TÉCNICO LISBOA

Photovoltaic production forecast at medium voltage distribution networks

Diogo Caneira Mendes

Thesis to obtain the Master of Science Degree in

Energy Engineering and Management

Supervisors: Prof. Pedro Manuel Santos de Carvalho
Prof. Hugo Gabriel Valente Morais

Examination Committee

Chairperson: Prof. Duarte de Mesquita e Sousa
Supervisor: Prof. Hugo Gabriel Valente Morais
Member of the Committee: Prof. Eduardo Manuel Godinho Rodrigues

July 2023

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

I would like to thank all my family and friends for their support throughout these years, for always believing in me and encouraging me to do more.

I would also like to thank my supervisor Prof. Hugo Gabriel Valente Morais for the support, availability, and knowledge that made this Thesis possible.

Last but not least to all the colleagues that helped me through this journey in IST and life. Thank you.

Abstract

This thesis addresses the critical need for accurate short-term solar forecasting in Portugal, driven by the growing adoption of solar energy as a sustainable power source and the inherent variability of solar power generation. Accurate short-term solar forecasting is crucial for efficient grid management and the integration of solar power into the existing energy infrastructure. Existing machine learning approaches often lack the integration of physical models and efficient optimization frameworks, which this work aims to fill.

The study focuses on 10 substations in Portugal, employing 2 years of data trained with XGBoost, TabNet, and NN. A physical model is integrated, destined to refine the model's predicted values by incorporating irradiation and temperature forecasting. *Optuna* is utilized for model optimization, providing the search parameters for each forecasting algorithm. Pre-processing is integrated to increase the model's accuracy and avoid measuring errors, involving customized techniques that assess the impact of data cleansing for each feature. In the proposed framework, XGBoost demonstrates superior performance and faster processing time, with an average increase of 14% from the benchmark forecast with low data processing until the more tailored approach that took all the preprocessing into consideration. The work reached in the end an average RRMSE of 0.1190 kW, which is a satisfactory result considering the limitations and time constraints. The use of a physical model did not perform as expected, being a path worth exploring in the future with data from other institutes to see its potential when joined with new ML algorithms.

Keywords

PV generation forecast, Machine learning, Extreme gradient boosting, Medium voltage distribution networks, Physical PV model

Resumo

Esta tese aborda a necessidade crítica de uma previsão solar precisa a curto prazo em Portugal, impulsionada pela crescente adoção da energia solar e pela variabilidade inerente de produção de energia solar. Uma previsão exata é crucial para uma gestão eficiente da rede e para a integração da energia solar na infraestrutura energética existente. As abordagens de aprendizagem automática existentes carecem da integração de modelos físicos e de uma optimização eficiente, o que este trabalho pretende colmatar.

O estudo centra-se em 10 subestações em Portugal, utilizando 2 anos de dados treinados com XGBoost, TabNet e redes neuronais. Um modelo físico é integrado, destinado a aperfeiçoar os valores previstos pelo modelo, incorporando a previsão de irradiação e temperatura. O *Optuna* é utilizado para a optimização do modelo, fornecendo os parâmetros de pesquisa para cada algoritmo. O pré-processamento é integrado para aumentar a precisão do modelo, envolvendo técnicas que avaliam o impacto para cada característica. Na estrutura proposta, o XGBoost demonstra um desempenho superior e um tempo de processamento mais rápido, com um aumento médio de 14% desde a previsão de referência com baixo processamento de dados até uma abordagem que levou em consideração todo o pré-processamento. O trabalho alcançou no final um RRMSE médio de 0,1190 kW, um resultado satisfatório considerando as limitações e restrições de tempo. A utilização de um modelo físico não teve o desempenho esperado, valendo a pena explorar no futuro com dados de outros institutos para ver o seu potencial quando conjugado com novos algoritmos.

Palavras Chave

Previsão de geração fotovoltaica, Aprendizagem automática, Reforço por gradiente extremo, Redes de distribuição em média tensão, Modelos físicos fotovoltaicos

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Networks
ArNN	Adaptive Recurrent Neural Network
BFS	Best first search
BGA	Binary genetic algorithm
BP	Back Propagation
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DRNN	Deep Recurrent Neural Networks
DTR	Decision Tree Regression
ELM	Extreme Learning Machine
EU	European Union
FNN	Feedforward neural network
FS	Feature selection
GBDT	Gradient Boosting Decision Tree
GHI	Global Horizontal Irradiance
GMT	Greenwich Mean Time
GPR	Gaussian process regression
GRU	Gated Recurrent Unit
GS	Genetic search
IST	Instituto Superior Técnico
KDE	Kernel Density Estimation

KPCA	Kernel Principal Component Analysis
LFS	Linear forward selection
LMS	Least median square
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
MSE	Mean Squared Error
NaN	Not a number
NN	Neural Networks
NWP	Numerical Weather Prediction
PCA	Principal Component Analysis
PCCS	Positive Correlation Coefficient Selection
PV	Photovoltaic
RES	Renewable Energy Sources
RF	Random Forest
RMSE	Root Mean Squared Error
RRMSE	Relative Root Mean Squared Error
SSFS	Subset size forward selection
SVM	Support Vector Machines
SVR	Support Vector Regression
TPE	Tree-structured Parzen Estimator
XGB	Extreme Gradient Boosting
XTNet	Extreme Gradient Boosting + TabNet + ResNet

1

Introduction

Contents

1.1 Motivation

In Europe, since 2012, the production volume of electricity from solar PV power in the EU has been steadily increasing. Eurostat states that in 2020, renewable energy sources made up 37% of gross electricity consumption in the EU in which solar power contributed with 14% of that share a really big increase when compared with 2008, where this renewable source only accounted for 1% [1]. Knowing the importance of solar generation the EU continues to fund research projects in order to find new materials and a better design for PV cells promoting more efficient solar panels and lower energy costs [2]. The funding opportunities provided by the European Commission address all the branches of renewable technologies, from efficiency to innovation, life-cycle assessment, and overall carbon mitigation.

A major concern surrounding solar power and its production is the variability and unpredictability of sunlight due to cloud cover and dust covering the cells along other conditions, adding up to this unpredictability. The biggest solar stations are typically located far from their final consumers adding an extra expense in the energy transportation. The variation of solar power also comes from the natural change of the sun's position related to the cells along the day and the year due to the relative position of Earth and Sun [3].

When considering the solar fluctuation along with the emerging RES share in Portugal and Europe's electricity consumption, the necessity to create a forecasting research regarding power production is an ongoing subject in the ML environment. A model's prediction, the result of applying a ML algorithm, can provide its user with the ability to detect solar patterns based on past data, evaluate the PV plant potential, and be aware of common uncertainties which ultimately helps in decision-making and therefore improves the stability of the system and its potential to grow [4] [5].

ML is a very large area in the AI world with multiple strategies available conducting numerous opportunities to develop different types of work and build upon older frameworks to extract the best of each one, called hybrid models. By joining the forecasting importance with the solar energy needs and ML capabilities a lot can be achieved and important conclusions drawn.

1.2 Objectives and contributions

The objective of the research pursued is to create a new methodology and estimate its performance regarding solar power forecasting in different solar sites in Portugal. In the present paper, the combination of recent ML algorithms and adequate pre-processing measures will be applied and optimized to ensure the highest accuracy possible while maintaining a framework that can be used in the future for similar works. Applying a constant optimization for each individual solar site and in each pre-processing stage will also be a contribution to the lack of optimization frameworks used. Using a hybrid model of both ML and a physical model for solar PV panel will also open a discussion regarding the implementation

of such procedure in a solar power forecasting research taking into account not only the power but also irradiation and temperature.

Since its an academic research, the processing time will also be examined to discuss how appropriate a learning algorithm is compared to commercial or more dense research due to time, computational constraints, and overall complexity.

1.3 Thesis Outline

The thesis is going to be divided into different chapters including Chapter 1 - Introduction where it is explained the motivation to do this research and an overview of the topic, the Chapter 2 - Background where the state-of-art of the technology used is review as well as direct analysis on previous works, the Chapter 3 - Methodology will include the algorithm development, the tuning done to pursue the optimization, verification, and validation of the models trained by the algorithms and the selection of a panel to create the physical model. The following chapters are Chapter 4 - Results, from the methodology applied, this section will demonstrate the results and discuss whether the approaches taken are advantageous or not. Finally, Chapter 5 - Conclusion and future work will reflect on the achievements made, compare them with the bibliography, and identify possible limitations and improvements for future works.

2

Background

Contents

2.1 State of the art

The rapid growth of PV systems calls for more accurate methods to forecast the performance and reliability due to investments and the reassurance of a stable energy mix [6]. Regarding the forecasting horizon, there are different strategies with distinct goals. Forecasts that reach a year ahead are often described as ”short-term”, whereas predicting over a year is usually described as ”long-term”. In [7], a review made on different time horizons saw a relation between the decreasing performance and the increasing forecasting horizon chosen. Even though, following multiple studies, it has a lower performance when compared to short-term, [8] and [9] associated the long-term prediction with a necessity in new PV installations to estimate the power produced in the life cycle of the system and to quantify degradation-influenced energy potentials from the thin film (a-Si) photovoltaic systems.

With respect to the forecasting method chosen, older forecasting techniques such as physical or statistical methods can provide an undesirable inaccuracy or need masses of historical data [3] [10] to achieve decent results. Nowadays, power forecasting ML methods stand as a more reliable approach being, in most comparisons, superior to previous methods. As an example in [11], the physical method applied to model the atmosphere as a fluid is more prone to uncertainties due to the complex data needed to work properly as well as the initial measurements of the atmospheric conditions [12]. The research of [13] reviewed traditional time-series forecasting along with up-to-date ML forecasting techniques, comparing both fields. Conclusions drawn verified, as in other research, that in PV systems traditional methods struggle to achieve satisfactory results compared to newer approaches such as ML.

ML is defined as a group of computational techniques utilizing the experience (historical data) to enhance performance or to achieve precise predictions. The experience denotes the previous information available to the learner that is naturally received from the electronic data recorded and made available for investigation [14]. One of the many advantages of ML is that it can work with big sets of data in order to learn and create its patterns but the way data is classified can vary with the chosen algorithm, the types of data are divided into supervised, unsupervised and reinforcement learning. Supervised learning refers to algorithms capable of following a general rule given by an input and a desired output. This is given by a set of training examples pairing the input data with the output, therefore producing an inferred function [15]. A supervised learning algorithm can be further divided into two categories: classification, where the output is a discrete value, categorizing a set of data into classes; and supervised regression, defined by an output with a continuous value that trains the model on past data and predicts future ones.

Unsupervised learning, opposite to supervised, analyses sets of data without the need of an expert working with less obvious patterns, and is mainly used in clustering and for feature reduction due to its ability to group data based on its similarities [16]. Reinforcement learning is the learning of a mapping from situations to actions so as to maximize a scalar reward or reinforcement signal. Oppose to supervised learning, the learner, instead of knowing which action to take, go through a trial-and-error search

in order to achieve the actions that return the highest reward the trial-and-error mechanism and the late reward are the two distinguishing features of this type of learning [17].

In [18] was developed an extensive analysis on multiple methods comparing the performance with linear regression algorithms while working with an increasing feature space where most regression algorithms tend to struggle. A new approach proposed in [19] to forecast short-term electricity demand applied FS methods based on a BGA-GPR approach that, when compared to other FS techniques, demonstrated better performance.

In [20] is reviewed in depth the use of ML in photovoltaic forecasting for very short and long-term time series. The work provides crucial information regarding the observations of past works creating a stronger starting point for future research by providing a discussion on the methods used, input parameters, and details in the algorithms used. Following [21], the review made on ML for renewable power forecasting, the conclusions drawn propose the existence of NWP to improve ML accuracy, the use of short-term forecasting when using ML techniques and try to combine the hybridization of multiple ML models with optimization techniques with an expected deceleration of the training process which ends up being a trade-off.

2.2 Related Work

In order to determine the optimal methodology for this work and offer a new perspective to the previously offered works in the machine learning/forecasting field, the purpose of this section is to evaluate earlier research on solar production forecast, stressing the approaches employed and the conclusions reached. This section breaks down the methodologies into three shorter sections to make it easier to understand previous research: data description and processing, where key preprocessing techniques are reviewed; feature exploration; and finally, an examination of the use of learning algorithms specifically in solar generation forecast.

2.2.1 Data processing and feature exploration

The utilization of big data sets is a typical practice in the forecasting field since it enables the forecasting algorithms to comprehend any missing data, anomalies, and trends which could possibly not be understood when considering a smaller percentage of data. It is also crucial to note that data sets can contain errors, especially if the data is provided by monitoring sensors or other devices that, due to external factors or poor handling, tend to lose precision. This has an impact on the learning algorithm's ability to anticipate outcomes in forecasting problems. Data processing might therefore be a crucial first step in order to avoid potential errors in following procedures.

Data preprocessing, which is commonly used in ML algorithms, is the adjustment of features in an understandable format to a posterior analysis or modeling. It entails a variety of techniques, most of which are focused on handling missing or inaccurate data, converting categorical into numerical data, and handling outliers [22]. [23] used multiple preprocessing methods to forecast meteorological comprising of:

- Removing gaps - When in the data file there is some NaN value caused by an error it can be removed to eliminate a possible noise in posterior processes;
- Rejecting night hours - With the aim of forecasting solar radiation the rows with 0 irradiation are deleted leaving only the daytime ones;
- Removing outliers - Outliers are values within a feature that stand out from the rest. The proposed method used to find these observations was the box plot being the outliers outside of it.

In [23] work, there was also training at each step to evaluate the performance from beginning to end, achieving a reduction of the error until the last step.

In [24] a time window was selected from 6 a.m. to 8 p.m. to exclude the irradiation values gathering only useful irradiance data. A more detailed technique in [25], considered removing irradiance data based on the sunrise and sunset of each day instead of assuming the same time window.

Data processing also means working on the selection of certain features that can enhance the study with proper information instead of analyzing a series of irrelevant features that may negatively impact the forecasting.

A feature is a quantifiable attribute or characteristic that may be analyzed and interpreted in machine learning. In solar forecasting, data sets with weather elements are often selected and include a large number of meteorological parameters. The use of many features can raise memory usage and processing expenses from a software perspective, which can also affect the model's capacity for pattern analysis and pattern interpretation as well as its rate of learning. A feature can demonstrate its irrelevance in two different ways: by being redundant, which means they don't add pertinent information when compared to the remaining features, or by being useless when it doesn't bring anything to the research by being trained. Researchers are unable to agree on which feature extraction and selection method performs better, so instead they evaluate each problem in detail, select an appropriate approach to address it, and train models with the optimal number of features in a general way, enough to avoid overfitting [26]. Feature extraction is preferred when working with input data that is not understandable to a learning algorithm; in contrast, feature selection maintains the physical meaning of the original features, facilitating the model's learning performance [27]. In [23] research it was eliminated linearly dependent characteristics as a feature extraction measure since they can produce extraneous information because one can be thought of as a linear combination of the other. In order to use this technique, pairs of features

were given a correlation index matrix, which was applied, dropping one if the pair had a high value of correlation, ranging from 0 if the features were independent to 1 if there was a substantial dependence between them.

Principal component analysis (PCA), another feature or dimension reduction strategy that is frequently used by researchers, analyzes a dataset's features with the aim of extracting the most crucial information and expressing it as a set of new orthogonal variables known as "principal components" [28]. When using the PCA, redundant data is eliminated, and the number of remaining principal components is always less than or equal to the number of original features. The first component will have the greatest variance possible, and the second component, which comes after the constraint of being orthogonal to the first component, will have the greatest variance possible [29]. In [30] is used KPCA an extension of PCA with the addition of kernel methods known for the ability to recognize and analyze patterns. The method works by projecting the original inputs into a high-dimensional feature space, making the data structure more linear. In this research, it is stated that this method is capable of achieving more reasonable results than other PCAs based methods. In [31] a series of feature selection methods were used such as, BFS, LFS, SSFS, Ranker, GS and PCCS concluding for the set tested that wrapper feature selection methods were better when compared with the absence of selection for different learning algorithms, therefore, improving the accuracy of the solar power prediction done.

2.2.2 Learning algorithms

In order to produce accurate forecasts, it is crucial to consider the use of a suitable learning algorithm, but it is also helpful to evaluate multiple methods in order to improve the research's accuracy. Depending on the issue, some works show an additional method for using ML algorithms, such as hybrid models, which combine the strengths of two distinct models to produce a more powerful one.

The analysis of [32] on multiple ML algorithms (ANN, RF, MLR, XGB, LSTM) showed good performances, and a comparison proposed in the work reflected a higher accuracy in the extreme gradient boosting, although it required higher expertise in the selection of parameters and additional computational techniques, with ANN the chosen method for future solar power output forecasting. In [33] is proposed a day-ahead solar irradiance forecast choosing the XGB as a learning algorithm and the KDE method to provide the probability density. After a comparison with other methods such as ELM, RF, and SVR, the deterministic forecasting results showed a much lower error as well as a lower training time being useful for other research, following the author's conclusions. In [34] the KPCA model mentioned in the feature selection section and the XGB algorithm were combined to provide a hybrid solution for short-term solar forecasting at 5 distinct locations. The results demonstrated that using this hybrid strategy increased the accuracy when compared to XGB alone, which was a good overall performance for short-term forecasting. Following the use of XGB, [35] compared this algorithm with DTR, LSTM, and

MLR to find the one that could perform better. The conclusions suggested that for the 2 years tested XGB was superior over the other models and all models presented a better performance in the global horizontal irradiance in comparison to diffuse irradiance. [36] For the purpose of hourly GHI forecasting for 3 different solar sites, a hybrid algorithm using the XGB forest and DNN outperformed individual state-of-the-art learning models (SVR, XGB, RF, and DNN). The hybrid model is more complex and time-consuming than other models but represents a trade-off that can be useful in some studies due to the improvement in the prediction error in the range of 33% to 40%.

In contrast to previous research, the forecasting of temperature and hourly irradiance from six different sites was done in [37] using four different models (CatBoost, NN, TabNet, and Naive Baseline), first individually where CatBoost showed a higher accuracy, and then using a proposed model composed of an amalgamation of CatBoost, NN, and TabNet. The results showed that when the three models are combined, the forecast is more accurate. Due to difficulties in the power grid caused by the impact of weather conditions on PV power generation, it is suggested in [38] that a hybrid model based on a NWP information be used, taking into account the models XGB and Tabnet separately and when combined to create XTNet, as well as using LSTM, SVM, MLP, and CNN as comparisons. In [39], the data given was classified as structural, time-series, or hybrid at a 24-hour horizon and 15-minute resolution using ANN and multiple regression models as forecasting methods. The results showed that ANN performed better for all three forms of data input, and the ANN with hybrid data had the lowest error recorded. The article also showed a substantial impact on the model in terms of the quality of data provided for the structural technique. In [40], the ArNN-BP, a feed-forward neural network incorporating Levenberg-Marquardt back-propagation and CNN, was implemented to estimate solar irradiance under various weather situations. The paper's conclusions reveal that the created model outperformed previous linear models in estimating solar energy for one-hour ahead, especially cloud concentration and higher GHI values which is the goal since lower ones are less relevant for power production.

In [41] a comparison is made in different meteorological stations in Kuwait to predict solar irradiation using ANN. The work compared the gradient descent method for ANN and the Levenberg-Marquardt algorithm with extensive testing for possible architectures as well as a third ANN with a high number of neurons. Conclusions indicate that, although being more precise in some areas, a high number of neurons tends to lose the required generalization, being the architecture that integrated the gradient descent method the one that showed the highest performance. The authors in [42] focused its work on DRNN to predict short-term solar irradiance. The method consisted of DRNN with LSTM units with two hidden layers and 35 hidden neurons compared with FNN trained using the back-propagation method and SVR. Results and conclusions demonstrated a big potential on the DRNN implementation in comparison to the other methods being a good predictive modeling alternative. An extensive study on DL algorithms for power load forecasting in [43] suggested that one of the ways that should be pursued for

future advancement was the use of hybrid algorithms, which achieved higher accuracy levels as well as a higher resilience to data. The test of hybridization as mentioned and the use of multiple models is always interesting as it is difficult for an algorithm to clearly perform better in the context of solar energy and load forecasting. Based on the physical and ML approaches, [44] developed a work where 14 PV plants were analyzed with 13 different algorithms expanding the relationship between the optimization of models, the hybrid models (ML and physical) versus ML models and the irradiance-to-power conversion. The methodology used with the prior referred approaches found a decrease in the MSE and MAE used as metrics for the calculations done when applying hybrid models. Such work can also be considered important due to the irradiance-to-power conversion methods, something that is not always studied in this field and rather forecasted separately. [45] By using feature selection methods with ML models such as (RF, SVR, CNN, LSTM, hybrid CNN-LSTM, SPAR and persistence) concluded that all can be considered acceptable to predict PV-performance data even though some are more memory and time consuming such as SVM that took from 30 to 50h and RF that needed close to 90GB of memory. Regarding the error metrics RF was the one that showed the highest accuracy. LMS, MLP and SVM were used to test feature selection methods to improve solar prediction in [31]. The methods were applied with default parameters and therefore didn't have the accuracy that one could expect if optimization and tuning were done as it occurs in the majority of works but following the results, LMS was the one with the best performance. By analyzing all of these researches new ML algorithms are expected to have a good performance, in particular, XGB and TabNet. It is clear that the use of hybrid models represents a clear advance in how the models are implemented with most of the papers concluding that it is the best approach for solar power forecasting. The feature selection, as mentioned, can greatly differ due to the type of work being developed but represents a necessity in multiple methodologies.

3

Methodology

Contents

3.1 Data description

Similar to other research mentioned in section 2.2.2, the data was gathered from a meteorological institute; therefore, its acquisition is subject to flaws and inconsistencies. This work analyzed 10 different substations, all located in Portugal, with recordings from the 1st of January 2020 until the 31st of December 2020. The recordings had a time difference of 15 minutes between each one.

The files were primarily .csv, therefore, the adopted method was to load all of them iteratively since the stations were numbered from 1 to 10 using Python's language for easier processing. The files can now be loaded and read efficiently, giving a first glance at how they were built. In Table 3.1 is a simple description of each feature, and Table 3.2 is an example of some of the substation 5 entries after they were loaded in Python.

Feature	Unit	Description
Power produced	kW	Power in kilowatts produced by the solar system
Irradiation	W/m^2	Irradiation in watts per square meter that reaches the solar panels
Temperature	K	Temperature in Kelvin registered

Table 3.1: Description of the data in each substation per timestamp

To get a better understanding of how it is presented in the given files, Table 3.2 demonstrates part of substation 5 entries.

Timestamp	Power produced (kW)	Temperature (K)	Irradiation W/m^2
28/07/2020 10:00	7700	298.0643	389.8923
28/07/2020 10:15	8070	298.7132	452.0959
28/07/2020 10:30	8410	299.362	514.2995
28/07/2020 10:45	8690	300.0108	576.5032
28/07/2020 11:00	8970	300.6597	638.7068
28/07/2020 11:15	9150	301.3085	700.9104
28/07/2020 11:30	9290	301.9573	763.114
28/07/2020 11:45	9430	302.6062	825.3177
28/07/2020 12:00	9520	303.255	887.5213
28/07/2020 12:15	9520	303.6071	963.5643
28/07/2020 12:30	9530	303.9592	1039.607

Table 3.2: Example of substation 5 entries.

For further analysis, a secondary group of information was given, shown in Table 3.3, referring to the year and month when each substation started operating, and the installed power in kVA.

Substation number	Installed Power (kVA)	Starting year	Starting month
1	2000	2014	7
2	40	2005	12
3	2038	2014	5
4	2038	2014	5
5	12254	2009	3
6	6000	2012	8
7	6000	2009	8
8	2000	2014	3
9	99	2010	9
10	8030	2014	12

Table 3.3: Details on each substation

With all the information described, the next stage is to process it in order to remove any errors or faults the sensors may have retrieved when gathering the data.

3.2 Pre-processing

3.2.1 Missing values

Missing values and gaps in the data can lead, as previously mentioned, to inaccuracy in the learning process, but more important than that, they can, in this type of data, indicate if a specific set is worth forecasting because the absence of a large number of entries will make the forecast meaningless.

3.2.1.A NaN values

The first approach was to confirm the overall number of NaN values as well as their distribution throughout the data selected (train, validation, and test). Since the data received was cleaned to a certain degree, it is not expected for the number to have an expression in the data, but filling in the missing entries with 0 could create unexpected noise. With the distribution known, it will be possible to obtain the percentage of NaN values in each data selection to carefully see if any section stands out in terms of the number of values it contains.

To handle the missing values, a threshold of 10% was chosen for each group of data used in the learning process (test, validation, and training data) in each substation. A percentage of NaN values superior to 10% in any of the three groups means that the substation is not eligible for training; otherwise, it would worsen the results. On the other hand, substations with less than 10% NaN values in any section will be trained, and the entries will be filled with 0 using Python's input commands. To summarize this, Figure 3.1 shows a diagram of the procedure.

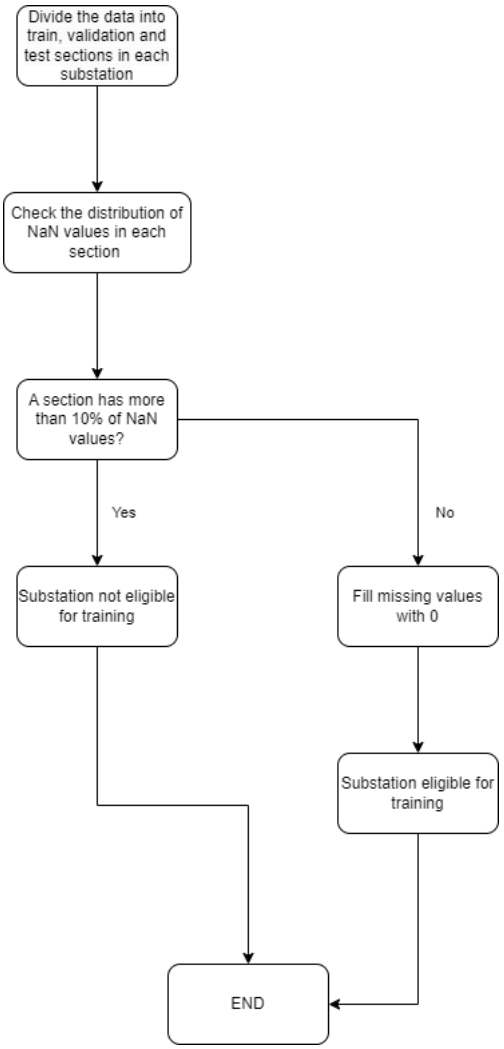

Figure 3.1: Diagram regarding the missing values handling.

3.2.1.B Data cleansing in power production

The absence of production days is a set of data that is either NaN, 0, or values really close to 0, therefore the methodology chosen was to select the consecutive days with those characteristics and if more than 3 days had no production all of them would be removed ensuring a more understandable learning pattern between the data without sudden breaks in production. Since this methodology removes all the rows in a day the training has to be done for all the desired features because the data set will be shortened. To improve the implementation of this process the NaN values will be filled with 0 to group the data and shorten the processing time. In Figure 3.2 is a diagram of the methodology to visualize how this was done.

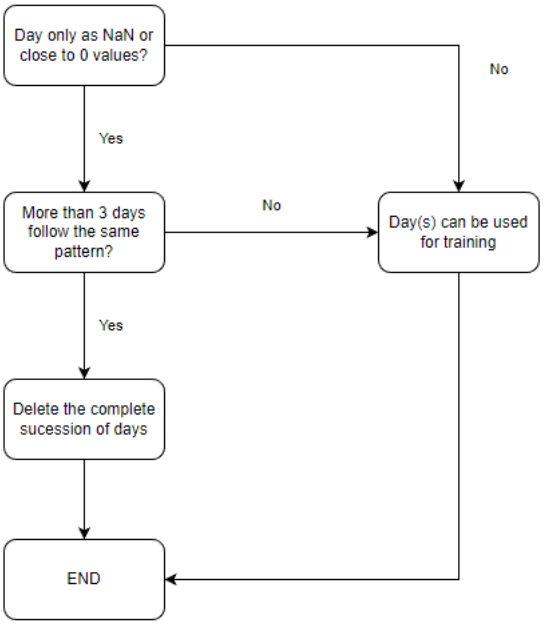

Figure 3.2: Diagram to implement the removal of days without production.

3.2.2 Cleansing of irradiation data

The cleansing of irradiation data was made with the use of *Astral*, a Python's package that calculates the position of the sun and moon and therefore the sunrise and sunset hours based on location. The methodology behind the use of *Astral* was to first find if the timestamp corresponds to an hour between the sunrise and sunset (based on each substation location) and if so maintain its value, otherwise the value should be changed to 0, being the entry considered an outlier. This approach also took into consideration daylight savings and GMT timezone where Portugal is inserted. The features involved in this pre-processing method were irradiation and the Power produced leaving the temperature values unchanged because its still measurable outside daytime. A diagram summarising the methodology is shown in Figure 3.3.

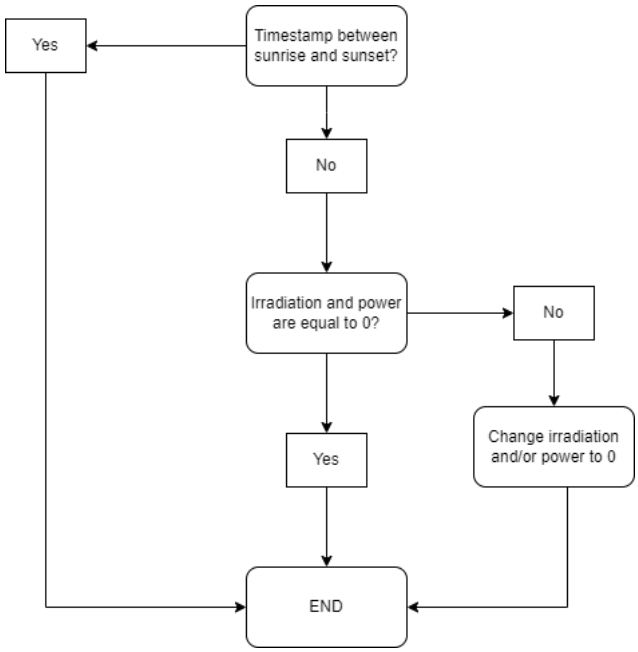

Figure 3.3: Implementation of *Astral* to remove outliers.

3.2.3 Feature selection and extraction

Feature selection in this work was a step that was avoided, as was feature extraction, due to the importance of the three main features described in Section 3.1. Discarding one would compromise further analyses. The models will be trained first regarding the power produced feature, then the temperature and irradiation will also be trained in order to produce a theoretical power by applying a final physical model.

3.3 Data Selection

To split the data under a ML algorithm, one of the most used strategies is the validation set approach. This technique starts by randomly dividing the data into three different sets (train, test, and validation). As this type of forecasting aims to project the data for future days, months, or years, the data was selected for training validation and testing following a chronological sequence, respectively.

The train data is defined by the data, also called samples, used to fit and train the model fitting the parameters of the classifier. The training sample used for the algorithms was the whole year of 2020.

The validation data is a set used to tune the model and give an early estimate of the model contributing to a decrease in the error rate. For this set, it was attributed to the first three months of 2021.

The test data corresponds to the last sample that will provide the performance of the final model created evaluating its error. This data corresponds to a set never seen by the model in order to have 0

information and provide an unbiased evaluation. In the literature, the validation set can also be called the test set [46]. Figure 3.4 shows how the data was divided.

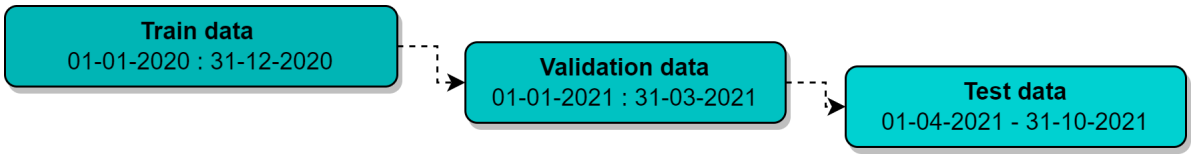

Figure 3.4: Test data, validation data, and training data diagram

3.4 Forecasting Algorithms

With the pre-processing and the data selected the next phase is the implementation of the learning algorithm, in this research three algorithms were used, XGB, Tabnet, and NN. The algorithms need certain parameters to run that are inherent to their architecture, therefore, in this section, each one will be described as their normal utilization practices. Parallel to the use of those learning algorithms, an optimization framework, will be accessed.

3.4.1 *Optuna* Framework

Optuna was designed for ML optimization, the objective is to be an easy-to-use and setup framework with an optimization of the hyperparameters in mind for better model performance. *Optuna* has a range of values for each hyperparameter and after this range of values, it returns the specific value for each hyperparameter that would have the best performance.

To achieve optimum results, *Optuna* has to be divided in 6 sequential parts:

- **Search space-** The search space of hyperparameters has to be defined first choosing for each hyperparameter a range of values that it can have. Since the learning algorithms differ, hyperparameters will also differ, therefore, they will be accessed in the methodology of each algorithm.
- **Objective function-** The objective function uses all the hyperparameters as inputs and returns performance of the model, the metric used to do so in this work was the MSE (discussed in section 3.5).
- **Study-** The study determines the number of times the framework will try different parameters, this part of the framework keeps track of the best set of hyperparameters found for the subsequent to be guided on.
- **Search algorithm-** The search algorithm for the framework was TPE because it has a good performance and is considered efficient. TPE compared to other methods can find an optimal solution

with fewer evaluations therefore, as it is expected a high computation time in some algorithms, this search algorithm is the more adequate to use.

- **Optimization-** Will represent the objective function direction, to achieve a high result the minimization of the objective is the logical option here. As the study, this information is passed across the trials to decrease the error.

- **Evaluation-** The evaluation gives the ability to examine the performance metric as well as see the best parameters found at the end of the trials.

Completed the implementation of the *Optuna* is now the opportunity to discuss the methodology of each learning algorithm and how it was integrated with this optimization framework.

3.4.2 Extreme Gradient Boost

XGB is short for extreme gradient boosting, a supervised learning algorithm that works on predicting the outcome of a certain set of variables, it is built based on GBDT because it uses a series of trees to build the final model.

The algorithm can be divided into two categories, classification (more suited to categorical data sets for example real or false, male or female, etc..) or regression (used in continuous data such as weather forecasting, prices, etc. . .) since the problem requires a regression approach because power forecasting is based on sets of values such as temperature, irradiation, and power, that will be the path taken in this work.

By using the regression classifier, the algorithm iteratively builds decision trees, with each subsequent tree correcting the residuals (differences between predicted and actual values) of the previous tree meaning that each tree reduces the overall residual of the model.

To measure the performance of the model is used an objective function, this function is composed of the loss function, here used MSE, and a regularization term responsible for the overfitting. The optimization of the objective function to increase accuracy is done by finding the points in the tree that when split provide the least MSE maximizing the gain of the function. This gain, which is a crescent during the process, can be considered the overall improvement of past nodes. The new trees created are built depth-wise and with specific pruning techniques that can remove unnecessary branches, decreasing processing time [47] [34].

3.4.2.A XGBoost - Optimization

As discussed in section 3.4.1 the optimization of each algorithm differs due to the architecture that translates here to the hyperparameters. In XGB the used hyperparameter optimization were the following:

- **Number of estimators-** The number of trees a model has, with the increase of this number the accuracy increases until it reaches an overfit where it starts to decrease;
- **Maximum depth-** The use of the maximum depth parameter will, as the name suggests, increase the size of the tree and its complexity;
- **alpha (α) and lambda (λ)-**Regulation parameters suited to avoid overfitting, these two parameters are similar and affect the loss term or loss functions;
- **Learning rate-** The amount that the weights are updated varying between 0 and 1, a higher value would mean a higher change in the weights, therefore, an early convergence, the opposite could get the algorithm stuck due to the minor changes it would provide. If one is not specified the default used by the algorithm is 0.3;

Table 3.4 demonstrates the range of values used in each hyperparameter for every substation. Making use of the fast processing time of XGB, it was built a training loop for the files in order to do the study of all substations iteratively and after, train them separately to make a prediction.

Hyperparameter	Range of values used
booster	gbtree
n_estimators	50 - 1000
max_depth	10 - 1000
reg_alpha	1×10^{-5} - 1×10^{-3}
reg_lambda	1×10^{-5} - 1×10^{-3}
learning rate	0.3

Table 3.4: Hyperparameters used for the optimization and consequent training using XGB.

3.4.3 TabNet

Tabnet is composed of a DNN architecture specific for tabular data. Similar to XGB, this learning algorithm uses a set of decision steps to learn upon the input features and then weigh the importance of each feature to make a prediction.

To explain how this algorithm will be operating on the data a step-wise procedure will be done similar to *Optuna*.

- **Model setup-** The first step is the model setup, because TabNet has a higher processing time, the model will be set up with a defined value for each hyperparameter used and only after the *Optuna* will be used as an optimization so there is a comparison between a first training and the other algorithms and after with TabNet optimized.
- **Feature transformation-** The feature transformation is an important component as it gives the ability to learn useful information for later predictions, it transforms the features optimizing them

with 3 sub-steps: linear transformation to improve stability, nonlinear transformation to improve stability and convergence in training and a feature masking to select the relevant features for each decision step.

- **Attention weighting-** This step will help the learning process regarding the most relevant features by selectively focusing on them. In the training process, high weights will be assigned to the most relevant features and low to the remaining all while being updated making it a dynamic method to capture patterns between features.
- **Training-** The training is similar to other algorithms and uses the prepared data and optimizes the parameters to minimize a loss function that will measure the difference between the predicted output by the model and the actual value. With training prior to the hyperparameters optimization Table 3.5 will show the first training done with general parameters before optimization as a benchmark for improvement.
- **Testing-** The testing will use the input features from the trained model and predict the output on unseen data.

Hyperparameter	Value
Number of layers	2
Number of regressors	1
Feature dimension	32
Number of features	7
Output dimensions	16
Number of decision steps	2
Number of groups	1

Table 3.5: Hyperparameters for TabNet training.

3.4.3.A TabNet- Optimization

The used parameters to process the optimization of the algorithm were the following:

Number of layers- The number of layers will affect the complexity of the TabNet model, increasing the layers will increase the model complexity which is suitable for big data, therefore, increasing the overall performance. In smaller sets, this value is typically smaller due to less complexity needed also avoiding overfitting.

Feature dimension- In this algorithm, the feature dimension is very important due to the improvement it can create. Specifically in data sets like the one in this work with a low number of features, the increase of this hyperparameter will allow to capture more complex patterns between features.

Output dimension- The output dimension depends on the ML task being performed, this value will correspond to the number of output nodes in the final layer of the neural network, in a forecasting

perspective this value would represent the time horizon trained, a high value of output dimension can result in better accuracy but is set to increase the computation resources as well as the training time.

Number of decision steps- In Tabnet, each decision step allows the model to learn a new feature mask (vector of binary values corresponding to an input feature) that will select the most relevant features in the current iteration. This set will be used in a neural network generating a new set of features to be used in the next iteration.

The optimized hyperparameters and the following range of values for TabNet are in Table 3.6

Hyperparameters	Range of values
Number of layers	1 - 20
Feature dimension	64 - 128
Output dimensions	4 - 32
Number of decision steps	1 - 10

Table 3.6: Hyperparameters for TabNet training optimization.

3.4.4 Neural Networks

The implementation of a NN is much like the methodology used in TabNet because this is already a DNN. In this section will be explained the architecture used that differs from the previous one. A neural network is composed of layers and these layers can be seen as building blocks so in the implementation they have to be applied sequentially. In the description below the layers are shown in the order they were coded to perceive the architecture completely.

- **Input layer** - Composed of the window size (sectioned data to help with processing) and the number of features for each step.
- **GRU layers** - These recurrent NN layers aim to extract from the data features that are the most influential through a sequence of vectors, the first layer creates an output that passes to the second one producing a sequence based on the previous.
- **Dense and Flatten layers** - Dense layer will use the previous layer that is fully connected as an input and the purpose is to learn a single scalar value that represents the output. The flatten layer will transform the output to a dimension linear tensor (compatible with the algorithm to be used in the learning) to be used by the evaluation metric that succeeds.
- **Dense layers with regularization** - The 2 dense layers are applied sequentially and the first will use the information from the flattened output passing it to the second one as input, these 2 layers will reduce the dimensionality of the input data and learn more complex representations. By using the regularization as seen in other algorithms it will prevent a possible overfit.

- **Reshape layer** - This will reshape the output to a tensor of (1,1) to in the end give the model with the specified layers, input, and output.

The last step is the training and testing of the models. As described in TabNet, the training was not iterative due to the expected processing time therefore each substation will be trained individually.

3.4.5 Power production forecast considering the physical model of the PV panel.

After applying the different forecasting algorithms to predict the power produced, this work will test the accuracy of using 2 of the 3 features (temperature and irradiation) through a physical model and compare with the already registered power in each substation. The model chosen is the JAM72S20 [48], a monocrystalline cell panel, in Table 3.7 are the electrical parameters and specifications found.

Specifications	Value
Length [m]	2.112
Width [m]	1.052
Cell	Mono
Weight [kg]	24.5
Nº of cells	144 (6*24)

Electrical parameters	Value
Rated Maximum Power (Pmax) [W]	455
Open Circuit Voltage (VOC) [V]	49.85
Maximum Power Voltage (Vmp) [V]	41.82
Short Circuit Current (Isc) [A]	10.88
Module Efficiency [$\%$]	20.5
Temperature Coefficient of Pmax (γ_Pmp)	-0.350%/C

Table 3.7: Solar panel specifications and electrical parameters for model (455/MR)

Computing the number of panels each substation needs will be done by finding the maximum power that was measured and dividing it by the predicted for 1 panel and rounding it as shown in equation 3.1. Equation 3.2 describes in detail each variable used to create the physical model.

$$np_n = round(\frac{P_{max_{yi},n}}{P_{max_{xi},n}}) \tag{3.1}$$

$$P_{x_i,n} = R_{x_i,n} * A * np_n * \eta * (1 + ((T_{x_i,n} - 298.15) * \gamma)) \tag{3.2}$$

Where:

- np - Number of solar panels;
- P - The power produced [kW];
- x_i - Corresponds to the i-th predicted point;
- R - Irradiation [$\frac{kW}{m^2}$];
- A - Area of a solar panel [m^2];

- η - Efficiency of the solar panel [$\%$];
- T - Temperature [K];
- γ - Temperature coefficient [$\%/C$];
- n - Substation n

3.5 Verification and Validation

In a work where learning algorithms are the base of the research, it is common practice to verify if an approach has good accuracy when completed by using a numerical model that translates performance to values such as RMSE, RRMSE, MAE and MAPE.

Validation is a guarantee, sometimes between different steps in the learning process, that the model is increasing in performance until the model reaches its desired threshold. After the validation the models can be tested with the test data and a new step will be needed which is the verification of how good the models are overall.

The verification in this work is related to the final outcome and how well the training went in the end, which was evaluated using the RMSE in features such as temperature and irradiation, while RRMSE is used for power. The results will be compared between the approaches made in terms of percentage to further evaluate and verify the overall increase through the research. That way, it is possible to answer some questions regarding the best methodologies and algorithms chosen.

3.5.1 Root mean squared error - RMSE

Root mean squared error is one of the most commonly used measures to evaluate how a model fits dictating the distance between the predicted values and the values that were tested. Due to their nature the lower it is the better it fits the dataset. This metric is calculated using the summation of the difference between the predicted values and the observed one squared divided by the total sampled size as shown in formula 3.3.

$$RMSE(y_i, x_i) = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2} \tag{3.3}$$

Where:

- n - Total sample size;
- y_i - The i-th measured point;
- x_i - The prediction of the i-th value

3.5.1.A Relative root mean squared error - RRMSE

The relative root mean squared error is a normalized version of the RMSE by dividing the calculated RMSE in each substation for its installed power. This way, it is possible to compare the results between the substations and the improvements made with the overall RRMSE.

$$RRMSE(y_i, x_i, n) = \frac{RMSE(y_i, x_i)}{S_n} \tag{3.4}$$

Where:

- y_i - The i-th measured point;
- x_i - The prediction of the i-th value;
- S_n - Installed power in substation n [kVA]

3.5.2 Mean absolute percentage error - MAPE and Mean absolute error - MAE

MAPE is the mean or average of the absolute percentage errors of forecasts. Error is defined as the actual or observed value minus the forecasted value. Percentage errors are summed without regard to sign to compute MAPE. The percentage when used with the absolute is considered an advantage due to the avoidance of negative errors, as shown in equation 3.5 [49]. MAE is very similar to MAPE but does not show the relative percentage since it is not divided by the actual value depicted in equation 3.6. This metric is applied in the compile stage of both TabNet and NN to evaluate the training and evaluation, this way the user can conveniently see, for example, how at the end of each epoch the model's performance and the model can determine through this error if it should continue searching for better parameters or stop when no improvements are being made.

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{At} \right| \tag{3.5}$$

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |A_t - F_t| \tag{3.6}$$

Where:

- n - The number of fitted points;
- A_t - The actual value;
- F_t - The forecasted value

4

Results and Discussion

Contents

This chapter will show the outcome of the proposed methods in the 3 - Methodology chapter. Starting with the data description, which will provide important information for the pre-processing stage, the data will be plotted throughout the day and in monthly periods for each feature. The section directly related to the pre-processing will show the removal of missing values, which includes different processes such as the distribution of NaN values to find ineligible substations, data cleansing in irradiation data, and finally the removal of missing values in production measurements. Section 4.3.1 begins with the results of training the eligible substations for the proposed algorithms (XGBoost, NN and Tabnet) as well as its processing time, which in the end will help to decide on the most viable algorithm for the remaining work. With the best algorithm chosen, section 4.3.2 demonstrates how the best algorithm performs with a new optimization regarding the cleansing in irradiation data. Section 4.3.3 will follow the optimizations done in advance and perform, with the chosen algorithm, the removal of missing values in production measurements providing, as in previous approaches, a plot of the features with their lowest and highest error in all substations trained.

Regarding the results obtained at the end of section 4.3.2 and 4.3.3, is done a power comparison between the power production and the temperature and irradiation features with an arbitrary solar panel and the physical model described in 3.4.5.

4.1 Data description

The results of the data description presented a variance between the 10 substations when loaded. In particular, substation 10 had 11 fewer entries than the remaining, which could be due to an error or fault in the sensors when recording values. The Table 4.1 encapsulates the most important information regarding this difference, and Figures 4.1, 4.3, 4.5 the distribution of all features along a certain time-frame.

Substation	Starting time	Ending time	Number of entries	Maximum Power registered (kW)	Installed power (kVA)
1	2020-01-01 00:00:00	2021-12-31 23:45:00	70176	1,909.00	2000
2	2020-01-01 00:00:00	2021-12-31 23:45:00	70176	43.00	40
3	2020-01-01 00:00:00	2021-12-31 23:45:00	70176	1,862.00	2038
4	2020-01-01 00:00:00	2021-12-31 23:45:00	70176	1,866.00	2038
5	2020-01-01 00:00:00	2021-12-31 23:45:00	70176	11,440.00	12254
6	2020-01-01 00:00:00	2021-12-31 23:45:00	70176	5,820.00	6000
7	2020-01-01 00:00:00	2021-12-31 23:45:00	70176	5,200.00	6000
8	2020-01-01 00:00:00	2021-12-31 23:45:00	70176	1,919.00	2000
9	2020-01-01 00:00:00	2021-12-31 23:45:00	70176	99.00	99
10	2020-01-01 00:00:00	2021-12-31 21:00:00	70165	7,800.00	8030

Table 4.1: Details of each substation.

4.1.1 Temperature

Observing the temperature distribution is possible to say that for those 2 random time-frames, there are no abnormal distributions, Figure 4.1 has its maximum between 12 and 15h and sees a more notable increase after the sunrise and decreases after sunset. In Figure 4.2 the plot shows a slow increase in temperature until the summer months achieving its higher record between June and July, after those months the temperature tends to decrease as expected as September starts.

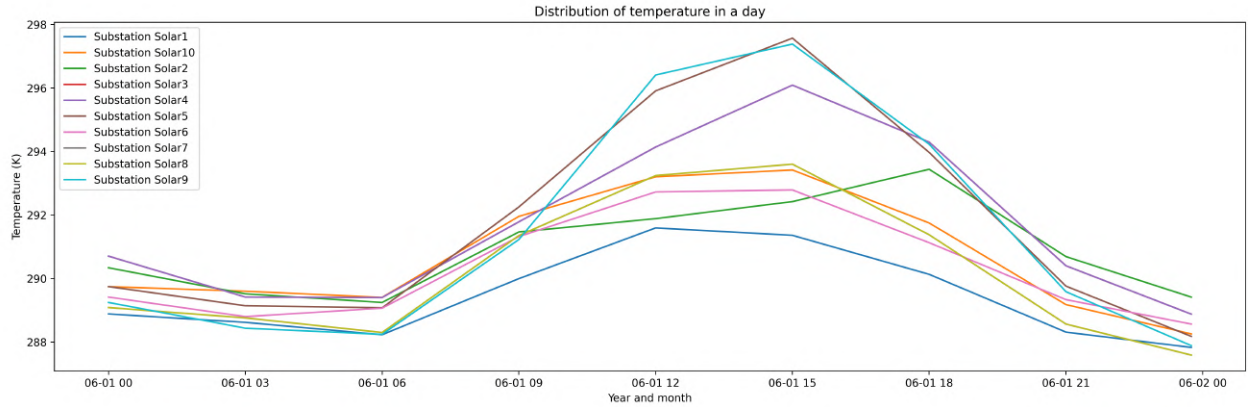

Figure 4.1: Substations temperature through the day.

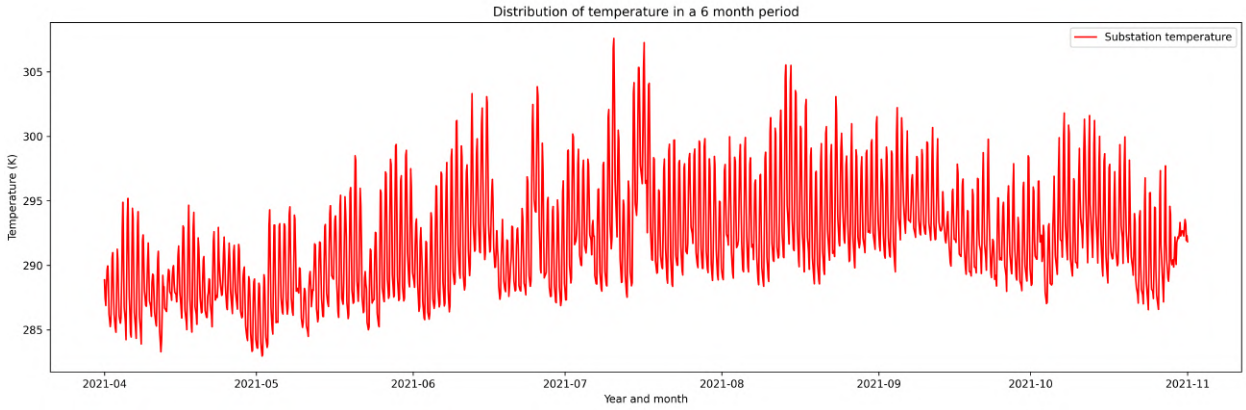

Figure 4.2: Substation 8 temperature in a 6-month period.

4.1.2 Irradiation

The distribution of irradiation through a day in Figure 4.3 shows, in a preliminary analysis, errors due to the sudden increase of irradiation close to midnight as well as a maximum that corresponds, in other features, to a low temperature and power production. The expected distribution should be close to the remaining features decreasing close to the sunset and above 0 close to the sunrise, being the time frame

00-3h and 21-00 imprecise. In a 6-month period, the maximum irradiation follows a similar pattern when compared with the temperature achieving its highest values between summer months and decreasing after August.

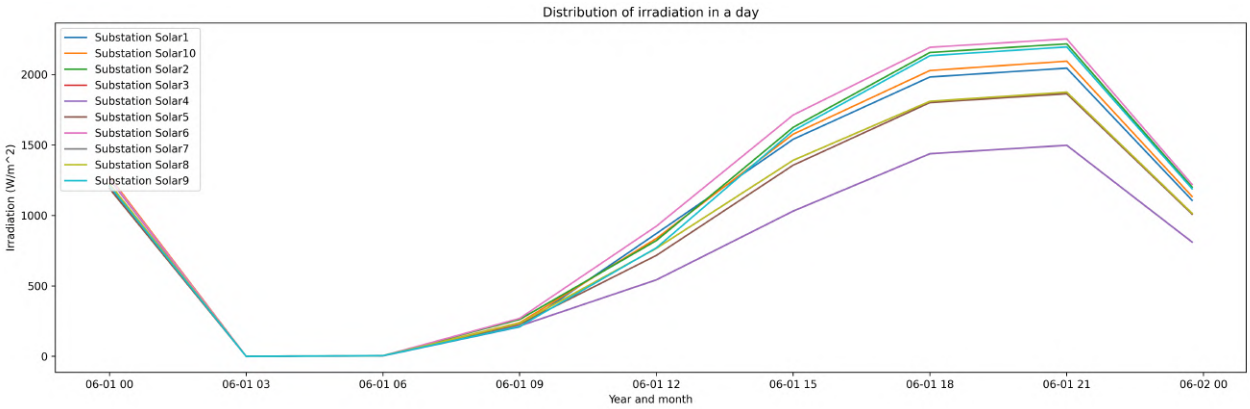

Figure 4.3: Substations irradiation through the day.

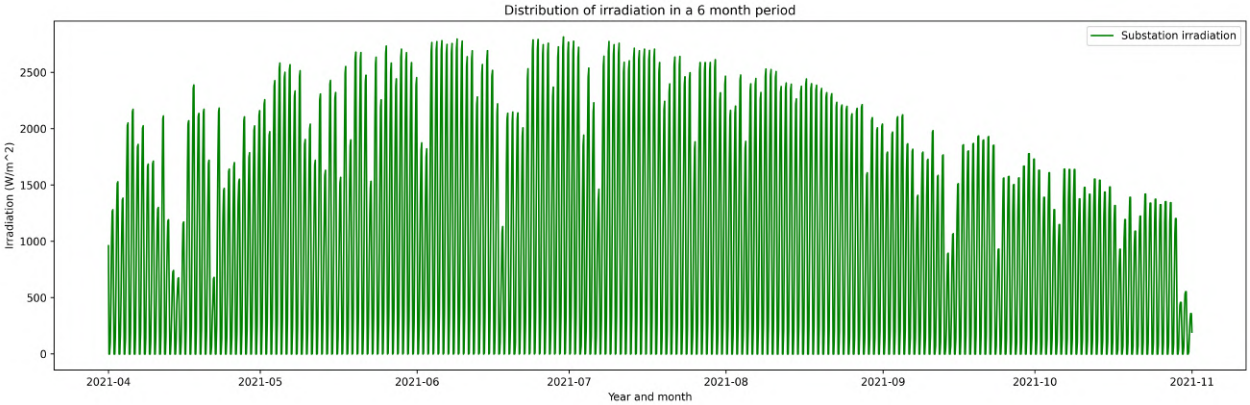

Figure 4.4: Substation 8 irradiation in a 6-month period.

4.1.3 Power produced

The power production plotted shows a ”bad” and ”good” day in the substations, Figure 4.5 represents a day with a lot of fluctuations between the day time not being ideal for forecasting because there is no clear pattern, fluctuations like that one can occur due to cloud opacity or other weather-related problems. The day in Figure 4.6 is a much more stable example of the substation’s behavior indicating a possible clearer sky since the sunrise with a high relation between the power production and the overall installed power.

In the description of the power production feature, the normalized power was used to demonstrate all the substations in the same graph because of the difference in the installed capacity. In both Figures 4.5

and 4.6 substations 3 and 4 demonstrate similar behavior and power distribution which could be related to the geographic proximity between them.

When plotted for these 2 different months is possible to say that both substations 2 and 10 do not show a sign of power production during the day, this can be an indicator of possible missing values or a sum of zeros during these days which can ultimately lead to an impossibility to use the learning algorithms in the referred substations.

To analyze these and other cases the next section will present the results of the pre-processing methodologies used with the described features.

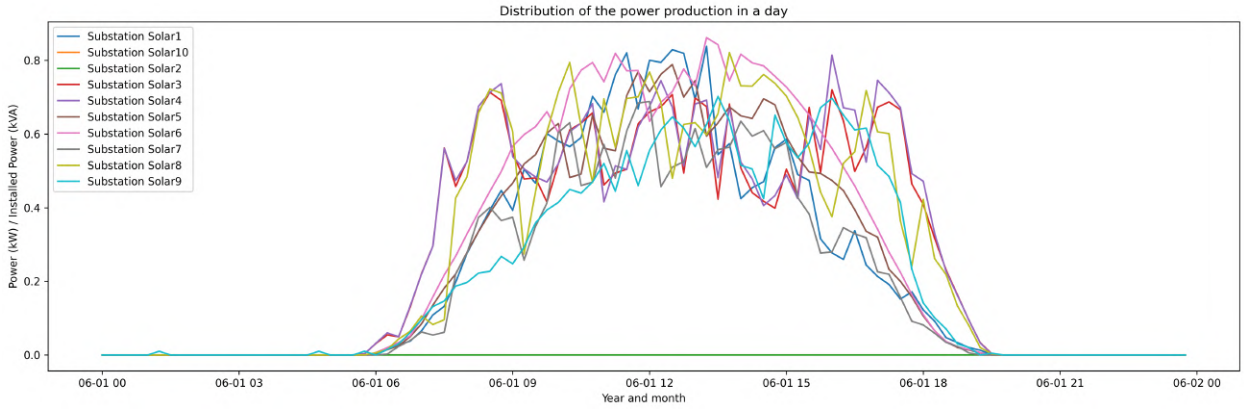

Figure 4.5: Substations power production through the day in June.

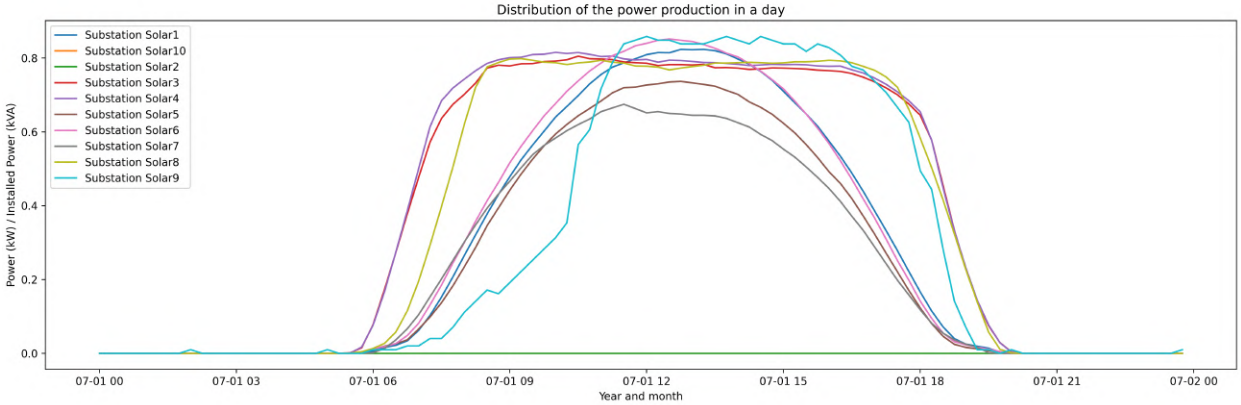

Figure 4.6: Substations power production through the day in July.

4.2 Pre-processing

4.2.1 Missing Values

The first step when entering the pre-processing stage was the sum of all the NaN entries for each feature, which is displayed in Table 4.2.

After this reading, a pattern in the temperature and irradiation data was found, and due to the size and frequency of the values, it was accessed in detail the position of such NaN entries. The result showed that 12 of the values were positioned at the beginning of the data set, from 00:00 to 02:45 of January 2020, and the other 11 at the end from 21:15 to 23:45 of December 2021. As expected and approached in Section 3.2.1, with this arrangement of data it was not used the threshold percentage rule to process the dataset; instead, these entries were removed and the dataset was shortened. The process was not expected to influence the results in a negative way because its during nigh-time therefore the irradiation and power production are 0.

When looking at the power production feature, substations 2,3,4,8, and 10 present a high number of missing values with no apparent pattern, therefore, the percentage of missing values when the data is split to training was calculated and is described in 4.3.

In conclusion, substations 2 and 10 were considered ineligible for training because there was 47.08% of data unavailable for testing in substation 2. In substation 10, 100% in both validation and test data were NaN values meaning an impossibility to test the data trained. Substations 3, 4, and 8 with a percentage below 10 were considered admissible.

Number of missing values registered				
Substation	**Temperature**	**Irradiation**	**Power production**	**Total number of entries**
1	23	23	0	70176
2	23	23	15518	70176
3	23	23	1945	70176
4	23	23	1945	70176
5	23	23	0	70176
6	23	23	0	70176
7	23	23	0	70176
8	23	23	937	70176
9	23	23	0	70176
10	12	12	35927	70165

Table 4.2: Number of missing values registered in each substation for temperature and irradiation features.

Number of missing values registered					
Substation	**Data split**	**Power production**	**% of decrease**	**Total number of entries**	**% of value NaN**
2	train	0	37.67%	35136	0.00%
	validation	0		8640	0.00%
	test	9673		20544	**47.08%**
3	train	1945	0.00%	35136	**5.54%**
	validation	0		8640	0.00%
	test	0		20544	0.00%
4	train	1945	0.00%	35136	**5.54%**
	validation	0		8640	0.00%
	test	0		20544	0.00%
8	train	937	0.00%	35136	**2.67%**
	validation	0		8640	0.00%
	test	0		20544	0.00%
10	train	898	16.27%	35136	**2.56%**
	validation	8640		8640	**100.00%**
	test	20544		20544	**100.00%**

Table 4.3: Distribution and percentage of NaN values in the substations.

4.2.2 Cleansing in irradiation data

Figure 4.3 made clear the necessity for an intervention regarding irradiation during night-time. The use of *Astral* demonstrated great results on the problem. Figures 4.7 and 4.8 show the changes in the data set changing the night-time values to 0 in substation 5 during 60 days. The remaining substations showed similar results confirming the need for this approach.

In Figures 4.7 and 4.8 its possible to notice a difference in the irradiation and confirm that, without the use of this windowing, the training could be subject to some errors due to the positive irradiation after the sunset and before sunrise. Regarding the power production in Figures 4.9 4.10 the results show a much closer relationship between the two which could indicate that the power production was measured with higher accuracy.

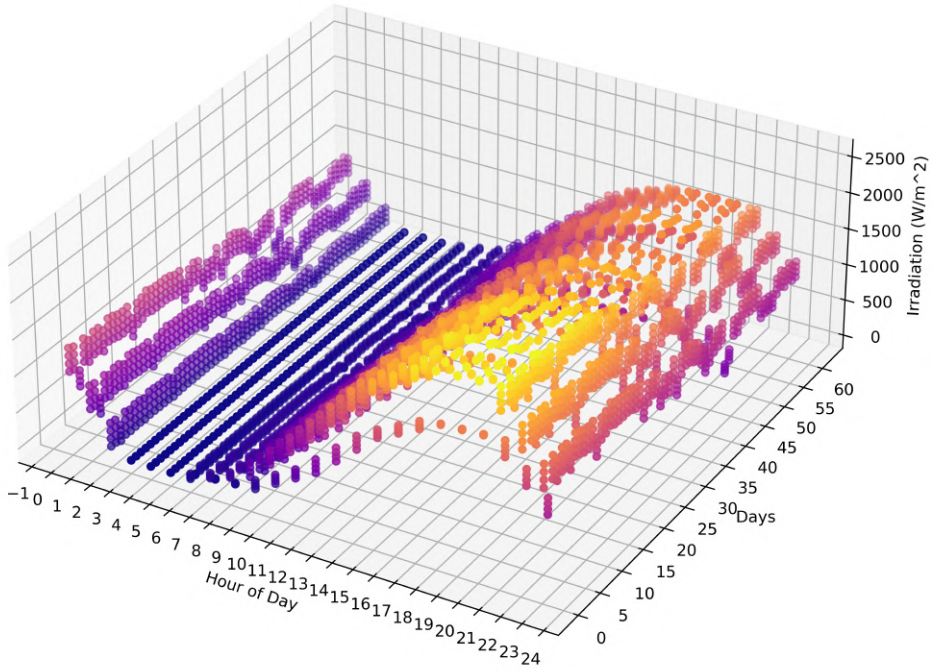

Figure 4.7: Substation 5 irradiation during 60 days **before** the cleansing.

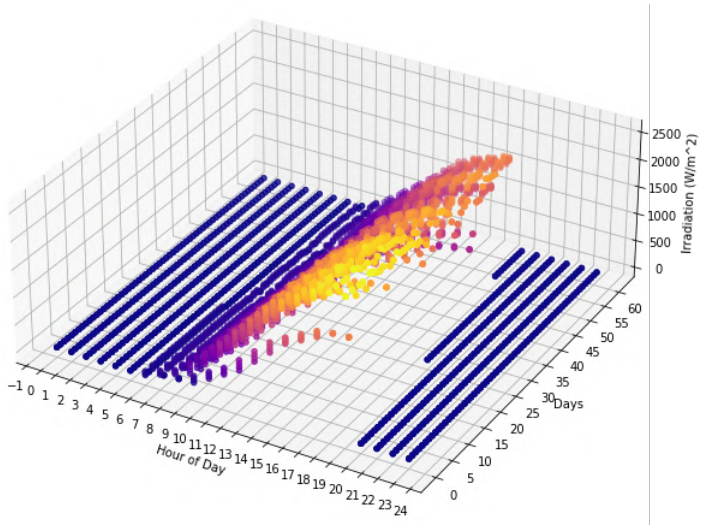

Figure 4.8: Substation 5 irradiation during 60 days **after** the cleansing.

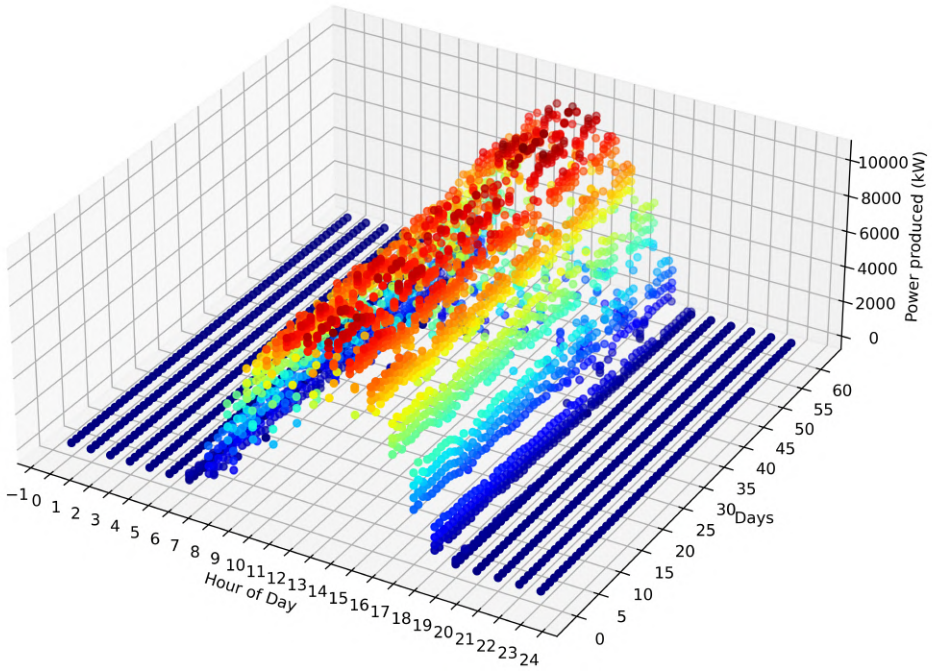

Figure 4.9: Substation 5 power production during 60 days **before** the cleansing.

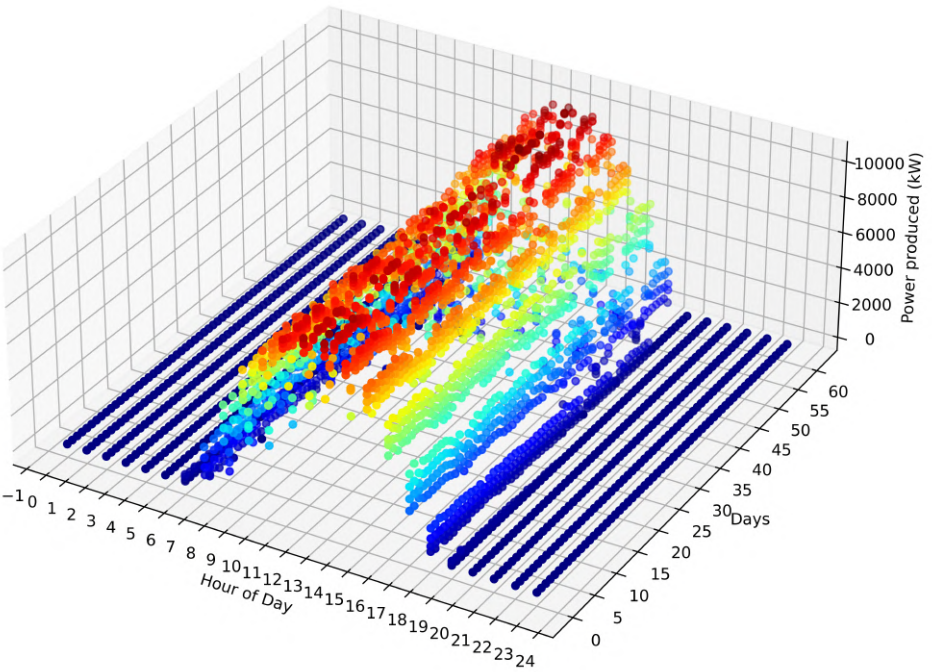

Figure 4.10: Substation 5 power production during 60 days **after** the cleansing.

4.2.3 Missing values in production measurements

The pre-processing for the 10 substations in relation to the production days found that only substations 1, 5, and 6 do not fulfill the requirements regarding the consecutive missing values measured. In Table 4.4 is the total days excluded in each one and is clear why is not possible, once again, to use substations 2 and 10 for training due to the lack of production in over 300 days. Contrary to substations 2 and 10, substations 1, 5, and 6 are expected to be the ones with the best performance since they do not have significant breaks in production.

Looking at Figures 4.12 and 4.11 by following the line that indicates the power production, after the peak on day 10, the production falls to 0, in the first picture using the irradiation cleansing the absence of production continues for over 4 days initiating again at sunrise in day 15. With the methodology chosen the window of data plotted can be pre-processed following the 3 consecutive days rule and the results in Figure 4.11 show that when the production falls to 0 and it reaches midnight the re-arranged data set starts at 00:00 of the next day with a positive production guarantying a more understandable set of data.

Substation	Number of total consecutive days without production
2	326 days
3	20 days
4	20 days
7	9 days
8	10 days
9	131 days
10	374 days

Table 4.4: Number of total consecutive days without production in each substation.

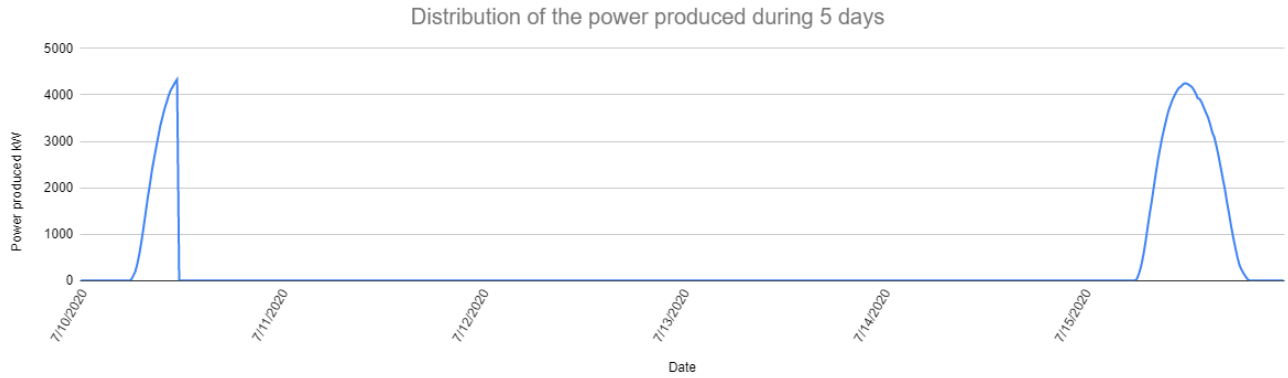

Figure 4.11: Substation 7 power production by only applying the cleansing on irradiation data.

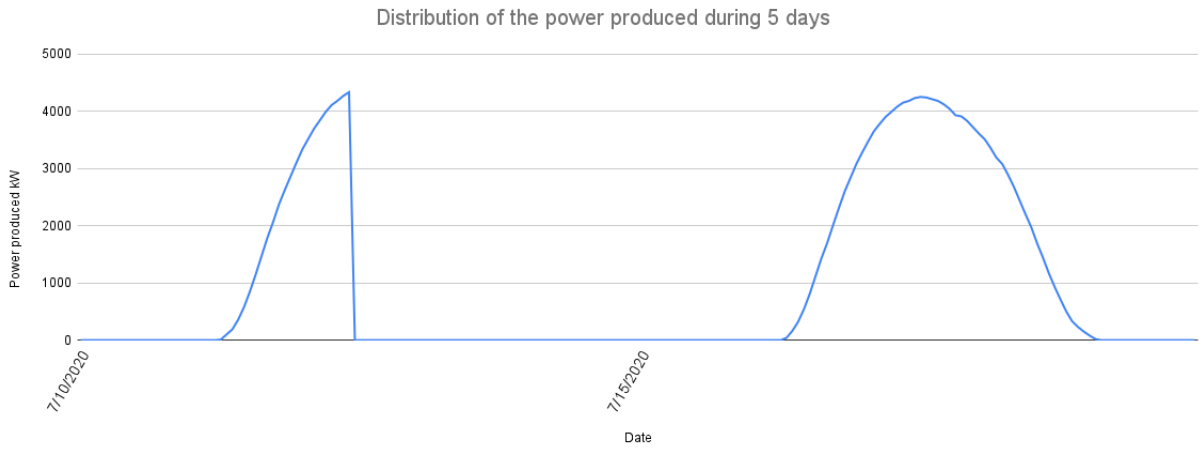

Figure 4.12: Substation 7 power production with both irradiation cleansing and non-productive days.

4.3 Results of forecasting

The results were evaluated following the metrics discussed in 3.5 and the outcome of each algorithm per substation is shown regarding the best model and the mean value of the errors calculated.

To discuss the results, the best method was to divide them through the approaches taken because the training of some learning algorithms might be done more than once and in different conditions. The use of *Optuna* to select the appropriate hyperparameters, was used in all the learning algorithms to further increase the forecasting capability therefore the total processing time of each one will include the optimization procedure along with training and predicting.

4.3.1 Forecast methods benchmark (Approach A)

The first approach is the foundation of this work, the results will demonstrate how each algorithm performs when only being removed the first and last entries of the data that had no values. All the algorithms were trained with the same data-splitting technique and the same rules of pre-processing.

4.3.1.A Power produced

The power produced feature, being the most important, was trained for the three algorithms, and the results are in Table 4.5, where it is shown not only the RRMSE for the best algorithm but also the average of all the sites and the average processing time for each forecasting method.

Substation	RRMSE(kW)		
	XGBoost	**NN**	**TabNet**
1	**0.1246**	0.2921	0.2556
3	0.1797	0.3137	**0.1746**
4	0.1831	0.3220	**0.1412**
5	0.0967	0.2712	**0.0736**
6	0.0954	0.3072	**0.0731**
7	**0.1297**	0.2453	0.3041
8	0.1740	0.3140	**0.1302**
9	**0.1837**	0.3100	0.3171
Average RRMSE(kW)	**0.1459**	0.2969	0.1837
Average processing time (hh:mm:ss)	**0:23:30**	2:52:23	136:51:43

Table 4.5: Power forecasting RRMSE for the eligible substations in the three algorithms.

The results show that XGB has better performance in 3 out of 8 substations; the RRMSE is lower for this algorithm considering the 8 substations analyzed; and the processing time is undeniably lower, meaning that a lot more can be done if XGB is adopted for further optimizations in this scenario. TabNet can, if seen individually, be a better approach to forecast power production, but in a work with multiple steps, the time and computational cost of using this learning algorithm as well as its optimization make it impractical. Another point that can justify the use of XGB, is how the training is much more consistent for all substations, with the highest being 0.1837 kW in substation 9 and the lowest being 0.0954 kW in substation 6, while TabNet shows in the same substations the highest at 0.3171 kW and the lowest at 0.0731 kW, creating in the end an average RRMSE 20% higher than XGB.

Using the TabNet technique, substations 3 and 4 with the same installed power displayed significantly comparable errors of 0.1746 kW and 0.1412 kW. This relationship is caused by the proximity of these substations geographically since 3 and 4 are situated close to one another the weather will be alike. A similar research was conducted on substations 5 and 6, which revealed considerably closer RRMSE in the three algorithms but was initially unable to establish a connection between the two due to their more than 100 km distance.

The best and worst days in terms of the RRMSE were presented for the test data to show how the XGB performed and how this error is relevant in terms of the verification of the work.

On the one hand, in Figure 4.13 is the maximum error registered during forecasting with a RRMSE of 0.4619 kW on substation 4, the predicted power has only a high a small peak of 200 kW at the beginning of the day followed by an absence of power production resulting in a high error. Here, the model did not produce a recognizable pattern for power production which resulted in 10 hours of sunlight but with close to 0 power production in normal lighting and atmospheric conditions. On the other hand, Figure 4.14 clearly shows an overlapping of both curves, resulting in the best possible scenario forecasted with a RRMSE of 0.006 kW.

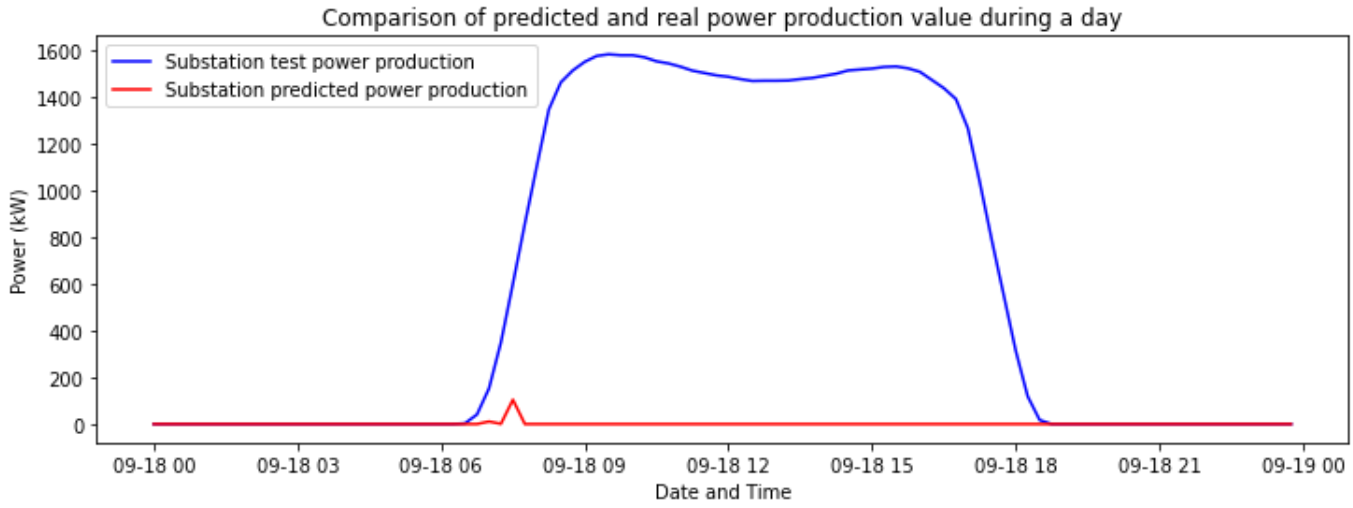

Figure 4.13: Substation 4 registered and predicted power during a day (Approach A).

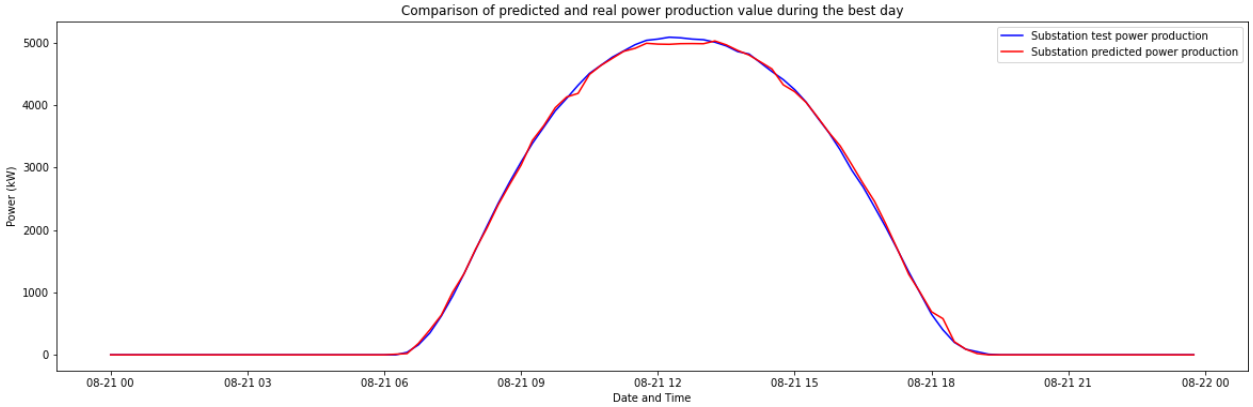

Figure 4.14: Substation 6 registered and predicted power during a day (Approach A).

4.3.1.B Irradiation and Temperature

After evaluating the performance of the three algorithms on the most important feature, it was performed similar training but for the other two features (irradiation and temperature).

Substation	XGBoost	
	RMSE	
	Temperature (K)	Irradiation (W/m^2)
1	2.4	299.2
3	3.7	319
4	3.7	317.3
5	3.7	286.1
6	2.3	250.2
7	3.3	314.2
8	3.1	304.6
9	3.8	284.4

Table 4.6: Forecasting of the temperature and irradiation with the best forecasting algorithm found.

The results of training both temperature and irradiation in Table 4.6 showed that the temperature can be easily predicted using the XGB as it was possible to achieve a RMSE of 3 K in the majority of the substations. Regarding the irradiation, this first training showed a RMSE close to 300 W/m^2 which is high due to greater contrast in values being more challenging for XGB to predict (when compared with temperature). The nature of the data retrieved also presented an unnatural distribution, as discussed before, only worsening the forecast. Because the temperature is always positive during the night, the poor projection of the irradiation led to the option in this approach of not applying the physical model because it would mislead the purpose of it by creating a number of time samples with positive power during the night hours affecting the training and consequent error between the power production and the power calculated. To perfect these results, a new approach will be made in Section 4.3.2.

4.3.2 Impact of the irradiation data cleansing in the forecast (Approach B)

In this section, the objective is to make use of *Astral* to see how the algorithms will train the models and understand if there is an improvement in the data. The results of this approach contain the comparison between the two features affected by the cleansing (irradiation and power produced) with previous training and the results of implementing the physical model. Since Approach A 4.3.1 showed a distinguishable performance by the XGB algorithm, it was adopted for this approach too.

Substation	XGBoost					
	RMSE			RRMSE		
	Irradiation (W/m^2)			Power produced(kW)		
	Approach A	Approach B	Accuracy improvement	Approach A	Approach B	Accuracy improvement
1	299.2	183.4	**38.71%**	0.1246	0.1224	**1.71%**
3	319	205.3	**37.66%**	0.1797	0.1787	**0.58%**
4	317.3	221.7	**36.85%**	0.1831	0.1850	-1.00%
5	286.1	178.9	**37.47%**	0.0967	0.0826	**14.50%**
6	250.2	148.6	**40.62%**	0.0954	0.0856	**10.27%**
7	314.2	227.3	**32.98%**	0.1297	0.1300	-0.19%
8	304.6	203	**35.23%**	0.1740	0.1526	**12.34%**
9	284.4	215.5	**39.85%**	0.1837	0.1750	**4.74%**
Average RRMSE(kW)				**0.1459**	**0.1390**	**4.72%**

Table 4.7: RMSE and RRMSE when applied the new approach in two features.

Table 4.7 shows the performance of Approach B 4.3.2, demonstrating that in all substations the irradiation ended with a lower error, and in substation 6 it reached an increase of 40.62% in performance when comparing with the error in Approach A 4.3.1. The power production forecast improved in 6 out of 8 substations with a maximum in substation 5 of 14.5%. Regarding the average RRMSE it also saw an increase of 4.72% from 0.1459 kW to 0.1390 kW. To visualize the forecasting errors the worst RRMSE found and the best will be compared in Figures 4.15 and 4.16

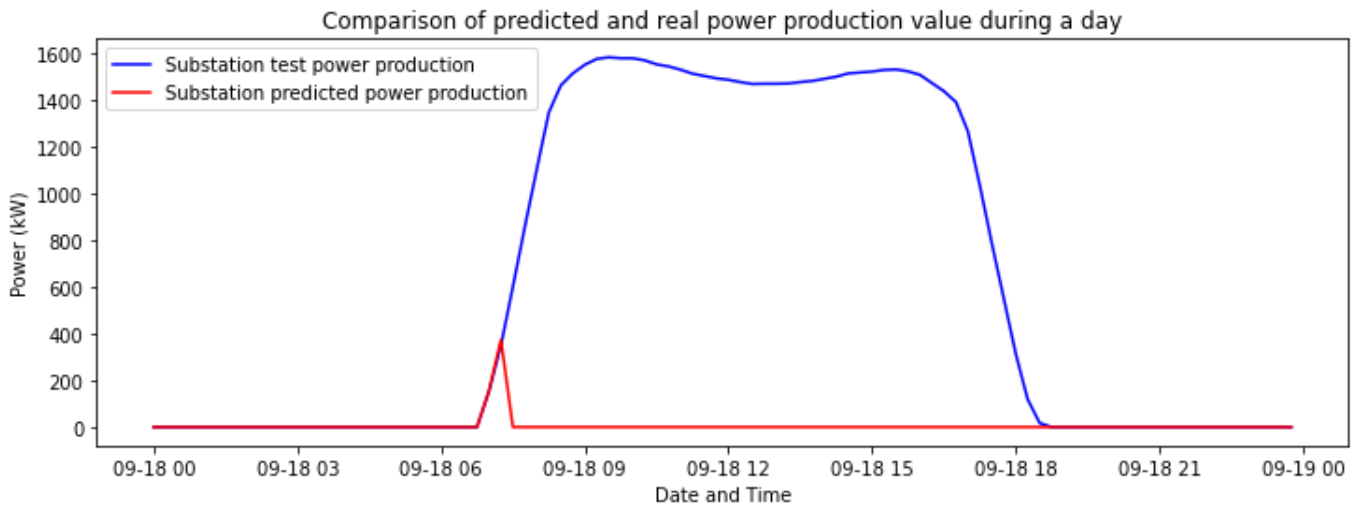

Figure 4.15: Substation 4 registered and predicted power during a day (Approach B).

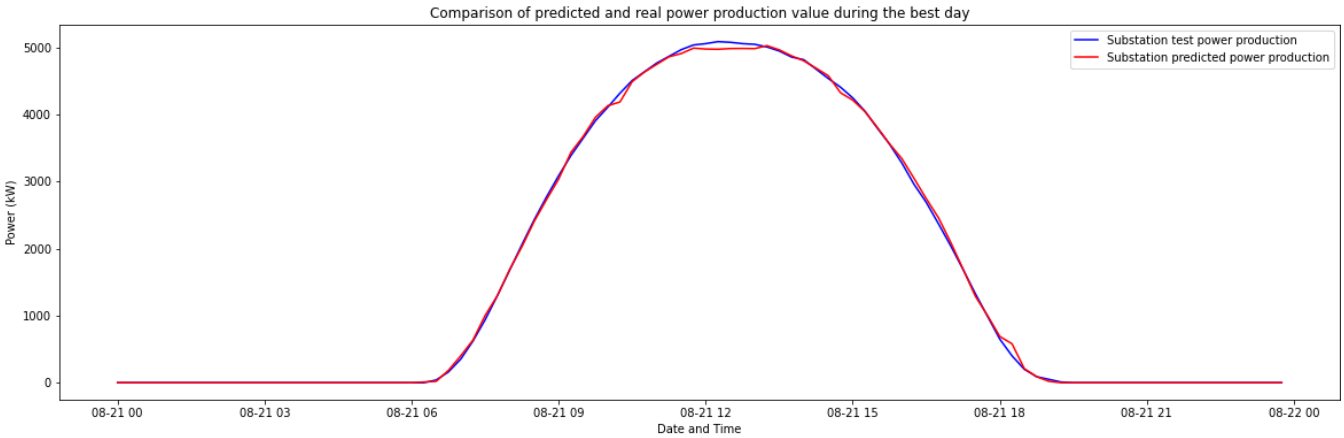

Figure 4.16: Substation 6 registered and predicted power during a day (Approach B).

By first looking only at Figure 4.15, it is possible to conclude the reason why its error is high: In the sunrise, the substation was able to predict a power that reached a value higher than 400 kW but then decreased and stopped, while in the test data, it had a normal behavior of a high increase during the sunrise while decreasing near the sunset. The pattern of Figure 4.16, contrary to what succeeded in substation 4 (RRMSE = 0.4619 kW), shows a close approximation between the test data and the predicted by the model for that specific day, overlapping it, as it already happened in the first forecast. The results indicate that the best and worst days are the same but also increased in performance in both ways, which indicates marginally improved accuracy by the model utilizing this preprocessing method.

A potential downside of this strategy is if the values are registered during the day but are not relevant, for example, on days with lower power production of 0 to 1 kW, the models will still be trained, resulting in a misleading gap. Another potential issue is the loss of some values that could have been neglected due to the ”real” dawn and sunrise being different from the estimated one, which means that if energy was created after the calculated sunset, that value would not be used.

Using Approach B 4.3.2, the irradiation is now within the same time frame as the power production; therefore, it is possible to infer over the use of a physical model also using the temperature that was forecasted previously.

4.3.2.A Forecast using the physical model (Approach B)

The first step of applying the physical model is the number of panels each substation has to have for the predicted values to match the registered, since the irradiation is given in W/m^2, it has to be multiplied by an array of panels to achieve the final power. With the equations 3.1 and 3.2 this number was achieved and following that the respective RMSE and RRMSE were computed. In 4.8 are the results of this study.

Substation	Number of panels	RRMSE(kW)
1	142	0.3023
3	145	0.2894
4	144	0.2951
5	840	0.2833
6	432	0.3227
7	374	0.2584
8	142	0.2920
9	8	0.3110
Average RRMSE(kW)		0.2943

Table 4.8: Number of panels per substation RRMSE calculation on Approach B.

The results are higher than previous forecasts with an increase in the average RRMSE from 0.1413 kW to 0.2942 kW. To have a better visualization of this error, two curves were plotted showing the, predicted, and computed power in the substations with the best and the worst RRMSE.

From Figures 4.18 and 4.17 is clear how the calculated power differs from the predicted one and how the peak is found closer to the night hours, sometimes after sunset. This uncommon distribution is affected by the behavior of the irradiation that hits the panels, which is, to a degree, the feature's reflection, and due to the peak irradiation, caused by possible measuring errors, it finds its higher values later than usual when compared with the temperature and power production.

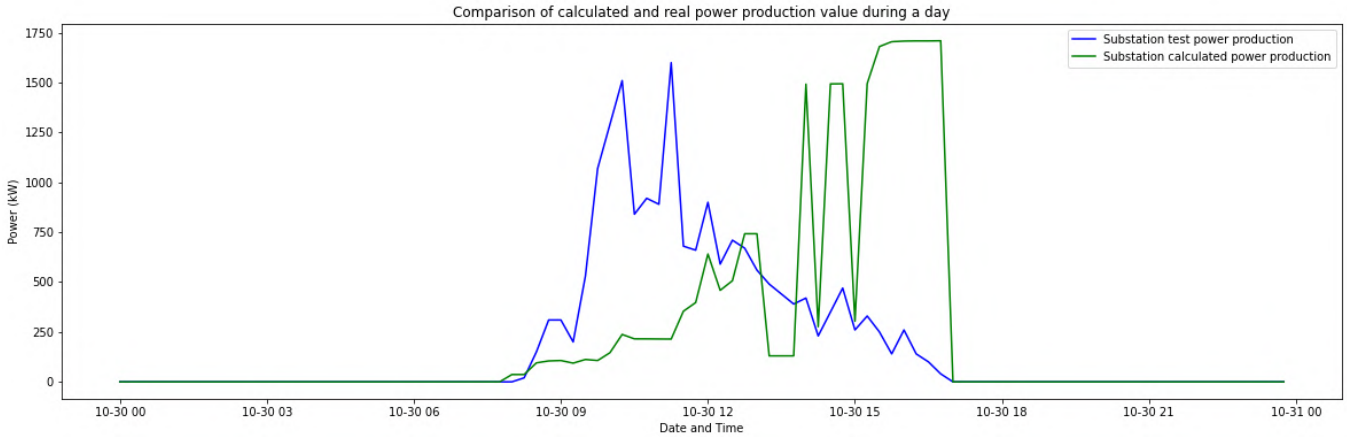

Figure 4.17: Substation 6 registered predicted and calculated power during a day.

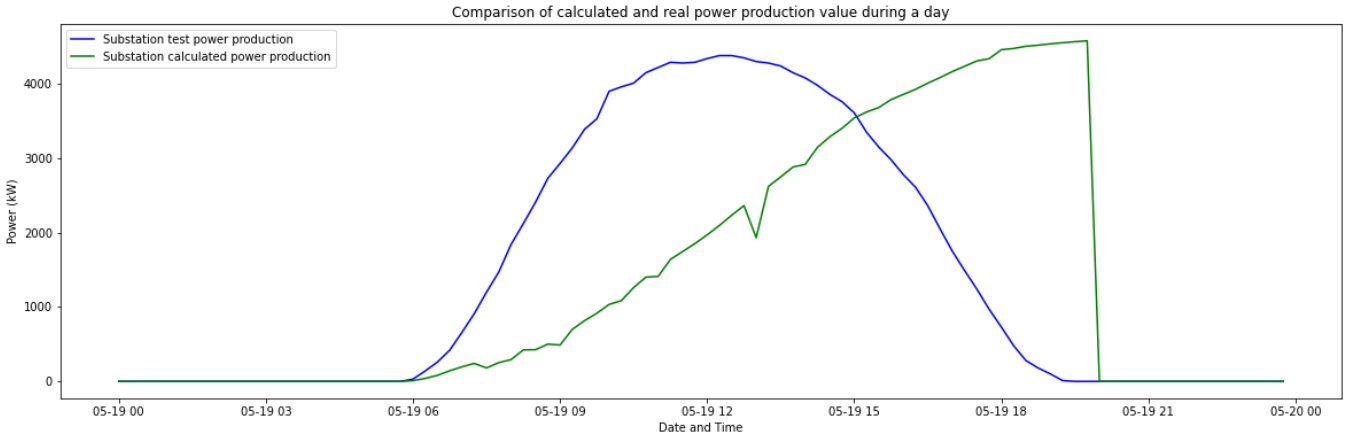

Figure 4.18: Substation 7 registered predicted and calculated power during a day (Approach B).

4.3.3 Impact of data cleansing in power production (Approach C)

In this section, the methods used were irradiation cleansing and data cleansing in power production. The goal of this approach is to create a data set that better explains the power production; In previous approaches, the algorithms could have predicted incorrect patterns in the data due to multiple days with 0 or close to 0 power production. Because the removed rows influenced the temperature and irradiation, they will also be trained in this method.

Substation	XGBoost					
	RMSE					
	Irradiation (W/m^2)			Temperature (K)		
	Approach B	Approach C	Accuracy improvement	Approach B	Approach C	Accuracy improvement
1	183.4	183.4	0.00%	2.4	2.4	0.00%
3	198.9	205.3	-3.26%	3.7	3.9	-7.15%
4	200.4	221.7	-10.65%	3.7	4	-7.42%
5	178.9	178.9	0.00%	3.7	3.7	0.00%
6	148.6	148.6	0.00%	2.3	2.3	0.00%
7	210.5	227.3	-7.98%	3.3	3.1	**4.01%**
8	197.3	203	-2.93%	3.05	3.1	-1.59%
9	171.1	215.5	-25.98%	3.81	3.83	-0.67%

Table 4.9: RMSE when applied the Approach C for temperature and irradiation.

Substation	XGBoost		
	RRMSE		
	Power produced (kW)		
	Approach B	Approach C	Accuracy improvement
1	0.1224	0.1224	0.00%
3	0.1787	0.1336	**25.25%**
4	0.1850	0.1349	**27.07%**
5	0.0826	0.0826	0.00%
6	0.0856	0.0856	0.00%
7	0.1300	0.1179	**9.31%**
8	0.1526	0.1348	**11.67%**
9	0.1750	0.1442	**17.63%**
Average RRMSE(kW)	**0.1390**	**0.1195**	**14.03%**

Table 4.10: RRMSE when applied Approach C in the power production feature.

The results in Table 4.9 show that removing several days without power in the power production feature had a great effect on the RRMSE. Compared with section 4.3.2, the power production forecast had, on average, an increase in accuracy of 14.03%, and 18.09% when compared with Approach A (the average RRMSE contains the forecasting of the remaining stations trained in previous approaches). By using this approach, the temperature and irradiation suffered an increase in error except for substation 7, which could be caused by the removal of a lot of rows that had data, creating gaps in the daily and monthly patterns of both features.

To plot the best and worst days in the substations, only the ones affected by the pre-processing were analyzed. Substation 9 registered the worst error in a day, and Substation 8 was the best. When looking at Figure 4.19 it is clear the difference between both predicted and test data, but when compared with the behavior of substation 4 in Approach A and B, here it follows a recognizable pattern regarding the power production: the power starts by increasing during the sunrise, reaches a maximum above 40 kW, and decreases closer to the sunset. Figure 4.20 shows an almost perfect prediction with an RRMSE of 0.0150 kW but with attention to an error in the prediction of the values between 3 a.m. and 6 a.m. that should be 0 as the test shows because there is 0 solar irradiation to produce power. Another point that could have increased the error compared with more accurate plots is the peaks between 9 a.m. and 12 a.m. as well as the one near 18 a.m. even though it's low. Training all three parameters means that a new power comparison should be done to evaluate the performance with the changes made.

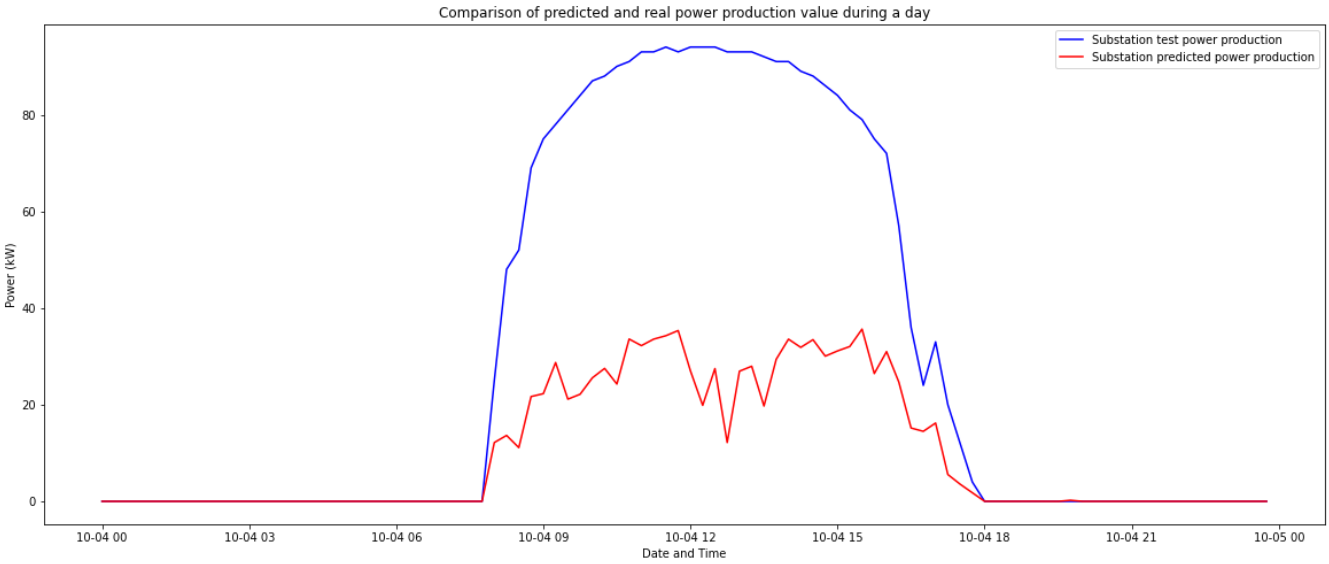

Figure 4.19: Substation 9 registered and predicted power during a day (Approach C).

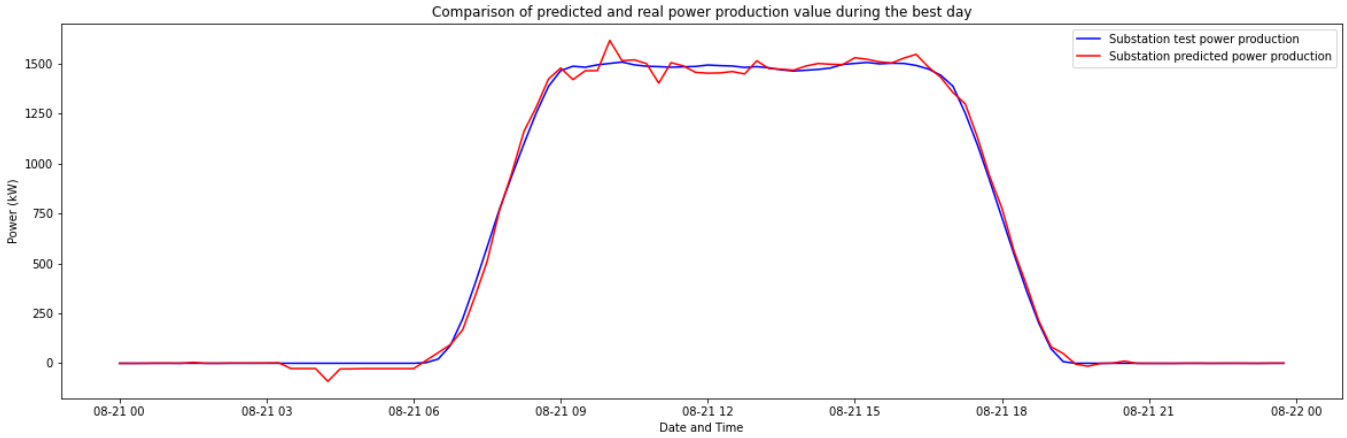

Figure 4.20: Substation 8 registered and predicted power during a day (Approach C).

4.3.3.A Forecast using the physical model (Approach C)

By performing this new power comparison and computing the difference with Approach B, the results show a close to 0 increase or decrease in error, as shown in Table 4.11.

Substation	Approach B		Approach C		
	Number of panels	RRMSE	Number of panels	RRMSE	Accuracy improvement
1	142	0.3023	141	0.3023	0.00%
3	145	0.2894	145	0.2901	-0.24%
4	144	0.2951	144	0.2952	0.00%
5	840	0.2833	840	0.2833	0.00%
6	432	0.3227	433	0.3227	0.00%
7	374	0.2584	375	0.2612	-1.07%
8	142	0.2920	142	0.2923	-0.11%
9	8	0.3110	8	0.3083	**0.89%**
Average RRMSE(kW)		0.2943		0.2944	-0.04%

Table 4.11: Number of panels per substation and RRMSE calculation on Approach C.

Registering a RRMSE of 0.4436 kW in substation 9 and 0.0371 kW in substation 4 Figures 4.21 and 4.22 shows the best and worst days regarding the both tested and calculated power production. The highest error found in substation 9 demonstrates once more how the influence of the irradiation greatly affects the computed power and the increase of error when compared with test data by following a pattern already seen in Figure 4.18. In detail, substation 4 had a similar distribution of calculated and tested power, but regarding the difference during peak hours, the comparison shows that they were captured by the training in temperature and irradiation features but could not reach the same values even though they are similar in the time scale. Analyzing the number of panels used for this last power comparison, it registered no significant difference in the substations which goes according to expectations since the highest values used to calculate this parameter are located close to 12 a.m. and will therefore be maintained in all the pre-processing methods applied here with a low probability of changing.

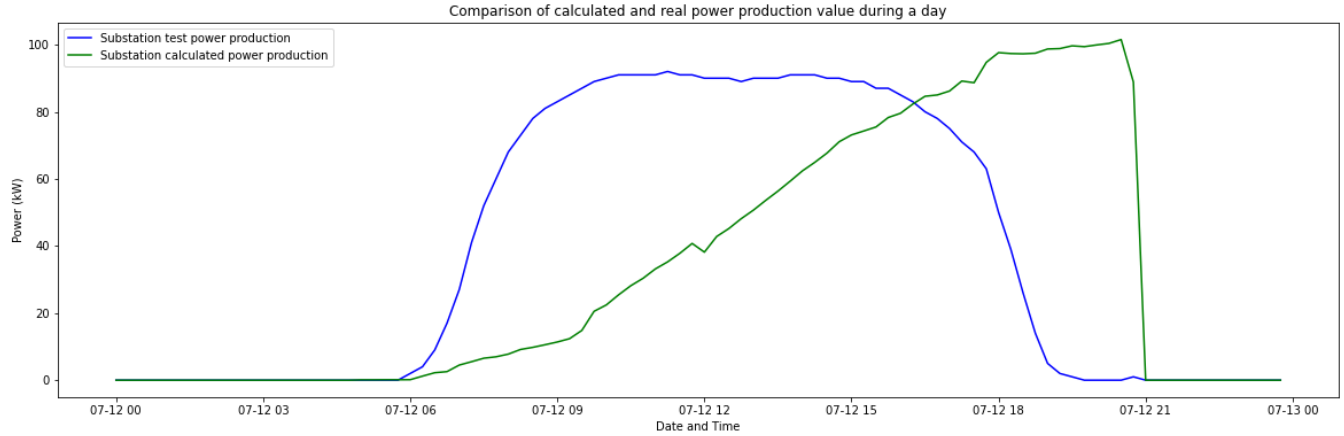

Figure 4.21: Substation 9 registered and predicted power during a day (Approach C).

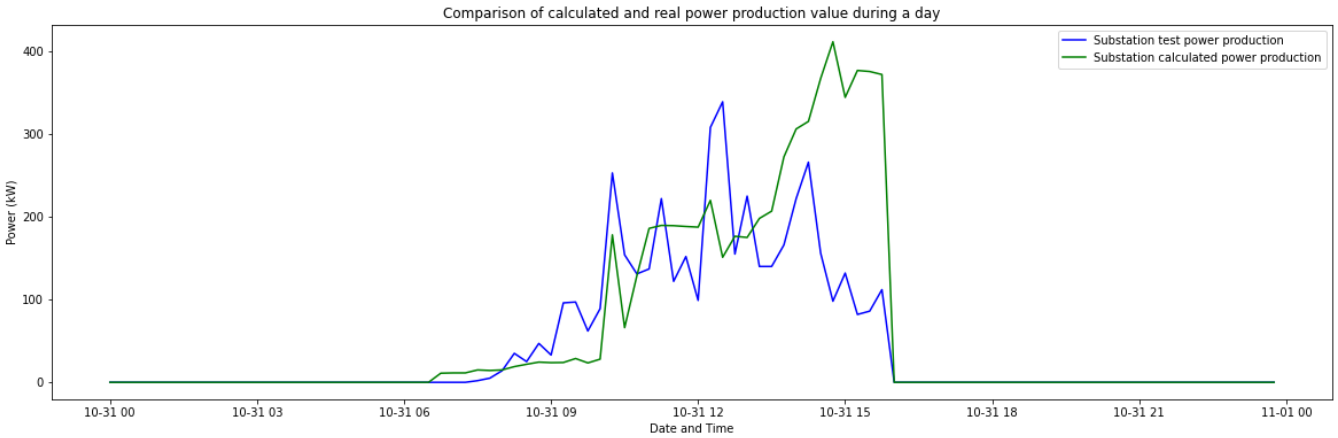

Figure 4.22: Substation 4 registered and predicted power during a day (Approach C).

5

Conclusions

Contents

5.1 Achievements

This thesis was developed with the goal of creating a new and improved methodology for solar power forecasting at different substations in Portugal. The research included 10 solar sites, each with a specific installed capacity (kVA).

The work reviewed different studies with respect to state-of-the-art procedures and found an opportunity to explore at an academic level the application of three ML algorithms (XGBoost, NN and TabNet), pairing them with *Optuna*. The tests revealed that XGBoost and TabNet have the best performance in the tested substations, with an average RRMSE of 0.1459 kW and 0.1837 kW, respectively. XGBoost also revealed a much lower average processing time training each substation in approximately 23 minutes.

Pre-processing techniques such as cleansing the data based on irradiation proved to be a great approach increasing the accuracy in all substations by 37% reaching a minimum RMSE of 171.0540 W/m^2 in substation 9. This impact also produced an improvement in the power produced, but not as high.

The approach solely focused on the impact of cleansing the power produced data showed an increase in the 5 substations trained that reached over 14% in the average RRMSE. Along with using XGBoost to predict the power produced, a physical model comparison was also applied to estimate the number of panels and the power produced in each solar site through temperature and irradiation forecasting. Using the physical model resulted in an RRMSE between actual values and predicted values of 0.2943 kW in Approach B and 0.2944 kW in Approach C, which is higher than by using just XGBoost.

As a comparison, in the literature, the work of [50] presented different algorithms for solar irradiation forecasting in complicated weather conditions which following the constraints in this research is a more suitable comparison. In [50] the best model found, for all weather, was LSTM with an RMSE of 66.69 W/m^2. [51] achieved an RMSE of 62.1618 W/m^2 for daily solar irradiance in Model II-BD based on LSTM. In [52] is conducted a temperature forecast for 12h that resulted in a RMSE of 1.5 K. Due to the correlation between longer forecasts and higher errors is possible to conclude through the literature that the accessory forecast of temperature used only in the physical model had great accuracy with an RMSE close to 3 K.

Concluding, the techniques proposed were all implemented, and the best method was XGBoost optimized by the *Optuna* Framework. All the pre-processing steps improved the predictions on solar power and should be reproduced in similar forecasting works. To forecast the irradiation, Approach C can be neglected since it reduced the accuracy due to the removal of data based on another feature. The temperature, due to the excellent measurements made, could have been forecasted without approaches B and C choosing only the algorithm through the benchmark forecast done in Approach A. The use of a physical model did not perform as expected, which will be explored in the limitations section 5.2.

5.2 Limitations and Future Work

As an academic work based on complex ML algorithms and the unpredictability of solar power production on multiple sites, this research found some limitations as well as some improvements for future work on the topic. In training, the algorithms chosen can require more time than expected to optimize, as well as high computational needs; therefore, choosing to optimize using a similar framework may be optional, and the tuning, made by an expert, can be less time-demanding. The results can also be improved by hybridizing the models with the best performance, for example, XGBoost and TabNet, or adopting more recent algorithms discussed in Chapter 2. All of these limitations and the use of such a long period of time between training data and tests also compromised the research in terms of validating the results with past papers because there is a small number of papers applying a closer methodology to compare all features correctly.

The data gathered due to their distribution and the missing data on some substations created a forecast that could not achieve the level of precision aimed at. For future work, the quality of the data should be higher, as should the number of points to use in training to guarantee that all sites are trained. Other adjustments that could also be made in future work could include testing with other sources of meteorological data, increasing the number of sites, or trying a moving window regarding the data splitting to create smaller differences between the trained and tested data.

Bibliography

[1] Eurostat, "Renewable energy on the rise: 37% of eu's electricity," 2022. [Online]. Available: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20220126-1

[2] E. Commission, "Why the eu supports solar energy research and innovation." [Online]. Available: https://research-and-innovation.ec.europa.eu/research-area/energy/solar-energy_en

[3] S. Pelland, J. Remund, J. Kleissl, T. Oozeki, and K. D. Brabanderel, "Photovoltaic and solar forecasting: State of the art," *IEA PVPS Task 14, Subtask 3.1*, October 2013.

[4] M. G. De Giorgi, P. Congedo, and M. Malvoni, "Photovoltaic power forecasting using statistical methods: Impact of weather data," *Science, Measurement Technology, IET*, vol. 8, pp. 90–97, 05 2014.

[5] K. Bakker, K. Whan, W. Knap, and M. Schmeits, "Comparison of statistical post-processing methods for probabilistic nwp forecasts of solar radiation," *Solar Energy*, vol. 191, pp. 138–150, 2019.

[6] I. Kaaya and J. Ascencio-Vásquez, "Photovoltaic power forecasting methods," in *Solar Radiation - Measurements, Modeling and Forecasting for Photovoltaic Solar Energy Applications*, D. M. Aghaei, Ed. Rijeka: IntechOpen, 2021, ch. 7. [Online]. Available: https://doi.org/10.5772/intechopen.97049

[7] P. Singla, M. Duhan, and S. Saroha, "A comprehensive review and analysis of solar forecasting techniques," *Frontiers in Energy*, pp. 1–37, 2021.

[8] M. Aslam, J.-M. Lee, H.-S. Kim, S.-J. Lee, and S. Hong, "Deep learning models for long-term solar radiation forecasting considering microgrid installation: A comparative study," *Energies*, vol. 13, no. 1, p. 147, 2019.

[9] N. M. Kumar and M. Subathra, "Three years ahead solar irradiance forecasting to quantify degradation influenced energy potentials from thin film (a-si) photovoltaic system," *Results in Physics*, vol. 12, pp. 701–703, 2019.

[10] H. Ye, B. Yang, Y. Han, and N. Chen, “State-of-the-art solar energy forecasting approaches: Critical potentials and challenges,” *Frontiers in Energy Research*, vol. 10, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fenrg.2022.875790

[11] M. Holmstrom, D. Liu, and C. Vo, “Machine learning applied to weather forecasting,” 15 2016.

[12] H. Ye, B. Yang, Y. Han, and N. Chen2, “State-of-the-art solar energy forecasting approaches: Critical potentials and challenges,” *Frontiers in Energy Research*, vol. 10, March 2022.

[13] B. Ramadevi and K. Bingi, “Chaotic time series forecasting approaches using machine learning techniques: A review,” *Symmetry*, vol. 14, no. 5, 2022. [Online]. Available: https://www.mdpi.com/2073-8994/14/5/955

[14] A. Subasi, *PRACTICAL MACHINEL EARNING FOR DATA ANALYSIS USING PYTHON*, A. Press, Ed. Elsevier, 202.

[15] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, “Machine learning methods for solar radiation forecasting: a review,” *Renewable Energy*, 2016.

[16] B. Mahesh, “Machine learning algorithms - a review,” *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, January 2020.

[17] R. S. Sutton, *Introduction: The Challenge of Reinforcement Learning*. Boston, MA: Springer US, 1992, pp. 1–3. [Online]. Available: https://doi.org/10.1007/978-1-4615-3618-5_1

[18] K. Christensen, M. Siggaard, and B. Veliyev, “A machine learning approach to volatility forecasting,” *Available at SSRN*, 2021.

[19] A. T. Eseye, M. Lehtonen, T. Tukia, S. Uimonen, and R. John Millar, “Machine learning based integrated feature selection approach for improved electricity demand forecasting in decentralized energy systems,” *IEEE Access*, vol. 7, pp. 91 463–91 475, 2019.

[20] P. Radzi, M. Akhtar, S. Mekhilef, and N. Mohamed Shah, “Review on the application of photovoltaic forecasting using machine learning for very short- to long-term forecasting,” *Sustainability*, vol. 15, p. 2942, 02 2023.

[21] H. Alkabbani, A. Ahmadian, Q. Zhu, and A. Elkamel, “Machine learning and metaheuristic methods for renewable power forecasting: a recent review,” *Frontiers in Chemical Engineering*, vol. 3, p. 665415, 2021.

[22] J. Ye, “Using machine learning for exploratory data analysis and predictive modeling,” 2015.

[23] A. Bramm, S. Eroshenko, and A. Khalyasmaa, “Effect of data preprocessing on the forecasting accuracy of solar power plant,” in *2021 XVIII International Scientific Technical Conference Alternating Current Electric Drives (ACED)*. IEEE, 2021, pp. 1–5.

[24] A. Alzahrani, “Short-term solar irradiance prediction based on adaptive extreme learning machine and weather data,” *Sensors*, vol. 22, no. 21, p. 8218, 2022.

[25] M. Alanazi and A. Khodaei, “Day-ahead solar forecasting using time series stationarization and feed-forward neural network,” in *2016 North American Power Symposium (NAPS)*, 2016, pp. 1–6.

[26] V. Kumar and S. Minz, “Feature selection: A literature review,” *Smart Comput. Rev.*, vol. 4, pp. 211–229, 2014.

[27] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, “Feature selection: A data perspective,” *ACM Comput. Surv.*, vol. 50, no. 6, dec 2017. [Online]. Available: https://doi.org/10.1145/3136625

[28] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[29] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, “Predicting solar generation from weather forecasts using machine learning,” in *2011 IEEE international conference on smart grid communications (SmartGridComm)*. IEEE, 2011, pp. 528–533.

[30] Q.-T. Phan, Y.-K. Wu, Q.-D. Phan, and H.-Y. Lo, “A novel forecasting model for solar power generation by a deep learning framework with data preprocessing and postprocessing,” *IEEE Transactions on Industry Applications*, vol. 59, no. 1, pp. 220–231, 2023.

[31] M. R. Hossain, A. M. T. Oo, and A. Ali, “The effectiveness of feature selection method in solar power prediction,” *Journal of Renewable Energy*, vol. 2013, 2013.

[32] Y. Essam, A. N. Ahmed, R. Ramli, K.-W. Chau, M. S. I. Ibrahim, M. Sherif, A. Sefelnasr, and A. El-Shafie, “Investigating photovoltaic solar power output forecasting using machine learning algorithms,” *Engineering Applications of Computational Fluid Mechanics*, vol. 16, no. 1, pp. 2002–2034, 2022.

[33] X. Lib, L. Maa, P. Chena, H. Xua, Q. Xinga, J. Yana, S. Lua, H. Fanb, L. Yangb, and Y. Chenga, “Probabilistic solar irradiance forecasting based on xgboost,” pp. 1087–1095, February 2022.

[34] Q.-T. Phan, Y.-K. Wu, and Q.-D. Phan, “Short-term solar power forecasting using xgboost with numerical weather prediction,” in *2021 IEEE International Future Energy Electronics Conference (IFEEC)*, 2021, pp. 1–6.

[35] O. Bamisile, C. J. Ejiyi, E. Osei-Mensah, I. A. Chikwendu, J. Li, and Q. Huang, “Long-term prediction of solar radiation using xgboost, lstm, and machine learning algorithms,” in *2022 4th Asia Energy and Electrical Engineering Symposium (AEEES)*, 2022, pp. 214–218.

[36] P. Kumari and D. Toshniwal, “Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance,” *Journal of Cleaner Production*, vol. 279, p. 123285, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959652620333308

[37] B. E. and C. Giannetti, “Short term load forecasting using tabnet: A comparative study with traditional state-of-the-art regression models,” 2021.

[38] Q. Wang, S. Chai, Y. Liu, and G. Wang, “Gtfd-xtnet: A tabular learning-based ensemble approach for short-term prediction of photovoltaic power,” 2022.

[39] M. Alshafeey, “Artificial intelligence forecasting techniques for reducing uncertainties in renewable energy applications,” Ph.D. dissertation, Budapesti Corvinus Egyetem, 2023.

[40] C. D. Crisosto González, “Solar irradiance forecast from all-sky images using machine learning,” 2023.

[41] M. Bou-Rabee, S. A. Sulaiman, M. S. Saleh, and S. Marafi, “Using artificial neural networks to estimate solar radiation in kuwait,” *Renewable and Sustainable Energy Reviews*, vol. 72, pp. 434–438, 2017.

[42] A. Alzahrani, P. Shamsi, C. Dagli, and M. Ferdowsi, “Solar irradiance forecasting using deep neural networks,” *Procedia Computer Science*, vol. 114, pp. 304–313, 2017.

[43] S. Aslam, H. Herodotou, S. M. Mohsin, N. Javaid, N. Ashraf, and S. Aslam, “A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids,” *Renewable and Sustainable Energy Reviews*, vol. 144, p. 110992, 2021.

[44] M. J. Mayer, “Benefits of physical and machine learning hybridization for photovoltaic power forecasting,” *Renewable and Sustainable Energy Reviews*, vol. 168, p. 112772, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032122006566

[45] D. V. Pombo, P. Bacher, C. Ziras, H. W. Bindner, S. V. Spataru, and P. E. Sørensen, “Benchmarking physics-informed machine learning-based short term pv-power forecasting tools,” *Energy Reports*, vol. 8, pp. 6512–6520, 2022.

[46] M. K. K. Johnson, *Applied Predictive Modeling*, Springer, Ed. Springer New York Heidelberg Dordrecht London.

[47] M. Castangia, A. Aliberti, L. Bottaccioli, E. Macii, and E. Patti, “A compound of feature selection techniques to improve solar radiation forecasting,” *Expert Systems with Applications*, vol. 178, p. 114979, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421004206

[48] E. Ltd., “Enf ltd..” [Online]. Available: https://www.enfsolar.com/pv/panel-datasheet/crystalline/46963

[49] A. de Myttenaere, B. Golden, B. L. Grand, and F. Rossi, “Mean absolute percentage error for regression models,” March 2016.

[50] Y. Yu, J. Cao, and J. Zhu, “An lstm short-term solar irradiance forecasting under complicated weather conditions,” *IEEE Access*, vol. 7, pp. 1–1, 10 2019.

[51] X. Huang, C. Zhang, Q. Li, Y. Tai, B. Gao, and J. Shi, “A comparison of hour-ahead solar irradiance forecasting models based on lstm network,” *Mathematical Problems in Engineering*, vol. 2020, pp. 1–15, 2020.

[52] B. Gong, M. Langguth, Y. Ji, A. Mozaffari, S. Stadtler, K. Mache, and M. G. Schultz, “Temperature forecasting by deep learning methods,” *Geoscientific Model Development*, vol. 15, no. 23, pp. 8931–8956, 2022. [Online]. Available: https://gmd.copernicus.org/articles/15/8931/2022/