# TIME SERIES PREDICTION OF SOLAR POWER GENERATION USING TREND DECOMPOSITION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÜRCAN KAVAKÇI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

FEBRUARY 2021

Approval of the thesis:

**TIME SERIES PREDICTION OF SOLAR POWER GENERATION USING TREND DECOMPOSITION**

submitted by **GÜRCAN KAVAKÇI** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** ______________

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** ______________

Assoc. Prof. Dr. Şeyda Ertekin
Supervisor, **Computer Engineering, METU** ______________

**Examining Committee Members:**

Prof. Dr. Nihan Kesim Çiçekli
Computer Engineering, METU ______________

Assoc. Prof. Dr. Şeyda Ertekin
Computer Engineering, METU ______________

Prof. Dr. Mehmet Reşit Tolun
Computer Engineering, Konya Food And Agriculture University ______________

Date: 10.02.2021

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Surname: Gürcan Kavakçı

Signature :

# ABSTRACT

## TIME SERIES PREDICTION OF SOLAR POWER GENERATION USING TREND DECOMPOSITION

Kavakçı, Gürcan
M.S., Department of Computer Engineering
Supervisor: Assoc. Prof. Dr. Şeyda Ertekin

February 2021, 66 pages

Accurate predictions are desirable in time series data due to the widespread usage of them in various domains. Each information in the data represents the characteristics of the time series. Making forecasting on data that has trend information is a complicated process. In this thesis, new methods are proposed to make better estimates on time series data which have trend information. In the first part of the study, features such as mean and trend were extracted from the history of the existing data by feature extraction methods and added to the data set as features. When machine learning algorithms were tested with this extended data set, better results were obtained compared to existing methods. In the second part of the study, trend decomposition was applied to the data. More stable data obtained after the decomposition was tested with the existing models, and then the final estimation was achieved by combining the decomposed trend data with the prediction results of the stable data. Higher performance results were observed than what was achieved by using the plain data and also the data with extended features. Then, in the third part of the study, linear estimation method was used to make predictions on the trend data as well. The final results were obtained by combining the predicted results of both the stable time series data and the trend. At each step, we demonstrate superior or competitive prediction performance than the previous step and the existing method in five different machine learning algorithms. Proposed methods are applied to the renewable energy domain and used in the forecasting of solar power generation in Turkey.

Keywords: Time Series Prediction, Trend Decomposition, Solar Power Generation Forecasting, Machine Learning Algorithms

# ÖZ

## TREND AYRIŞTIRMASI KULLANARAK GÜNEŞ ENERJİSİ ÜRETİMİNİN ZAMAN SERİSİ TAHMİNİ

Kavakçı, Gürcan
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Doç. Dr. Şeyda Ertekin

Şubat 2021 , 66 sayfa

Zaman serisi verileri çeşitli alanlarda yaygın olarak kullanıldıkları için doğru tahminlenmeleri önemlidir. Verilerdeki her bilgi, zaman serilerinin özelliklerini temsil eder. Trend bilgisine sahip veriler üzerinde tahmin yapmak karmaşık bir süreçtir. Bu tezde, trend bilgisine sahip zaman serisi verileri üzerinde daha iyi tahminler yapabilmek için yeni yöntemler önerilmiştir. Çalışmanın ilk bölümünde, özellik çıkarma yöntemleri ile mevcut verilerin geçmişinden ortalama ve eğilim gibi özellikler çıkarılmış ve özellik olarak veri setine eklenmiştir. Genişletilmiş bu veri seti ile makine öğrenimi algoritmaları test edildiğinde, mevcut yöntemlere göre daha iyi sonuçlar elde edilmiştir. Çalışmanın ikinci bölümünde verilere trend ayrıştırması uygulanmıştır. Ayrışmadan sonra elde edilen daha kararlı veriler, mevcut modellerle test edilmiş ve ardından ayrıştırılmış trend verileri ile kararlı verilerin tahmin sonuçları birleştirilerek nihai tahmin gerçekleştirilmiştir. Düz veriler ve ayrıca genişletilmiş özelliklere sahip veriler kullanılarak elde edilen sonuçlardan daha yüksek performans sonuçları gözlemlenmiştir. Daha sonra çalışmanın üçüncü bölümünde, trend verileri üzerinde de tahminlerde bulunmak için doğrusal tahmin yöntemi kullanılmıştır. Nihai sonuçlar, hem kararlı zaman serisi verilerinin hem de trendin tahmin edilen sonuçları ile birleştirilerek elde edilmiştir. Her adımda, beş farklı makine öğrenimi algoritmasında önceki adımdan ve mevcut yöntemden daha üstün veya rekabetçi tahmin performansı sergiliyoruz. Önerilen yöntemler yenilenebilir enerji alanına uygulanmıştır ve Türkiye'de güneş enerjisi üretiminin tahmin edilmesinde kullanılmıştır.

Anahtar Kelimeler: Zaman Serisi Tahminleme, Trend Ayrıştırma, Güneş Enerjisi Üretim Tahmini, Makine Öğrenimi Algoritmaları

I dedicate this thesis to my beloved family and those who read this page.

# ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor Assoc. Prof. Dr. Şeyda Ertekin for her support, guidance, insights through this journey and feedback.

I would like to thank my thesis committee members, Prof. Dr. Nihan Kesim Çiçekli and Prof. Dr. Mehmet Reşit Tolun for their valuable feedback on this thesis.

I also wish to thank Dr. Ümit Çavuş Büyükşahin for valuable discussion about my thesis.

I also want to thank my company Esen Sistem Entegrasyon and all my managers.

I would like to thank my parents Mehmet and Meryem Kavakçı, as well as Mustafa and Hatice Irmak, for their incredible support.

Last but not least, my dearest appreciations go to my beloved wife Ayşen, who always stands back of me and provides full support with all her heart.

# TABLE OF CONTENTS

## LIST OF TABLES

TABLES

## LIST OF FIGURES

FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| ANN | Artificial Neural Network |
| RF | Random Forest |
| SVR | Support Vector Regression |
| SVM | Support Vector Machines |
| GBM | Gradient Boosting Machine |
| LGBM | Light Gradient Boosting Machine |
| KNN | K-Nearest Neighbors |
| GHI | Global Horizontal Irradiance |
| DNI | Direct Normal Irradiance |
| DHI | Diffuse Horizontal Irradiance |
| MAE | Mean Absolute Error |
| RMSE | Root Mean Square Error |
| WAM | Weighted Arithmetic Mean |
| MW | Megawatt |
| API | Application Programming Interface |
| ACA | Ant Colony Algorithm |
| OLS | Ordinary Least Squares |
| MA | Moving Average |
| SPO | Solar Power Output |
| ARIMA | Autoregressive Integrated Moving Average |
| MLR | Multivariate Linear Regression |

# CHAPTER 1

# INTRODUCTION

Renewable energy became more popular than ever because of its widely usage. Solar power generation has started to be used frequently by large and small businesses and home users. Variable factors in solar energy generation cause different generation capacities over time. Even if solar panels are installed in the same place, location, weather conditions, time zone, and amount of daylight cause different amounts of energy generation. Electricity produced from solar energy is sold to the market by both small and large operators. Operators have to decide when to sell and store electricity due to the variable market policy. Besides, small producers and consumers (prosumers) want to know how much energy they will produce in the short term because of the constant need for energy usage. Therefore, small and large producers need to estimate how much energy they will produce in the next period. In this way, they know how much energy they need to produce from fossil fuel or how much energy they should procure if solar energy is not enough. It is not easy to make a standard estimate, as inputs that influence generation are very variable. Hence, created models with machine learning algorithms are used for accurate forecasts in the market.

In the solar power domain, there are time series obtained with timestamp and various solar data. Time series is the observation of events occurring within and associating with a fixed period with a time stamp. Time series analysis is the analysis of these events that occur within the period for both prediction and historical cause and consequence relation. Thanks to the developing technology, large amounts of time series data are continuously produced and stored in many areas. Movement information from many sensors in spacecraft in space technology, hourly and daily market data in the economy, data collected from devices connected to patients in the health sector, sales prices in trade, generation and consumption values in the energy sector, application usage activities of computer users are examples of time series data type. Time series prediction aims to make predictions in these areas. For example, it tries to predict which page computer users will visit in the next step from the application usage history. Extraordinary situations and dangers are tried to be predicted from the sensor data in spacecraft. In the energy sector, the amount of energy that will be demanded in the near future and will be supplied at the same time is tried to be predicted. In the economy, it is tried to predict the monetary policies to be presented in the future according to the market situation. Making accurate predictions in each area is vital to making important decisions in that area.

## 1.1 Problem Definition

It is possible to make highly accurate predictions when we have a sufficient amount of data and when the data has a specific pattern for a long time. However, it becomes difficult to make highly accurate prediction if the data format does not show a specific pattern and changes continuously. The main problem with solar prediction is that the power output is directly dependent on natural environmental factors in the relevant time period. For example, solar energy has a specific pattern throughout the year. Solar power generation increases in summer and decreases in winter. Simultaneously, if weather forecast data are also provided, solar energy generation is more accurately predicted.

Due to the nature of machine learning models, the models use the data structure learned from old data to make predictions. For accurate estimation, the characteristics of the old data and the data to be estimated must be similar. However, generation of solar power is very erratic due to it is direct dependence on weather conditions, seasonal changes, location, time of day, and positioning of solar panels. Thus forecasting methods can not produce uniformly efficient results for all regions.

Due to the increasing number of installed solar energy generation facilities, the generation amount increases day by day. Although solar power generation shows ups and downs seasonally, it has an increasing trend. For this reason, differences occur in the same period of consecutive years. Even if the data structure is determined seasonally, the increasing trend of solar power generation disrupts the determined structure and prevents high-accuracy predictions since current models do not take this upward trend into account.

### 1.2 Problem Solution

Considering that solar generation shows typical behavior patterns with time-varying, seasonal and trend patterns and working on stable data is the strength of existing models, instead of training the models with the increasing trend of the data, it is crucial shaping existing data to a stable form. Because mathematical and statistical models forecast the possible generation based on past statistics. For this, unstable data can be divided into two components as trend and seasonal. Thus, two stable data are obtained from unstable data. Since trend data does not contain any seasonal variations, the future value of trend data is predicted with high accuracy by using different methods suitable for linear estimation. Due to seasonal data being stable, the future value of seasonal data is predicted with high accuracy by using various machine learning algorithms suitable for nonlinear estimation. Thus, a final highly accurate forecast is obtained by combining two predicted results.

### 1.3 Motivation & Objectives

Prediction methods of solar photovoltaic power plant generations have been investigated by many studies in order to find those that are prone to this domain. Previous research and literature reviews have shown that various machine learning algorithms, such as ANN [1], SVR [2], Random Forest [3], KNN [4] and LGBM [5], have been used to predict solar energy. In these studies, various data have been studied with existing algorithms or hybrid models. These algorithms work with the data as it is, or they try to add new features into the data, but they do not incorporate the trend feature of the data properly.

Since solar power generation data is time series data, this problem is a time series problem. Time series are divided into two categories as linear and nonlinear data. While better results are obtained by linear estimation methods on linear data, non-linear methods are not efficient in such data. On the contrary, complex data is not predicted sufficiently accurate by linear methods.

It is observed from the literature survey that there is a need for improved models for solar power prediction. In this thesis, a solution is sought to the trend problem. In this problem, we propose different ways of incorporating trend features of data to the machine learning algorithms. Therefore, trend data and current seasonal data are separated to avoid this problem. It is aimed to predict trend data in linear form by linear estimation methods and seasonal data by nonlinear methods. In this way, the strengths of each method are used for both data structures.

By solving the trend problem, a day-ahead forecast will be more accurate. Thus, consumers and producers will use more accurate predictions to predict a day-ahead generation and make more accurate forward-looking decisions. For example, small producers will check whether sufficient generation will be provided the next day and plan to produce or purchase energy from different sources. Besides, major manufacturers will be able to decide how much storage capacity they will need for the next day based on the next day's forecasting.

In addition, in solar power community, system operators need accurate predictions for correct system operations such as monitoring, maintenance, dispatching, and also for generation scheduling. They can compare actual and predicted solar power generation to see if there is any problem with the system, or they can plan for future supply. They can be prepared for the load demand without overestimating or underestimating the generation of distributed sources.

## 1.4 The Outline of The Thesis

In the first part of the thesis, the problem is defined, and the solution is proposed. Work motivation is mentioned, and the objective proposal is summarized. In Chapter 2, the time series concept and its patterns are explained. Inner features of time series like trend and seasonality are described. In addition, the chapter discusses the usage area of time series. In chapter 3, previous studies and their limits about solar power generation and forecasting are presented. Chapter 4 discusses the content of the data set and how it is collected. Besides, how the data set prepared to be used in the experiments is explained. In chapter 5 proposed model is mentioned. Also, intermediate methods that contribute to the final solution are discussed. The MAE and RMSE performance results of the developed models are presented with the existing models and intermediate models. As a result, conclusions are given in Chapter 6.

# CHAPTER 2

# TIME SERIES

Time series are discussed in this section. In the first part, information about the time series is explained. In the second part, time series data structure is mentioned. Finally, data set examples and their properties are summarized.

## 2.1 Introduction to Time Series

Time series is a collection of statistical data usually collected at regular intervals. In the time series data, each row of data is ordered according to time. Hence, each data has a definite relationship with the time it is obtained, as shown in Table 2.1. Measurements in time series can be taken as hourly, daily, monthly, yearly or any regular interval. If a future value is unknown until a certain date, new data at the related time is estimated with time series [6]. The data pattern is an important factor in understanding what kind of time series the data was in the past because it is expected to behave same in the future. The pattern of data is used to select an appropriate prediction method.

## 2.2 Time Series Data Structure

Time series consist of a combination of one or more data with a characteristic structure. These characteristics are;

- Trend
- Seasonality

Table 2.1: Structure of Time Series

| Time | Observation |
|---|---|
| time 1 | observation in time 1 |
| time 2 | observation in time 2 |
| time 3 | observation in time 3 |
| time 4 | observation in time 4 |
| time 5 | observation in time 5 |

- Noise

The trend is the perception of an increase or decrease in the data that is observed continuously. The trend is not a mandatory characteristic of data. It is typical for the trend to decrease in one part of the data and increase in the other part. For example, it is observed that the daily mobile shopping volume increases due to the widespread usage of technology [7]. Likewise, an increasing trend occurs in the total amount of energy generation since the number of installed solar panels increases day by day. While electricity usage decreases in summer, it has an increasing trend in winter due to the necessity for heating.

Seasonality is a pattern that repeats over a period of time. It is not a mandatory characteristic of data like trend. For example, the increase in electricity usage in the winter, the decrease in the summer, and the continuation of this cycle every year show that electricity usage has an annual pattern. Likewise, another example is that user visits to a website increases in the morning but decreases at night. In this example, it is seen a daily repeating pattern. Patterns in time series can be seconds, minutes, or even years, depending on the frequency of data collection. Hence seasonality is significant in predicting the future value of the time series.

Noise is such of anomaly caused by unpredictable factors that remain when the trend and seasonality parts are removed from data. For example, even if gold prices rise every year during the summer months, in some specific cases, such as economic crises, the fact that gold prices fall during that summer can be considered as noise.

The observed data can be explained with these three structures as follows:

$$O(t) = S(t) + T(t) + N(t) \tag{2.1}$$

where *O(t)* is observed data; *S(t)*, *T(t)*, *N(t)* are seasonality, trend and noise in time *t*, respectively. These structures are usually analyzed in time series forecasting.

Time series data are classified as univariate and multivariate. In the univariate dataset, there is only data observed at the related time. In the multivariate dataset, there is ancillary information obtained simultaneously with the observed data. For example, solar energy generation values on their own are univariate data. This dataset includes time and observed data. However, a multivariate dataset is obtained by including other data that contribute directly or indirectly to the observed data at the same time. For example, to obtain multivariate data, weather information, cloud rate, rain rate and power generation locations are added to the data in addition to the solar generation value. Univariate datasets are more straightforward to visualize than multivariate ones.

## 2.3 Time Series as Examples

We provide here four well-described datasets in order to describe seasonality, trend and noise, and to visualize.

### 2.3.1 Minimum Daily Temperatures Data

The dataset shown in Figure 2.1 is from Australia. This dataset, consisting of 3650 measurements, is a smooth seasonality data and no noticeable trend. It is a proper dataset to use in research. It has high resolution and a uniform seasonality [8].

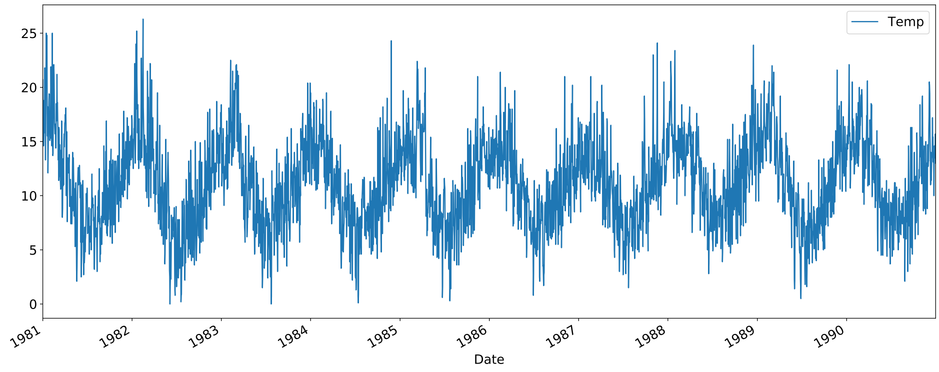


Figure 2.1: Minimum Daily Temperatures Data

### 2.3.2 Sunspot Data

Sunspot dataset is astronomy data measured monthly in 235 years between 1749-1983. As can be seen from Figure 2.2, it has seasonality, but there are many noise data [8].

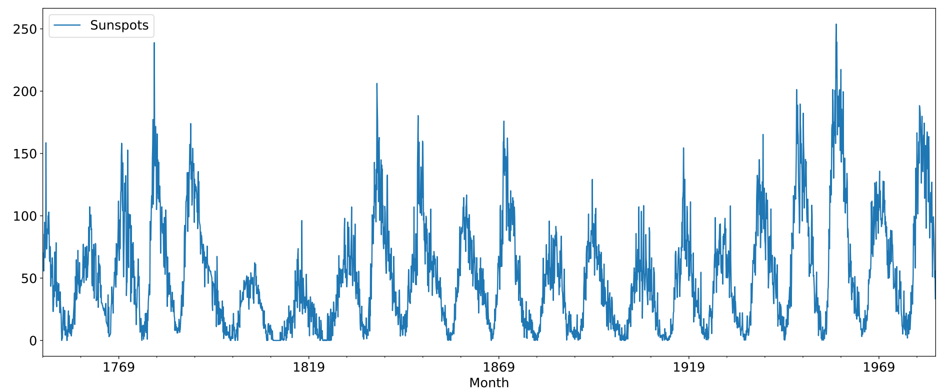


Figure 2.2: Sunspot Data

### 2.3.3 Daily Female Births Data

It is the daily number of girls born in California in 1959 with 365 measurements. The dataset does not contain a significant trend or seasonality, as in Figure 2.3. In this dataset, forecasting is more complicated than other datasets [8].

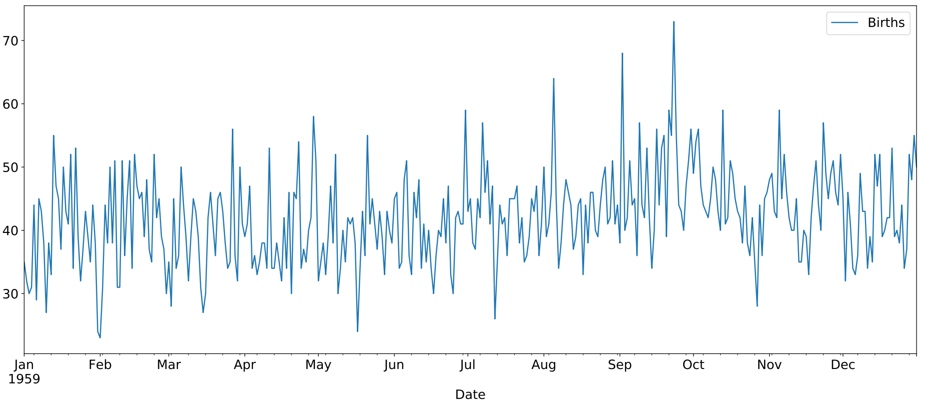


Figure 2.3: Daily Female Births Data

### 2.3.4 Shampoo Sales Data

It is monthly shampoo sales data for three years. It has a total of 36 observations in 3 years. As can be seen from Figure 2.4, it is a data set with an increasing trend. It also includes some seasonalities in it [8].

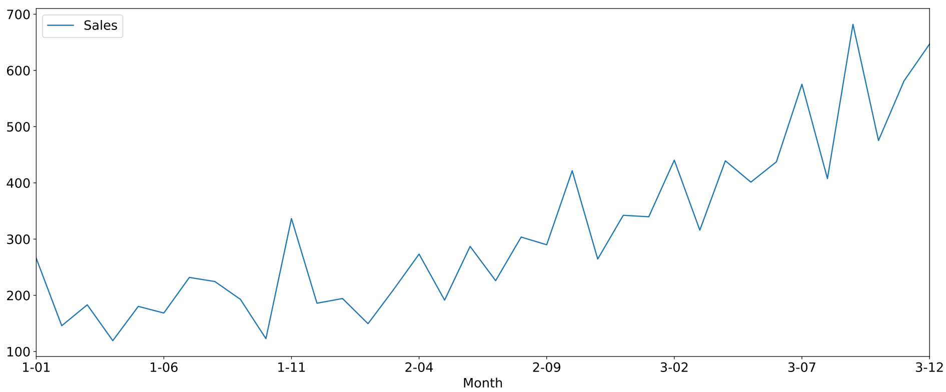


Figure 2.4: Shampoo Sales Data

# CHAPTER 3

# RELATED WORK

In solar power generation, forecasting using time series has become topic. In this domain, in order to improve predictions, machine learning algorithms are used commonly. Therefore, many researchers around the world have a number of approaches corresponding with this problem in the literature [9] [10] [11][12]. Physical models, statistical models, adaptive and trained models, and combined forecasting models are traditional methods for solar power generation estimations [13] [14] [15] [16].

Machine learning algorithms use two main groups of model-based methods. These models are multivariate model-based methods [17] [18] [12] [19] and univariate model-based methods [9]. Multivariate model-based methods use multiple input values to estimate solar power prediction such as temperature, cloudiness and clearness indices, solar irradiance, relative humidity, wind speed. In contrast, univariate methods include just past values of generation. Methods that are univariate model-based can be classified as linear and nonlinear models. Linear models are autoregressive and autoregressive moving average models [20]. Additionally, nonlinear models are neural networks [21], support vector machine [22], decision tree [23], k-nearest neighbors [23].

In statistical models, evaluations are made by establishing a correlation between the collected time series numeric data for long-term estimates. Adaptive and trained models use artificial intelligence (AI) algorithm that is trained on historical data. A relation is set between the solar forecast information and the output produced power. Unlike statistical models, complicated non-linear connections are created. Available historical data are used for both statistical and trained models.

Artificial intelligence methods learn the relationship between the output power and the data that directly or indirectly contributes to the output power of solar power plants. Unlike statistical algorithms, AI algorithms are more powerful because AI algorithms create a complex and non-linear model between input and output data. The precondition for both AI and statistical algorithms is gathering time series data that include historical measured data and power outputs with high-quality [24] [25].

Especially, Artificial Neural Network (ANN) is a commonly used algorithm in time series. To calculate the effect of each input on the output in machine learning models is a powerful feature of neural networks. This feature of the neural network provides more accurate results than standard algorithms. Also, in order to produce successful results for ANN, the parameters should be well tuned [26]. ANNs depend on resolution, scale and prediction variables commonly [27].

It is impractical and difficult to determine whether time series data are generated linearly or nonlinearly. Also, it is not easy to determine that one model performs better than another model in time series data. Hence choosing the right technique for each experiment is a challenge. A suitable method with high accuracy can be found for any existing situation, but it is not easy to know which is best for future studies. Because changes may occur on the data, or it is also possible that the data do not have a fixed structure. The difficulty of model selection can be alleviated by combining more than one method. Real-world data are not always linear or nonlinear. They can often show different variations. In such cases, neither linear nor nonlinear prediction algorithms are suitable for the data. Because linear methods are not suitable for nonlinear data, and nonlinear methods cannot predict linear and nonlinear data with the same accuracy.

Denton [28] showed that neural networks give better results than linear models if there are outliers or multicollinearity in the data. Markham and Rakes [29] showed that ANN's performance on linear problems is dependent on sample size and noise level. Therefore, it has shown that ANN cannot be blindly applied to all kinds of data. Consequently, the data can be modeled with high accuracy by combining linear and nonlinear models. There have been studies showing that there is no single method to deal with all kinds of structures in the literature [30] [31]. The reason for this, the real-world data are more complex than just linear or nonlinear data. Various researches exist in the literature to find a solution to this problem [32][33] [34]. It was shown in [35] that using a hybrid model in time series reduces the uncertainty of the model because of the possible unstable or changing patterns in the data.

Zhang [32] proposed a hybrid method using ANN and autoregressive integrated moving average (ARIMA). These are successful algorithms in their linear and nonlinear domains. However, none of them is a universal solution to handle all situations. The ARIMA method is not suitable for complex nonlinear data. ANN is not suitable for linear problems. The hybrid system proposed by Zhang [32] consists of two stages. In the first step, the ARIMA model is used to analyze the linear part of the data. In the second step, the neural network model is used to model the residual of the ARIMA model. Hence, linear and nonlinear parts can be modeled separately with different models, and the results can be combined. Combined methods are an effective way to improve forecasting performance. The experiments showed that the hybrid method obtained better results than the results obtained by applying each component model separately.

Khashei and Bijari [33] also proposed a hybrid method to yield more general and more accurate model than traditional models. Khashei and Bijari [33] considered time series as a function of linear and nonlinear components. In the first step, ARIMA is used to recognize the linear structure of the data. In the second step, multi-layer perceptrons are used as neural networks to describe preprocessed data. The proposed method's suitability and efficiency are shown on time series data sets. In this study, they guaranteed that the performance of the hybrid method they proposed would not be worse than ARIMA and ANN. Khashei and Bijari's model [33] is shown to be a more general and more accurate method than Zhang's model [32].

The hybrid method proposed by Babu and Reddy [34] is suitable for both one-step ahead and multi-step ahead predictions. In the proposed method, the time series data is decomposed into two components using the moving average (MA) filter method. While the less volatile part is the trend which is modelled with ARIMA, the high volatility which is modelled with ANN is the residual after the less volatile part is removed from the data. As a result, the final prediction is obtained by adding the two results. The proposed method's accuracy is much better than Zhang's [32] and Khashei-Bijari's [33] method in terms of MAE and MSE.

The new hybrid method proposed by Büyükşahin and Ertekin [36] does not use the assumptions used in previous studies and solves the performance degeneration problem that occurs in adverse situations. Firstly MA method is used to characterize nonlinearity. Then ARIMA is applied to the linear component. In the final stage, ANN is used to combine the result of ARIMA, the nonlinear component and the original data. While extracting or modeling components, assumptions are not made such as being the linear components are the output of ARIMA, modelling residual data with ANN, or the results are a combination of linear and nonlinear structure. In addition, the hybrid method is improved with the EMD technique. The proposed hybrid method, which does not make any assumptions, creates a more general model and shows that it gives remarkably superior results than previous studies [32] [33] [34].

Majidpour et al.[9] aim to make super short-term fast prediction using four well-known algorithms (ARIMA, kNN, SVR, RF) with univariate data. The purpose of this study is to make one step ahead prediction of solar-generated power in dynamic control systems. Since the fast result is the first objective of this study, the algorithm only uses the history of the data. As a result, it is observed that each algorithm gives different results for different error criteria. For example, while kNN is the best for SMAPE, RF is the best for MAE. In this study, machine learning-based algorithms (SVR, RF and kNN) outperformed traditional ARIMA method. It has also been shown that RF and kNN are better than SVR. It is also shown that RF and kNN are better than SVR.

In some researches, it is studied to improve solar power prediction accuracy [10] [11]. Zheng et al.[10] indicate that the research on multi-region solar power is rare. Therefore, this study contributes to solar power prediction by using LSTM and particle swarm optimization algorithms. Particle swarm optimization is used to optimize parameters to improve prediction. This study aims to predict the total solar power output (SPO) of power systems in multi-regions. In the proposed method, multi-region solar power plant's time series data and measurement points are collected in the database. Next, the proposed SPO algorithm optimizes the parameters for LSTM.

After obtaining optimal parameters such as learning rate, time step, batches, number of neural units, the results of different LSTM structures are compared. The appropriate LSTM is selected according to training loss. Then the selected LSTM is trained. The proposed method [10] was tested with eight solar power plants in five different regions. As a result, four situations, including different seasons, show that the proposed method is superior to other methods.

Sharadga et al.[11] compared different time series forecasting models for PV output power prediction. The studied methods are statistical and artificial intelligence methods. ARMA, ARIMA and SARIMA are tested as statistical models, BI-LSTM, LSTM, fuzzy c-mean clustering, LR, NN, MLP and feed-forward NN are tested as neural networks. In this study, BI-LSTM algorithm, which increases power prediction accuracy for large-scale power plants, is proposed. Besides, different NN and statistical approaches are compared for time series forecasting of large-scale PV systems. This study shows that NNs are more accurate than the suggested statistical models. However, it is indicated that both neural networks and statistical models can be used for one-hour ahead prediction where solar irradiance and weather measurements are not available. So time series forecasting for PV plants are reliable only for one hour ahead.

Prema and Rao [37] studied on time series of solar power to find the best time horizon to get the most accurate predictions. They used time series models such as moving average models, exponential smoothing models and decomposition models. Multiplicative decomposition was selected due to seasonal value changes over time. By using these techniques, it was found that the decomposition model works well with a max data duration of 2 months and the prediction can be extended up to a week without much increase in the error in short-term predictions of solar irradiance.

Solar power prediction is an important research area. There are various researches in the literature for more accurate prediction of future solar power generation from renewable energy plants [38] [39] [40] [41] [42]. In [43] it is shown that SVR based model outperforms a reference autoregressive (AR) model by using historical data of atmospheric transmissivity. İzgi et al. [44] show that ANN gives the best solar power prediction in 5 min time horizon for short term and 35 min for medium term for solar PV data. In [45] daily solar power prediction using ANN, kNN, SVM and MLR are studied as data driven approaches on two inputs which are included and excluded meteorological parameters. It is showed that none of them outperform others in all prediction scenarios. Lorenz et al. [46] present an approach to predict regional photovoltaic power output based on forecasts up to three days. For regional forecasts, they show that accuracy is increasing depending on the size of the region. It is also shown that the accuracy of the GHI prediction is the determining factor for the accuracy of the power forecast. A software tool is developed to get the best solar power management in [47] . Results show a daily solar radiation prediction with high accuracy, with correlation coefficients of 0.89. A hybrid model that combines ML methods with theta statistical method is proposed by Alkandari and Ahmad [48]. Two diversity techniques, i.e. structural diversity and data diversity, are employed. Results show that a hybrid model combining ML methods and statistical method outperforms a hybrid model that only combines ML models without statistical method. Gensler et al. [49] introduced powerful deep learning algorithms such as Deep Belief Networks, AutoEncoder, and LSTM in the field of energy power forecasting. In [49] they use combinations of these deep learning algorithms to take strengths compared to standard multilayer perceptron and physical forecasting model and show that superior forecasting performance compared to ANN and other referenced physical models are achieved.

In the method we proposed, our aim is to show that unlike other studies, the decomposed data can be predicted with higher accuracy with any machine learning algorithm. Although ANN is generally used in the literature, predictors in real life might prefer using different ML algorithms due to various reasons. Predictors may not be able to change the algorithm on their side. In this case, they can improve their predictions by integrating our proposed method into their existing structures. The method we proposed is a general method that works with ML algorithms and can be applied to any time series that has a trend. Thus, highly accurate predictions can be achieved by using different ML algorithms after decomposition.

# CHAPTER 4

# SOLAR POWER DATA

The total installed power is 1362.60 megawatts (MW) in Turkey by 1644 solar power plants. Solar energy systems work in two different concepts. The first one is the photovoltaic system, and the second one is the thermal system. Radiation transmitted through the sun in the photovoltaic system is converted to energy owing to solar panels. Thus, generated energy becomes suitable for the inverter. In the thermal system, solar energy is carried to a certain point by the help of mirrors, and a particular substance (oil, water, etc.) is heated at this point. Energy is obtained with the heated substance by using steam pressure [50].

In Turkey, 564 stations are registered, and the generation capacity is 5095 megawatt-electrical (MWe). Annual electricity generation is approximately 2433 gigawatt-hours (GWh).

Furthermore, 23 cities have above 10 MW of installed power in Turkey. Konya has the highest installed power with a ratio of 0.29. The total installed power of 23 cities that produce most is 1499 MW.

Turkey solar power generation data obtained from Enerji Piyasaları İşletme Anonim Şirketi (EPİAŞ) is used for experiments in this thesis. The obtained data is a time series data. Solar power generation data consists of hourly data from October 2017 to October 2019 as Figure 4.1. The data contains 24 hours time slots which are derived both during the day and at night when generation is low. Figure 4.2 is a hourly data example of a day of winter, and figure 4.3 is a hourly data example of a day of summer.

Night generation is deficient and may be ignored. Besides, there is sufficient generation in summer between 6-7 a.m. and 5-6 p.m., but not in winter.

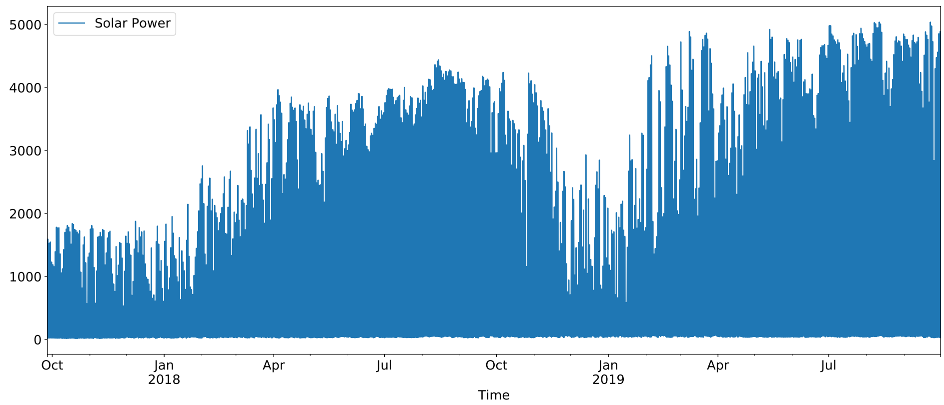


Figure 4.1: Solar Data From October 2017 to October 2019

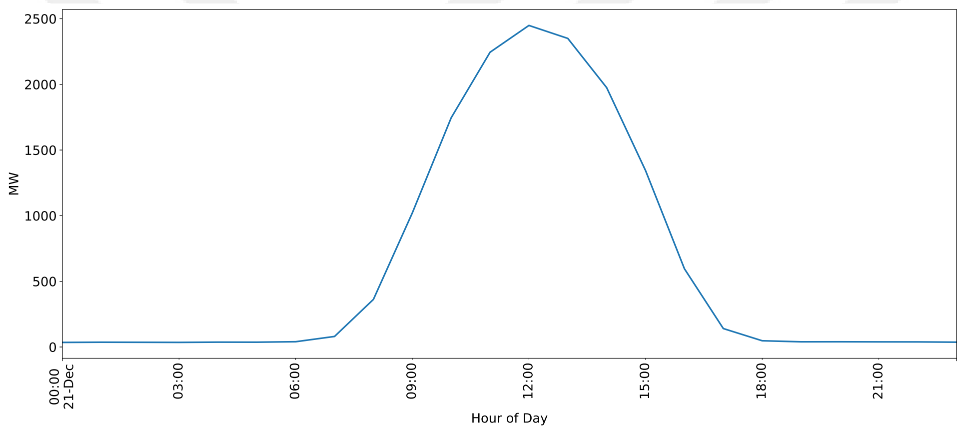


Figure 4.2: Solar Power Generation on 21/12/2018

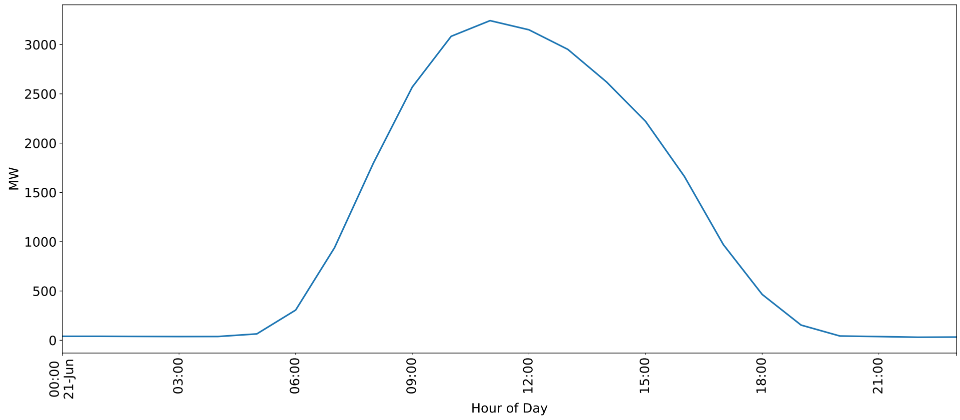


Figure 4.3: Solar Power Generation on 21/06/2018

## 4.1 Prediction on Solar Power Data

Solar power data is also a time series. Making accurate predictions on solar power data is important for system operations. The advantages of accurate predictions on solar power are as follows:

Dispatchability: The daily operations of reliable energy systems depend mainly on a day ahead distribution of the power plant [51]. Meaningful plans for the next day are only possible with correct a day ahead predictions. For this, both generation and consumption must be predicted correctly.

Efficiency: Output power fluctuations are caused by interruptions in sources due to the frequency and voltage fluctuations in the system. Some countries penalize producers for this situation. Therefore, it is not efficient for producers to underestimate to avoid penalties.

Monitoring: Sometimes, it may indicate an error in the system if there is more than expected difference between the predicted value and the produced value. It is essential to be able to take action before it is too late.

## 4.2 Data Preparation

We aim to estimate the total generation of the next day with the time series data we have. For this purpose, we have the generations of previous days. Since generation is not linear as Figure 4.4 and many factors affect solar power generation, we add new features to the data to make the data more meaningful. Two methods are introduced below for adding new features.

### 4.2.1 Historical Data

Two kinds of historical data is used in data preparation. For the data considering the last three days, we add the solar power generation data belonging to 48 and 72 hours ago to the data set. For the data considering the last seven days, we add also the generation data belonging to 48, 72, 96, 120, 144 and 168 hours ago to the data

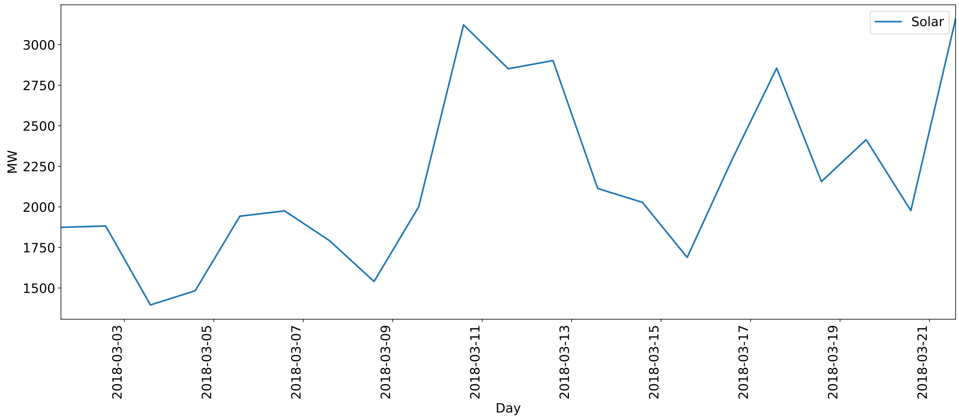


Figure 4.4: Solar Power Generation in March 2018 at 02:00 pm

set. Also we append the month and quarter information to both data set. Solar power generation data belonging to 24 hours ago is not used because it is not provided yet.

### 4.2.2 Solar Irradiance Indices

Irradiance is solar power measured in a range at a particular time. Its unit is watts per square meter. Irradiance is divided into different components such as DNI, DHI, GHI by different measurements [52].

#### 4.2.2.1 DNI

Direct Normal Irradiance (DNI) is the amount of light reaching perpendicular to the surface [53]. Direct irradiance can be calculated as subtracting extraterrestrial irradiance far away from the atmosphere from atmospheric absorption and scattering by clouds. Losses are related directly to the time of day, cloud and moisture [54]. DNI is the most important part of global irradiance.

#### 4.2.2.2 DHI

Diffuse Horizontal Irradiance (DHI) is terrestrial irradiance received by a horizontal surface scattered or emitted by the atmosphere [52]. It is not arrive on a direct path from the sun. DHI is a part of GHI. Pyranometer is used to measure DHI. Unlike DNI, DHI is measured with blocking direct light of sun. The pyranometer is also be shaded [55].

#### 4.2.2.3 GHI

Global Horizontal Irradiance (GHI) is the amount of terrestrial irradiance that falls on the earth's surface horizontally [52]. GHI is total solar radiation. The global irradiance is calculated with both direct and diffuse irradiance as:

$$GHI = DNI \times \cos\theta_z + DHI \tag{4.1}$$

where $\theta_z$ is solar zenith angle [56].

Two kinds of irradiance indices are used in second part of data preparation.

The first dataset which we call *combined* dataset is Turkey's solar index. Because we merge solar indexes of all cities by using weighted arithmetic mean (WAM) as

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \tag{4.2}$$

where $x_i$ in non-empty finite multiset of data $\{x_1, x_2, ..., x_n\}$, with corresponding non-negative weights $\{w_1, w_2, ..., w_n\}$ [57]. As a result, we have just one solar index group, which includes GHI, DNI, and DHI.

The second dataset which we call *separated* dataset is solar indexes of Turkey's cities. Because we use all cities solar indexes separately in the dataset. We take into account the first 23 cities' solar indexes. First 23 cities generate more than 10 MW of energy, as shown in Table 4.1.

For 23 cities, we calculate irradiance indexes and separately add calculated irradiance indexes to feature sets for each solar index (GHI, DNI, DHI). As a result, we have 69 irradiance indexes for all influential cities.

Table 4.1: Generation Amount and Rates of Cities Above 10 MW

| City | Power (MW) | Weight (%) |
|---|---|---|
| KONYA | 440.154 | 29.36 |
| KAYSERİ | 263.85 | 17.60 |
| ANKARA | 114.851 | 7.66 |
| AFYONKARAHİSAR | 89.31 | 5.96 |
| GAZİANTEP | 85.16 | 5.68 |
| NİĞDE | 71.71 | 4.78 |
| DENİZLİ | 50.297 | 3.35 |
| BALIKESİR | 44.514 | 2.97 |
| ANTALYA | 38.85 | 2.59 |
| BURDUR | 38.65 | 2.58 |
| MERSİN | 33.883 | 2.26 |
| ELAZIĞ | 32.75 | 2.18 |
| ISPARTA | 25.35 | 1.69 |
| NEVŞEHİR | 23.43 | 1.56 |
| VAN | 20.4 | 1.36 |
| KAHRAMANMARAŞ | 20.156 | 1.34 |
| MANİSA | 19.14 | 1.28 |
| AYDIN | 18.5 | 1.23 |
| İZMİR | 16.104 | 1.07 |
| ADANA | 15.897 | 1.06 |
| AKSARAY | 13.2 | 0.88 |
| ERZURUM | 12.18 | 0.81 |
| SİVAS | 11.06 | 0.74 |

### 4.2.3 Parts of Data

The data used in the thesis consists of two parts. The first part is the original data. The second one is the prepared part as computing with both original data and external data. The architect of proposed data preparation method is shown in Figure 4.5.

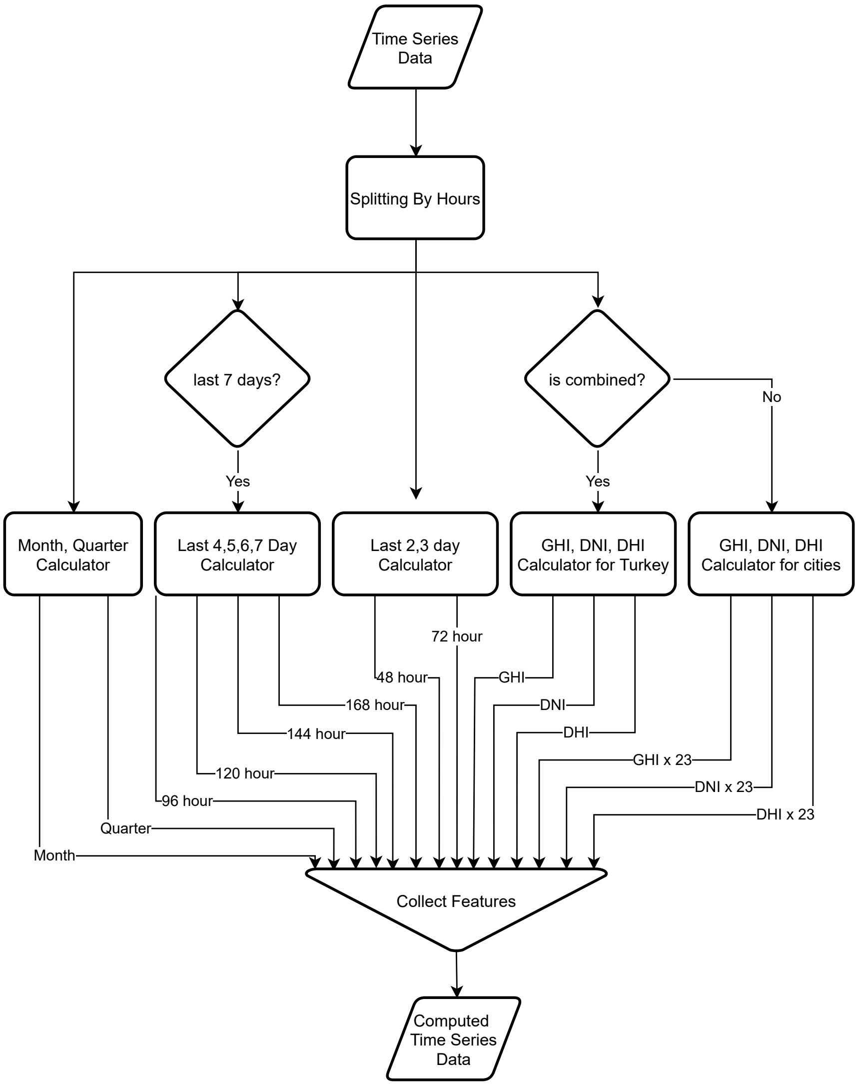


Figure 4.5: Proposed Data Preparation Method

#### 4.2.3.1 Original Data

The original data is the collected data that has been observed by the systems and available for users. In the original data, there are two features that are shown as

Table 4.2: Original Data from Solar Power Generation on 21/06/2019

| Date | Generation |
|---|---|
| 21/06/2019 0:00 | 39.13 |
| 21/06/2019 1:00 | 38.62 |
| 21/06/2019 2:00 | 38.97 |
| 21/06/2019 3:00 | 38.03 |
| 21/06/2019 4:00 | 37.54 |
| 21/06/2019 5:00 | 68.34 |
| 21/06/2019 6:00 | 362.67 |
| 21/06/2019 7:00 | 1112 |
| 21/06/2019 8:00 | 2132.96 |
| 21/06/2019 9:00 | 3047.79 |
| 21/06/2019 10:00 | 3645.16 |
| 21/06/2019 11:00 | 3876.84 |
| 21/06/2019 12:00 | 3814.45 |
| 21/06/2019 13:00 | 3701.97 |
| 21/06/2019 14:00 | 3412.16 |
| 21/06/2019 15:00 | 2921.15 |
| 21/06/2019 16:00 | 2225.35 |
| 21/06/2019 17:00 | 1388.03 |
| 21/06/2019 18:00 | 640.73 |
| 21/06/2019 19:00 | 197.38 |
| 21/06/2019 20:00 | 54.95 |
| 21/06/2019 21:00 | 47.86 |
| 21/06/2019 22:00 | 45.17 |
| 21/06/2019 23:00 | 44.78 |

follows:

- **DateTime**: Information when the data was obtained.
- **Solar**: The corresponding total generation data in MW. It was obtained overall Turkey at the related time.

Table 4.2 shows time and solar generation on 21/06/2019.

#### 4.2.3.2 Computed Data

Four kinds of the dataset are prepared with original data and external information. When preparing the first type of data set, the criterion used is the data belongs to how many days before. The criteria are the last three days and the last seven days. The criterion used in preparing the second type of data set is whether solar indices will be calculated on a country basis or city basis. Firstly each cities solar indexes are considered separately, and it is called a 'separate dataset'. Secondly, Turkey's combined solar index, called 'combined dataset', is calculated. The combination of the 4 data sets is as follows:

- Combined last 3 days
- Combined last 7 days
- Separated last 3 days
- Separated last 7 days

These four datasets which are *combined last 3 days*, *combined last 7 days*, *separated last 3 days*, *separated last 7 days* are combination of two datasets and will be used with name *Feature Set 1*, *Feature Set 2*, *Feature Set 3*, *Feature Set 4* respectively. Features of these datasets are;

- For all datasets
  - Month
  - Quarter
  - Solar power of 48 hours before
  - Solar power of 72 hours before
- For last 7 days datasets
  - Solar power of 96 hours before
  - Solar power of 120 hours before
  - Solar power of 144 hours before

    - Solar power of 168 hours before

- For combined datasets
    - GHI
    - DNI
    - DHI
- For separated datasets
    - GHI x 23
    - DNI x 23
    - DHI x 23

Table 4.3 and Table 4.4 shows a sample of computed dataset on 15/06/2019 to 21/06/2019 at 12:00 pm for combined and separated datasets respectively.

## 4.3 Characteristics of Data

The data has some characteristic features. Since solar energy generation is directly connected to solar power, the solar energy panels' positions according to the sun affect the generation. While there is minimal generation at night (probably due to moonlight), the amount of daily generation increases in a daylight, and it reaches the peak at noon, as shown in Figure 4.2 and 4.3. If the hourly generation capacity is examined, much generation is observed in the day, and it becomes close to 0 at night. Besides the day and night relationship, the generation amount is different in specific periods of the year, whether it is a month or a season. In summer, generation is much higher than in winter, as Figure 4.7. In addition, the generation is high since the days are longer in the summer months, as Figure 4.3. In the winter months, the sun's visibility is less and causes generation to decrease, as Figure 4.2. The boxplot of the generation amount by month is shown in Figure 4.6

Another factor affecting energy generation is the geographic location where the solar panels are located because the amount and the angle of the sun coming to the re-

Table 4.3: Computed Combined Data from Solar Power Generation on 15/06/2019 to 21/06/2019 at 12:00 pm

| Features | 15/06/2019 | 16/06/2019 | 17/06/2019 | 18/06/2019 | 19/06/2019 | 20/06/2019 | 21/06/2019 |
|---|---|---|---|---|---|---|---|
| Solar | 3950.44 | 3958.87 | 4186.84 | 3554.57 | 3451.14 | 3317.6 | 3814.45 |
| Solar_prev_48 | 3873.92 | 3860.94 | 3950.44 | 3958.87 | 4186.84 | 3554.57 | 3451.14 |
| Solar_prev_72 | 3836.58 | 3873.92 | 3860.94 | 3950.44 | 3958.87 | 4186.84 | 3554.57 |
| Solar_prev_96 | 4076.55 | 3836.58 | 3873.92 | 3860.94 | 3950.44 | 3958.87 | 4186.84 |
| Solar_prev_120 | 4090.75 | 4076.55 | 3836.58 | 3873.92 | 3860.94 | 3950.44 | 3958.87 |
| Solar_prev_144 | 3854.24 | 4090.75 | 4076.55 | 3836.58 | 3873.92 | 3860.94 | 3950.44 |
| Solar_prev_168 | 4765.76 | 3854.24 | 4090.75 | 4076.55 | 3836.58 | 3873.92 | 3860.94 |
| month | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| quarter | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| ghi | 863.788912 | 864.093288 | 864.3549 | 864.573297 | 864.747985 | 864.878423 | 864.964023 |
| dni | 817.322525 | 816.576963 | 815.822139 | 815.057878 | 814.283991 | 813.500274 | 812.706508 |
| dhi | 171.471629 | 171.941546 | 172.406316 | 172.86588 | 173.320176 | 173.769135 | 174.212684 |

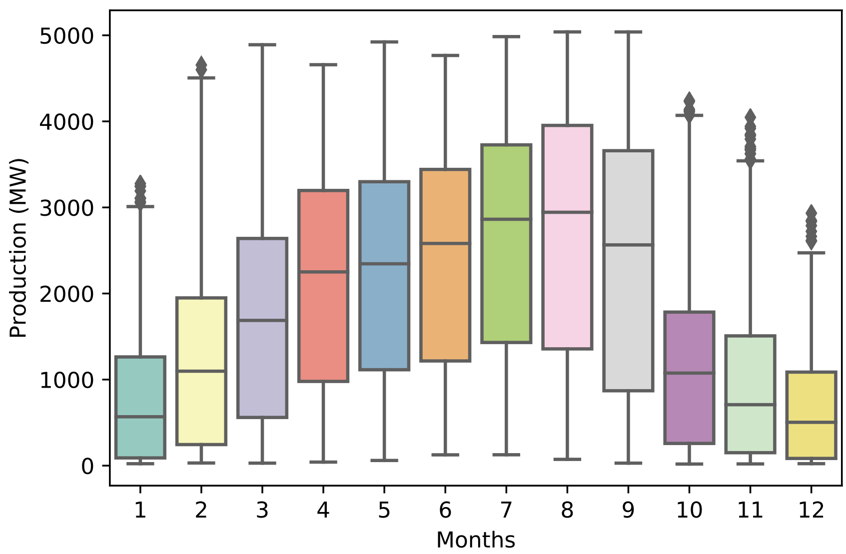


Figure 4.6: Box Plot of Generation by Month

Table 4.4: Computed Separated Data from Solar Power Generation on 15/06/2019 to 21/06/2019 at 12:00 pm

| Features | 15/06/2019 | 16/06/2019 | 17/06/2019 | 18/06/2019 | 19/06/2019 | 20/06/2019 | 21/06/2019 |
|---|---|---|---|---|---|---|---|
| Solar | 3950.44 | 3958.87 | 4186.84 | 3554.57 | 3451.14 | 3317.6 | 3814.45 |
| Solar_prev_48 | 3873.92 | 3860.94 | 3950.44 | 3958.87 | 4186.84 | 3554.57 | 3451.14 |
| Solar_prev_72 | 3836.58 | 3873.92 | 3860.94 | 3950.44 | 3958.87 | 4186.84 | 3554.57 |
| Solar_prev_96 | 4076.55 | 3836.58 | 3873.92 | 3860.94 | 3950.44 | 3958.87 | 4186.84 |
| Solar_prev_120 | 4090.75 | 4076.55 | 3836.58 | 3873.92 | 3860.94 | 3950.44 | 3958.87 |
| Solar_prev_144 | 3854.24 | 4090.75 | 4076.55 | 3836.58 | 3873.92 | 3860.94 | 3950.44 |
| Solar_prev_168 | 4765.76 | 3854.24 | 4090.75 | 4076.55 | 3836.58 | 3873.92 | 3860.94 |
| month | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| quarter | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| KONYA_ghi | 891.190194 | 891.344192 | 891.456255 | 891.525946 | 891.552783 | 891.536235 | 891.475725 |
| KONYA_dni | 853.720563 | 852.568529 | 851.409046 | 850.241914 | 849.066921 | 847.88384 | 846.692428 |
| KONYA_dhi | 161.007825 | 161.667151 | 162.321545 | 162.970939 | 163.615265 | 164.254446 | 164.888403 |
| KAYSERI_ghi | 847.784897 | 848.522185 | 849.21426 | 849.860623 | 850.46073 | 851.013991 | 851.51977 |
| KAYSERI_dni | 803.087736 | 803.301214 | 803.501365 | 803.688089 | 803.861274 | 804.020787 | 804.166484 |
| KAYSERI_dhi | 179.752251 | 179.827376 | 179.897872 | 179.963693 | 180.024788 | 180.0811 | 180.132567 |
| ANKARA_ghi | 820.826308 | 821.637413 | 822.402345 | 823.120664 | 823.791884 | 824.415475 | 824.990866 |
| ANKARA_dni | 730.089271 | 730.571349 | 731.037969 | 731.489053 | 731.924506 | 732.344218 | 732.748061 |
| ANKARA_dhi | 202.859845 | 202.839525 | 202.813764 | 202.782527 | 202.745777 | 202.70347 | 202.655557 |
| AFYONKARAHISAR_ghi | 883.928052 | 884.477517 | 884.983428 | 885.445381 | 885.862929 | 886.235574 | 886.562776 |
| AFYONKARAHISAR_dni | 809.224368 | 809.131211 | 809.026248 | 808.909382 | 808.780499 | 808.639472 | 808.486157 |
| AFYONKARAHISAR_dhi | 183.730479 | 183.919152 | 184.103471 | 184.283392 | 184.458864 | 184.629832 | 184.796238 |
| GAZIANTEP_ghi | 849.145594 | 849.211768 | 849.236699 | 849.219863 | 849.160692 | 849.058568 | 848.91283 |
| GAZIANTEP_dni | 857.822053 | 856.321157 | 854.814474 | 853.301737 | 851.782667 | 850.256965 | 848.724322 |
| GAZIANTEP_dhi | 144.234193 | 145.006069 | 145.772691 | 146.533986 | 147.289877 | 148.040281 | 148.78511 |
| NIGDE_ghi | 885.58292 | 885.975983 | 886.325439 | 886.630806 | 886.891548 | 887.107085 | 887.27679 |
| NIGDE_dni | 858.483331 | 857.76329 | 857.033848 | 856.294835 | 855.546065 | 854.787338 | 854.018441 |
| NIGDE_dhi | 164.727011 | 165.218942 | 165.706257 | 166.188888 | 166.666765 | 167.139813 | 167.60795 |
| DENIZLI_ghi | 839.108905 | 839.036891 | 838.926228 | 838.776565 | 838.587505 | 838.358609 | 838.089392 |
| DENIZLI_dni | 739.949361 | 738.772243 | 737.586261 | 736.391235 | 735.186966 | 733.973242 | 732.749834 |
| DENIZLI_dhi | 189.811685 | 190.387152 | 190.957602 | 191.522967 | 192.083173 | 192.63814 | 193.187786 |
| BALIKESIR_ghi | 830.944744 | 830.566287 | 830.146239 | 829.684292 | 829.180093 | 828.633249 | 828.043324 |
| BALIKESIR_dni | 724.981188 | 723.22751 | 721.466571 | 719.698138 | 717.921963 | 716.137782 | 714.345317 |
| BALIKESIR_dhi | 194.253788 | 195.039513 | 195.818391 | 196.59034 | 197.355274 | 198.113097 | 198.863712 |
| ANTALYA_ghi | 884.810308 | 884.718497 | 884.589909 | 884.424157 | 884.220809 | 883.979387 | 883.69937 |
| ANTALYA_dni | 854.063759 | 852.619818 | 851.171262 | 849.717882 | 848.259455 | 846.795748 | 845.326514 |
| ANTALYA_dhi | 141.885324 | 142.59307 | 143.296859 | 143.996643 | 144.69237 | 145.383984 | 146.071426 |
| BURDUR_ghi | 891.113253 | 891.16979 | 891.185736 | 891.1607 | 891.094241 | 890.985875 | 890.835069 |
| BURDUR_dni | 821.925327 | 820.706451 | 819.480029 | 818.24587 | 817.003765 | 815.753493 | 814.494818 |
| BURDUR_dhi | 175.936285 | 176.613599 | 177.285721 | 177.952582 | 178.61411 | 179.27023 | 179.920857 |
| MERSIN_ghi | 849.812571 | 848.825076 | 847.80229 | 846.743765 | 845.649011 | 844.517491 | 843.348626 |
| MERSIN_dni | 836.846741 | 833.223837 | 829.611632 | 826.009591 | 822.41717 | 818.833815 | 815.25896 |
| MERSIN_dhi | 144.354656 | 145.936184 | 147.504963 | 149.06095 | 150.604101 | 152.13436 | 153.651669 |
| ELAZIG_ghi | 843.16817 | 843.54792 | 843.881021 | 844.166918 | 844.405009 | 844.594644 | 844.735131 |
| ELAZIG_dni | 871.846258 | 870.927362 | 869.999086 | 869.061215 | 868.113515 | 867.155738 | 866.187621 |
| ELAZIG_dhi | 142.504635 | 143.060414 | 143.611447 | 144.157661 | 144.69898 | 145.235322 | 145.766602 |
| ISPARTA_ghi | 892.05806 | 892.076303 | 892.053646 | 891.989692 | 891.883995 | 891.736065 | 891.545363 |
| ISPARTA_dni | 818.996455 | 817.64473 | 816.285836 | 814.919565 | 813.545696 | 812.16399 | 810.774196 |
| ISPARTA_dhi | 180.894582 | 181.645644 | 182.391025 | 183.130655 | 183.864456 | 184.592346 | 185.314234 |
| NEVSEHIR_ghi | 875.706294 | 876.330672 | 876.909796 | 877.44318 | 877.930289 | 878.370541 | 878.763308 |
| NEVSEHIR_dni | 835.593813 | 835.496463 | 835.387156 | 835.265775 | 835.13219 | 834.986254 | 834.827807 |
| NEVSEHIR_dhi | 175.827818 | 176.037594 | 176.242815 | 176.443427 | 176.639369 | 176.830578 | 177.016985 |
| VAN_ghi | 820.889463 | 821.484963 | 822.030776 | 822.526249 | 822.970684 | 823.363334 | 823.703404 |
| VAN_dni | 827.459598 | 826.773807 | 826.074796 | 825.362332 | 824.636161 | 823.896013 | 823.141597 |
| VAN_dhi | 183.903036 | 184.454955 | 185.000762 | 185.540355 | 186.073626 | 186.600462 | 187.120741 |
| KAHRAMANMARAS_ghi | 827.131964 | 827.799073 | 828.42425 | 829.006977 | 829.546695 | 830.042796 | 830.494627 |
| KAHRAMANMARAS_dni | 820.362172 | 820.418284 | 820.462688 | 820.495272 | 820.515906 | 820.524447 | 820.520736 |
| KAHRAMANMARAS_dhi | 151.130977 | 151.2483 | 151.362186 | 151.472594 | 151.579475 | 151.68278 | 151.782452 |
| MANISA_ghi | 827.276359 | 827.559597 | 827.803034 | 828.006349 | 828.169178 | 828.291113 | 828.371706 |
| MANISA_dni | 704.296215 | 703.900178 | 703.491795 | 703.070952 | 702.637523 | 702.191361 | 701.732304 |
| MANISA_dhi | 204.297199 | 204.573541 | 204.845794 | 205.113912 | 205.377842 | 205.637526 | 205.892903 |
| AYDIN_ghi | 831.241871 | 831.222837 | 831.165896 | 831.070723 | 830.936951 | 830.764168 | 830.551919 |
| AYDIN_dni | 711.387717 | 710.39158 | 709.38562 | 708.369677 | 707.343579 | 706.307134 | 705.260137 |
| AYDIN_dhi | 201.856843 | 202.358828 | 202.85618 | 203.348835 | 203.836723 | 204.31977 | 204.797897 |
| IZMIR_ghi | 825.571825 | 825.911084 | 826.211216 | 826.471906 | 826.692793 | 826.873475 | 827.013507 |
| IZMIR_dni | 693.319372 | 693.049325 | 692.766709 | 692.471421 | 692.163341 | 691.842333 | 691.508245 |
| IZMIR_dhi | 210.628368 | 210.858901 | 211.085426 | 211.3079 | 211.526272 | 211.740492 | 211.950499 |
| ADANA_ghi | 845.266994 | 844.411744 | 843.520044 | 842.591438 | 841.625421 | 840.621447 | 839.578925 |
| ADANA_dni | 840.055515 | 836.690061 | 833.3327 | 829.98293 | 826.640237 | 823.304098 | 819.973977 |
| ADANA_dhi | 141.73504 | 143.203378 | 144.660443 | 146.106184 | 147.540548 | 148.963474 | 150.374894 |
| AKSARAY_ghi | 866.644677 | 867.172434 | 867.656642 | 868.096835 | 868.492501 | 868.843081 | 869.147971 |
| AKSARAY_dni | 838.332068 | 838.078 | 837.812972 | 837.536861 | 837.249528 | 836.950822 | 836.640574 |
| AKSARAY_dhi | 159.798403 | 160.050657 | 160.298964 | 160.543275 | 160.783535 | 161.019684 | 161.251659 |
| ERZURUM_ghi | 870.985813 | 871.670659 | 872.302649 | 872.881169 | 873.405556 | 873.8751 | 874.289046 |
| ERZURUM_dni | 917.33606 | 916.849849 | 916.352321 | 915.843301 | 915.3226 | 914.790011 | 914.245315 |
| ERZURUM_dhi | 151.435996 | 151.884104 | 152.326634 | 152.763506 | 153.194636 | 153.619936 | 154.039311 |
| SIVAS_ghi | 833.487777 | 834.207102 | 834.877768 | 835.499251 | 836.070983 | 836.59235 | 837.062694 |
| SIVAS_dni | 776.710254 | 776.811589 | 776.897598 | 776.96816 | 777.023136 | 777.06237 | 777.085686 |
| SIVAS_dhi | 198.352699 | 198.502765 | 198.647246 | 198.786083 | 198.919211 | 199.046558 | 199.168049 |

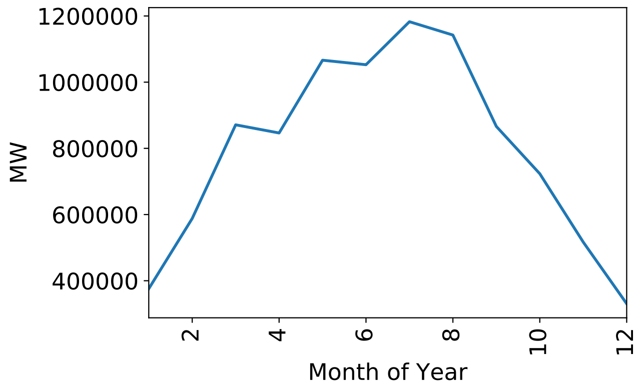


Figure 4.7: Solar Power Generation by Month

gion are different in various locations of the world. Therefore irradiance indexes are calculated according to the region where each solar panel is located.

Air conditions in the day are also the reasons that affect generation. For example, generation is high in sunny weather and low in cloudy or rainy weather. However, we will not consider weather conditions in our experiments.

## 4.4 Training and Validation Set

The part of the data set from October 2017 to October 2019 has been used in the study. Noteworthy generation data begin on October 2017. The 20-month section from October 2017 to June 2019 has been used for model development, and the 4-month section from June 2019 to October 2019 has been used for cross-validation of the model.

# CHAPTER 5

# EXPERIMENTS AND RESULTS

In this thesis, we propose a novel method that decomposes the data and calculates separately for time series forecasting, which aims to overcome the instability caused by the trend. In the first part of the study, we extract features such as mean and trend from the history of the existing data by feature extraction methods and add to the data set as features. In the second part of the study, we apply trend decomposition to the data, and we run machine learning models with separated stable data. Then, in the third part of the study, we use the linear estimation algorithm to make predictions on the trend data as well. We obtain the final results by combining the predicted results of both the stable time series data and the trend.

## 5.1 Time Series Forecasting Algorithms

This section gives a detailed explanation of the various time series models. In this thesis, we experiment with five well-known machine learning algorithms with our proposed methods.

### 5.1.1 ANN

ANN is implemented in many applications with approaches such as supervised, unsupervised, and reinforcement learning. ANN learns from the data to create a relevant connection between input and output in the supervised learning approach [58]. Backpropagation algorithms are used for training purposes so that ANNs learn a specific pattern. In the backpropagation algorithm, while the information flows between the nodes, the error information flows in the reverse direction. It is in use to determine the strength of the connections between nodes with the weight parameter. Weights are adjusted using some methods like gradient descent, and the error is reduced to the lowest level. After being trained in this way, the model runs with the least error. When new data is provided to the model, the result is obtained most accurately by calculating each input's effect on the output using the previously set weight parameters. In researches, different kinds of optimization methods are proposed for training ANN to improve the performance [59]. ANN models map past observations to future value. The output $y_t$ and the input $(y_{t-1}, y_{t-2}, ..., y_{t-p})$ can be represented as:

$$y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j g \left( \beta_{0j} + \sum_{i=1}^{p} \beta_{ij} y_{t-i} \right) + \varepsilon_t, \tag{5.1}$$

$$g(x) = \frac{1}{1 + exp(-x)}, \tag{5.2}$$

where $\alpha_j (j = 0, 1, 2, ..., q)$ and $\beta_{ij} (i = 0, 1, 2..., p; j = 1, 2, ..., q)$ are the model parameters called connection weights; q is the hidden nodes number and p is the inputs number [32].

### 5.1.2 kNN

K-nearest neighbor method (kNN), which is called lazy learner, is trained on run time in the machine learning environment. kNN is also one of the most straightforward and most uncomplicated methods since the datasets' classification is based on their nearest neighbors' class. As a result, the datasets are allocated to a more related class. The number k is used to determine how close it will be the class. Number k usually has small values. The smaller k means the closer data sets are classified to their neighbors [4].

The Euclidean distance used in kNN is the commonly used one. The distance is calculated between test and train data. The Euclidean distance between sample $x_i$

and $x_k$ (k=1,2,...,*n*) is identified as:

$$d(x_i, x_k) = \sqrt{(x_{i1}, x_{k1})^2 + (x_{i2}, x_{k2})^2 + .. + (x_{ir}, x_{kr})^2} \tag{5.3}$$

where $x_i$ is an input sample with *r* features ($x_{i1}$, $x_{i2}$, ..., $x_{ir}$), *n* is the total number of sample (*i*=1,2,...,*n*) and r is the number of feature(*j*=1,2,...,*r*) [4].

### 5.1.3 LGBM and RF

The Light Gradient Boosting Machine (LGBM) and Random Forest (RF) are two common decision tree-based group algorithms [60].

For the classification problem, boosting algorithms are offered initially. The main idea is getting more accurate predictions by joining many single models that are weaker to achieve more robust models. Gradient boosting machines (GBM) is offered by Friedman by extending the boosting [61]. GBM minimizes the loss function as a numerical optimization algorithm. Behind the scene, this algorithm uses new decision trees at each step to decrease loss function. Firstly the model starts with any guess, then each round new decision trees are provided for residual parts, and the previous one renews. Iteration continues until the given k number. By adding new decision trees to residual parts, the model is being improved where it does not work correctly [62].

Gradient boosting regression tree (GBRT) is another type of decision tree as a regression model. GBRT is different from RF by iterations. GBRT improves its inner trees by creating a new model to minimize previous improper training parts. By using an iterative approach, the accuracy of the model regularly improves. LightGBM is a distributed kind of classic GBRT. Using LGBM on parallel operations, computational costs decrease significantly [63].

The feasibility of the multi-stage model based on Random Forest optimized by Ant Colony Algorithm was verified by predicting solar radiation that is monthly sampled and gathered in three locations in Australia [64].

### 5.1.4 SVR

Support Vector Machines (SVM) is one of the selected methods in this thesis to predict solar power generation. Vapnik is the first one who proposed SVM as a new machine learning algorithm. [65]. Nowadays, satisfactory forecasting results are obtained at an electricity market price by employing a support vector machine (SVM) [66]. The simplest form of SVM classification is the maximal margin classifier. It is used to solve the most straightforward classification problem that is the binary classification case with linear separable training data [67].

Support Vector Regression (SVR) is extended from support vector machines, a popular machine learning tool. Time series predictions created using SVM, which are explored in [68], include the load forecasting.

## 5.2 Error Metrics

We use mean absolute error (MAE) and root mean square error (RMSE) for performance measurement mainly as (5.4) and (5.5). For MAE $\widehat{P}(i)$ is predicted value of solar power at the i'th hour, $P_m(i)$ is the actual value of solar power at the i'th hour.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{P}(i) - P_m(i)\right| \tag{5.4}$$

$$nRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\widehat{P}(i) - P_m(i)}{P_n(i)}\right)^2}.100 \tag{5.5}$$

Indices (5.5) and (5.6) are recently recommended by the European and International Energy Agency (IEA) for reporting irradiance model accuracy [69]

$$nMBE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\widehat{P}(i) - P_m(i)}{P_n(i)}\right).100 \tag{5.6}$$

## 5.3 Existing Algorithms

As a first step, we use existing models to determine the current performance of ANN, RF, SVR, kNN, and LGBM. We have four kinds of computed dataset which is extended from existing solar power generation dataset.

Our goal is to predict the next day's generation. For this, we need the generation values of the previous days. However, since today's generation has not been completed yet, there is no data for today. So there is a day without data between the day we try to predict and the days of the old data we have. Hence there are two more generation values that belong to 48 and 72 hours before for the last three days' solar power generation dataset. There are six more generation values that belong to 48, 72, 96, 120, 144, and 168 hours before for the last seven days' solar power generation dataset, and there is a day without data between predicted and available days again. The other two datasets have irradiance indexes (GHI, DNI, DHI) based on the calculation method, which are a country basis or city basis.

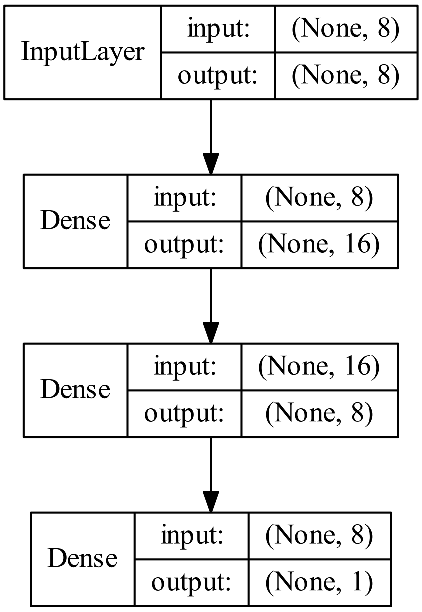


Figure 5.1: ANN Model

We use five machine learning algorithms, which are ANN, RF, SVR, kNN and LGBM, in this thesis. As the first step of the study, we run existing models on four datasets and keep the results to compare with the methods we propose. Results of existing models' error metrics are base for the proposed methods. We experiment with ANN by using a different combination of layers and input dimensions. We decide to use the ANN, whose configuration (p,q) is 3-layer with eight input dimensions (8,3) and one output dimension. The model can be seen in Figure 5.1.

ANN model is trained with 50 iterations and 1000 epochs. The Random Forest model is trained with 50 iterations, but others are trained with one iteration because the Random Forest model has randomness, but others do not have. MAE and RMSE scores of the experiments are shown in Table 5.1. Figure 5.2 shows trained and predicted part of solar power dataset. Structure of the solar power data can be seen in Figure 5.3

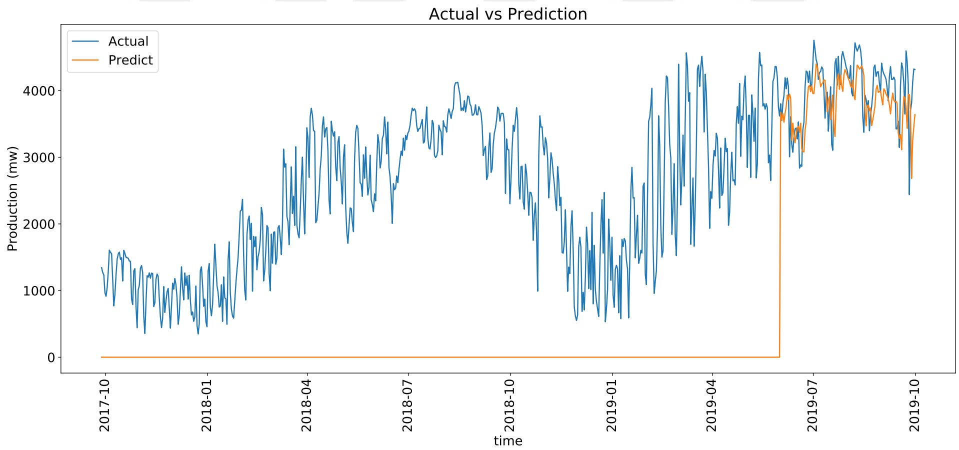


Figure 5.2: Trained and Predicted Part of Solar Power Data

## 5.4 Feature-based Modelling

In [70] it is aimed to increase the forecasting performance by using extra statistical and structural features that summarize characteristics of the time series. The proposed method tries to predict more accurate results by using determined extra characteristic features.

In this section, we study both mean and trend features mentioned in [70]. By adding

Table 5.1: MAE and RMSE of Plain Training of Machine Learning

| Model | Feature Set | MAE | RMSE |
|---|---|---|---|
| ANN | Feature Set 1 | 260.991 | 322.391 |
| | Feature Set 2 | 250.033 | 307.310 |
| | Feature Set 3 | 242.073 | 302.787 |
| | Feature Set 4 | 241.232 | 311.084 |
| RF | Feature Set 1 | 314.521 | 363.293 |
| | Feature Set 2 | 336.623 | 389.941 |
| | Feature Set 3 | 318.953 | 365.312 |
| | Feature Set 4 | 326.654 | 376.990 |
| SVR | Feature Set 1 | 229.834 | 294.076 |
| | Feature Set 2 | 213.264 | 275.314 |
| | Feature Set 3 | 244.837 | 301.686 |
| | Feature Set 4 | 220.272 | 278.124 |
| kNN | Feature Set 1 | 277.628 | 345.090 |
| | Feature Set 2 | 299.35 | 352.512 |
| | Feature Set 3 | 285.811 | 343.341 |
| | Feature Set 4 | 311.543 | 371.734 |
| LGBM | Feature Set 1 | 268.933 | 323.424 |
| | Feature Set 2 | 246.864 | 309.658 |
| | Feature Set 3 | 259.608 | 320.170 |
| | Feature Set 4 | 243.526 | 304.118 |

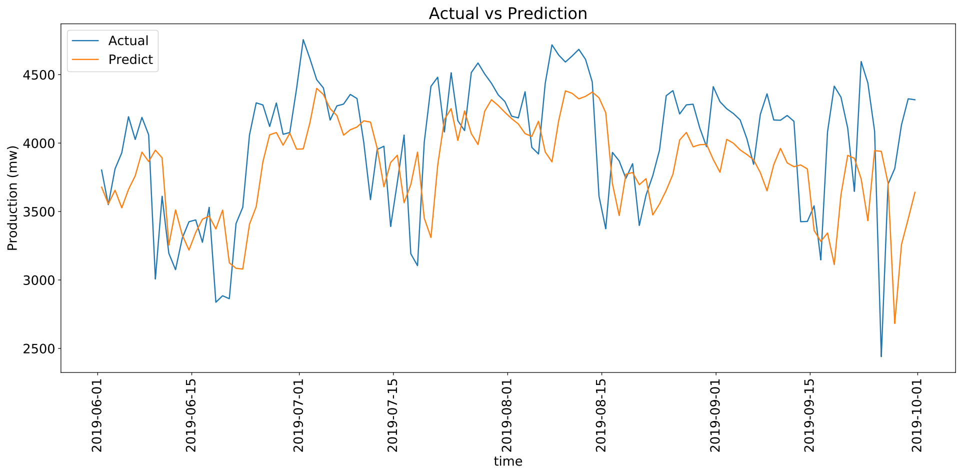


Figure 5.3: Structure of Predicted Solar Power Data

these new parameters to the multivariate dataset as a new feature, the model can detect characteristics of the data more strongly. Mean feature is calculated as

$$Mean = \frac{1}{N}\sum_{i=1}^{N} xi, \tag{5.7}$$

where N is certain length of frame and $xi$ is actual value for the i'th observation

Trend feature is calculated with Ordinary Least Squares (OLS) method of linear regression. To calculate $a$ as a trend feature, S is minimized as

$$\hat{y} = ax, \tag{5.8}$$

$$S = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - ax)^2 = \sum_{i=1}^{n}(\hat{\epsilon}_i)^2 = min, \tag{5.9}$$

where $\hat{y}_i$ is predicted value for the i'th observation, $y_i$ is actual value for the i'th observation , $\epsilon_i$ is error for the i'th observation, $n$ is total number of observations.

To calculate new features, that are mean and trend, we use $k$ as window size of the last days' values of dataset. $k$ indicates how many days of data have been used. $k$ value is variable for each of both the method and dataset and should be tuned well. We determine $k$ value by trying many combinations. Table 5.2 and Table 5.3 show comparisons of prediction results between feature based models and existing models. Also, Table 5.2 and Table 5.3 show determined *k* values for each method and dataset. Table 5.2 and Table 5.3 show that feature based models give more accurate results compared to existing models.

Feature extracted model experiments show that the dataset's mean and trend characteristics are too valuable to ignore. These results direct us to focus on the trend characteristic of solar power data. Since Turkey's capacity is increased by establishing new solar plants every day, the trend is increasing at the same time. To have increasing time series trend data prevents forecasting consistent generation information. Even if machine learning models are trained with existing solar power generation, future generation is more than now because of installing more generation capacity with new power plants.

## 5.5 Trend Decomposition

Time series are composed of a linear and nonlinear components. That is,

$$y_t = L_t + N_t, \tag{5.10}$$

where $L_t$ is linear component and $N_t$ is nonlinear component. The trend decomposition approach aims to estimate the linear part of data as a trend to separate this component from data. Stable data, which is the nonlinear part of data, can be calculated by subtracting the determined linear part of data from the original part as

$$s_t = y_t - L_t, \tag{5.11}$$

where $s_t$ denotes stable nonlinear data.

Our proposed method, which aims to solve increasing or decreasing trend problems, can make more consistent and high accuracy estimations in time series by separating trend data from original data. The trend decomposition method provides separation of trend data and stable data from original data. In this way, high accuracy results are

Table 5.2: MAE Comparison Between Existing and Feature Based Models

| Model | Feature Set | Existing Model | Feature Based Model | |
|---|---|---|---|---|
| | | | Mean (k) | Trend (k) |
| ANN | Feature Set 1 | 260.991 | 251.139 (3) | **245.787** (3) |
| | Feature Set 2 | 250.033 | **228.342** (9) | 230.092 (9) |
| | Feature Set 3 | 242.073 | **219.062** (15) | 236.925 (11) |
| | Feature Set 4 | 241.232 | 241.893 (3) | **227.404** (3) |
| RF | Feature Set 1 | 314.521 | 312.304 (3) | **308.126** (3) |
| | Feature Set 2 | 336.623 | 334.275 (3) | **320.154** (11) |
| | Feature Set 3 | 318.953 | 316.592 (3) | **314.466** (3) |
| | Feature Set 4 | 326.654 | 324.716 (3) | **320.364** (11) |
| SVR | Feature Set 1 | 229.834 | 226.056 (3) | **225.992** (3) |
| | Feature Set 2 | 213.264 | 213.076 (9) | **212.63** (9) |
| | Feature Set 3 | 244.837 | **234.33** (3) | 237.471 (3) |
| | Feature Set 4 | 220.272 | **218.454** (7) | 219.81(7) |
| kNN | Feature Set 1 | 277.628 | 280.738 (11) | **275.725** (9) |
| | Feature Set 2 | 299.35 | **296.494** (7) | 298.203 (11) |
| | Feature Set 3 | 285.811 | 263.967 (15) | **284.396** (15) |
| | Feature Set 4 | 311.543 | **308.749** (7) | 310.558 (3) |
| LGBM | Feature Set 1 | 268.933 | **247.0** (15) | 256.927 (3) |
| | Feature Set 2 | 246.864 | **242.772** (3) | 244.777 (3) |
| | Feature Set 3 | 259.608 | 269.254 (3) | **248.055** (3) |
| | Feature Set 4 | 243.526 | 245.635 (3) | **239.353** (3) |

Table 5.3: RMSE Comparison Between Existing and Feature Based Models

| Model | Feature Set | Existing Model | Feature Based Model | |
|---|---|---|---|---|
| | | | Mean (k) | Trend (k) |
| ANN | Feature Set 1 | 322.391 | 310.056 (3) | **306.943** (3) |
| | Feature Set 2 | 307.310 | **287.924** (9) | 290.907 (9) |
| | Feature Set 3 | 302.787 | **286.528** (15) | 298.729 (11) |
| | Feature Set 4 | 311.084 | 310.250 (3) | **302.679** (3) |
| RF | Feature Set 1 | 363.293 | 359.933 (3) | **356.919** (3) |
| | Feature Set 2 | 389.941 | 387.219 (3) | **373.643** (11) |
| | Feature Set 3 | 365.312 | 361.445 (3) | **361.078** (3) |
| | Feature Set 4 | 376.990 | 373.929 (3) | **371.094** (11) |
| SVR | Feature Set 1 | 294.076 | **286.876** (3) | 287.054 (3) |
| | Feature Set 2 | 275.314 | 274.518(9) | **273.603** (9) |
| | Feature Set 3 | 301.686 | 292.489 (3) | **289.638** (3) |
| | Feature Set 4 | 278.124 | 277.894 (7) | **276.524** (7) |
| kNN | Feature Set 1 | 345.090 | 338.553 (11) | **335.715** (9) |
| | Feature Set 2 | 352.512 | **348.674** (7) | 350.905 (11) |
| | Feature Set 3 | 343.341 | **327.522** (15) | 342.090 (15) |
| | Feature Set 4 | 371.734 | **368.041** (7) | 370.498 (3) |
| LGBM | Feature Set 1 | 323.424 | **308.012** (15) | 315.826 (3) |
| | Feature Set 2 | 309.658 | **305.669** (3) | 305.983 (3) |
| | Feature Set 3 | 320.170 | 322.446 (3) | **308.918** (3) |
| | Feature Set 4 | 304.118 | 302.714 (3) | **298.583** (3) |

obtained from predictions with existing machine learning models on stable data.

**Algorithm 1** Trend decomposition and forecasting

**procedure** TIMESERIESFORECAST($time\_series$)<br>
    $trend, resid \leftarrow$ TRENDDECOMPOSE($time\_series$)<br>
    $stable \leftarrow$ SUBTRACTOR($time\_series, trend$)<br>
    $predicted\_stable \leftarrow$ PREDICTWITHML($stable$)<br>
    $predicted\_trend \leftarrow trend$<br>
    **if** $isExtrapolatedPrediction = True$ **then**<br>
        $predicted\_trend \leftarrow$ PREDICTWITHLINEAR($trend$)<br>
    **end if**<br>
    $prediction \leftarrow$ MERGE($predicted\_stable, predicted\_trend$)<br>
    **return** $prediction$<br>
**end procedure**

The architect of proposed method is shown in Figure 5.4. The pseudo-code of the proposed method is indicated in Algorithm 1. In the trend decomposition function, the moving average method is used to calculate the trend part of the data. The moving average is the main structure of the decomposition methods. Moving average defines the trend cycle procedure. Each computation is done by replacing the oldest observation with the next observation. Moving average of order $p$ is defined as the average consisting of observation and the $k$ points on either side. The formula for moving average can be seen in equation 5.12.

$$\hat{T}_t = \frac{1}{2k+1} \sum_{i=-k}^{k} y_{t+i}, \tag{5.12}$$

$$p = 2k + 1 \tag{5.13}$$

where $\hat{T}_t$ is a trend at time $t$ within $p$ periods of $t$. 365 is chosen as a $p$ for experiments. In decomposition tasks, an appropriate moving average period determination is an important task. Time series data is mainly the composition of seasonal and trend patterns. The trend decomposition method decomposes original data to trend and resid data. Two kinds of decomposition models exist, such as additive and multiplicative models.

1. *Additive decomposition*: Observed data is taken as a sum of the decomposed patterns
$$O(t) = S(t) + T(t) + N(t), \tag{5.14}$$

2. *Multiplicative decomposition*: Observed data is taken as a product of the decomposed patterns
$$O(t) = S(t) * T(t) * N(t), \tag{5.15}$$

where *O(t)* is observed data; *S(t)*, *T(t)*, *N(t)* are seasonality, trend and noise in time *t*, respectively.

The additive decomposition method is selected when the seasonal data does not have peak values. However multiplicative decomposition method is selected when the seasonal data do vary much. We used the additive decomposition method because solar power data has smooth increasing over the years.

In pseudo-code, to get stable (de-trended) data, trend data is subtracted from original data as

$$StableData = OriginalData - TrendData. \tag{5.16}$$

Original, trend and stable data samples at 02:00 pm can be shown in Figure 5.5, 5.6, 5.7 respectively.

In our method, trend data, which is extracted from the original data, is used in two different ways, which will be explained in following subsections.

### 5.5.1 Using Maximum Value of Trend Data

In this method, prediction is not applied to trend data. As shown in Figure 5.4, trend data is used as-is. Stable data is applied to existing machine learning algorithms:

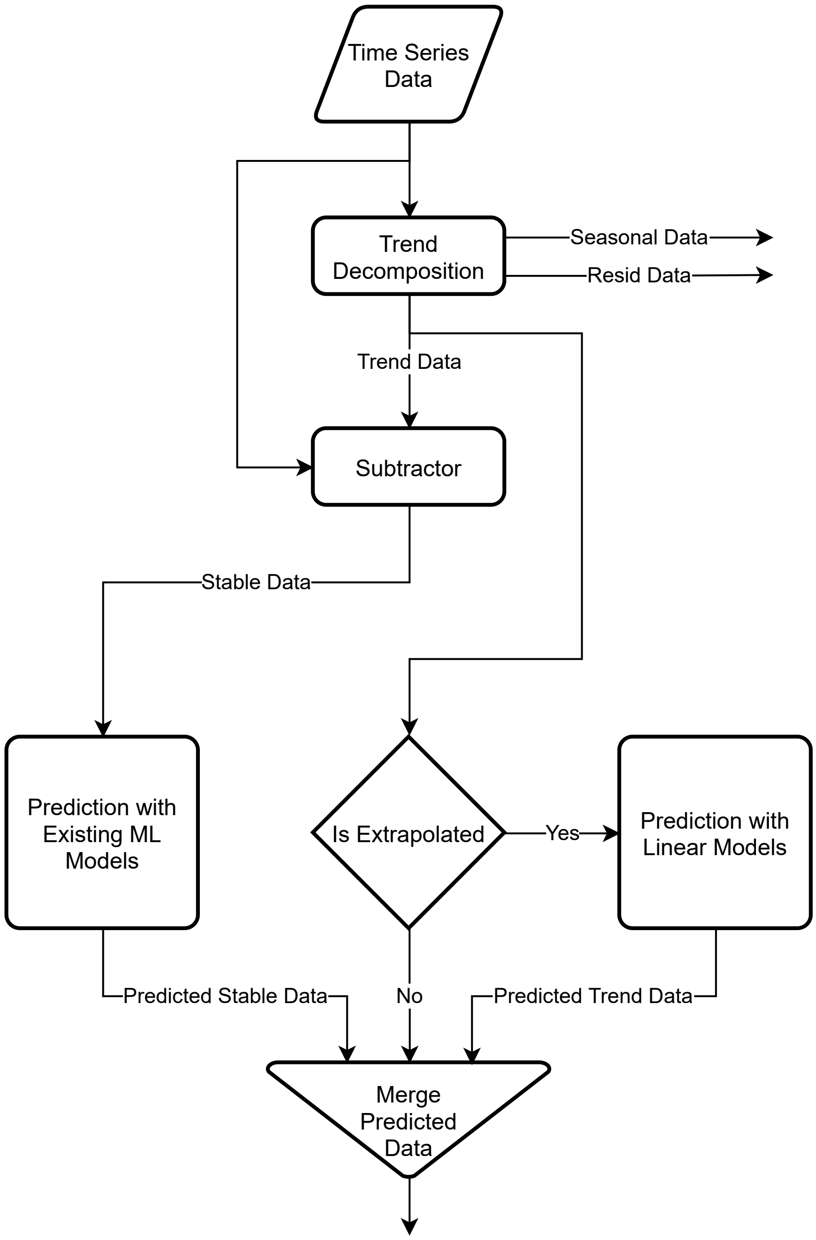


Figure 5.4: Proposed Trend Decomposition Method

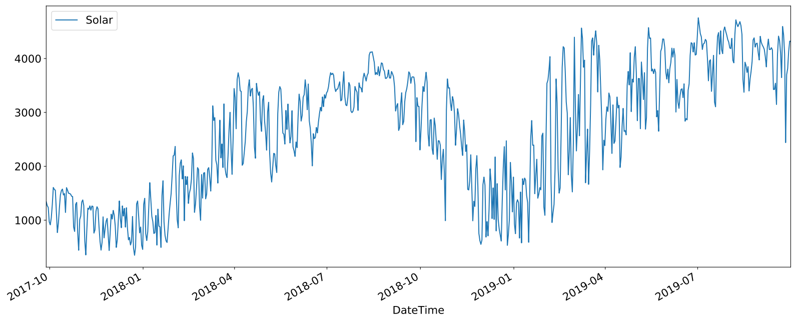


Figure 5.5: Sample of Original Data

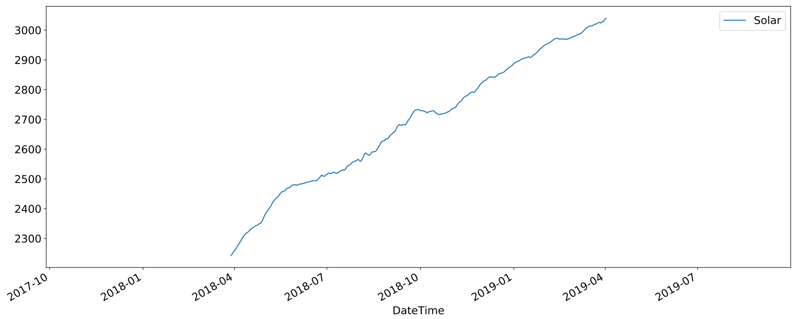


Figure 5.6: Sample of Trend Data

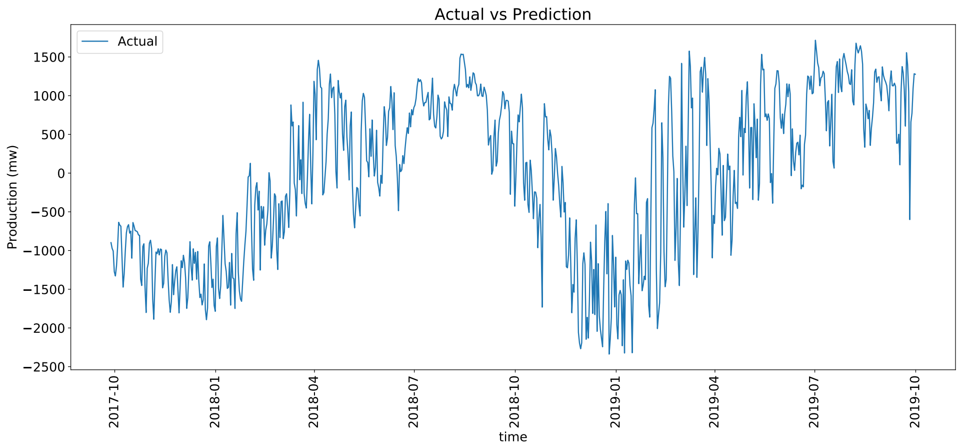


Figure 5.7: Sample of Stable Data

ANN, RF, SVR, kNN, and LGBM. After the prediction is completed, results from machine learning algorithms and decomposed trend data are summed to get the final prediction. Since no prediction is made with trend data, the maximum value is used in the merging phase. Used sample trend data is shown in Figure 5.8.

Table 5.4 and Table 5.5 show MAE and RMSE metric results of proposed trend decomposition predictions without extrapolation. Our experimental observations have shown that the trend decomposition method improves the accuracy of prediction, as shown in Tables 5.4 and 5.5.

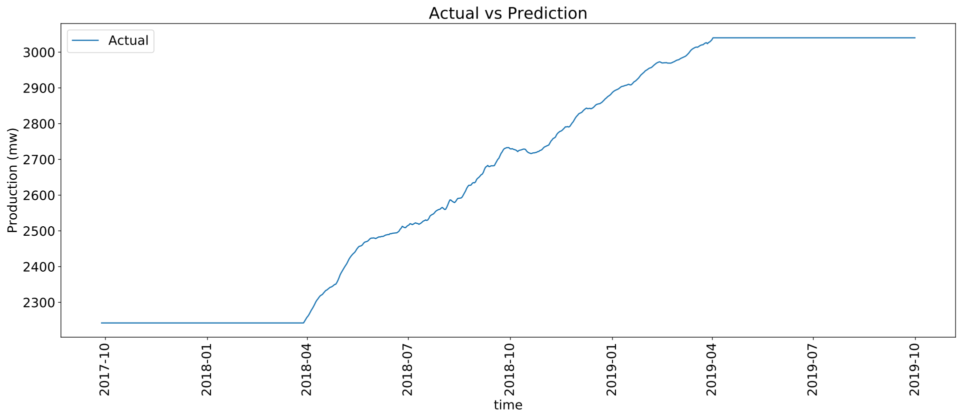


Figure 5.8: Sample Trend Data Without Extrapolation

### 5.5.2 Extrapolation on Trend Data

In Turkey, solar power installed capacity increases every day. Hence trend data in solar power generation data increases as well. It is necessary to estimate future trend data to get more accurate predictions. Although various machine learning algorithms are used to predict stable data, linear regression algorithms are used in trend estimation because linear estimation algorithms are more robust than nonlinear estimation algorithms on linear data. After stable data is predicted by various machine learning algorithms and trend data is predicted by linear prediction algorithms, the final results were obtained by combining the predicted results of both the stable time series data and the trend. Ordinary Least Squares (OLS) method is used to predict future value of trend data as in 5.8. Extrapolated sample trend data can be seen in Figure 5.9. Proposed trend decomposition method is tested on four datasets which are *combined* *last 3 days*, *combined last 7 days*, *separated last 3 days*, and *separated last 7 days*. Table 5.4 and Table 5.5 show that the trend decomposition method with extrapolation improves both standard decomposition and existing algorithm results.

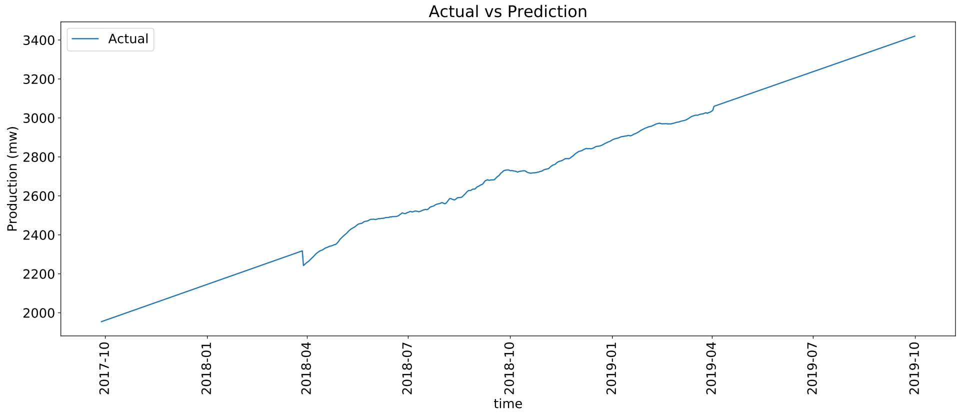


Figure 5.9: Sample Trend Data With Extrapolation

## 5.6 T-test Results

When comparing two sets of results, we can only prove how similar they are by using error metrics, and it is not possible to prove that these sets are different. The student's t-test, also known as the t-test, is the most common test for comparing samples [71].

In student's t-test, comparing the two samples (X and Y), can be calculated as follow.

$$t = \frac{m_X - m_Y}{\sqrt{\frac{S^2}{n_X} + \frac{S^2}{n_Y}}} \tag{5.17}$$

where, *mX* and *mY* represent the mean value of the group X and Y, respectively. *nX* and *nY* represent the sizes of the group X and Y, respectively. $S^2$ is an estimator of the pooled variance of the two groups. It can be calculated as follow :

$$S^2 = \frac{\sum(x - m_X)^2 + \sum(x - m_Y)^2}{n_X + n_Y - 2} \tag{5.18}$$

with degrees of freedom (df): $df = n_X + n_Y - 2$ [72].

Table 5.4: MAE Result of Trend Decomposed Model

| Model | Feature Set | Existing Model | Feature Based Model | | Trend Decomposed Model | | Improvement |
|---|---|---|---|---|---|---|---|
| | | | Mean (k) | Trend (k) | Normal | Extrapolated | (%) |
| ANN | Feature Set 1 | 260.991 | 251.139 (3) | 245.787 (3) | 220.234 | **200.58** | 23.15 |
| | Feature Set 2 | 250.033 | 228.342 (9) | 230.092 (9) | 229.33 | **215.295** | 13.89 |
| | Feature Set 3 | 242.073 | 219.062 (15) | 236.925 (11) | 218.053 | **207.816** | 14.15 |
| | Feature Set 4 | 241.232 | 241.893 (3) | **227.404** (3) | 255.973 | 237.59 | 1.51 |
| RF | Feature Set 1 | 314.521 | 312.304 (3) | 308.126 (3) | 233.044 | **202.337** | 35.67 |
| | Feature Set 2 | 336.623 | 334.275 (3) | 320.154 (11) | 246.919 | **203.746** | 39.47 |
| | Feature Set 3 | 318.953 | 316.592 (3) | 314.466 (3) | 232.023 | **199.538** | 37.44 |
| | Feature Set 4 | 326.654 | 324.716 (3) | 320.364 (11) | 240.268 | **198.183** | 39.33 |
| SVR | Feature Set 1 | 229.834 | 226.056 (3) | 225.992 (3) | 205.075 | **191.272** | 16.78 |
| | Feature Set 2 | 213.264 | 213.076 (9) | 212.63 (9) | 201.997 | **193.035** | 9.49 |
| | Feature Set 3 | 244.837 | 234.33 (3) | 237.471 (3) | 205.993 | **192.469** | 21.39 |
| | Feature Set 4 | 220.272 | 218.454 (7) | 219.81(7) | 200.897 | **191.138** | 13.23 |
| kNN | Feature Set 1 | 277.628 | 280.738 (11) | 275.725 (9) | 235.327 | **216.763** | 21.92 |
| | Feature Set 2 | 299.35 | 296.494 (7) | 298.203 (11) | 238.798 | **221.396** | 26.04 |
| | Feature Set 3 | 285.811 | 263.967 (15) | 284.396 (15) | 227.226 | **210.207** | 26.45 |
| | Feature Set 4 | 311.543 | 308.749 (7) | 310.558 (3) | 233.117 | **210.434** | 32.45 |
| LGBM | Feature Set 1 | 268.933 | 247.0 (15) | 256.927 (3) | 218.506 | **202.567** | 24.68 |
| | Feature Set 2 | 246.864 | 242.772 (3) | 244.777 (3) | 220.969 | **203.447** | 17.59 |
| | Feature Set 3 | 259.608 | 269.254 (3) | 248.055 (3) | 212.529 | **196.929** | 24.14 |
| | Feature Set 4 | 243.526 | 245.635 (3) | 239.353 (3) | 212.783 | **193.902** | 20.38 |

Table 5.5: RMSE Result of Trend Decomposed Model

| Model | Feature Set | Existing Model | Feature Based Model | | Trend Decomposed Model | | Improvement |
|---|---|---|---|---|---|---|---|
| | | | Mean (k) | Trend (k) | Normal | Extrapolated | (%) |
| ANN | Feature Set 1 | 322.391 | 310.056 (3) | 306.943 (3) | 285.218 | **267.692** | 16.97 |
| | Feature Set 2 | 307.310 | 287.924 (9) | 290.907 (9) | 229.329 | **286.713** | 6.7 |
| | Feature Set 3 | 302.787 | 286.528 (15) | 298.729 (11) | 285.311 | **280.619** | 7.32 |
| | Feature Set 4 | 311.084 | 310.250 (3) | **302.679** (3) | 343.120 | 320.033 | 2.88 |
| RF | Feature Set 1 | 363.293 | 359.933 (3) | 356.919 (3) | 287.058 | **269.425** | 25.84 |
| | Feature Set 2 | 389.941 | 387.219 (3) | 373.643 (11) | 300.190 | **267.233** | 31.47 |
| | Feature Set 3 | 365.312 | 361.445 (3) | 361.078 (3) | 283.714 | **266.698** | 26.99 |
| | Feature Set 4 | 376.990 | 373.929 (3) | 371.094 (11) | 291.695 | **263.754** | 30.04 |
| SVR | Feature Set 1 | 294.076 | 286.876 (3) | 287.054 (3) | 271.9 | **262.535** | 10.73 |
| | Feature Set 2 | 275.314 | 274.518(9) | 273.603 (9) | 266.001 | **261.874** | 4.88 |
| | Feature Set 3 | 301.686 | 292.489 (3) | 289.638 (3) | 267.601 | **263.198** | 12.76 |
| | Feature Set 4 | 278.124 | 277.894 (7) | 276.524 (7) | 263.142 | **262.341** | 5.67 |
| kNN | Feature Set 1 | 345.090 | 338.553 (11) | 335.715 (9) | 306.507 | **293.687** | 14.9 |
| | Feature Set 2 | 352.512 | 348.674 (7) | 350.905 (11) | 305.220 | **283.457** | 19.59 |
| | Feature Set 3 | 343.341 | 327.522 (15) | 342.090 (15) | 291.172 | **280.059** | 18.43 |
| | Feature Set 4 | 371.734 | 368.041 (7) | 370.498 (3) | 295.819 | **274.430** | 26.18 |
| LGBM | Feature Set 1 | 323.424 | 308.012 (15) | 315.826 (3) | 282.012 | **274.994** | 14.97 |
| | Feature Set 2 | 309.658 | 305.669 (3) | 305.983 (3) | 285.027 | **273.745** | 11.6 |
| | Feature Set 3 | 320.170 | 322.446 (3) | 308.918 (3) | 273.735 | **269.036** | 15.97 |
| | Feature Set 4 | 304.118 | 302.714 (3) | 298.583 (3) | 274.525 | **264.552** | 13.01 |

Table 5.6: P-Value of T-Test Results

| Model | Feature Set 1 | Feature Set 2 | Feature Set 3 | Feature Set 4 |
|---|---|---|---|---|
| ANN | 0.0129 | 0.018 | 0.0146 | 0.0165* |
| RF | 5.55E-05 | 7.73E-06 | 4.03E-06 | 2.77E-06 |
| SVR | 0.0245 | 0.0111* | 0.0032 | 0.0569 |
| kNN | 0.032 | 0.00234 | 0.001 | 0.000256 |
| LGBM | 0.00143 | 0.0277 | 0.00151 | 0.00882 |

* It is calculated with predictions between 08:00 am - 05:00 pm.

If the p-value is inferior or equal to the significance level of 0.05, it can be concluded that predicted values are significantly different from the existing results. Table 5.6 shows the p-value of t-test results of existing algorithms and proposed methods.

P-values in Table 5.6 shows that new results from proposed methods are significantly different from existing algorithms except for two results, which are ANN with Feature Set 4 and SVR with Feature Set 2. P-value of ANN with Feature Set 4 and SVR with Feature Set 2 are greater than 0.05. Solar power generation is close to zero in the early morning and late evening hours. Thus both existing and predicted generation values are very close. Having too close values makes them look like there is not enough difference between the results. Thus predicted values between 08:00 am - 05:00 pm are used to calculate p-values for these two model-feature set duals. As a result it shows that there is a significant difference between the existing and proposed method results.

# CHAPTER 6

# CONCLUSION AND FUTURE WORK

In this thesis, we proposed a novel time series prediction methods that significantly improve prediction performance using the underline trend information in time series data. We show that making predictions in time series data after trend decomposition improves prediction performance regardless of the machine learning algorithms used for prediction. The study can be categorized into three main groups.

In the first part of the study, we experimented featured-based model, which extracts extra descriptive features of a given time series with solar power data. We decided that the most appropriate feature to solve the trend problem mentioned in the thesis would be the trend feature. While using the trend feature, we specified a different window size for each model and data set. We tested many variations to tune window size well. In our experiments, we showed that the feature-based method, which added new features to the data set, is more successful than the current methods.

In the second part, we decomposed time series to get trend part of the data. We calculated stable data by subtracting trend data from the original data. We trained the existing machine learning models with the remaining stable data. Then, we obtained the final result by combining the maximum value of the trend data we have extracted with the prediction results. We concluded that the results obtained provide more accurate results compared to both the existing methods and the feature-based method.

In the last part of the study, we also made predictions on the trend in addition to the second part. Since trend data is linear data, we used linear estimation methods. Thus, we trained the models on both stable data and trend data, and used the strengths of each model on the relevant piece of data. The proposed method's performance was compared with ANN, kNN, SVR, RF and LGBM using MAE and RMSE. We confirmed that the method we proposed provides the most accurate result with the least error on the results obtained.

Random forest is the most improving machine learning algorithm by the proposed trend decomposition method for both error metrics. The improvements vary between 25% and 39% for both error metrics. However, SVR gives the best prediction results by the proposed trend decomposition method with the least error results. Although ANN is generally preferred for solar power prediction in the literature, we show that SVR is also good candidate in solar power generation when the data has a trend. T-test results show that the proposed method's forecasted values are significantly different from the existing method's forecasted values.

Time series data might have different components such as trend and seasonality. The methods we proposed in this thesis are general methods that can be applied to any type of time series that has a trend because our proposed methods use trend decomposition. Since Turkey's solar power data is a kind of time series, we applied our proposed methods to Turkey's solar power data. We show that our proposed methods give more accurate results than existing algorithms. All of the machine learning algorithms mentioned in this thesis are improved with proposed methods. The improvements vary between 1% and 39% for both error metrics.

In the solar power domain, hourly one-day ahead predictions are essential since making accurate predictions on solar power data is important for system operations such as dispatchability, efficiency and monitoring. Decision-makers need to know how much energy they will get for next day because they have to decide if they need to produce extra energy from other sources or they need extra storage capacity due to over generation. The proposed method helps them to forecast one-day ahead prediction for solar power generation regardless of which machine learning method they use. The solar power community can easily integrate the method we proposed to get more accurate predictions without having to change their existing solutions.

Besides, in the literature, studies have been conducted using the last day's data for the next day's predictions. However, when we predict the next day in real-world applications, the last day's generation information has not reached us yet. Hence our work actually estimates two days later. Our study differs from other studies in the literature with trend decomposition on multivariate data and two-day ahead predictions with various machine learning algorithms.

In the current study, the trend component of the time series is decomposed and predicted with linear prediction methods. A future research direction is to decompose seasonality and to predict seasonality separately. Another future research plan is to improve the prediction accuracy of decomposed trend and also to test more machine learning algorithms with proposed methods. In this thesis, the number 365 is selected as a period value for trend decomposition. Comparing different numbers as a period to find more accurate predictions on time series for trend decomposition can be future work as well.

In this thesis, we present a highly accurate time series prediction algorithm and demonstrate its performance in predicting solar power generation. We expect that our approach will provide practitioners, engineers, system operators and decision-makers in solar power industry, an effective way to make decisions based on good predictions.

# REFERENCES


[1] B. Sivaneasan, C. Yu, and K. Goh, "Solar forecasting using ann with fuzzy logic pre-processing," *Energy Procedia*, vol. 143, pp. 727–732, 2017. Leveraging Energy Technologies and Policy Options for Low Carbon Cities.

[2] A. Alfadda, R. Adhikari, M. Kuzlu, and S. Rahman, "Hour-ahead solar pv power forecasting using svr based approach," in *2017 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pp. 1–5, 2017.

[3] C.-F. Yen, Y. Hsieh, K.-W. Su, M.-C. Yu, and J.-S. Leu, "Solar power prediction via support vector machine and random forest," *E3S Web of Conferences*, vol. 69, p. 01004, 01 2018.

[4] N. A. Ramli, M. F. Abdul Hamid, and N. H. Azhan, "Solar power generation prediction by using k-nearest neighbor method," vol. 2129, 07 2019.

[5] S. R. Chowdhury, S. Mishra, A. O. Miranda, and P. K. Mallick, "Energy consumption prediction using light gradient boosting machine model," in *Advances in Power Systems and Energy Management* (N. Priyadarshi, S. Padmanaban, R. K. Ghadai, A. R. Panda, and R. Patel, eds.), (Singapore), pp. 413–422, Springer Singapore, 2021.

[6] G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. USA: Holden-Day, Inc., 1990.

[7] S. Nasir and B. Kurtuluş, *Technology Is Transforming Shopping Behavior: In-Store Mobile Technology Usage*, pp. 1508–1529. 06 2017.

[8] J. Brownlee, "7 Time Series Datasets for Machine Learning." `https://machinelearningmastery.com/time-series-datasets-for-machine-learning/`, Dec 2020. [Online; accessed 18-January-2021].

[9] M. Majidpour, H. Nazaripouya, P. Chu, H. Pota, and R. Gadh, "Fast univariate time series prediction of solar power for real-time control of energy storage system," *Forecasting*, vol. 1, pp. 107–120, 09 2018.

[10] J. Zheng, H. Zhang, Y. Dai, B. Wang, T. Zheng, Q. Liao, Y. Liang, F. Zhang, and X. Song, "Time series prediction for output of multi-region solar power plants," *Applied Energy*, vol. 257, p. 114001, 2020.

[11] H. Sharadga, S. Hajimirza, and R. S. Balog, "Time series forecasting of solar power generation for large-scale photovoltaic plants," *Renewable Energy*, vol. 150, pp. 797–807, 2020.

[12] L. Gigoni, A. Betti, E. Crisostomi, A. Franco, M. Tucci, F. Bizzarri, and D. Mucci, "Day-ahead hourly forecasting of power generation from photovoltaic plants," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 831–842, 2018.

[13] D. Heinemann, E. Lorenz, and M. Girodo, "Forecasting of solar radiation," *Solar Energy Resource Management for Electricity Generation from Local Level to Global Scale*, 01 2006.

[14] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, no. 10, pp. 1772 – 1783, 2009.

[15] C. Chen, S. Duan, T. Cai, and B. Liu, "Online 24-h solar power forecasting based on weather type classification using artificial neural network," *Solar Energy*, vol. 85, no. 11, pp. 2856 – 2870, 2011.

[16] E. Kochneva, "Solar power generation short-term forecasting model's implementation experience," *MATEC Web of Conferences*, vol. 208, p. 04005, 01 2018.

[17] Y. Kassa, J. Zhang, and D. Zheng, "Pv power forecasting using an integrated ga-pso-anfis approach and gaussian process regression based feature selection strategy," *CSEE Journal of Power and Energy Systems*, vol. 4, 06 2018.

[18] N. Tang, S. Mao, Y. Wang, and R. M. Nelms, "Solar power generation forecasting with a lasso-based approach," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1090–1099, 2018.

[19] A. Prasad and M. Kay, "Assessment of simulated solar irradiance on days of high intermittency using wrf-solar," *Energies*, vol. 13, p. 385, 01 2020.

[20] Rui Huang, T. Huang, R. Gadh, and Na Li, "Solar generation prediction using the arma model in a laboratory-level micro-grid," in *2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*, pp. 528–533, 2012.

[21] A. Asrari, T. X. Wu, and B. Ramos, "A hybrid algorithm for short-term solar power prediction—sunshine state case study," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 2, pp. 582–591, 2017.

[22] S. Belaid and A. Mellit, "Sarima-svm hybrid model for the prediction of daily global solar radiation time series," pp. 712–717, 11 2016.

[23] C. Voyant, F. Motte, G. Notton, A. Fouilloy, M.-L. Nivet, and J.-L. Duchaud, "Prediction intervals for global solar irradiation forecasting using regression trees methods," *Renewable Energy*, vol. 126, pp. 332 – 340, 2018.

[24] M. Lange and U. Focken, *Physical Approach to Short-Term Wind Power Prediction*. Springer, 2006.

[25] B. Ernst, B. Oakleaf, M. L. Ahlstrom, M. Lange, C. Moehrlen, B. Lange, U. Focken, and K. Rohrig, "Predicting the wind," *IEEE Power and Energy Magazine*, vol. 5, pp. 78–89, Nov 2007.

[26] D. O'Leary and J. Kubby, "Feature selection and ann solar power prediction," *Journal of Renewable Energy*, vol. 2017, pp. 1–7, 11 2017.

[27] Y. Zhang, M. Beaudin, H. Zareipour, and D. Wood, "Forecasting solar photovoltaic power production at the aggregated system level," in *2014 North American Power Symposium (NAPS)*, pp. 1–6, Sep. 2014.

[28] J. W. Denton, "How good are neural networks for causal forecasting," *The Journal of Business Forecasting Methods & Systems*, vol. 14, p. 17, 1995.

[29] I. S. Markham and T. R. Rakes, "The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression," *Comput. Oper. Res.*, vol. 25, pp. 251–263, 1998.

[30] C. Chatfield, “What is the ‘best’ method of forecasting?,” *Journal of Applied Statistics*, vol. 15, no. 1, pp. 19–38, 1988.

[31] S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler, “The accuracy of extrapolation (time series) methods: Results of a forecasting competition,” *Journal of Forecasting*, vol. 1, no. 2, pp. 111–153, 1982.

[32] G. Zhang, “Time series forecasting using a hybrid arima and neural network model,” *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[33] M. Khashei and M. Bijari, “A novel hybridization of artificial neural networks and arima models for time series forecasting,” *Applied Soft Computing*, vol. 11, pp. 2664–2675, 03 2011.

[34] C. N. Babu and B. E. Reddy, “A moving-average filter based hybrid arima–ann model for forecasting time series data,” *Applied Soft Computing*, vol. 23, pp. 27–38, 2014.

[35] C. Chatfield, “Model uncertainty and forecast accuracy,” *Journal of Forecasting*, vol. 15, no. 7, pp. 495–508, 1996.

[36] Ümit Çavuş Büyükşahin and Şeyda Ertekin, “Improving forecasting accuracy of time series data using a new arima-ann hybrid method and empirical mode decomposition,” *Neurocomputing*, vol. 361, pp. 151–163, 2019.

[37] P. V and U. Rao, “Development of statistical time series models for solar power prediction,” *Renewable Energy*, vol. 83, 11 2015.

[38] M. Lotfi, M. Javadi, G. Osório, C. Monteiro, and J. Catalão, “A novel ensemble algorithm for solar power forecasting based on kernel density estimation,” *Energies*, vol. 13, 01 2020.

[39] J. Torres, A. Troncoso, I. Koprinska, Z. Wang, and F. Martínez-Álvarez, “Big data solar power forecasting based on deep learning and multiple data sources,” *Expert Systems*, vol. 36, p. e12394, 03 2019.

[40] J.-P. Lai, Y.-M. Chang, C.-H. Chen, and P.-F. Pai, “A survey of machine learning

models in renewable energy predictions," *Applied Sciences*, vol. 10, p. 5975, 08 2020.

[41] M. H. ÖZDEMİR, M. İNCE, B. L. AYLAK, O. ORAL, and M. A. TAŞ, "Installed solar power prediction for turkey using artificial neural network and bidirectional long short-term memory," *Business & Management Studies: An International Journal*, vol. 8, pp. 4047–4068, Dec. 2020.

[42] P. Dash, I. Majumder, N. Nayak, and R. Bisoi, "Point and interval solar power forecasting using hybrid empirical wavelet transform and robust wavelet kernel ridge regression," *Natural Resources Research*, vol. 29, 02 2020.

[43] J. Zeng and W. Qiao, "Short-term solar power prediction using a support vector machine," *Renewable Energy*, vol. 52, pp. 118–127, 2013.

[44] E. İzgi, A. Öztopal, B. Yerli, M. K. Kaymak, and A. D. Şahin, "Short–mid-term solar power prediction by using artificial neural networks," *Solar Energy*, vol. 86, no. 2, pp. 725 – 733, 2012.

[45] H. Long, Z. Zhang, and Y. Su, "Analysis of daily solar power prediction with data-driven approaches," *Applied Energy*, vol. 126, pp. 29–37, 2014.

[46] E. Lorenz, J. Hurka, D. Heinemann, and H. G. Beyer, "Irradiance forecasting for the power prediction of grid-connected photovoltaic systems," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 2, no. 1, pp. 2–10, 2009.

[47] M. Millán, J. Lobato, P. Cañizares, and M. Rodrigo, "Prediction and management of solar energy to power electrochemical processes for the treatment of wastewater effluents," *Electrochimica Acta*, vol. 335, p. 135594, 2020.

[48] M. Alkandari and I. Ahmad, "Solar power generation forecasting using ensemble approach based on deep learning and statistical methods," *Applied Computing and Informatics*, 2020.

[49] A. Gensler, J. Henze, B. Sick, and N. Raabe, "Deep learning for solar power forecasting — an approach using autoencoder and lstm neural networks," pp. 002858–002865, 10 2016.

[50] "Güneş Enerji Santralleri." https://www.enerjiatlasi.com/gunes/, 2020. [Online; accessed 29-December-2020].

[51] M. Marinelli, F. Sossan, G. Costanzo, and H. Bindner, "Testing of a predictive control strategy for balancing renewable sources in a microgrid," *I E E E Transactions on Sustainable Energy*, vol. 5, no. 4, pp. 1426–1433, 2014.

[52] "Irradiance And Insolation." https://pvpmc.sandia.gov/modeling-steps/1-weather-design-inputs/irradiance-and-insolation-2/. [Online; accessed 31-January-2021].

[53] S. Singh, "Solar Irradiance Concepts: DNI, DHI, GHI, GTI." https://www.yellowhaze.in/solar-irradiance/, Jan 2020. [Online; accessed 31-January-2021].

[54] Wikipedia, "Weighted arithmetic mean — Wikipedia, the free encyclopedia." https://en.wikipedia.org/wiki/Solar_irradiance, 2021. [Online; accessed 18-February-2021].

[55] "Diffuse horizontal irradiance."

[56] E. Cogliani, "The role of the direct normal irradiance (dni) forecasting in the operation of solar concentrating plants," *Energy Procedia*, vol. 49, pp. 1612 – 1621, 2014. Proceedings of the SolarPACES 2013 International Conference.

[57] Wikipedia, "Weighted arithmetic mean — Wikipedia, the free encyclopedia." http://en.wikipedia.org/w/index.php?title=Weighted%20arithmetic%20mean&oldid=996434502, 2020. [Online; accessed 29-December-2020].

[58] M. Abuella and B. H. Chowdhury, "Solar power forecasting using artificial neural networks," *2015 North American Power Symposium (NAPS)*, pp. 1–5, 2015.

[59] J. Kleissl, *Solar energy forecasting and resource assessment*. Academic Press, 2013.

[60] M. Gong, Y. Bai, J. Qin, J. Wang, P. Yang, and S. Wang, "Gradient boosting machine for predicting return temperature of district heating system: A case study

for residential buildings in tianjin," *Journal of Building Engineering*, vol. 27, p. 100950, 2020.

[61] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[62] S. Touzani, J. Granderson, and S. Fernandes, "Gradient boosting machine for modeling the energy consumption of commercial buildings," *Energy and Buildings*, vol. 158, 11 2017.

[63] S. Zhang, Y. Wang, Y. Zhang, D. Wang, and N. Zhang, "Load probability density forecasting by transforming and combining quantile forecasts," *Applied Energy*, vol. 277, p. 115600, 2020.

[64] R. Prasad, M. Ali, P. Kwan, and H. Khan, "Designing a multi-stage multivariate empirical mode decomposition coupled with ant colony optimization and random forest model to forecast monthly solar radiation," *Applied Energy*, vol. 236, pp. 778 – 792, 2019.

[65] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, Sep 1995.

[66] JunHua Zhao, ZhaoYang Dong, and Xue Li, "Electricity price forecasting with effective feature preprocessing," in *2006 IEEE Power Engineering Society General Meeting*, pp. 8 pp.–, 2006.

[67] J. H. Zhao, Z. Y. Dong, X. Li, and K. P. Wong, "A framework for electricity price spike analysis with advanced data mining methods," *IEEE Transactions on Power Systems*, vol. 22, no. 1, pp. 376–385, 2007.

[68] N. I. Sapankevych and R. Sankar, "Time series prediction using support vector machines: A survey," *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 24–38, 2009.

[69] R. H. Inman, H. T. Pedro, and C. F. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in Energy and Combustion Science*, vol. 39, no. 6, pp. 535 – 576, 2013.

[70] U. Ç. Büyükşahin and Ş. Ertekin, "Tek değişkenli zaman serileri tahmini için öznitelik tabanlı hibrit arima-ysa modeli," *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, vol. 35, pp. 467 – 478, 2019.

[71] T. D. Gauthier and M. E. Hawley, "Chapter 5 - statistical methods," in *Introduction to Environmental Forensics (Second Edition)* (B. L. Murphy and R. D. Morrison, eds.), pp. 129 – 183, Burlington: Academic Press, second edition ed., 2007.

[72] "T-TEST ESSENTIALS: DEFINITION, FORMULA AND CALCULATION." `https://www.datanovia.com/en/lessons/types-of-t-test/unpaired-t-test/students-t-test/`, 2021. [Online; accessed 02-January-2021].

## Appendix A

## SAMPLE ACTUAL & PREDICTED RESULTS OF RANDOM FOREST MODEL

Table A.1: Actual and Predicted Values of Random Forest at 01/06/2019

| Date | Hour | Actual | Predicted |
|---|---|---|---|
| 01/06/2019 | 07:00 | 1223.88 | 1132.185 |
| | 08:00 | 1176.66 | 1115.664 |
| | 09:00 | 1237.6 | 1146.12 |
| | 10:00 | 1278.82 | 1106.853 |
| | 11:00 | 1281.55 | 1153.602 |
| | 12:00 | 1171.78 | 1148.28 |
| | 13:00 | 1280.62 | 1132.102 |
| | 14:00 | 1284.31 | 1081.956 |
| | 15:00 | 1184.95 | 1147.596 |
| | 16:00 | 916.27 | 1143.455 |
| | 17:00 | 1176.15 | 1043.022 |
| | 18:00 | 1180.95 | 1063.949 |
| | 19:00 | 1087.47 | 1097.297 |
| | 20:00 | 1152.8 | 1066.178 |

Table A.2: Actual and Predicted Values of Random Forest at 30/09/2019

| Date | Hour | Actual | Predicted |
|---|---|---|---|
| 30/09/2019 | 07:00 | 34.31 | 57.40948 |
| | 08:00 | 35.14 | 53.23641 |
| | 09:00 | 53.45 | 49.97518 |
| | 10:00 | 57.87 | 50.0012 |
| | 11:00 | 61.66 | 51.95372 |
| | 12:00 | 43.48 | 49.19227 |
| | 13:00 | 32.36 | 56.72669 |
| | 14:00 | 35.67 | 56.20644 |
| | 15:00 | 35.96 | 53.48064 |
| | 16:00 | 39.09 | 50.15729 |
| | 17:00 | 40.93 | 50.1833 |
| | 18:00 | 46.74 | 51.02836 |
| | 19:00 | 40.17 | 52.44637 |
| | 20:00 | 35.4 | 49.48167 |