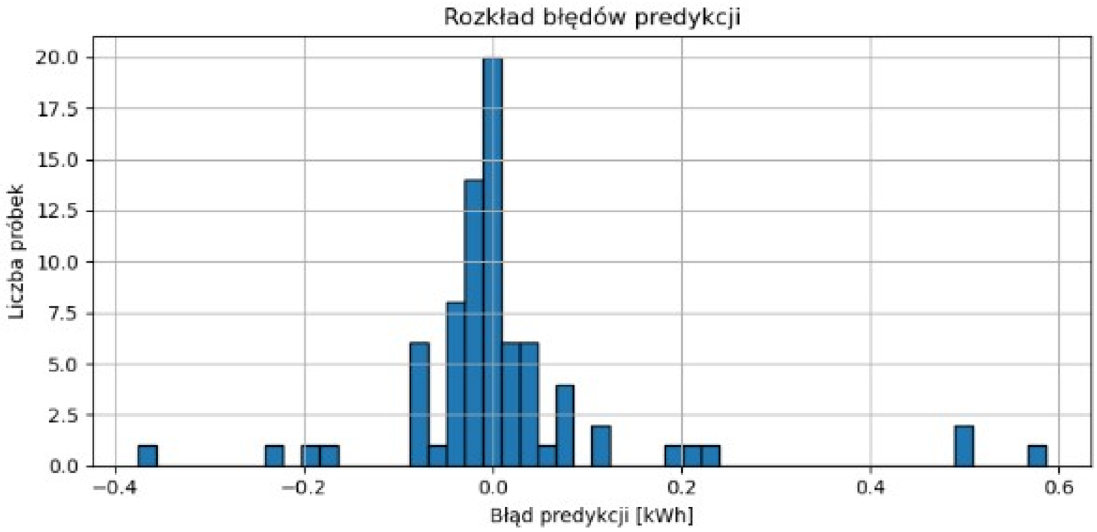


Figure 3. Prediction Error Distribution

Figure 3 shows a histogram illustrating the distribution of the regression model's prediction error—that is, the difference between predicted and actual energy production (in kWh).

Insights from the chart analysis:

1. **Concentration around zero**: Most errors are centered around zero, indicating that the model predicts energy values with high accuracy in many cases.
2. **Moderately symmetric distribution**: The error distribution is moderately symmetric, although some outliers are visible on both the underestimation and overestimation sides.
3. **High precision in typical conditions**: The majority of samples fall within the ±0.05 kWh error range, which demonstrates the model's high precision under typical conditions.
4. **Larger errors above 0.4 kWh**: These are likely due to an insufficient number of training samples recorded under high irradiance (GHI) values, limiting the model's ability to accurately represent system behavior in strong sunlight conditions.

The chart confirms that the model generally exhibits good prediction quality and stability, although further improvement and data expansion—especially under high irradiance conditions—are recommended.

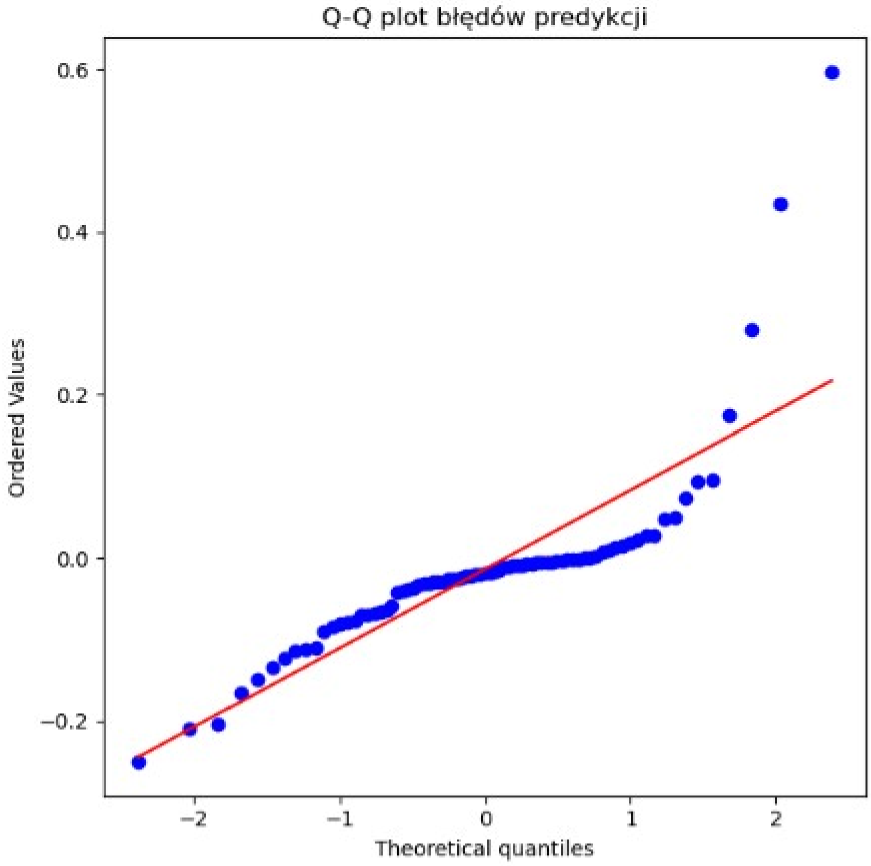


Figure 4. Q-Q Plot of Prediction Errors

Figure 4 presents a Q-Q plot used to compare the empirical distribution of the model's prediction errors with the theoretical normal distribution.

Insights from the chart analysis:

1. **Alignment along the reference line**: Most points lie along the red reference line, suggesting that the error distribution closely resembles a normal distribution in the central range (i.e., for smaller errors).
2. **Deviations in the tails**: Departures at both ends of the plot (distribution tails) indicate the presence of outliers—rare, larger prediction errors.
3. **Points above the line on the right**: These confirm several instances of significant overestimation of energy production by the model.
4. **Possible causes of outliers**:
   - Too few examples with high irradiance (GHI),
   - Irregularities in the input data (e.g., sudden irradiance changes not captured by the model),
   - Underestimation of model complexity needed in extreme conditions.

In summary, the model demonstrates acceptable statistical properties, but its error distribution is not perfectly normal—suggesting that further optimization or dataset expansion could improve the stability of predictions, particularly in extreme cases.

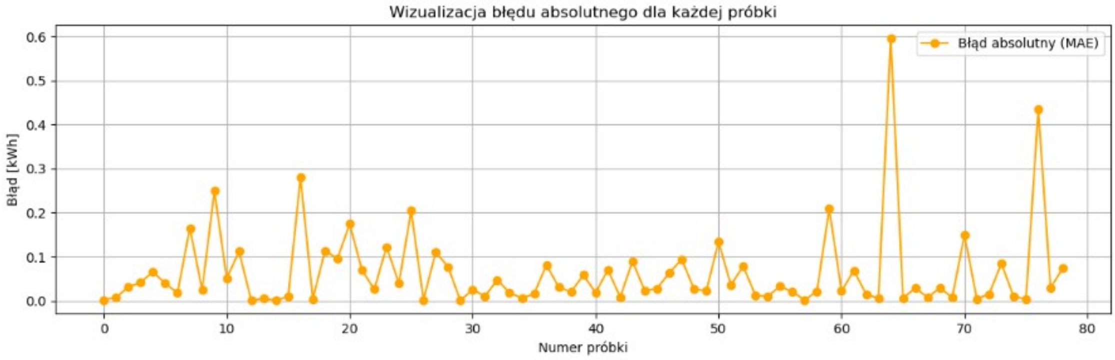


Figure 5. Visualization of Absolute Error (MAE) per Sample

Figure 5 shows the absolute error (Mean Absolute Error, MAE) for each data sample. The X-axis represents the sample index, and the Y-axis indicates the error value in kWh.

Insights from the chart analysis:

1. **Low error for most samples**: The majority of samples exhibit low error (below 0.1 kWh), confirming the generally good prediction quality.
2. **Isolated error peaks**: A few prominent peaks (e.g., around samples 60 and 75) exceed 0.4 kWh, indicating instances where the model significantly under- or overestimated energy production.
3. **Possible causes of high errors in selected samples**:
   - Sudden changes in solar irradiance during the day (e.g., passing clouds),
   - Too few similar examples in the training dataset,
   - Atypical atmospheric conditions not well captured by the model.

Summary: The model demonstrates stable and low prediction errors in most cases. However, for a few samples, further analysis or expansion of the training dataset (particularly with high GHI days) may be necessary to improve accuracy under extreme conditions.

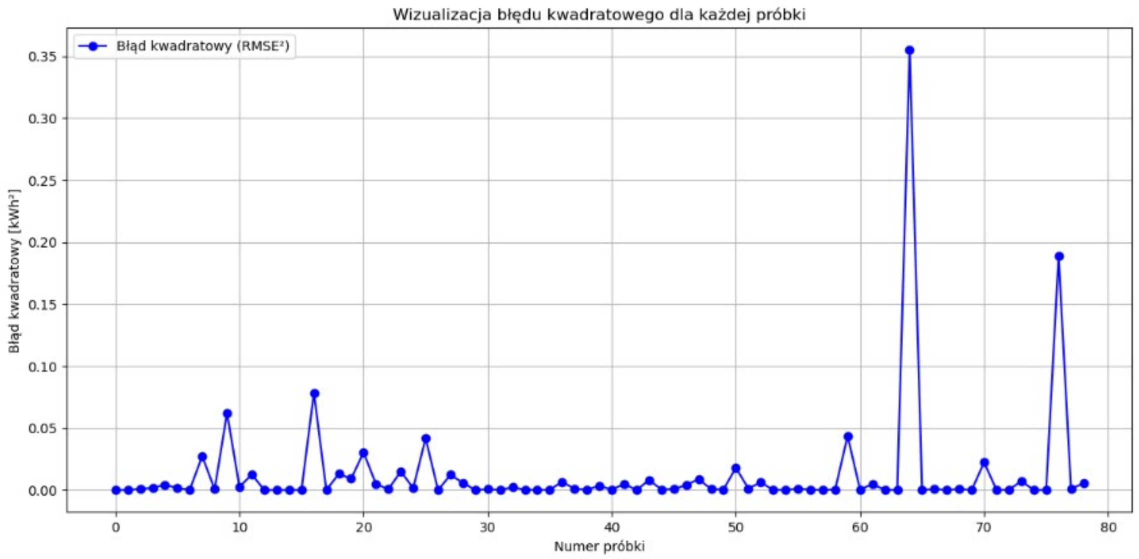


Figure 6. Visualization of Squared Error (RMSE²) per Sample

Figure 6 presents the squared error (RMSE²) values for each validation data sample. The X-axis shows the sample index, and the Y-axis represents the squared error in kWh².

Insights from the chart analysis:

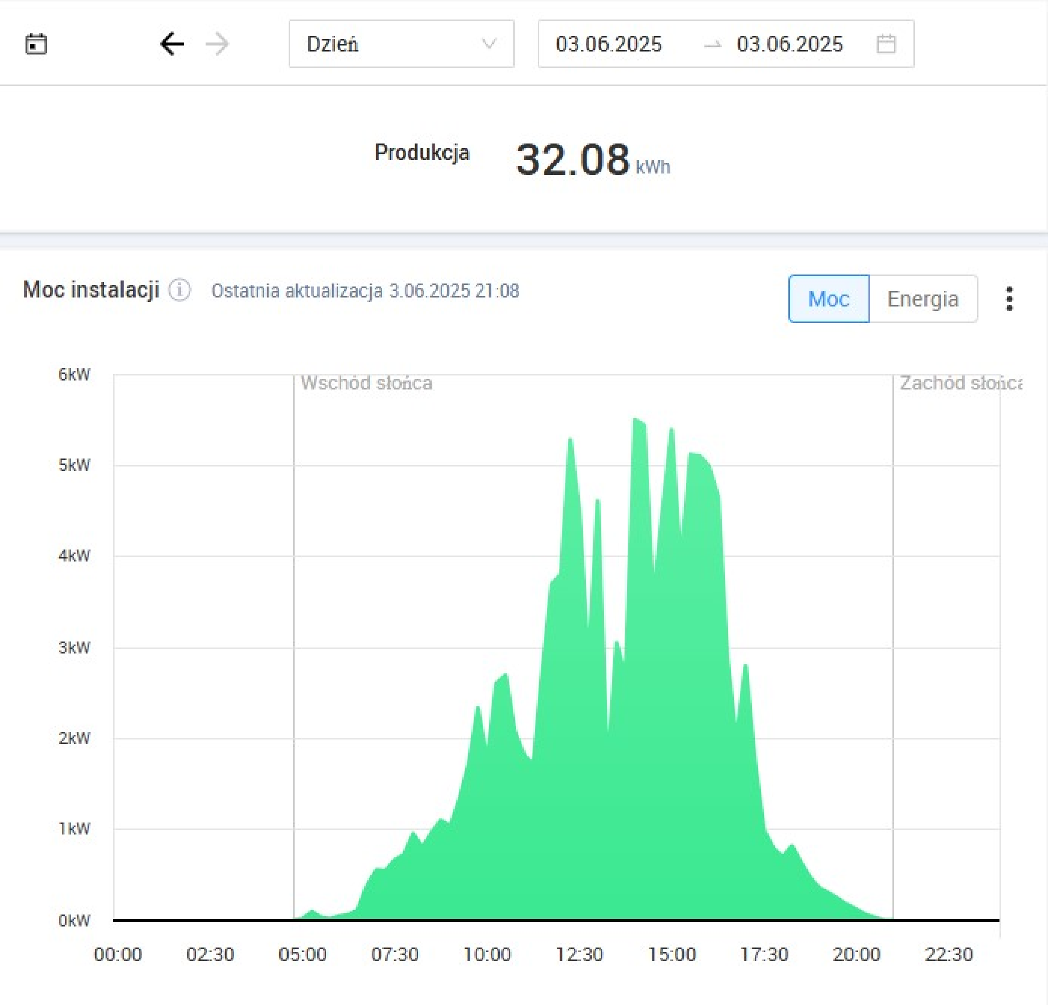
1. **Low squared error in most cases**: The majority of samples have very low squared error values, indicating high prediction quality and low deviation between actual and predicted values.
2. **Two prominent error spikes**: Notable spikes appear around samples 61 and 75, significantly exceeding other values. This means that larger prediction errors occurred in those instances.
3. **High RMSE² values suggest**:
   - Large deviation of predictions from actual values,
   - Presence of extreme weather conditions,
   - Possibly insufficient number of representative training examples, especially under high GHI conditions.

Summary: The squared error chart confirms that the model generalizes well for most cases. The largest deviations are rare and may be reduced by enriching the dataset with high-irradiance days or through further hyperparameter tuning.

Preliminary analysis indicates relatively low MAE and RMSE values and a high R² score, suggesting that the model effectively captures the variability of actual energy production. Final metric values will be provided after complete evaluation.

3.3 Validation and Visualization of Results

3.3.1 Comparison of Daily Prediction with SolarEdge Data



Dzień

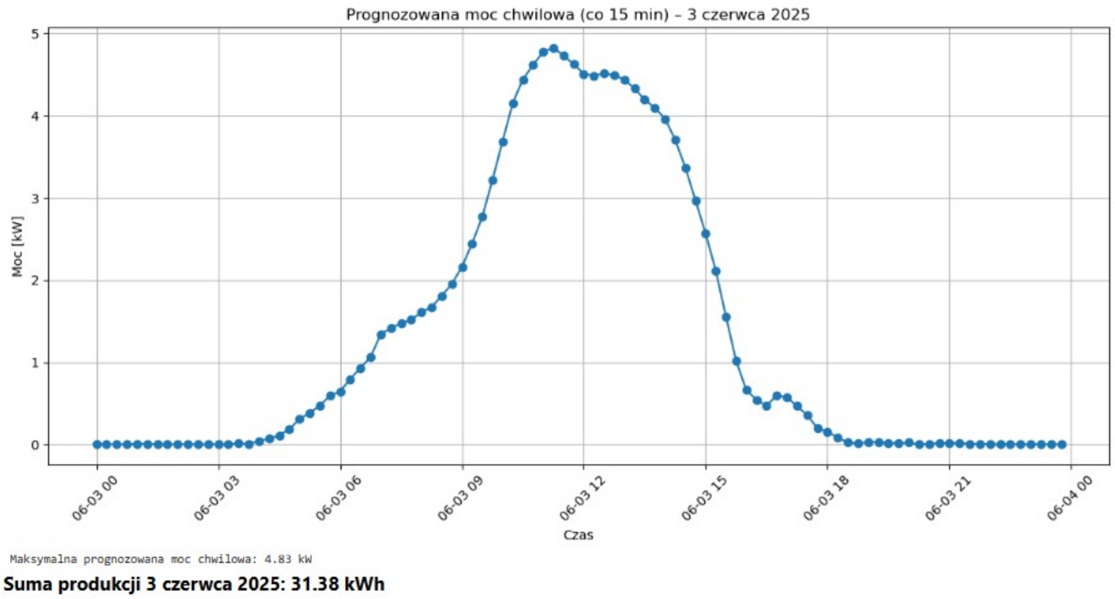
03.06.2025 → 03.06.2025

Produkcja 32.08 kWh

Moc instalacji ⓘ Ostatnia aktualizacja 3.06.2025 21:08

Figure 7. Daily Energy Production Chart from SolarEdge System

Figure 7 illustrates the daily electricity production generated by the photovoltaic installation, based on monitoring data from the SolarEdge system.



Maksymalna prognozowana moc chwilowa: 4.83 kW

Suma produkcji 3 czerwca 2025: 31.38 kWh

Figure 8. Forecasted Energy Production Chart by the Predictive Model

Figure 8 presents the forecasted electricity production generated by the predictive model, represented as the sum of 15-minute interval predictions aggregated into hourly cycles. Although a precise comparative analysis would require inverter data with high time resolution, the presented charts allow for an approximate comparison of total production and daily trends. This enables the assessment of the model's consistency with the actual energy production profile.

Findings from the comparison of actual and forecasted energy production charts:

1. **Daily trend alignment:** Both charts show a similar daily pattern—production increases from morning hours, peaks around noon, and gradually decreases toward the evening. This indicates that the model captures the general trend driven by solar movement and daily irradiance.
2. **Close match in total daily production:** The actual production on June 3, 2025, was 32.08 kWh, while the model predicted 31.38 kWh—a difference of only 0.7 kWh, which corresponds to less than 2.2%, indicating strong predictive performance.
3. **Differences in instantaneous power profiles:** The model produces a smooth, symmetric power curve, whereas the actual SolarEdge chart exhibits significant fluctuations due to transient cloud cover, shading, and other atmospheric factors. The model does not capture such micro-disturbances, which is a natural limitation when high-resolution cloud cover data is unavailable.
4. **Comparative limitations:** The model chart shows aggregated 15-minute power values, while the SolarEdge data includes real-time measurements at shorter intervals. Thus, point-by-point comparison is limited, but the graphs are sufficiently accurate for evaluating total production and daily patterns.
5. **Impact of data resolution:** With 5-minute input and output data intervals (instead of 15), the predictive model could better reflect short-term power variations, increasing accuracy. Higher temporal resolution allows more precise tracking of irradiance changes caused by cloud movement.
6. **Practical conclusions:** The predictive model is suitable for planning daily energy output, forecasting yields, and supporting energy management in photovoltaic

systems. Its current limitations mainly stem from the absence of high-frequency atmospheric data and lower sampling resolution.

4. Conclusions and Future Development Directions

4.1 Summary of Obtained Results

This study developed and tested a neural network model for forecasting photovoltaic energy production. The model was based on real-world data and irradiance forecasts from the SOLCAST platform, showing satisfactory predictive properties and operational stability.

The model utilized both historical and real-time inverter data, including voltage, current, instantaneous power values, and total produced energy, recorded every 5 minutes. Each record also contained a timestamp, which improved the model's ability to represent actual operating conditions and allowed partial production to be calculated as the difference between subsequent cumulative readings.

Input variables included global horizontal irradiance (GHI), ambient temperature, and time-transformed features such as sin_hour and cos_hour to represent the diurnal cycle. Although other variables like day of the year or diffuse irradiance (DHI) were considered, DHI was unintentionally omitted at this stage. Future iterations may incorporate it.

It was also observed that model performance could improve with a shorter input interval (e.g., 5 minutes instead of 15) to better capture brief irradiance fluctuations—such as clouds passing over the PV array.

It is important to note that due to the diversity of PV system configurations (tilt angles, orientations, shading), each installation may require a dedicated model. Nevertheless, a neural network trained on inverter historical data shows potential for adaptation to individual installations if trained on sufficient data.

4.2 Final Conclusions

Using artificial neural networks to forecast energy production in small-scale PV systems proved effective and feasible under real-world conditions. The results indicate that even a simple input set (global irradiance, temperature, time features) yields sufficiently accurate predictions.

The model successfully learned the relationship between weather variables and energy production, confirmed by performance metrics (MAE, RMSE, R^2) and visual chart analysis. Appropriate architecture and basic hyperparameter tuning (epochs, neuron count) yielded strong alignment between predicted and actual values.

Experiments confirm that machine learning—and neural models in particular—can support PV system management by optimizing energy use, battery charging, and grid export. Future research may extend the dataset (e.g., with DHI, humidity, cloud cover) and explore alternative architectures or automated hyperparameter tuning.

4.3 Future Research and Improvement Opportunities

Potential directions for further development of the predictive model include expanding the input dataset and applying more advanced machine learning techniques. Specifically:

1. **Adding new input features**, such as:
   - DHI (Diffuse Horizontal Irradiance), providing extra insight under cloudy conditions;
   - PV module temperature (e.g., measured locally), affecting conversion efficiency;
   - Air humidity, possibly correlating with irradiance-limiting conditions;
   - Seasonal variables reflecting the yearly cycle (e.g., day of year as sin_day and cos_day), helping the model adapt to solar position and daylight length variations;
   - Installation-specific features like azimuth, tilt, and shading. Note that some of these may already be indirectly captured by time variables (sin_hour, cos_hour), which encode the angle of solar incidence and can reflect site-specific shading patterns.

   If sufficient historical data exists, the model can learn that, e.g., production is lower at 8:00 a.m. despite high GHI due to shading. Such patterns, if seen before, will inform future predictions. Thus, sin_hour and cos_hour convey both time and site-specific irradiance profile information. However, rare events (e.g., occasional morning shading) require more training examples for the model to generalize well.

2. **Using recurrent neural networks (e.g., LSTM or GRU)**, tailored for time-series data, to better capture hourly and daily production continuity and weather variation trends.
3. **Employing hyperparameter tuning tools**, such as GridSearch, RandomSearch, or Optuna, to optimize neuron count, layer structure, learning rate, and epoch number.
4. **Integrating real-time weather data**, via APIs like SOLCAST, to generate forecasts using the most current meteorological conditions.
5. **Developing a web application or energy management system**, using model predictions for decision support—e.g., energy storage timing, heat pump control, or smart load balancing.
6. **Training dedicated models for each installation**, taking into account unique characteristics. However, a model based on inverter historical data can adapt and scale if trained with sufficient local examples.

Incorporating these improvements would yield a more universal, accurate, and practical tool for forecasting PV energy production across diverse sites, system sizes, and weather conditions.