A Deep Learning approach for Time Series Imputation on Photovoltaic data

Nicolò Vescera

November 9th, 2023

Università degli Studi di Perugia

Intro

Problem Introduction

The growing need for generating clean energy from renewable sources has resulted in extensive data collection.

However, these data often contain gaps and deficiencies.

Accurate imputation of these gaps is essential to ensure the reliability of analyses and predictions based on this data.

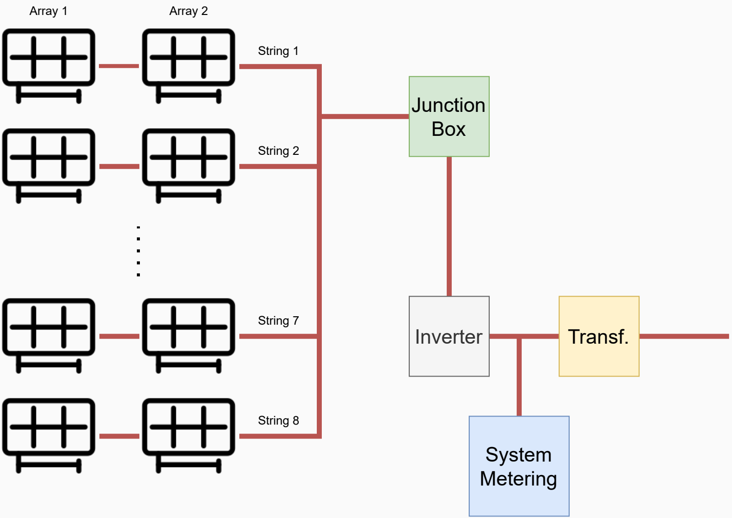
Problem Definition

Given a set of N time series $S = \{S_1, S_2, \ldots, S_N\}$, where each $S_i = (t_i, v_i)$, represents the moment t_i when v_i was recorded,

$S^* \in S$, a specific target series that represent the *Total generated energy*,

The objective of the **imputation problem** is to **estimate eventual missing values** $v_j, \ldots, v_{j+h} \in S^*$, using other time points temporally adjacent to t_j and t_{j+h}.

Photovoltaic Implant



Original Dataset

Files

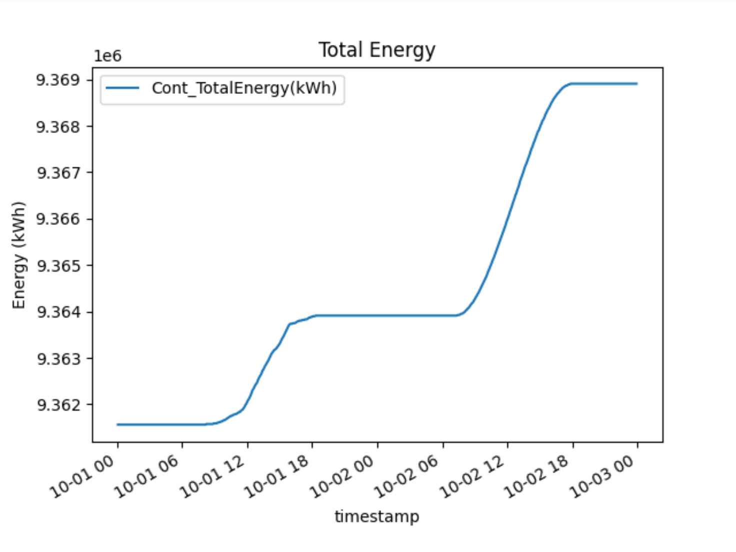
- 2022_02_02_inverter.csv
- 2022_02_02_meter.csv
- 2022_02_02_platDevice.csv
- 2022_02_02_stringbox.csv

⋮

- 2023_06_16_inverter.csv
- 2022_06_16_meter.csv
- 2022_06_16_platDevice.csv
- 2022_06_16_stringbox.csv

Start Date	Feb. 02, 2022
End Date	June 16, 2023
Files	2,814
File Type	6
Size	54.3 MB

Target Feature



Data Preprocessing

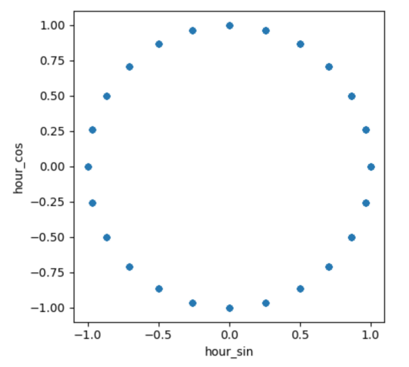
Monolithic Table

timestamp	$\text{DEV.NAME}_1\text{\_FEAT}_1$	...	$\text{DEV.NAME}_n\text{\_FEAT}_n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
01/10/2022 10:00	...	...	...
01/10/2022 10:05	...	...	...
01/10/2022 10:10	...	...	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$

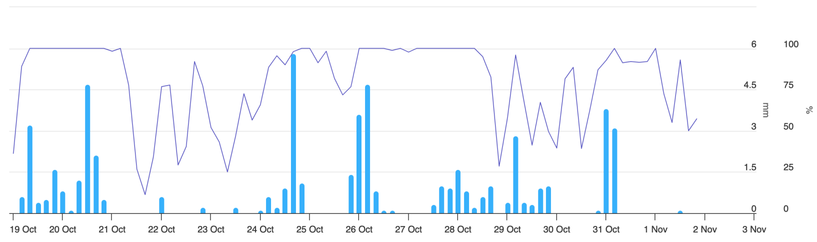
Rows (#)	143,430
Columns (#)	733
Size (MB)	463.9

Timestamp Cyclical Encoding

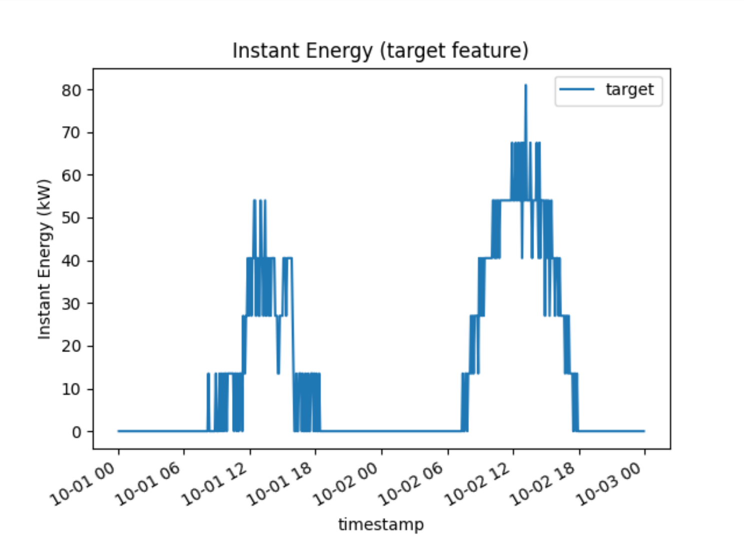
We cyclically encoded timestamps using sine and cosine values to capture time patterns for minutes, hours, days, and years.



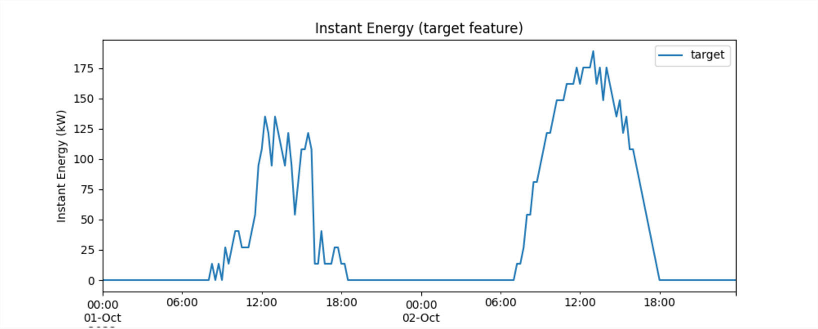
Historical Weather



Target feature



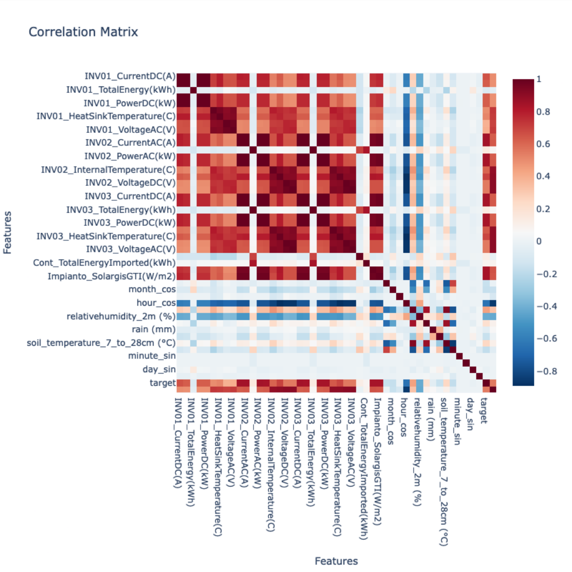
Re-sampling



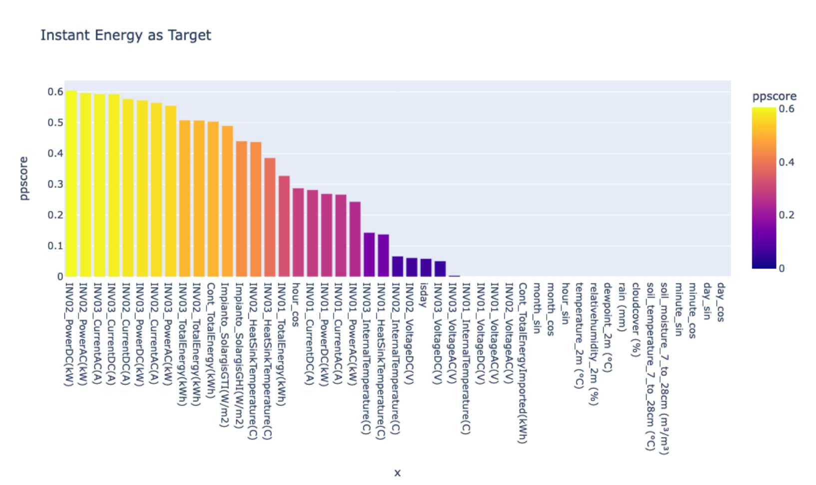
Instant Energy (target feature)
target
Instant Energy (kW)
175
150
125
100
75
50
25
0
00:00
01-Oct
06:00
12:00
18:00
00:00
02-Oct
06:00
12:00
18:00

Feature Selection

Correlation Matrix



Power Predictive Score (PPS)



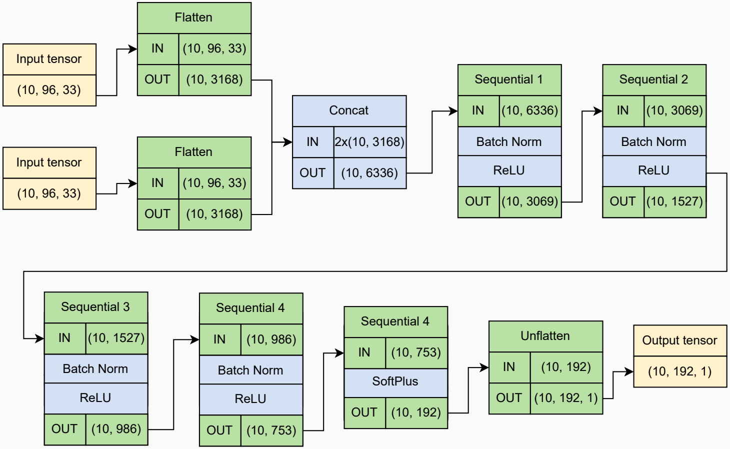
Resulting Dataset

	Resulting Dataset	*Pre-processed Dataset*
Rows (#)	46,537	143,430
Columns (#)	33	763
Size (MB)	14.5	543.7

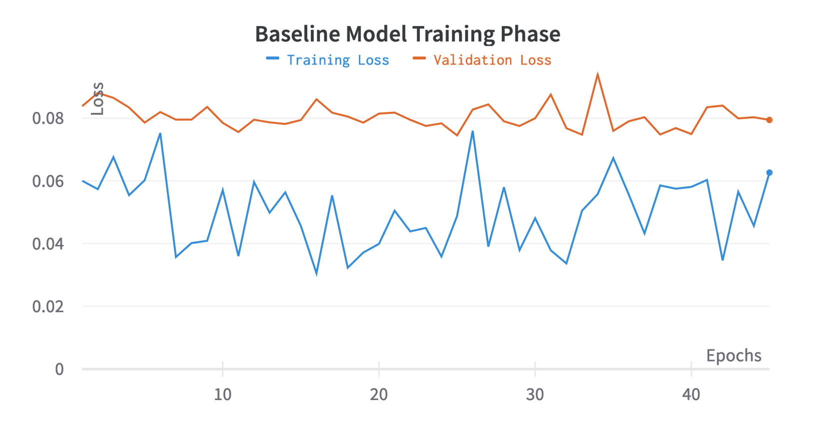
	Start	End	Rows
Train	June 01, 2022	February 28, 2023	24,864
Val	March 01, 2023	March 31, 2023	2,880
Test	April 01, 2023	April 30, 2023	2,880

Deep Learning Models

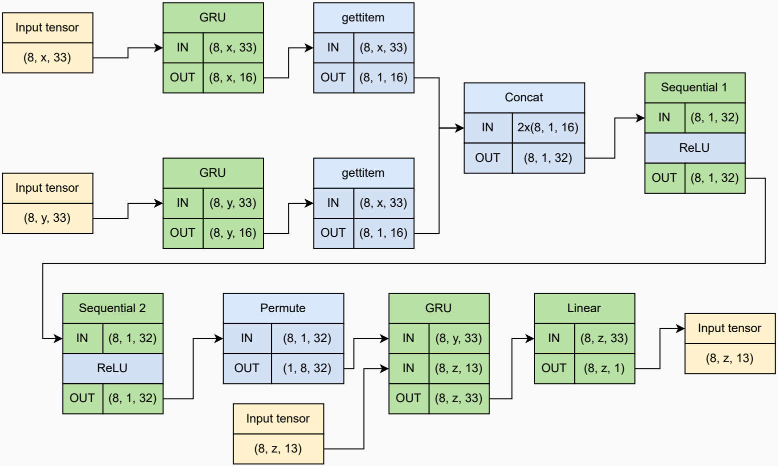
MLP-based Model Architecture



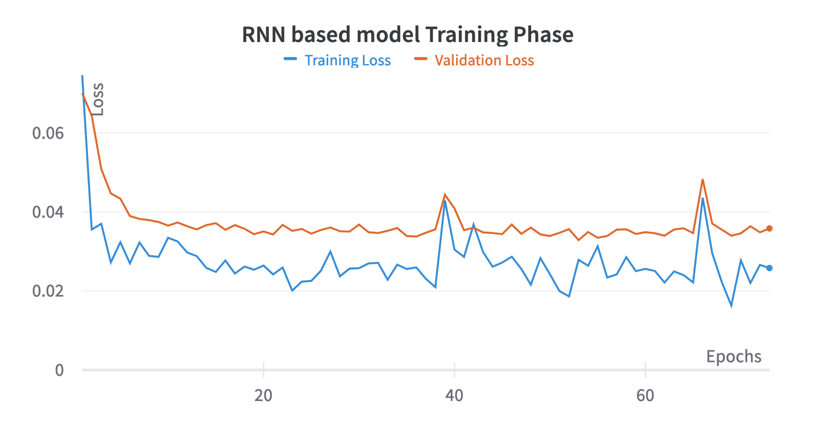
Training Phase



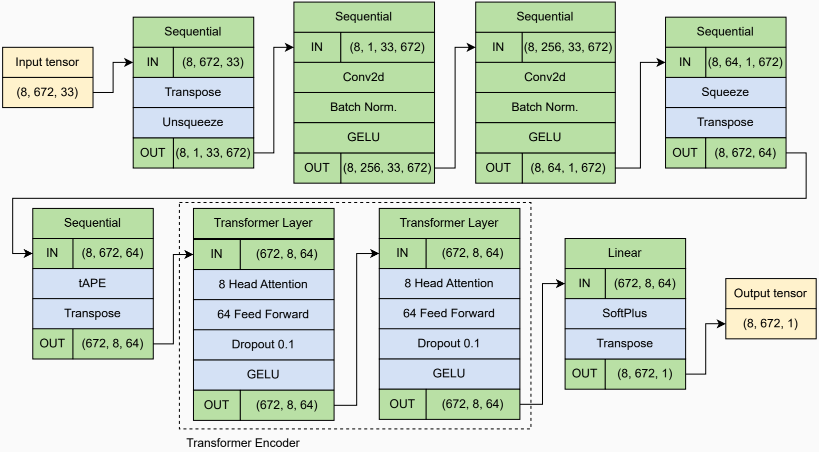
RNN-based Model Architecture

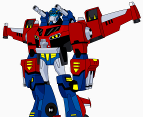


Training Phase

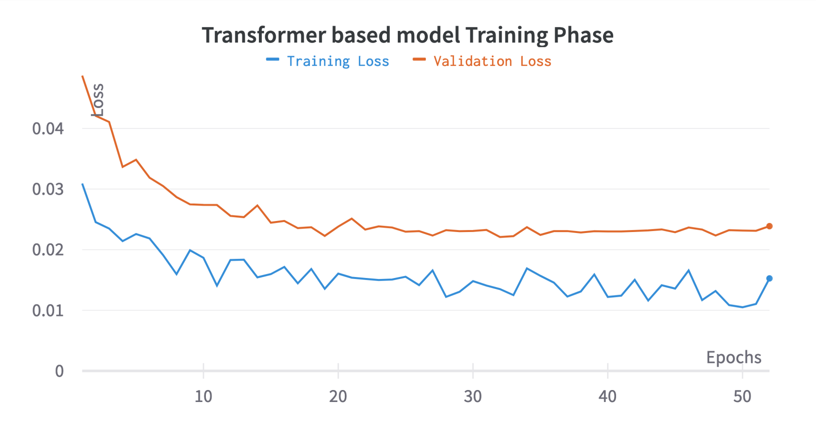


Transformer-based Model Architecture

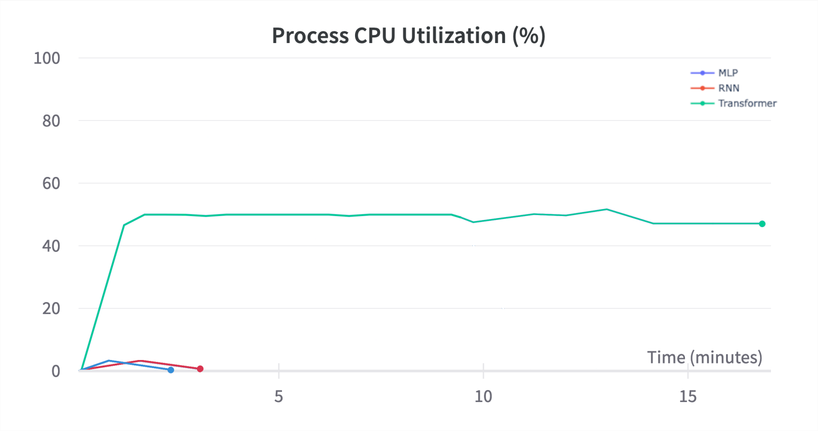




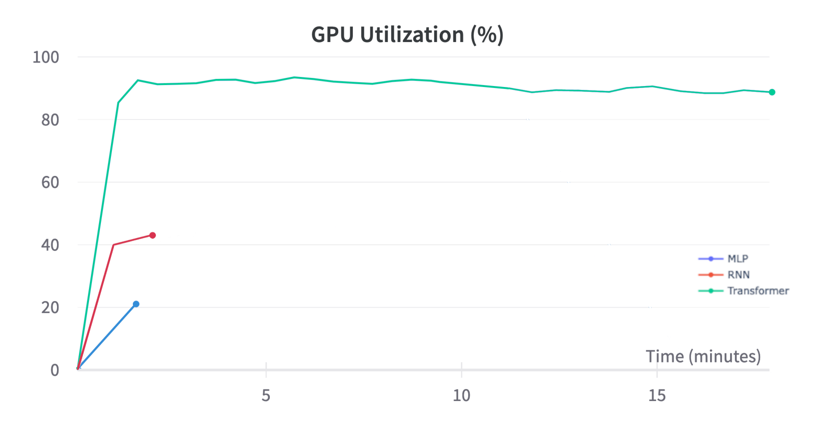
Training Phase



CPU Utilization

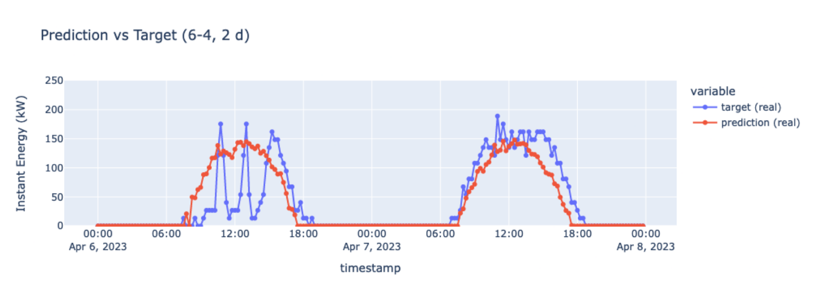


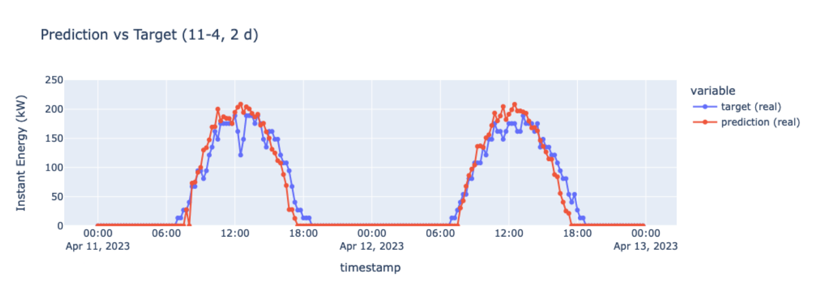
GPU Utilization



Model Evaluations & Comparisons

MLP-based Model

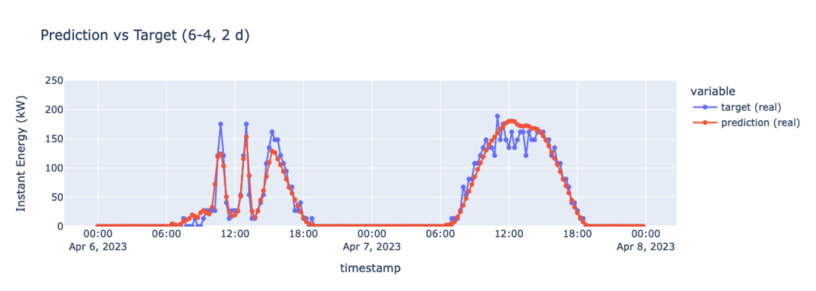


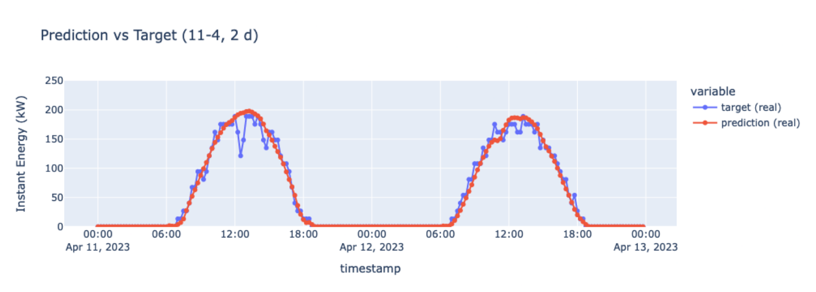


MLP-based Model

AVG MAE	(kW)	14.11	±	3.81
AVG MAPE	(%)	70.98	±	27.99
AVG MAPE@20	(%)	53.01	±	20.10
AVG R^2		0.69	±	0.17

RNN-based Model

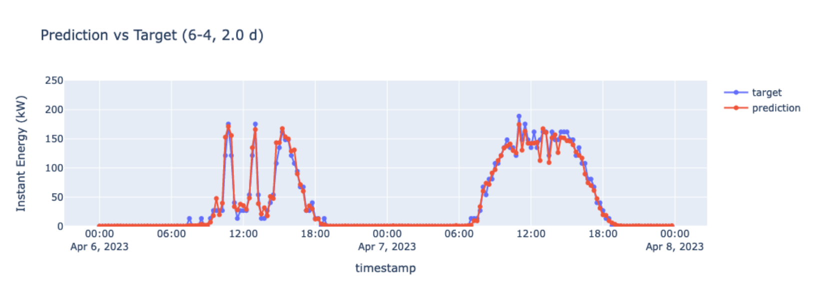


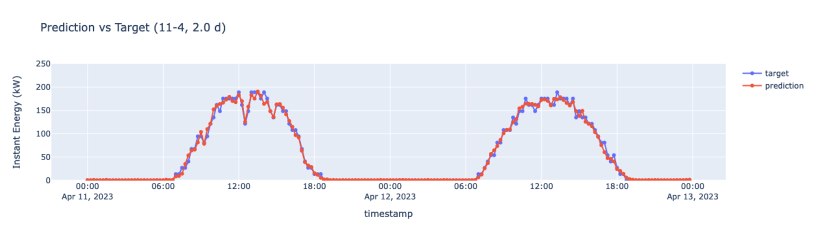


RNN-based Model

AVG MAE	(kW)	6.86	$\pm$	1.87
AVG MAPE	(%)	28.83	$\pm$	10.02
AVG MAPE@20	(%)	21.02	$\pm$	8.67
AVG R^2		0.92	$\pm$	0.06

Transformer-based Model

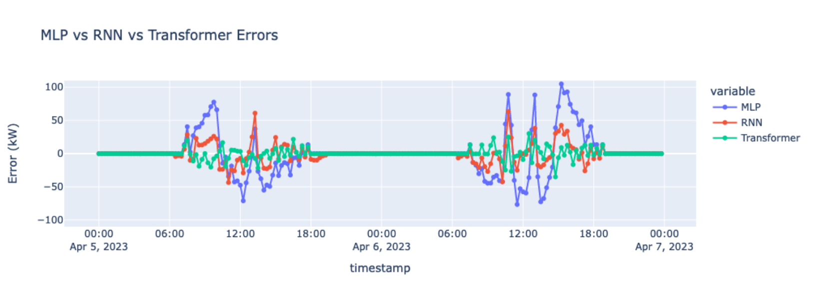


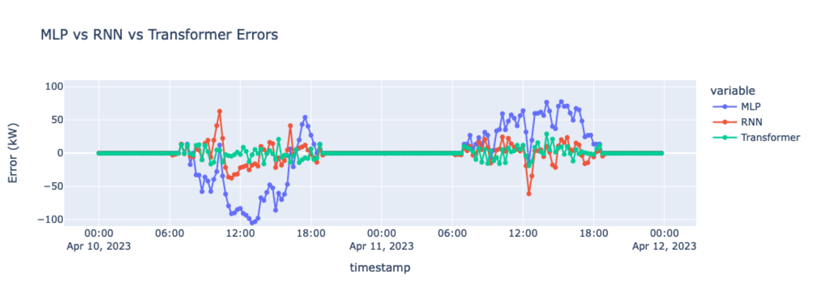


Transformer-based Model

AVG MAE	(kW)	3.76	$\pm$	0.39
AVG MAPE	(%)	18.14	$\pm$	6.76
AVG MAPE@20	(%)	11.18	$\pm$	3.33
AVG R^2		0.98	$\pm$	0.02

Comparisons





Comparisons

	Gain (%)	
	MLP vs RNN	*RNN vs Trans.*
MAE	51.38	45.18
MAPE	59.38	30.07
MAPE@20	60.34	46.81
R^2	25.00	6.12
AVG	*49.03*	*32.05*

Final Conclusions

In this thesis, we introduced 3 different deep learning models to address the **time series imputation problem** on photovoltaic data.

1. **Transformer**:

   PRO: Better performance overall
   CONS: Computationally complex training phase

2. **RNN**:

   PRO: Able to train on low-performing machines
   CONS: Quite high errors

3. **MLP**:

   CONS: Not able to solve the problem, very high errors

The end