Photovoltaic production forecast at medium voltage distribution networks

Diogo Caneira Mendes

***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.

***Index Terms*—PV generation forecast, Machine learning, Extreme gradient boosting, Medium voltage distribution networks, Physical PV model**

I. Introduction

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.

A. *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.

B. *Thesis Outline*

The thesis is going to be divided into different chapters including Chapter I - Introduction where it is explained the motivation to do this research and an overview of the topic,

the Chapter II - Background where the state-of-art of the technology used is review as well as direct analysis on previous works, the Chapter III - 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 the Chapter IV - Results, from the methodology, applied this section will demonstrate the results and discuss whether the approaches taken are advantageous or not. Finally, Chapter V - Conclusion and future work will reflect on the achievements made, compare them with the bibliography, and identify possible limitations and improvements for future works.

II. Background

A. 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.

B. Related Work

In order to determine the optimal methodology for this work and offer a fresh 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.

*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 [14]. [15] used multiple preprocessing methods to forecast meteorological comprising of removing gaps in the data, rejecting night hours and outliers. The sequence of processing the data achieved a continuous reduction of error in training from beginning to end.

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 [16]. 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 [17]. In [15] 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.

C. Learning algorithms

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 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 [18] 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 being ANN the chosen method for future solar power output forecasting. In [19] is proposed work for a day-ahead solar irradiance choosing the XGB as a learning algorithm and the KDE method to provide the probability density. After a comparison with other methods like ELM, RF, and SVR, the deterministic forecasting results showed a much lower error as well as a lower training time being very applicable in actual engineering practice as following the author's conclusions. In [20] 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 utilizing this hybrid strategy increased accuracy when compared to using XGB alone, which was a good overall performance for short-term forecasting. Following the use of XGB, [21] 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. [22] For 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 a trade-off that can be useful in some studies is that the improvement reached a prediction error in the range of 33% to 40%.

An extensive study on DL algorithms for power load forecasting in [23] 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, [24] 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. 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.

III. METHODOLOGY

Similar to other research mentioned in section II-C, 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 data showed 3 different features, commonly used in solar power forecasting that were: temperature in Kelvin, irradiation in watts per square meter that reaches the solar panels, and the power in kilowatts produced by the solar system.

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

A. *Pre-processing*

*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.

*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.

*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.

*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.

*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 for this research. 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.

*4) **Data Selection**:* To split the data under a machine learning 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 zero information and provide an unbiased evaluation. In the literature, the validation set can also be called the test set [25].

B. 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.

*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.

*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 [26] [20].

*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.
- Testing- The testing will use the input features from the trained model and predict the output on unseen data.

The optimization of this algorithm was made by choosing the following hyperparameters:

- 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 machine learning 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.

*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.

*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 simple physical model and compare with the already registered power in each substation. The model chosen is the JAM72S20 [27], a monocrystalline cell panel with an area of 2.222 m^2 and a maximum power of 455 W.

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 one panel and rounding it as shown in equation 1. Equation 2 describes in detail each variable used to create the physical model.

$$np_n = round(\frac{P_{max_{yi},n}}{P_{max_{xi},n}}) \tag{1}$$

$$P_{x_i,n} = R_{x_i,n} * A * np_n * \eta * (1 + ((T_{x_i,n} - 298.15) * \gamma)) \tag{2}$$

Where: np - Number of solar panels; P - The power produced in kW; x_i - Corresponds to the i-th predicted point; R - Irradiation $\frac{kW}{m^2}$; A - Area of a solar panel in m^2; η - Efficiency of the solar panel $\%$; T - Temperature in K; γ - Temperature coefficient $\%/C$; n - Substation n.

C. 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.

*1) **Root mean squared error - RMSE and Relative root mean squared error - RRMSE**:* 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. 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. 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.

$$RMSE(y_i, x_i) = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2} \tag{3}$$

$$RRMSE(y_i, x_i, n) = \frac{RMSE(y_i, x_i)}{S_n} \tag{4}$$

Where: n - Total sample size; y_i - The i-th measured point; x_i - The prediction of the i-th value; S_n - Installed power in substation n kVA.

*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 5 [28]. 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 6. This metric is applied in the compile stage of both TabNet and NN to evaluate the training and it is part of the evaluation step.

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{At}\right| \tag{5}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t| \tag{6}$$

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

IV. Results and Discussion

The Results and Discussion chapter will, as best as possible, show the outcome of the proposed methods in the III - Methodology section.

A. *Pre-processing*

*1) **Missing Values**:* The first step when entering the pre-processing stage was the sum of all the NaN entries for each feature.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 III-A1, 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.

Concerning the power produced feature, substations 2,3,4,8 and 10 presented 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.Results showed that substations 2 and 10 can be 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.

*2) **Cleansing in irradiation data**:* The cleansing of irradiation data was done using the *Astral* package and had great results in the data changing all values outside of sunrise and sunset to 0 for each day. Since in a first glance, the irradiation data demonstrated an abnormal distribution (positive values of irradiation during night-time and peak irradiation close to sunset) this cleansing was expected to have a great performance in the forecasting stage. The implementation of the same cleansing in power production was less notable which is an indicator of a better distribution of points throughout the day.

*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 I 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.

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 I
NUMBER OF TOTAL CONSECUTIVE DAYS WITHOUT PRODUCTION IN EACH SUBSTATION.

B. *Results of forecasting*

The results were evaluated following the metrics discussed in III-C 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.

*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.

*a) **Power produced**:*

The power produced feature, being the most important of this work, was trained for the three algorithms, and the forecast results showed 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. Results showed that XGB had better performance in 3 out of 8 substations; the RRMSE was lower for this algorithm considering the 8 substations analyzed; and the processing time is undeniably lower with XGBoost taking an average of 23:30 minutes to be optimized and trained while TabNet took an average of 136 hours, meaning that a lot more can be done if the XGB is adopted for further optimizations in this scenario. TabNet can, if seen individually, be a better approach to forecasting the power produced, 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. Figure 1 shows the best prediction achieved for a day with an RRMSE of 0.006 kW with a clear overlap of both curves throughout the whole day.

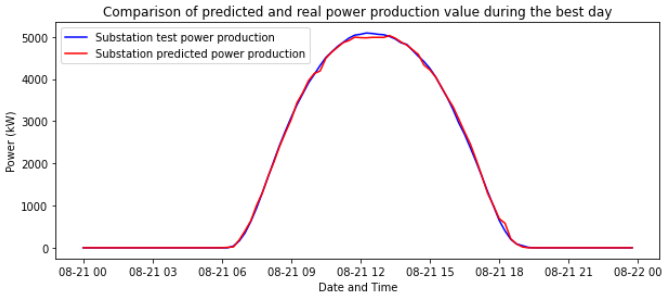

Fig. 1. Substation 6 registered predicted and calculated power during a day

*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).

The results of training both temperature and irradiation showed that the temperature can be easily predicted using the XGB without many optimizations needed 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, this high error can occur due to a high variation of values which is more challenging for the model to predict (when compared with the temperature), and the nature of the data that present an unnatural distribution as discussed before. Because the temperature is always positive during the night, the poor projection of the irradiation led to the option in this approach of not using the physical model because it would mislead the work by creating a number of time samples with positive power during the night hours worsening the training and consequent error between the power produced and the power calculated. In a way of trying to perfect these results, a new approach will be made in Section IV-B2.

*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 IV-B1 showed a distinguishable performance by the XGB algorithm, it was adopted for this approach too.

The results from Approach B showed a distinguished increase in performance in both irradiation and the power produced, the first feature had an increase of over 30% with substation 6 reaching 40% improvement, and the power also saw an increase in some substations but lower with the new average RRMSE reaching 0.1390 kW, 4% lower than Approach A. Contrary to Approach A, the plot shown demonstrates the worst day forecasted. Figure 2 demonstrates that the model used in this substation was able to predict the power during sunrise that reached a value higher than 400 kW but then decreased drastically and stopped, while in the test data, it had a normal behavior of a high increase during the sunrise while decreasing near the sunset. When comparing the same day in both approaches, Approach B still presented a lower error with an RRMSE = 0.4619 kW for this day.

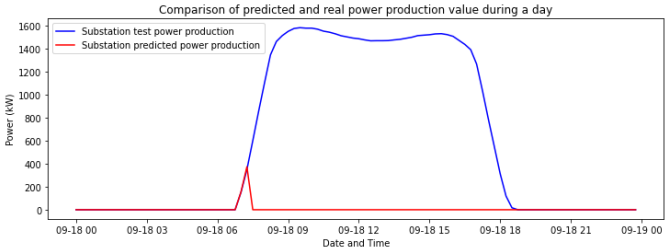

Fig. 2. Substation 4 registered and predicted power during a day

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 IV-B2, the irradiation is now within the same time frame as the power produced; therefore, it is possible to infer over the use of a physical model also using the temperature that was forecasted previously.

*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 1 and 2 this number was achieved and following that the respective RMSE and RRMSE were computed. In Table II are the results of this study. The results are higher than previous forecasts with an increase in the average RRMSE from 0.14134 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.

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 II
NUMBER OF PANELS PER SUBSTATION RRMSE CALCULATION ON APPROACH B.

In Figure 3 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.

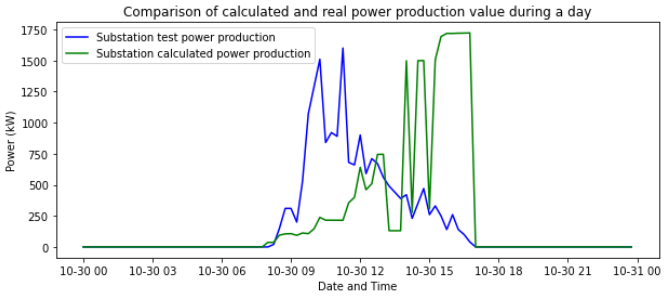

Fig. 3. Substation 6 registered predicted and calculated power during a day

*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 was 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.

The results in Table III show that removing several days without power in the power production feature had a great effect on the RRMSE. Compared with section IV-B2, 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). Concerning the remaining features, this approach lead to a decrease in performance in the majority of substations 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.

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 III
RRMSE WHEN APPLIED APPROACH C IN THE POWER PRODUCED FEATURE.

Figure 4 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.

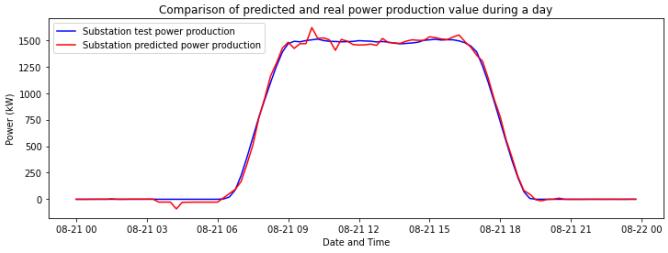

Fig. 4. Substation 8 registered and predicted power during a day

*a) **Forecast using the physical model (Approach C):***

By performing this new power comparison and computing the difference with approach B, the results do not show a significant change in the errors.

Registering a RRMSE of 0.4436 kW in substation 9, Figure 5 shows the best and worst days regarding 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 previous approaches.

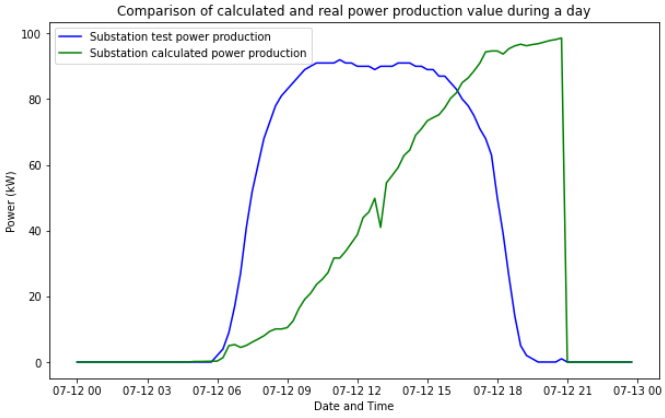

Fig. 5. Substation 9 registered and predicted power during a day

V. Conclusion

A. 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 [29] presented different algorithms for solar irradiation forecasting in complicated weather conditions which following the constraints in this research is a more suitable comparison. In [29] the best model found, for all weather, was LSTM with an RMSE of 66.69 W/m^2. [30] achieved an RMSE of 62.1618 W/m^2 for daily solar irradiance in Model II-BD based on LSTM. In [31] 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 V-B.

B. 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 II. 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.

ACKNOWLEDGMENT

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.

REFERENCES

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

[15] A. Bramm, S. Eroshenko, and A. Khalyasmaa, "Effect of data pre-processing 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.

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

[17] 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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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

[23] 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.

[24] 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

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

[26] 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

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

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

[29] 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.

[30] 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.

[31] 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/