Facultad de Ciencias Económicas y Empresariales

Wind Power Generation Forecasting Using Transformer-based Time Series Models

Author: Teresa Oriol Guerra
Supervisor: Jenny Alexandra Cifuentes Quintero

MADRID | June 2025

Abstract

The growing penetration of renewable energy sources, particularly wind power, into modern energy systems has heightened the need for accurate and reliable forecasting techniques. Effective short-term wind power prediction is essential for ensuring grid stability, optimizing dispatch strategies, and reducing operational costs. Over the past decade, researchers have progressively moved from traditional statistical and machine learning methods to deep learning approaches capable of modeling nonlinear and temporal dependencies. Among these, Transformer-based architectures have gained increasing attention for their ability to handle long-range temporal correlations, previously unattainable with conventional recurrent models.

In this study, a Transformer-based model is proposed for short-term wind power forecasting using only historical generation data. The time series is segmented into overlapping sequences of fixed lengths (12, 24, and 36 hours) using a sliding window approach. Different Transformer configurations were evaluated across different input lengths, varying hyperparameters such as model dimension, number of attention heads, encoder layers, dropout rate, and batch size. Model performance was assessed on both training and test sets using MSE, MAE, and MAPE as error metrics. The methodology also included a hyperparameter optimization process and normalization procedures to enhance training stability and generalization.

The results indicate that the best performance was obtained using a 24-hour input window, achieving a test MAPE of 0.09%. This configuration outperformed models trained with longer input sequences, suggesting that shorter historical contexts are sufficient for accurate short-term forecasting in this dataset. Furthermore, when compared against traditional approaches such as LSTM and GRU architectures under optimized settings, the best Transformer model showed superior predictive accuracy. However, this improvement came at the cost of higher computational complexity. These findings support the effectiveness of Transformer models for wind power forecasting and highlight opportunities for future research involving spatial modeling, transfer learning, and multistep prediction.

Keywords: wind power forecasting, Transformer Model, time series prediction, deep learning, renewable energy

Resumen

El aumento en la integración de fuentes de energía renovable, especialmente la energía eólica, en los sistemas energéticos modernos ha intensificado la necesidad de técnicas de predicción precisas y fiables. Una predicción efectiva de la energía eólica a corto plazo es esencial para garantizar la estabilidad de la red, optimizar las estrategias y reducir los costes operativos. En la última década, los investigadores han pasado progresivamente de métodos estadísticos tradicionales y de aprendizaje automático a enfoques de *deep learning*, capaces de modelar dependencias no lineales y temporales. Entre estos, las arquitecturas Transformer han ganado atención por su capacidad para captar correlaciones temporales de largo alcance, antes inalcanzables con modelos recurrentes.

En este estudio, se propone un modelo Transformer para la predicción de la energía eólica a corto plazo utilizando únicamente datos históricos de generación. La serie temporal se segmenta en secuencias superpuestas de longitudes fijas (12h, 24h y 36h) mediante un enfoque de ventana deslizante. Se evaluaron distintas configuraciones del modelo Transformer en función de diferentes longitudes de entrada, variando hiperparámetros como la dimensión del modelo, el número de cabezas de atención, capas del codificador, *dropout* y *batch size*. El rendimiento del modelo se evaluó tanto en los conjuntos de entrenamiento como de prueba utilizando el MSE, el MAE y el MAPE como métricas de error. La metodología incorporó optimización de hiperparámetros y normalización para mejorar la estabilidad y la generalización del modelo.

Los resultados indican que el mejor rendimiento se obtuvo utilizando una ventana de entrada de 24h, alcanzando un MAPE de prueba del 0,09%. Esta configuración superó a los modelos entrenados con secuencias de entrada más largas, lo que sugiere que contextos históricos más cortos son suficientes para una predicción precisa a corto plazo en este conjunto de datos. Además, al compararse con enfoques tradicionales como las arquitecturas LSTM y GRU en condiciones optimizadas, el mejor modelo Transformer mostró una mayor precisión predictiva. Sin embargo, esta mejora implicó un mayor coste computacional. Estos hallazgos respaldan la eficacia de los modelos Transformer para la predicción de energía eólica y destacan oportunidades para futuras investigaciones en modelado espacial, aprendizaje por transferencia y predicción multietapa.

Palabras clave: energía eólica, modelo Transformer, series temporales, *deep learning*

Contents

List of Figures

List of Tables

Acronyms

AR	Auto-regressive
ANN	Artificial Neural Network
ARIMA	AutoRegressive Integrated Moving Average
CNN	Convolutional Neural Network
DL	Deep Learning
EDA	Exploratory Data Analysis
EEMD	Ensemble Empirical Mode Decomposition
ENTSO-e	European Network of Transmission System Operators for Electricity
EU	European Union
f-ARIMA	fractional AutoRegressive Integrated Moving Average
GRU	Gated Recurrent Unit
IF	Isolation Forest
IQR	Interquartile Range
LSTM	Long Short-Term Memory
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
MW	Megawatts
RED II	Renewable Energy Directive
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Networks
WPF	Wind Power Forecasting

Chapter 1

Introduction

1.1 Motivation

Wind energy stands out as one of the most promising and rapidly expanding renewable energy sources globally. According to the International Renewable Energy Agency, the installed capacity of wind energy has grown exponentially in recent decades and is expected to remain a central element in the global transition toward a more sustainable energy system (IRENA, 2023). This growth is driven by several key factors, including wind energy's high efficiency, cost-effectiveness and its ability to significantly reduce greenhouse gas emissions compared to fossil fuel-based energy sources (Hanifi, Liu, Lin, & Lotfian, 2020).

In addition to reducing dependence on fossil fuels, wind energy offers several other significant benefits. It is a clean, abundant power source that produces zero direct greenhouse gas emissions, which add a big contribution to climate change. This positions wind energy as a relevant solution for meeting global carbon reduction targets and minimizing environmental impact. Furthermore, wind energy contributes to energy security by diversifying the energy mix and reducing reliance on imported fuels. It also serves as a scalable solution to meet the growing global energy demand (Hassan, Algburi, Sameen, Salman, & Jaszczur, 2023). As geopolitical uncertainties continue to affect global energy supplies, the ability of wind energy to generate domestic power becomes increasingly valuable for countries striving for greater energy independence. Additionally, advancements in technology have made wind energy more cost-competitive, with prices now comparable to or even lower than conventional energy sources such as coal and natural gas. This economic viability, combined with its environmental benefits, underscores the importance of wind energy and highlights the need for continued investment, particularly in regions with slower adoption rates (Hassan et al., 2023).

As illustrated in Figure 1.1, the cumulative installed wind energy capacity has increased significantly across major regions. Asia has shown the most pronounced growth trajectory, surpassing Europe and North America, which have also demonstrated steady increases over

the years. In contrast, regions like South America, Oceania, Africa and the Middle East have been slower to adopt wind energy, often due to differences in infrastructure, funding and policy support. This gap between the rapid advances in some parts of the world and the slower pace in others highlights the need for stronger global efforts to make renewable energy more accessible everywhere.

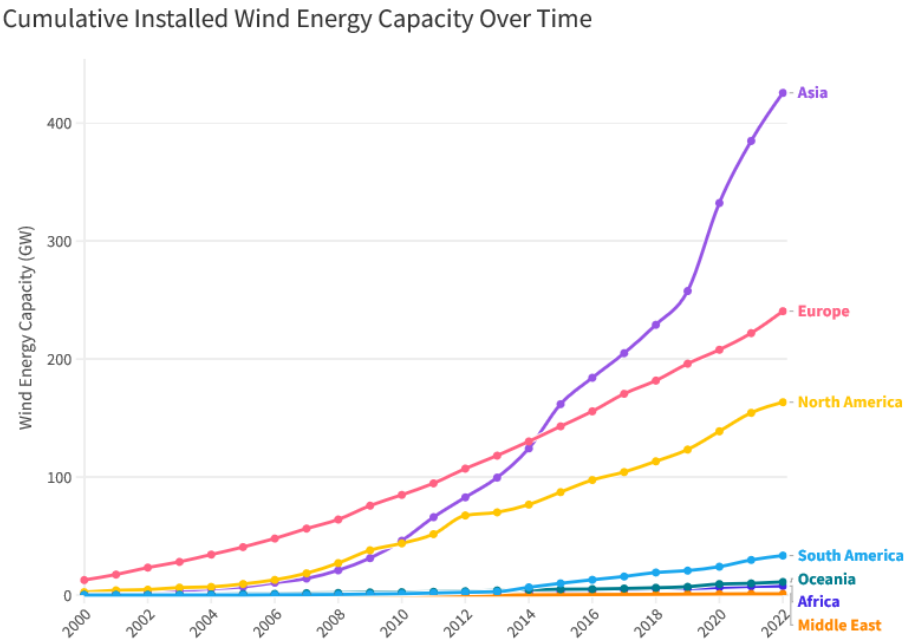


Figure 1.1: Cumulative Installed Wind Energy Capacity by Region
Source: (Our World in Data, 2024). Self-Elaboration

Within the European context, wind energy plays a key role in achieving the European Union's (EU) ambitious regulatory framework for reducing carbon emissions and transitioning to a sustainable energy system. The EU's Green Deal and Fit for 55 package set clear targets with the aim of reducing net greenhouse gas emissions by at least 55% by 2030 (Commission, 2020b, 2021). As one of the fastest-growing renewable sources, wind energy is required to enable Member States to meet their renewable energy goals, such as those outlined in the Renewable Energy Directive (RED II). This EU directive mandates that at least 32% of the EU's energy consumption come from renewables by 2030 (Commission, 2020c). Moreover, the EU's carbon pricing mechanisms, such as the Emissions Trading System, further incentivize the shift away from fossil fuels by putting a price on carbon emissions, which makes wind energy a more attractive option (Commission, 2020a). Expanding wind energy capacity not only reduces Europe's reliance on imported fossil fuels but also strengthens its position as a global leader in the renewable energy transition, setting a benchmark for sustainable development worldwide.

Spain, as shown in Figure 1.2, stands out as one of the leading countries in wind energy generation. With a well-established infrastructure and a strong commitment to renewable

energy, Spain has consistently ranked among the top wind energy producers in the European region. The country has successfully integrated wind power into its energy mix, contributing significantly to EU renewable energy targets. By 2022, Spain's cumulative wind energy capacity had reached levels comparable to other global leaders such as Germany and India. This sustained growth underscores Spain's role in advancing the EU's renewable energy objectives and highlights its potential as a model for other regions seeking to scale their wind energy efforts.

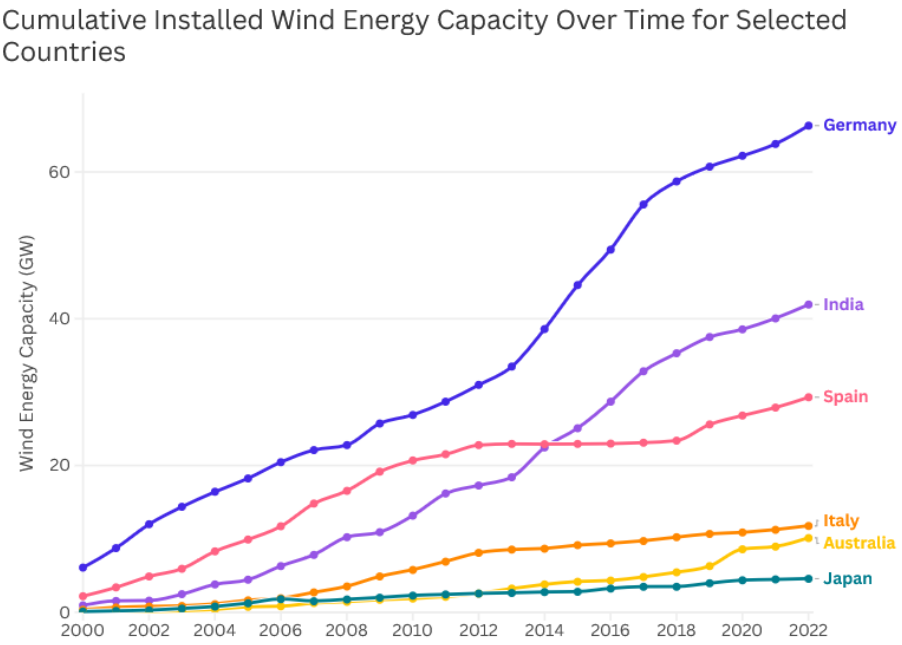


Figure 1.2: Cumulative Installed Wind Energy Capacity by Country
Source: (Our World in Data, 2024). Self-Elaboration

While wind energy has many advantages, one of its biggest challenges is its unpredictable nature. Wind power is inherently variable because of its dependence on meteorological and geographic factors. This variability makes it harder to smoothly integrate wind energy into the power grid. When forecasts are inaccurate, they can lead to grid instability, a greater reliance on fossil-fuel backup systems and higher operating costs. Moreover, unreliable forecasting also makes it more difficult to plan and manage energy use effectively, which reduces the overall performance of wind systems. This is why accurate forecasting is so important. It helps stabilize the grid, reduce operational inefficiencies and minimize the environmental and financial costs associated with backup power systems (Hanifi et al., 2020).

The importance of wind energy forecasting varies depending on the time horizon considered. Short-term forecasting, typically ranging from minutes to a few hours ahead, is particularly critical for grid operations and energy trading. Accurate short-term predictions allow grid operators to balance supply and demand in real time, ensuring system reliability and avoiding costly imbalances. Furthermore, in competitive electricity markets, precise

short-term forecasts enable wind farm operators to optimize their bids, reducing penalties associated with under or overestimation of generation (Ahmed, Muhammad, Abbas, Aziz, & Mahmood, 2024). Medium and long-term forecasts, on the other hand, are important for capacity planning, infrastructure development, and policy-making, but their operational impact is often less immediate (Shan, Niu, Chai, & Gu, 2024). Therefore, while all forecasting horizons are important, short-term forecasting is required in the day-to-day integration of wind energy into power grids, where small errors can have significant economic and operational consequences.

In this context, various forecasting strategies have been proposed to address the challenges posed by the variability of wind energy generation. Traditional statistical models, such as AutoRegressive Integrated Moving Average (ARIMA), have been widely used for time series forecasting due to their simplicity and interpretability. However, these models often struggle to capture the non-linear and complex patterns inherent in wind power generation data. As the field has evolved, more advanced Machine Learning (ML) approaches, such as recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks, have gained prominence. These methods are better suited for sequential data and have demonstrated improved performance in capturing temporal dependencies. Despite their advantages, RNNs and LSTMs face limitations, including the vanishing gradient problem and high computational costs, particularly when dealing with long sequences or large datasets.

To overcome these limitations, Transformer-based models have emerged as a powerful alternative in time series forecasting. Originally developed for natural language processing tasks, Transformers leverage self-attention mechanisms to model relationships across all positions in a sequence simultaneously. This capability allows Transformers to capture long-range dependencies more effectively than traditional RNN-based architectures, which process sequences sequentially. Additionally, Transformers enable parallelized computations, resulting in faster training times and scalability to large datasets. These advantages make Transformer-based models particularly well-suited for addressing the challenges associated with wind energy forecasting, especially in short-term horizons where high accuracy and computational efficiency are crucial (Brownlee, 2020).

This study builds upon the growing body of research exploring the application of Transformers in time series forecasting, with a specific focus on enhancing the precision of short-term wind energy predictions. Short-term forecasting, which typically spans from minutes to several hours ahead, is important in real-time grid operations, energy market optimization and system reliability. Given the increasing importance of wind energy, improving forecasting accuracy is required for mitigating its inherent variability and ensuring its seamless integration into modern power systems. By leveraging the strengths of Transformer architectures, this study not only contributes to more reliable short-term wind energy predictions but also holds potential for broader applications, such as optimizing wind energy utilization and encouraging its adoption in regions where its development lags behind.

1.2 Objectives

1.2.1 General Objective

The general objective of this research is to develop and evaluate a Transformer-based model for forecasting wind power generation using historical time series data. The model will be applied to a case study using wind power generation data from Spain. Its performance will be compared against other forecasting techniques to assess its accuracy and effectiveness.

1.2.2 Specific Objectives

The general objective is broken down into specific objectives, which will allow for the analysis and extraction of the necessary conclusions during the development of this work.

- ***To contextualize the importance of wind energy forecasting by analyzing its role in addressing the variability of wind power generation and its integration into modern power systems.*** This will involve a detailed discussion on the relevance of accurate forecasting for grid stability, operational efficiency and energy market optimization.
- ***To identify and analyze the techniques used for short-term wind energy forecasting in recent years through a comprehensive review of the literature***. This review will focus on traditional statistical methods, ML approaches and advancements in Deep Learning (DL) techniques, highlighting their advantages and limitations.
- ***To implement a Transformer-based architecture for short-term wind energy forecasting, leveraging its self-attention mechanism to address the challenges associated with capturing temporal dependencies and variability in wind power generation.*** The model will be designed to handle historical time series data and optimize forecasting accuracy.
- ***To apply the Transformer-based model to a case study using historical wind power generation data from Spain and compare its performance against other forecasting techniques, such as LSTM and Gated Recurrent Unit (GRU).*** This comparison will assess the model's accuracy and effectiveness in a real-world context, providing insights into its practical applicability.

1.3 Structure of the Document

The present document is structured into five main chapters. The first chapter provides an introduction to the study, outlining the motivation behind the research, its relevance in the

context of wind energy forecasting and the primary objectives. Chapter 2 reviews the evolution of wind energy forecasting methods over the years, with a focus on traditional statistical approaches, advanced ML techniques and their respective applications and limitations. Chapter 3 provides an explanation of Transformer models, emphasizing their architecture and suitability for time series forecasting. This chapter also describes the methodology adopted in this research, including the processes for data selection, preprocessing, descriptive analysis and the implementation of the Transformer-based model for short-term wind energy prediction. Chapter 4 presents the results of the study, offering a detailed analysis of the model's performance. A comparison is also made between the proposed Transformer-based approach and other forecasting techniques, to evaluate their accuracy and effectiveness. Finally, Chapter 5 concludes the study, summarizing the key findings, discussing their implications and proposing potential directions for future research.

Chapter 2

From Traditional Methods to Transformer-based Time Series Models in Wind Power Forecasting: State of the Art

The increasing integration of wind energy into power grids has heightened the need for accurate short-term forecasting, as the inherent variability of wind power generation poses significant challenges for grid stability and operational planning (Hanifi et al., 2020; Wang et al., 2011). Short-term wind energy forecasts, typically ranging from minutes to several hours ahead, is required to balance supply and demand, optimizing energy dispatch and reducing the reliance on backup power sources (Hanifi et al., 2020; Liu & Zhang, 2024; Wang et al., 2011). In electricity markets, accurate short-term predictions allow wind farm operators to participate more effectively, minimizing financial penalties due to forecast errors. Moreover, as renewable energy penetration continues to grow, improving the precision of short-term wind power forecasting (WPF) has become relevant for ensuring the reliability and efficiency of modern power systems.

To enhance the accuracy of WPF, predictive techniques have evolved significantly in recent years, incorporating more sophisticated methodologies to capture the complex temporal and spatial patterns in wind generation data. Initially, traditional approaches were developed to provide wind power forecasts based on historical data and meteorological variables, relying on relatively simple modeling techniques. These early models primarily consisted of physical and statistical methods, which, despite their computational efficiency, struggled to account for the complexity and variability of wind behavior. According to (Hanifi et al., 2020), WPF methods can be categorized into three main approaches: physical models, statistical models and hybrid approaches that integrate ML. The first WPF models were developed using physical and statistical approaches independently, along with simple regression models. These

methods generated forecasts based on historical wind data, physical conditions or statistical relationships, offering basic yet useful predictions. Although computationally inexpensive and effective for very short-term forecasting (minutes to hours ahead), these models were inherently limited in accuracy, as they failed to capture complex wind dynamics and rapid meteorological changes.

Among these approaches, the physical model has been widely used to improve the accuracy of wind power forecasts by incorporating meteorological and physical principles. As detailed by (Lange & Focken, 2006), physical models emphasize the role of meteorology, fluid dynamics and energy economics in understanding fluctuations in wind power generation. These models integrate atmospheric flows, boundary-layer meteorology

and thermal stratification to provide a more comprehensive representation of the factors affecting wind power output. By incorporating a better understanding of terrain effects, thermal gradients and large-scale meteorological patterns, physical models aim to refine the accuracy of wind power predictions. However, despite these advancements, physical models still face limitations in achieving high precision. Errors often arise due to imperfect modeling of atmospheric conditions and sudden, unpredictable weather changes. Even minor inaccuracies in numerical weather predictions can lead to significant discrepancies in WPF. Furthermore, the relationship between wind speed and power output is nonlinear, making it difficult for physical models to generalize across different meteorological conditions. Forecasting errors tend to vary depending on wind regimes, atmospheric stability and terrain characteristics, requiring continuous improvements and validation to enhance prediction accuracy. These challenges have driven the exploration of alternative approaches, including statistical and ML-based models, which offer greater adaptability in handling complex wind behavior and improving short-term forecast reliability.

To address the limitations of physical models, time series-based statistical approaches gained prominence as an alternative for WPF. Among these, the ARIMA model has been widely adopted due to its effectiveness in capturing short-term linear dependencies and patterns in historical data. ARIMA models analyze past observations to identify autoregressive (AR) and moving average (MA) components while adjusting for trends and seasonal fluctuations, enabling improved forecasting accuracy compared to simpler approaches such as persistence models. Unlike persistence models, which assume that future wind power generation will resemble the most recent observations, ARIMA accounts for temporal dependencies, making it more suitable for short- and medium-term predictions. For instance, (Kavasseri & Seetharaman, 2009) demonstrated the advantages of ARIMA in predicting wind power generation trends and seasonal variations, showing that it consistently outperformed persistence models in medium-term forecasting.

Despite these advantages, ARIMA models face inherent limitations when dealing with long-range dependencies and complex temporal structures in wind power generation data. Traditional ARIMA assumes integer differencing, which can be restrictive when modeling

time series with long memory effects. This limitation prevents ARIMA from fully capturing the gradual decay of past influences on future values, which is particularly relevant for wind power generation, where long-term dependencies often emerge due to seasonal and climatic cycles (Chen, Pedersen, Bak-Jensen, & Chen, 2009). To handle these constraints, fractional-ARIMA (f-ARIMA) models were introduced as an extension of the standard ARIMA framework. By allowing the differencing parameter to assume fractional values, f-ARIMA effectively models long-range correlations in time-series data, offering greater flexibility in capturing persistent dependencies. Studies have shown that f-ARIMA can significantly enhance forecasting accuracy, with one study reporting a 42% improvement in Mean Squared Error (MSE) for hourly wind speed predictions compared to traditional ARIMA (Kavasseri & Seetharaman, 2009). This enhanced capability makes f-ARIMA particularly useful for applications requiring a more nuanced understanding of long-term patterns, improving the reliability of wind power forecasts over extended time horizons.

Although f-ARIMA enhances forecasting accuracy by capturing long-range dependencies, certain challenges remain that limit its effectiveness in WPF. The assumption of stationarity restricts its applicability in scenarios where non-linear or chaotic dynamics dominate, such as abrupt wind speed fluctuations caused by atmospheric turbulence. Additionally, parameter selection, determining the optimal AR, differencing (I), and MA terms, requires significant manual intervention, reducing its scalability for real-time applications. While f-ARIMA improves trend and seasonality modeling, its performance weakens when confronted with sudden changes in wind speed or complex meteorological interactions, which are common in short-term forecasting (Kavasseri & Seetharaman, 2009).

Given these constraints, ARIMA-based models remain useful for short- and medium-term forecasting, particularly when capturing periodic patterns in historical data. However, to overcome their limited ability to model non-linear dependencies, recent studies have explored hybrid approaches that integrate ARIMA with advanced predictive techniques. These methods combine ARIMA's statistical strengths with ML or DL architectures, enhancing forecasting accuracy by incorporating data-driven pattern recognition. The adoption of hybrid models reflects an effort to develop forecasting techniques that balance interpretability, adaptability, and computational efficiency, addressing the increasing complexity of wind power prediction (Kavasseri & Seetharaman, 2009)

Building upon these advancements, ML has emerged as a transformative tool in WPF, offering a powerful means to model the nonlinear and highly dynamic nature of wind energy generation. Unlike traditional statistical models, ML techniques can automatically learn patterns from large datasets, capturing intricate relationships between meteorological variables and wind power output. Among the most widely applied ML methods in this domain are Random Forests and Gradient Boosting, which have demonstrated strong predictive capabilities in WPF (Hanifi et al., 2020). These approaches often benefit from advanced data preprocessing techniques, such as feature engineering and dimensionality reduction, which

enhance model robustness and generalization. Additionally, ML-based methods can adapt to changes in data distribution more effectively than traditional approaches, making them particularly valuable for short-term forecasting where wind variability is more pronounced.

Beyond standalone ML models, hybrid approaches that integrate ML algorithms with statistical decomposition methods have demonstrated superior predictive accuracy. These models decompose the original time series into different frequency components, enabling the ML algorithms to capture distinct temporal patterns more effectively. For instance, the combination of ML techniques with decomposition methods such as Ensemble Empirical Mode Decomposition (EEMD) has proven highly effective in handling the chaotic and non-stationary nature of wind power data. By isolating different signal components, these hybrid models improve prediction accuracy for both short-term and long-term variations, addressing key challenges in WPF. Empirical studies have shown that ML-based hybrid models consistently outperform traditional approaches by leveraging both statistical rigor and the flexibility of data-driven learning (Wu, Meng, Fan, Zhang, & Liu, 2022).

Despite their strengths, ML models also present notable challenges that can impact their applicability in WPF. One of the primary concerns is their reliance on large amounts of high-quality data, as inadequate or biased datasets can significantly degrade model performance. Additionally, training complex ML models requires substantial computational power, particularly when dealing with high-dimensional feature spaces and real-time forecasting applications. The optimization process, including hyperparameter tuning, is another critical limitation, as selecting the appropriate configuration for a model often demands extensive experimentation and domain expertise. Moreover, interpretability remains an ongoing issue in ML-based forecasting. Many of these models operate as "black boxes", making it difficult to understand the rationale behind their predictions. This lack of transparency can hinder trust in the model's outputs, especially in energy markets and grid operations where decision-making requires explainable and reliable forecasts (Hanifi et al., 2020).

To address some of these challenges and further improve predictive accuracy, DL, a subset of ML, has emerged as a powerful alternative. Unlike traditional statistical approaches, DL methods automatically learn complex patterns from large datasets, capturing intricate relationships between meteorological variables and wind power generation. Various artificial neural network architectures have been applied in this field, including Multilayer Perceptrons (MLPs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs), each contributing unique advantages depending on the forecasting objectives and data characteristics.

MLPs, widely recognized for their simplicity and adaptability, have been employed to map input features such as wind speed, wind direction, and atmospheric pressure to power output predictions. While their lack of temporal structure limits their standalone application in time-series forecasting, they remain useful as components in hybrid models, where they complement other architectures by learning non-linear feature mappings (Wang et al.,

2011). Beyond fully connected networks, CNNs, originally developed for image processing, have been adapted for WPF, particularly in hybrid frameworks. By leveraging convolutional layers, these models effectively extract spatial correlations from multivariate meteorological data, enhancing forecasting accuracy. CNNs have demonstrated strong performance when combined with RNN-based architectures, as they preprocess spatial dependencies before sequential modeling. For instance, hybrid CNN-LSTM models have been successfully employed to simultaneously capture spatial and temporal patterns, leading to improved multi-step forecasting performance. Studies have shown that these models outperform standalone LSTMs in scenarios where meteorological features exhibit strong spatial dependencies, such as offshore wind farms (Hanifi et al., 2020; Wang et al., 2011).

Despite these advancements, traditional DL architectures still face limitations in handling long-range dependencies efficiently, often requiring extensive computational resources for training and hyperparameter tuning. Additionally, their reliance on sequential processing constrains scalability, particularly for large-scale wind power datasets. To address these challenges, recent research has explored the application of Transformer-based models, which leverage self-attention mechanisms to improve forecasting performance. Transformer models have demonstrated remarkable accuracy, particularly in multi-step forecasting tasks, where capturing complex dependencies over extended horizons is critical. By integrating advanced signal processing techniques such as EEMD, transformers effectively reduce noise and extract meaningful patterns from wind speed data, leading to more reliable predictions. Empirical studies have shown that transformer-based models significantly outperform traditional RNN and CNN architectures in terms of key evaluation metrics, such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Their scalability and ability to process variable-length input and output sequences make them particularly well-suited for real-world WPF applications (Wu et al., 2022).

Furthermore, the transformer-based encoder-decoder framework, when combined with EEMD, has achieved state-of-the-art accuracy in multi-step wind speed forecasting. The self-attention mechanism enables these models to selectively focus on different temporal patterns within the input sequence, enhancing their ability to capture intricate dependencies with greater precision. This capability is particularly valuable in WPF, where fluctuations are influenced by a multitude of interdependent factors, such as meteorological conditions and terrain variations. These advancements underscore the potential of ML to continue evolving and overcoming current forecasting limitations through the integration of novel methodologies, setting new benchmarks in predictive accuracy and operational reliability (Wu et al., 2022).

Table 2.1 provides a comparative summary of the most relevant WPF techniques that have been employed in the recent literature. The table reveals a methodological evolution from traditional statistical models, to more advanced ML and DL frameworks, ANNs, and Transformer-based architectures. The literature suggests a trend toward increasingly

data-driven, hybrid and deep architectures that not only capture complex temporal dynamics but also enhance generalization across different sites and conditions. These methods have demonstrated measurable improvements in forecasting accuracy, often outperforming traditional models.

Table 2.1: Summary of Wind Power Forecasting Methods: From Traditional Approaches to Transformer-based Time Series Models.

Source	Objective	Input variables	Temporary Stage	Used techniques	Main results
(Kavasseri & Seetharaman, 2009)	Improve day-ahead and two-day-ahead wind speed forecasts using fractional time series models	Hourly average wind speed (2 s samples averaged to 10 min, then hourly)	4 weeks per site across four different wind monitoring sites in North Dakota (May, Dec, Mar, Oct); Forecast Horizon: Day-ahead (24 hours) and in some cases up to two-day-ahead (48 hours)	ARIMA, f-ARIMA (36 models tested, selected via AIC, parameters estimated using Exact Maximum Likelihood via Ox-ARFIMA)	f-ARIMA outperformed ARIMA and persistence: 42% average FMSE improvement; 95.3% avg. correlation with actual; better robustness in volatile regimes; wind power forecasts derived using turbine power curve
(Lin & Liu, 2020)	Forecast wind power using decision trees combined with DL.	SCADA (Supervisory Control and Data Acquisition) features such as wind speed and ambient temperature from a turbine in Scotland	12 months, 1-second sampling	DL + Isolation Forest (IF)	IF provides a more robust preprocessing step for forecasting, especially when the input data deviate from a normal distribution, unlike traditional methods such as Elliptic Envelope
(Lima, Guetter, Freitas, Panetta, & de Mattos, 2017)	Apply a boosting strategy by combining numerical weather prediction with statistical filtering to improve the accuracy of regression-based wind power forecasts	Atmospheric global-scale forecasts from Brazil	7 and 12 months, sampling rate of 10 minutes; Forecast Horizon: 72h	Mix of physical and statistical model: Kalman filter and regression	RMSE down to 100.51 using cubic regression; Kalman filter reduced bias and error
(Pelletier, Masson, & Tahan, 2016)	Improve site-specific wind turbine power curve modeling using Artificial Neural Network (ANN)	Wind speed, air density, turbulence intensity, etc. from 140 Nordic turbines	12 months, 10-min avg (from 1 Hz data)	Multi-stage MLP (2-layer ANN with 6 inputs)	MAE 15.3–15.9; outperforming IEC, parametric and non-parametric models; scalable to more inputs
(Zhang, Yan, Infield, Liu, & Lien, 2019)	Use LSTM network for wind power production prediction compared to other models such as Radial Basis Function and Back Propagation.	Wind speed from a northern chinese wind farm	3 months, sampling rate of 15 min; Forecast horizon:48h	LTSM and Gaussian Mixture Model	RMSE: 6.37%, best accuracy; LSTM outperformed all benchmarks; Gaussian Mixture Model gave most reliable confidence intervals
(Wu et al., 2022)	To develop a multistep short-term wind speed forecasting model using a Transformer architecture	Wind speed at various heights, temperature, pressure, humidity, wind direction, turbulence, etc. (from NWTC-M2 tower dataset).	19 years (2002–2020), 10-min intervals; 2020 (1 year) used for testing; Forecast horizon: 3h, 6h, 12h and 24h	EEMD + Transformer (encoder–decoder structure)	MAE: 0.243–0.453, RMSE: 0.326–0.651

Chapter 3

Methodological Framework and Transformer Model Implementation

This chapter presents the methodological framework adopted to conduct the quantitative analysis developed in this Bachelor's thesis. The methodology is structured into a series of stages, each designed to ensure the appropriate preparation, modeling, and evaluation of historical wind power generation data using Transformer-based techniques. These stages are supported by Python libraries such as Pandas, NumPy, Matplotlib, Scikit-learn and TensorFlow, which enable efficient data handling, visualization, model implementation and performance evaluation. The methodological process is summarized in Figure 3.1 and consists of the following key stages:

1. **Data preparation:** the first stage involves collecting and integrating historical data on wind power generation from multiple sources covering the period 2020-2024. After integration, the data undergoes a cleaning process to handle missing values and correct inconsistencies. A key aspect of this stage involves restructuring the dataset into overlapping subsequences, a necessary transformation when working with time series forecasting models. This process, known as sliding window creation, segments the continuous time series into fixed-length windows, where each window serves as an input sequence used to predict the immediate next time step.

2. **Exploratory data analysis:** This stage focuses on examining the structure and characteristics of the data using visual and statistical techniques. The goal is to identify trends and patterns, providing initial insights into the temporal behavior of wind power generation.

3. **Modeling:** The modeling phase centers on implementing a transformer-based architecture for short-term forecasting of wind power. The model is trained on the prepared dataset to capture both short-term fluctuations and longer-term dependencies,

with hyperparameter selection and model tuning conducted to optimize predictive performance. In this part, the Train-Test split will be carried out in order to assess the generalization capacity of the forecasting model. Given the sequential nature of the data, this division will be performed chronologically to ensure that training data precede the test set, simulating the real-world conditions under which the model will be applied.

4. **Evaluation:** The final stage evaluates the model's performance using established error metrics, including MAE, MSE and Mean Absolute Percentage Error (MAPE). The results are compared with alternative forecasting approaches, such as LSTM and GRU models, to assess the relative accuracy and suitability of the Transformer model for short-term wind power prediction.

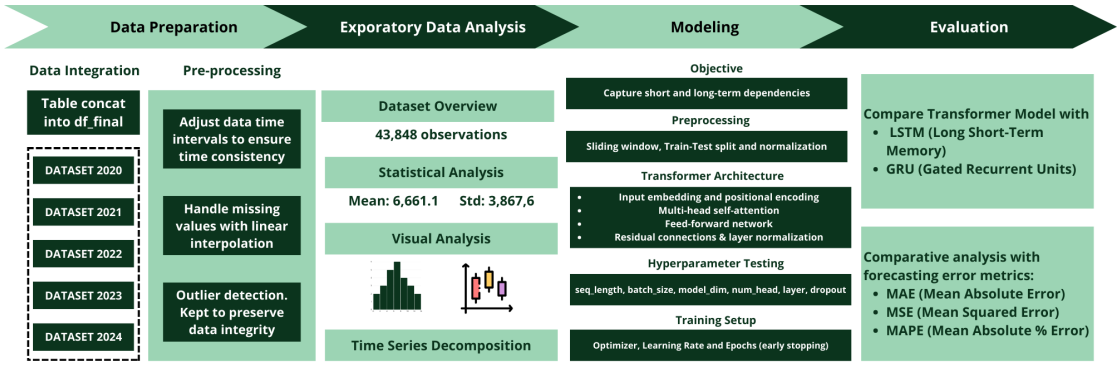


Figure 3.1: Methodological framework of the study.
Source: Self-Elaboration

3.1 Data Preparation

The first stage of the analysis consists of data preparation, beginning with the integration of historical wind power generation data obtained from the European Network of Transmission System Operators for Electricity (ENTSO-e), covering the period from 2020 to 2024 (ENTSO-E, 2025). These hourly records for Spain are merged into a single, chronologically ordered dataset to ensure temporal continuity and facilitate further analysis. During this process, inconsistencies in data formatting are identified and corrected to ensure uniformity across the five-year dataset.

It was observed that, starting on May 24, 2022, the recording frequency of the data source changed from hourly to 15-minute intervals. To maintain consistency with the rest of the dataset and ensure compatibility with the Transformer-based modeling approach, these quarter-hourly values are aggregated into hourly observations by calculating the average wind power generation within each hour. This resampling step guarantees the homogeneity of the

time series, which is important for both exploratory analysis and model training. This integration process ensures consistency and completeness across different time periods by merging data sources while preserving the chronological order of observations. During this stage, column names, data formats and variable types are standardized to prevent inconsistencies and potential errors in downstream processes, particularly during model training.

Once the data integration is complete, the next step involves addressing missing values within the dataset. Initially, a thorough check is performed to identify the number of missing values in each column. Upon detecting the missing values, the specific timestamps where data was absent are listed. It was identified that the missing values coincided with daylight saving time adjustments, which caused certain hours to be skipped entirely. This observation confirmed that the missing values did not result from data collection errors but rather from time shifts that led to non-existent timestamps. To address these gaps, linear interpolation was applied. This method estimates the missing values using a weighted average of the nearest available data points, specifically the observations immediately before and after the missing entry. Linear interpolation was chosen because it maintains a smooth transition between known values, making it easier to preserve the dataset's temporal consistency. This approach is particularly suitable when the missing intervals are short and when maintaining continuity in the time series is relevant for accurate modeling, which is especially necessary when using a Transformer Model.

Additionally, outlier detection and treatment is performed using the Interquartile Range (IQR) method to identify anomalous values that could distort the modeling process. Outliers in wind power generation data may result from sensor errors, system failures or atypical meteorological events. Methodologically, outliers cannot be automatically removed, as some extreme values may reflect genuine variability in wind conditions. Instead, outliers are flagged and reviewed within the context of meteorological conditions at the time. If a clear justification exists (e.g., erroneous sensor readings), the outlier is replaced using local interpolation techniques. Whereas if the extreme value reflects actual wind variability, it is retained to preserve the true distribution of the data.

3.2 Exploratory data analysis

After completing the data preparation process, the next stage involves conducting Exploratory Data Analysis (EDA). This phase aims to gain a comprehensive understanding of the dataset, examining statistical properties, visual patterns and temporal structures within the wind power generation data.

The first step in the EDA process involves examining the shape of the dataset, focusing on the number of observations and the dimensional structure of the wind power generation data. Understanding the dataset's size and temporal resolution helps determine whether the data

is sufficient for developing a reliable forecasting model. Since time series models require a consistent and adequately large sample to capture both short-term fluctuations and long-term patterns, this assessment ensures that the data meets these requirements. Following this initial examination, a statistical analysis is performed to explore the main characteristics of the wind power generation variable. Descriptive statistics, including the mean, median, variance and standard deviation, are calculated to summarize the central tendency and variability of the data. These metrics facilitate the identification of the distribution and dispersion of wind power values, allowing for a better understanding of variations in generation levels over time.

In addition to computing numerical statistics, it is useful to visualize the distribution of wind power generation to detect irregularities or patterns that may not be apparent through numerical summaries alone. A histogram can illustrate the frequency of different power generation levels, revealing potential asymmetry or unusual distributions. To further examine variability and detect potential outliers, boxplots are constructed for each year, allowing for a comparative analysis of power generation across different periods. Identifying outliers at this stage helps determine whether specific years exhibit abnormal patterns, guiding the decision on whether transformations or adjustments are required.

The EDA also includes a time series decomposition to break down the wind power generation data into its primary components: trend, seasonality and residual. The trend component highlights long-term changes in power generation, while the seasonal component captures repeating patterns associated with daily, weekly or annual cycles. The residual component represents irregular variations that do not follow the trend or seasonal patterns. Decomposing the time series helps identify consistent patterns and fluctuations, which is important for selecting features that enhance the model's predictive performance.

3.3 Modeling

The modeling phase in this analysis focuses on implementing a Transformer-based architecture for short-term WPF. The objective is to leverage the Transformer's ability to capture both short-term fluctuations and long-term dependencies in sequential data, enabling accurate predictions of wind power generation. Unlike traditional recurrent models, Transformers are well-suited for handling long-range temporal dependencies without relying on sequential processing. This characteristic makes them particularly effective for time series forecasting, where capturing patterns at different temporal scales is fundamental.

3.3.1 Fundamentals of Transformers

Transformers are a type of DL model originally introduced for natural language processing tasks but have since demonstrated remarkable performance in time series forecasting. The

core innovation of the Transformer model lies in its self-attention mechanism, which allows the model to weigh the relevance of different parts of the input sequence independently of their position (Vaswani et al., 2017). This approach overcomes the limitations of traditional recurrent architectures, such as RNNs and LSTM networks, which process sequences in a step-by-step manner, leading to challenges when capturing long-range dependencies.

The Transformer architecture consists of two main components: the encoder and the decoder, each composed of multiple layers that stack self-attention and feed-forward neural networks. The encoder processes the input sequence and generates a set of contextual embeddings, which capture the relationships between different parts of the sequence. The decoder then uses these embeddings to produce the output sequence. Although the original architecture was designed for sequence-to-sequence tasks, the encoder component alone is often used for time series forecasting, as it efficiently captures temporal patterns without the need for autoregressive decoding (Li & Law, 2024). The traditional block diagram of a Transformer model in Figure 3.2 illustrates its key components. The input embeddings are the first stage, where each element of the input sequence is transformed into a continuous representation. These embeddings are combined with positional encodings to retain information about the order of the sequence, as Transformers do not inherently encode positional information. This integration of positional encodings is essential to ensure that the model can distinguish between elements based on their positions within the sequence, preserving the temporal context.

Next, the model processes the embedded inputs through multi-head self-attention layers. In these layers, the model learns to identify which parts of the input sequence are most relevant to each prediction, regardless of their position within the sequence. The self-attention mechanism dynamically adjusts the focus of the model, allowing it to weigh different parts of the input based on their relevance to the task. This enables the model to capture both local and long-range dependencies within the data, facilitating the recognition of complex temporal patterns. The multi-head approach further enhances this capability by enabling the model to simultaneously attend to multiple aspects of the input, improving its ability to learn diverse patterns from the data (Bu & Cho, 2020). This parallel attention mechanism allows the model to process information more efficiently and effectively, particularly when dealing with intricate or long-term temporal relationships.

Following the self-attention layer, a feed-forward network is applied to each position independently. This network consists of two fully connected layers with a non-linear activation function in between, which allows the model to transform the attended representations into more complex features. Residual connections and layer normalization are incorporated after both the self-attention and feed-forward layers, ensuring stable training and preventing gradient vanishing. The use of residual connections helps maintain the flow of gradients through deeper networks, while layer normalization standardizes the output, contributing to more robust and faster convergence. The encoder stack can be repeated multiple times to

increase the model's capacity to learn complex representations. The output of the final encoder layer is then used as the input to the subsequent processing stages, either directly in the case of forecasting or passed to a decoder if a sequence-to-sequence approach is required. This modular structure not only facilitates capturing long-term dependencies but also makes the model highly parallelizable, significantly reducing training time compared to recurrent models (Vaswani et al., 2017).

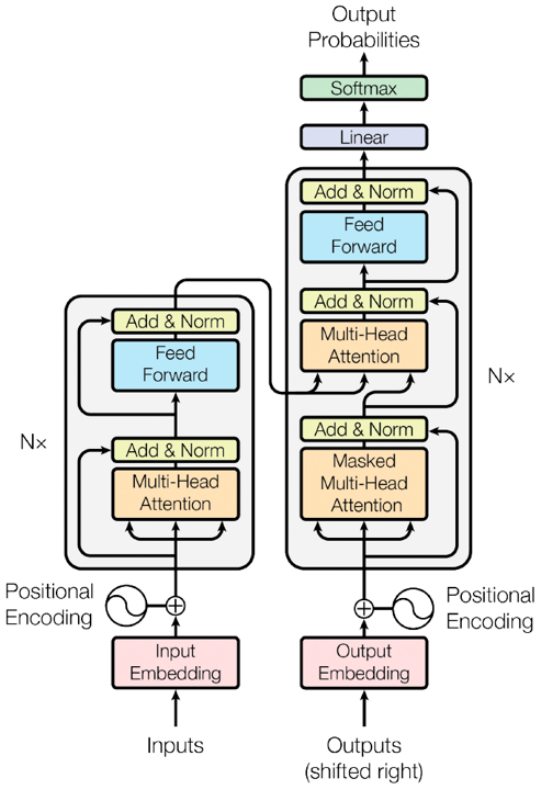


Figure 3.2: The Transformer Model Architecture
(Vaswani et al., 2017)

3.3.2 Implementation Details

Once the fundamental elements of the Transformer model have been established, the next step involves describing the methodological process for its implementation. The first task is to restructure the time series data into subsequences using the sliding window technique. This approach segments the continuous time series into overlapping windows of a fixed length, where each window represents a set of past observations used to predict the subsequent value. The length of the window is initially chosen based on domain knowledge and exploratory analysis, but it will later be fine-tuned as a hyperparameter to optimize model performance.

After generating the subsequences, the dataset is divided into training and testing sets to evaluate the model's generalization ability. To preserve the temporal integrity of the time

series data, the first 70% of the sequences are allocated to the training set, while the remaining 30% constitute the test set. This division reflects the sequential nature of the data, where older observations are used to train the model and more recent data is reserved for validation. This method of splitting ensures that the model does not have access to future information during training, therefore it maintains the realistic forecasting scenario where past data is used to predict future values.

After restructuring the time series data, the next step is to normalize the wind power generation values. Normalization is applied to ensure that all input features are on a comparable scale, which facilitates model convergence and enhances numerical stability during training. In time series forecasting, unnormalized data can lead to large gradient updates, making the training process unstable and slowing down convergence. Therefore, z-score scaling is employed to help the Transformer model learn more efficiently by reducing the impact of disparate value ranges. Once the data has been normalized, the implementation of the Transformer model begins. To achieve optimal model performance, various combinations of hyperparameters are tested. This process involves systematically adjusting key model parameters, training the model with each combination and evaluating its performance using error metrics that will be discussed in a subsequent section. The objective is to minimize both training and testing errors, thereby ensuring that the model generalizes well to new data.

To optimize the Transformer's performance, several hyperparameters are defined and fine-tuned:

- **Sequence length (seq_length)**: This parameter determines the number of time steps considered in a single input sequence. The choice of sequence length directly influences the model's ability to capture historical patterns and forecast future values. To adapt the sequential nature of wind power data to the Transformer model, the sliding window technique segments the time series into fixed-length windows. Each window serves as an input sequence used to predict subsequent values. Choosing the window size requires balancing the need to capture sufficient historical context while maintaining computational efficiency. In this study, sequence lengths of 12, 24 and 36 are tested, corresponding to half-daily (12 hours), daily (24 hours) and one-and-a-half-daily (36 hours) observations. These values are chosen to explore the model's performance across different temporal resolutions.

- **Batch size:** This parameter specifies the number of sequences processed simultaneously during training. A smaller batch size generally results in more stable training but at the cost of longer training time. Conversely, a larger batch size accelerates training but may impair generalization, as the model might overfit by memorizing patterns rather than learning them. In this study, the batch size will be adjusted by testing values such as 16, 32 and 64.

- **Model dimension (model_dim):** This parameter defines the size of the learned representation for each input token (time step). A smaller dimension reduces computational cost but may limit the model's ability to capture complex dependencies. In contrast, a larger model dimension enhances representation capacity but increases the risk of overfitting, particularly when the dataset size is limited. This hyperparameter is tested with model dimensions of 16 and 32, which are expected to provide sufficient capacity to capture meaningful patterns within the time series data.

- **Number of heads (num_heads):** This parameter indicates the number of attention heads used within the multi-head attention mechanism. A low number of heads results in a simpler model but may struggle to capture diverse relationships within the data. Increasing the number of heads allows the model to learn different attention patterns, improving its ability to process complex temporal relationships. In this study, 2, 4 and 8 heads are tested, ensuring that the model dimension (model_dim) is divisible by the number of heads (num_heads).

- **Number of layers (num_layers)**: This parameter defines the number of stacked Transformer encoder layers. Increasing the number of layers enhances the model's capacity to learn complex hierarchical representations. However, an excessive number of layers may lead to overfitting or significantly increase computational time. Given the nature of the task and the relatively limited size of the dataset, it is generally sufficient to employ 2 to 3 layers to achieve an appropriate balance between model capacity and generalization performance.

- **Dropout rate:** This regularization parameter randomly disables a proportion of neurons during training to mitigate overfitting. A moderate dropout rate prevents the model from becoming overly dependent on specific neurons. However, an excessively high dropout rate may result in underfitting, as the model may fail to capture essential patterns. In this study, dropout rates of 0.1, 0.2 and 0.3 are tested to evaluate their impact on model performance.

Moreover, other hyperparameters have been set based on values reported in the literature to ensure consistency with proven practices in time series forecasting. This approach leverages prior knowledge from previous studies, providing a foundation for model configuration. In the training process, the standard Adam optimizer is chosen for its adaptive learning rate mechanism and well-established effectiveness in training DL models across a variety of tasks (Wu et al., 2022). Adam adjusts learning rates individually for each parameter, making it particularly suitable for complex architectures and moderate-sized datasets. In this study, the default learning rate of 0.001 was used, as it offers a practical balance between convergence speed and stability. During model training, the number of epochs of training epochs is dynamically controlled by an early stopping function, which monitors validation loss and halts

training if no improvement is observed over a defined patience period. This prevents unnecessary computation in cases of early convergence, while still allowing for additional training iterations if the model has not fully converged.

3.4 Evaluation

Once the optimal Transformer model configuration has been selected, its performance will be compared with that of traditional DL strategies commonly used for time series forecasting, such as LSTM and GRU. These models have been widely employed in the field of wind power prediction due to their ability to capture temporal dependencies and model non-linear patterns within sequential data (Liu & Zhang, 2024). However, despite their proven effectiveness, LSTM and GRU networks often face challenges related to long-range dependencies, especially in cases where the input sequence length is substantial. These models process data sequentially, which can lead to inefficiencies when modeling long-term patterns (Liu & Zhang, 2024).

In contrast, the Transformer model addresses these challenges by utilizing self-attention mechanisms. This architectural advantage positions Transformers as a potentially superior alternative for time series forecasting, particularly when the data exhibits complex temporal relationships. However, given that LSTM and GRU models remain state-of-the-art in forecasting applications, it is important to conduct a comparative analysis to objectively assess whether the Transformer model offers a significant improvement . In this way, the comparative analysis will employ several error metrics that are widely used in time series forecasting. These metrics are chosen to assess not only the accuracy of the predictions but also the generalization ability of each model. The selected metrics will include MAE, MSE and MAPE, as they provide complementary perspectives on prediction accuracy.

Among the selected evaluation metrics, the MAE is particularly useful for measuring the average magnitude of errors between predicted and actual values, regardless of their direction (Wu et al., 2022). MAE quantifies the average absolute difference between the predicted values ($\hat{y}_t$) and the actual observed values (y_t) over a given period. This metric provides an intuitive interpretation of the model's accuracy, as it directly represents the average error in the same units as the data. The MAE is mathematically expressed in Equation 3.1.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t| \tag{3.1}$$

In this equation, n denotes the total number of observations, while $\hat{y}_t$ and y_t represent the predicted and actual values at time t, respectively. The absolute value operator, denoted by $|\cdot|$, ensures that negative differences do not offset positive ones, thus reflecting the magnitude of the prediction errors without considering their direction. MAE is particularly advanta-

geous when the objective is to understand the average prediction error without being overly sensitive to outliers.

Another key evaluation metric is the MSE, expressed in Equation 3.2, which computes the average of the squared differences between the actual values and the predicted values. This metric is especially effective when the objective is to penalize larger errors more severely, as the squaring process increases the impact of greater deviation (Wu et al., 2022).

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2 \tag{3.2}$$

In this equation, n represents the total number of observations, $\hat{y}_t$ and y_t are the predicted and actual values at time t, respectively. The squaring ensures all error terms are positive and magnifies larger errors. While MSE offers a powerful measure of overall accuracy, it is also highly sensitive to outliers. This means that a few large errors can disproportionately influence the final value, which requires careful tuning of the prediction model to avoid overfitting or misinterpretation of model performance.

The MAPE is a widely used metric that expresses prediction errors as a percentage of the actual values, offering a scale-independent view of model accuracy (Wu et al., 2022). It is defined as follows in Equation 3.3:

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{3.3}$$

Here, y_t is the actual observed value, $\hat{y}_t$ is the predicted value, and n is the total number of observations. The result is scaled by 100 to express the error as a percentage. MAPE is particularly useful for comparing model performance across datasets with different units or magnitudes, as it standardizes the error. However, it can become unreliable when actual values are close to zero, since small denominators can produce extremely large or undefined percentage errors. Therefore, while MAPE offers clear interpretability, it should be used cautiously in datasets with values near zero.

Chapter 4

Experimental Results

This chapter presents the results obtained from the implementation and evaluation of the Transformer-based model for short-term WPF. The analysis focuses on assessing the model's predictive performance, comparing it with more traditional forecasting techniques: LSTM and GRU models. The evaluation metrics used include MAE, MSE and MAPE, providing a comprehensive assessment of the model's performance. To ensure transparency and reproducibility, all code developed throughout this Bachelor's thesis has been made publicly available on GitHub (Oriol, 2025). The repository contains the complete set of scripts and notebooks used for data preprocessing, model training, hyperparameter tuning and result evaluation.

4.1 Preprocessed Data and Descriptive Insights

Following the data preparation and descriptive analysis methodology outlined in Chapter 3, the dataset was read and processed to ensure consistency. The dataset consists of 43,848 observations, covering the period from 2020 to 2024. The key variables include "Fecha y Hora", representing the timestamp of the recorded data, as well as the separate variables "Fecha" and "Hora" for more granular temporal analysis. The main variable of interest is "Generation", which measures wind power generation in megawatts (MW). During the data cleaning process, a total of 6 missing values were identified within the "Generation" variable. According to the proposed methodology, linear interpolation was applied to fill these gaps, as it provides a smooth transition between known values, preserving the temporal consistency of the dataset. This approach was particularly suitable given the relatively short nature of the missing intervals. The resulting time series after the interpolation process can be visualized in Figure 4.1.

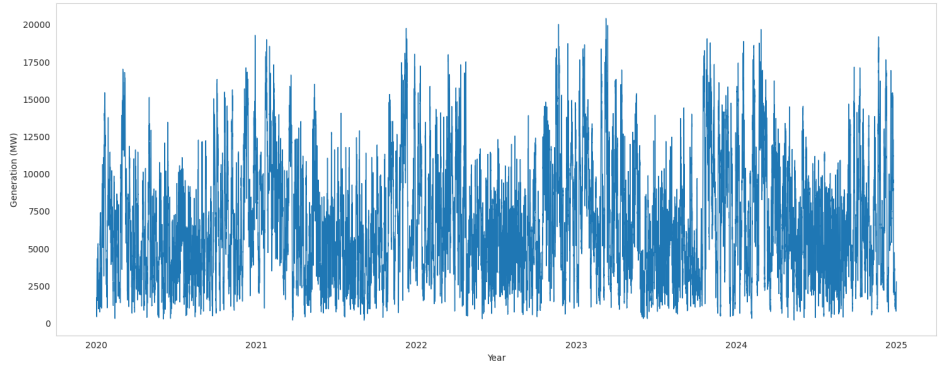


Figure 4.1: Evolution of Wind Power Generation in Spain from 2020 to 2025
Self-Elaboration

The descriptive statistics of the "Generation" variable, as shown in Table 4.1, reveal characteristics of wind power generation from 2020 to 2024. The mean value of approximately 6,661.10 MW indicates the average wind power output during the observed period, while the standard deviation of 3,867.57 MW highlights significant variability, which is typical in wind energy data due to fluctuating meteorological conditions. The minimum recorded generation value of 196 MW and a maximum of 20,321 MW demonstrate the wide range of wind power outputs observed. Additionally, the median value of 5,943 MW suggests that half of the observations fall below this point, indicating a slight right skew in the data distribution. The IQR, between the first quartile (3,587.44 MW) and the third quartile (9,119 MW), further confirms the presence of variability, with a substantial proportion of values concentrated in the mid-range.

Statistics	**Generation**
Count	43,848
Mean	6,661.10
Standard Deviation	3,867.57
Minimum value	196
25%	3,587.44
50%	5,943
75%	9,119
Maximum Value	20,321

Table 4.1: Statistics of Dataset Variables

To gain a deeper understanding of the distribution and variability of wind power generation over the study period, a histogram of the generation values was constructed. Figure 4.2 shows that the distribution of wind power generation is right-skewed, with most observations concentrated between 2,500 MW and 7,500 MW. The presented distribution suggests that moderate levels of power generation are more common, while higher values are progres-

sively less frequent. The tail on the right side of the histogram reflects sporadic high-output events, which are less frequent but significantly impact the overall distribution. To further investigate the variation in wind power generation over the years, a boxplot was constructed to represent the distribution for each year from 2020 to 2024 (Figure 4.3). The boxplot displays the median generation level for each year, along with the IQR and any potential outliers. The median values remain relatively stable across the years, which shows consistency in typical wind power output. However, the presence of some outliers above the upper whisker is notable. This might suggest that high wind power events occur, maybe during certain seasons. These outliers, although visually prominent, do not drastically affect the median or the central distribution of the data.

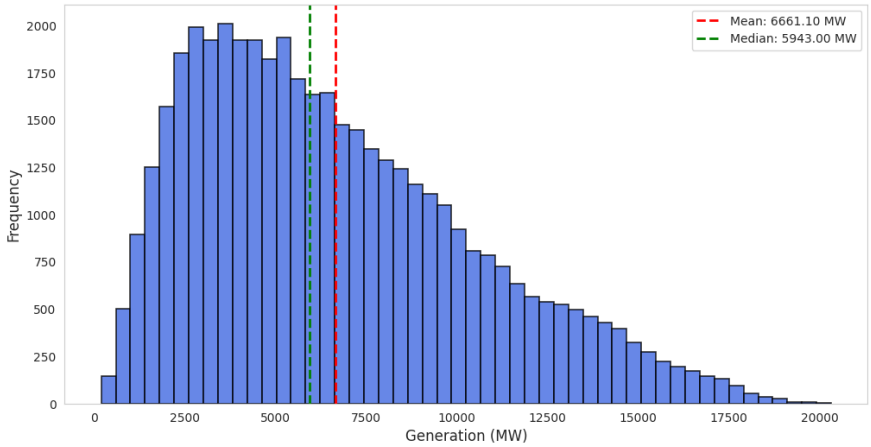


Figure 4.2: Distribution of wind power generation in Spain from 2020 to 2025.
Self-Elaboration

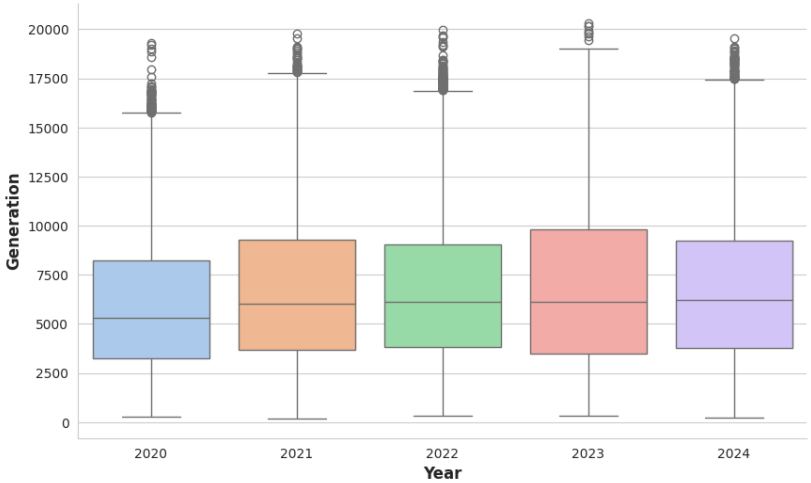


Figure 4.3: Boxplot of the distribution of wind power generation in Spain from 2020 to 2025
Self-Elaboration

To complement the previous exploratory analysis, an outlier detection procedure was carried out to characterize atypical high-generation events in the dataset. The IQR method was applied, using $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$ as thresholds to identify extreme values. A total of 257 outliers were detected across the 2020–2024 period, as illustrated in Figure 4.4. The year 2022 stands out with the highest number of outliers (86), likely reflecting exceptional meteorological conditions that led to significant deviations from standard generation patterns. In contrast, 2023 shows the lowest count (8), indicating greater output stability. The remaining years exhibit intermediate figures, with 2020 and 2024 recording 69 and 61 outliers respectively and 2021 a total of 30. These year-to-year differences underscore the natural variability of wind power generation and its dependence on atmospheric dynamics. According to the methodology described in Section 3.1, no modifications were applied to these extreme values in the modeling phase. Given that the outliers aligned with plausible meteorological phenomena and showed no indication of data recording errors, their retention was deemed appropriate. Preserving these observations ensured that the dataset reflected the full variability of real-world wind power generation, which is essential for building models capable of generalizing under diverse conditions.

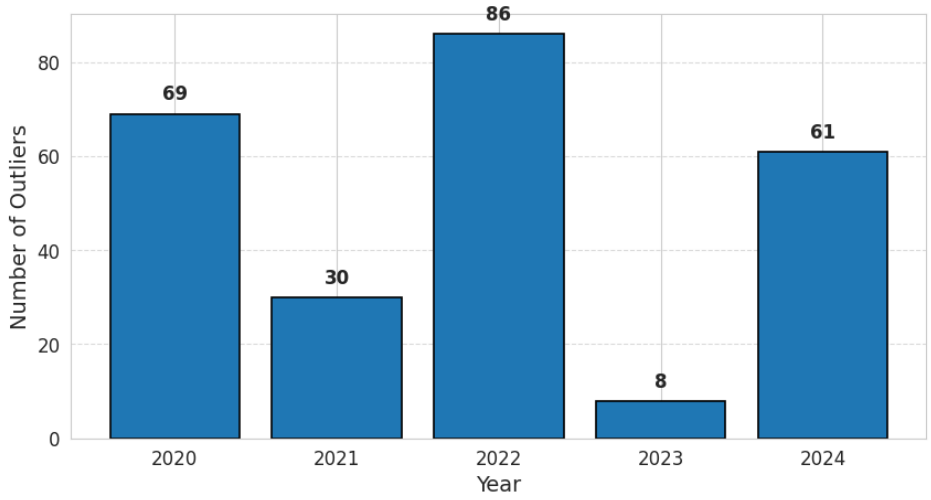


Figure 4.4: Outlier Count - Wind Power Generation in Spain from 2020 to 2025
Self-Elaboration

To analyze the temporal structure of wind power generation, a time series decomposition was performed to isolate its key components: trend, seasonality and noise. Figure 4.5 displays the decomposition of the original wind power generation series from 2020 to 2024. The trend component, shown in the second plot, indicates a gradual increase in average wind power generation over the years, with notable fluctuations around specific periods. This upward trend may reflect advancements in wind energy infrastructure or improved generation efficiency. The seasonal component exhibits repeating cycles, likely due to daily and yearly wind variations. Strong periodic fluctuations suggest predictable wind patterns that can be

leveraged for forecasting. The noise component, shown in the fourth plot, captures the residual variations not explained by the trend or seasonal patterns. These irregular fluctuations indicate the presence of unpredictable changes in wind power generation, likely influenced by short-term meteorological events or operational inconsistencies.

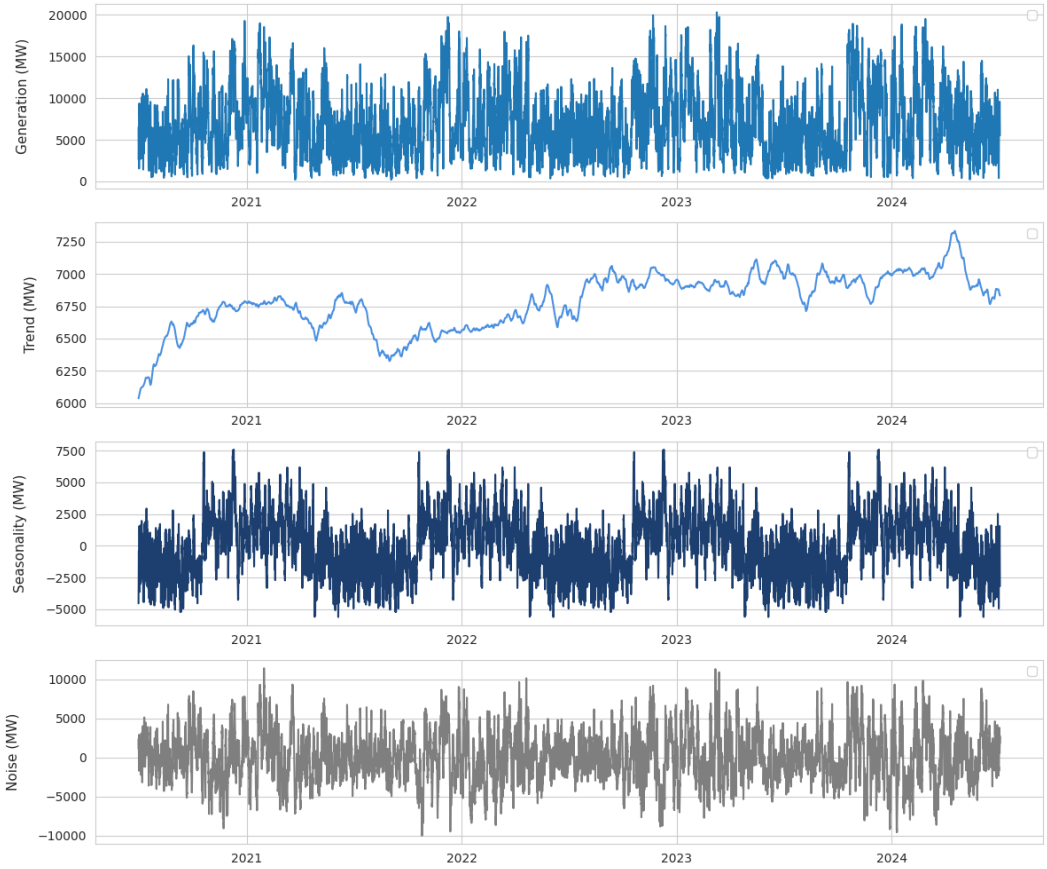


Figure 4.5: Decomposition of Wind Power Time Series
Self-Elaboration

4.2 Transformer Model Performance Evaluation

As described in the methodological framework in Chapter 3, a Transformer-based architecture was implemented for short-term WPF using historical generation data. The continuous time series was segmented into fixed-length overlapping windows using the sliding window technique, with the goal of capturing relevant temporal dependencies. For each configuration, a normalization process was applied to stabilize the learning dynamics, followed by model training using an optimized set of hyperparameters. The performance was evaluated on both training and test sets using three error metrics: MSE, MAE and MAPE.

To explore the model's sensitivity to the temporal resolution of input data, three different sequence lengths were considered: 12, 24, and 36 hours. This section begins by reporting

the results corresponding to the 12-hour sequence, which allows the model to use half-daily information to predict subsequent values. A total of twenty Transformer models were trained under this configuration, each combining different values of dimensionality, number of attention heads, encoder layers, dropout rate and batch size. The results are summarized in Table 4.2. As highlighted in the table, Model_1 yields the best predictive performance, with an MSE of 147.09, MAE of 8.06, and MAPE of 0.25% on the test set. This model was configured with a dimensionality of 32, 4 attention heads, 3 encoder layers, a dropout rate of 0.1 and a batch size of 16. Also, it displayed a low train-test performance gap, which confirms its generalization capacity. On the other hand, models such as Model_9 and Model_18 underperformed significantly, with MSE values exceeding 4000 and 10000 on the test set, respectively. These outcomes suggest that certain combinations of hyperparameters, particularly involving larger batch sizes and higher dropout rates, may degrade the model's ability to learn the underlying dynamics effectively.

Model	Set	Dim	Heads	Layers	Dropout	Batch	MSE	MAE	MAPE
Model_1	**Train**	**32**	**4**	**3**	**0.1**	**16**	**137.97**	**7.50**	**0.24%**
	Test	**32**	**4**	**3**	**0.1**	**16**	**147.09**	**8.06**	**0.25%**
Model_2	Train	16	2	3	0.2	64	1593.55	29.58	0.89%
	Test	16	2	3	0.2	64	1751.18	30.94	0.91%
Model_3	Train	16	2	2	0.2	16	698.27	21.11	0.46%
	Test	16	2	2	0.2	16	740.01	21.59	0.47%
Model_4	Train	32	8	2	0.2	32	1320.39	28.62	0.74%
	Test	32	8	2	0.2	32	1389.44	29.51	0.75%
Model_5	Train	16	4	3	0.3	64	1395.83	32.49	0.96%
	Test	16	4	3	0.3	64	1463.32	33.22	0.98%
Model_6	Train	16	4	3	0.2	32	472.19	13.17	0.38%
	Test	16	4	3	0.2	32	455.34	13.77	0.39%
Model_7	Train	16	4	3	0.1	32	2689.30	31.28	0.59%
	Test	16	4	3	0.1	32	3231.18	34.75	0.62%
Model_8	Train	16	2	3	0.3	64	2270.38	27.83	0.60%
	Test	16	2	3	0.3	64	2697.18	30.70	0.62%
Model_9	Train	16	2	3	0.2	32	3858.70	48.24	1.31%
	Test	16	2	3	0.2	32	4202.52	50.30	1.34%
Model_10	Train	32	4	3	0.2	64	1205.48	22.97	0.92%
	Test	32	4	3	0.2	64	1291.69	24.45	0.94%
Model_11	Train	32	2	3	0.3	16	783.88	26.40	0.73%
	Test	32	2	3	0.3	16	771.03	26.04	0.72%
Model_12	Train	16	2	3	0.1	64	1311.40	25.37	0.46%
	Test	16	2	3	0.1	64	1319.10	25.40	0.46%
Model_13	Train	32	4	2	0.2	16	763.29	16.58	0.22%
	Test	32	4	2	0.2	16	863.80	17.98	0.23%
Model_14	Train	32	2	2	0.1	16	334.28	11.65	0.41%
	Test	32	2	2	0.1	16	366.71	12.03	0.41%
Model_15	Train	16	2	2	0.2	32	489.42	16.09	0.51%

Model	Set	Dim	Heads	Layers	Dropout	Batch	MSE	MAE	MAPE
	Test	16	2	2	0.2	32	549.21	16.78	0.52%
Model_16	Train	16	4	3	0.1	16	301.78	15.47	0.46%
	Test	16	4	3	0.1	16	300.16	15.40	0.46%
Model_17	Train	16	4	3	0.1	64	2662.94	38.90	0.76%
	Test	16	4	3	0.1	64	2816.11	40.18	0.76%
Model_18	Train	32	2	3	0.1	32	9069.65	60.36	0.93%
	Test	32	2	3	0.1	32	10917.67	66.23	0.96%
Model_19	Train	32	4	2	0.2	64	2903.67	46.27	1.02%
	Test	32	4	2	0.2	64	3081.01	46.93	1.02%
Model_20	Train	32	2	3	0.3	64	764.87	16.05	0.41%
	Test	32	2	3	0.3	64	920.97	17.96	0.44%

Table 4.2: Train and Test performance for Transformer models for 12h sequence input

Following the evaluation of the 12-hour sequence length, this section presents the performance results of Transformer models trained with a 24-hour input window. This configuration allows the model to leverage an entire day's worth of historical wind power generation data to predict the following value, potentially capturing daily periodicities and broader temporal patterns that may not be evident in shorter sequences. Table 4.3 summarizes the performance of twenty Transformer configurations under the 24-hour input setting. Each model varies in its architectural parameters, including dimensionality, number of attention heads, encoder layers, dropout rate and batch size. As shown in the table, Model_3 achieved the best overall results, with a test MSE of 40.76, MAE of 4.24, and MAPE of only 0.09%. This model utilized a relatively simple architecture: a dimensionality of 16, 2 attention heads, 2 encoder layers, a dropout rate of 0.2 and a batch size of 16. The compact design of this model appears to strike a favorable balance between learning capacity and generalization, particularly for daily input sequences.

By contrast, models with more complex configurations, such as Model_2 and Model_8, showed poor performance on the test set, with MSE values exceeding 9,000 and 17,000 respectively, and MAPEs above 1.3%. These results suggest that certain combinations of high dropout rates, large batch sizes or excessive attention heads may hinder learning efficiency and increase forecasting error. Notably, Model_3 stood out not only as the best within the 24-hour group but also across all tested configurations, outperforming even the top-performing 12-hour model in every metric. Its exceptionally low MAPE of 0.09% highlights the model's capacity to exploit daily patterns effectively when architectural choices remain appropriately constrained. Furthermore, additional models within this group, such as Model_14 and Model_15, also demonstrated solid performance, with test MSE values of 81.01 and 136.32, and MAPEs of 0.17%, suggesting that the 24-hour input length, when well-optimized, is particularly suited to capturing the temporal dynamics of wind power generation.

Model	Set	Dim	Heads	Layers	Dropout	Batch	MSE	MAE	MAPE
Model_1	Train	32	4	3	0.1	16	849.35	21.95	0.71%
	Test	32	4	3	0.1	16	896.18	22.64	0.73%
Model_2	Train	16	2	3	0.2	64	7806.57	43.59	1.28%
	Test	16	2	3	0.2	64	9302.29	48.93	1.34%
Model_3	**Train**	**16**	**2**	**2**	**0.2**	**16**	**35.43**	**3.94**	**0.08%**
	Test	**16**	**2**	**2**	**0.2**	**16**	**40.76**	**4.24**	**0.09%**
Model_4	Train	32	8	2	0.2	32	576.75	18.15	0.35%
	Test	32	8	2	0.2	32	629.87	18.87	0.38%
Model_5	Train	16	4	3	0.3	64	4414.55	45.53	1.16%
	Test	16	4	3	0.3	64	4875.95	46.88	1.16%
Model_6	Train	16	4	3	0.2	32	210.50	9.21	0.30%
	Test	16	4	3	0.2	32	169.09	9.60	0.32%
Model_7	Train	16	4	3	0.1	32	597.42	17.66	0.66%
	Test	16	4	3	0.1	32	608.63	18.25	0.67%
Model_8	Train	16	2	3	0.3	64	16023.74	97.47	2.83%
	Test	16	2	3	0.3	64	17986.81	103.38	2.92%
Model_9	Train	16	2	3	0.2	32	3212.99	26.60	0.72%
	Test	16	2	3	0.2	32	3533.37	27.78	0.73%
Model_10	Train	32	4	3	0.2	64	4625.24	61.52	1.50%
	Test	32	4	3	0.2	64	4536.86	60.83	1.49%
Model_11	Train	32	2	3	0.3	16	767.07	15.67	0.40%
	Test	32	2	3	0.3	16	945.64	17.04	0.41%
Model_12	Train	16	2	3	0.1	64	3454.56	46.00	1.09%
	Test	16	2	3	0.1	64	3758.52	48.12	1.13%
Model_13	Train	32	4	2	0.2	16	420.53	15.87	0.44%
	Test	32	4	2	0.2	16	463.14	16.72	0.46%
Model_14	Train	32	2	2	0.1	16	75.01	7.60	0.16%
	Test	32	2	2	0.1	16	81.01	7.77	0.17%
Model_15	Train	16	2	2	0.2	32	115.05	7.66	0.16%
	Test	16	2	2	0.2	32	136.32	8.13	0.17%
Model_16	Train	16	4	3	0.1	16	310.17	12.94	0.38%
	Test	16	4	3	0.1	16	354.73	13.52	0.39%
Model_17	Train	16	4	3	0.1	64	719.70	13.37	0.38%
	Test	16	4	3	0.1	64	840.52	14.42	0.39%
Model_18	Train	32	2	3	0.1	32	2760.29	44.77	0.85%
	Test	32	2	3	0.1	32	2875.77	45.45	0.86%
Model_19	Train	32	4	2	0.2	64	776.46	25.03	0.48%
	Test	32	4	2	0.2	64	784.31	25.13	0.49%
Model_20	Train	32	2	3	0.3	64	1446.21	19.62	0.27%
	Test	32	2	3	0.3	64	1799.35	21.48	0.29%

Table 4.3: Train and Test performance for Transformer models for 24h sequence input

Finally, the evaluation of Transformer models trained with a 36-hour input sequence is presented in Table 4.4. This configuration provides the model with an extended temporal context. However, increasing the sequence length also introduces additional complexity and may result in diminishing returns if the added temporal information is not relevant for the target prediction. Due to the significantly longer training time associated with this configuration, only fifteen models were trained instead twenty, as was done for the other other setups.

The results show that Model_9 delivered the best performance for the 36-hour configuration, achieving a test MSE of 123.38, MAE of 8.78 and a remarkably low MAPE of 0.17%. This model was configured with a dimensionality of 32, 4 attention heads, 2 encoder layers, a dropout rate of 0.2 and a batch size of 16.

In contrast, several models with high model dimensions, large batch sizes or excessive attention heads suffered from poor generalization. For instance, Model_2 and Model_6 yielded extremely high error values, with test MSEs exceeding 400,000 and 200,000 respectively and MAPEs well above 5%. These outcomes indicate that increasing model complexity beyond a certain point, particularly in combination with long input sequences, can significantly degrade forecasting performance. When comparing across the three input configurations, it shows that the 36-hour window does not lead to a systematic improvement over shorter sequences. Although the best-performing models in this group achieved strong results, they did not outperform the top configurations from the 24-hour setup.

In fact, the 24-hour sequence, led by Model_3, consistently achieved the lowest error metrics among all models tested. This outcome suggests that, for the present wind power dataset, shorter historical windows are sufficient to capture the relevant temporal patterns required for accurate forecasting in this specific case.

Model	Set	Dim	Heads	Layers	Dropout	Batch	MSE	MAE	MAPE
Model_1	Train	16	4	2	0.3	32	426.78	18.22	0.37%
	Test	16	4	2	0.3	32	439.82	18.49	0.37%
Model_2	Train	32	8	2	0.1	32	411671.68	493.15	11.36%
	Test	32	8	2	0.1	32	478868.84	529.65	11.73%
Model_3	Train	32	4	3	0.1	64	19296.40	113.18	2.45%
	Test	32	4	3	0.1	64	21004.06	117.92	2.50%
Model_4	Train	32	4	3	0.3	64	1164.76	31.19	0.77%
	Test	32	4	3	0.3	64	1199.37	31.42	0.79%
Model_5	Train	32	2	3	0.3	16	451.19	17.64	0.45%
	Test	32	2	3	0.3	16	442.77	17.43	0.45%
Model_6	Train	16	8	3	0.3	64	175468.64	284.88	4.98%
	Test	16	8	3	0.3	64	217620.47	312.00	5.18%
Model_7	Train	16	4	3	0.1	32	1122.83	26.56	0.52%
	Test	16	4	3	0.1	32	1147.60	26.89	0.52%
Model_8	Train	32	2	2	0.2	16	125.83	9.39	0.29%
	Test	32	2	2	0.2	16	128.81	9.40	0.29%

Model	Set	Dim	Heads	Layers	Dropout	Batch	MSE	MAE	MAPE
Model_9	**Train**	**32**	**4**	**2**	**0.2**	**16**	**117.60**	**8.55**	**0.17%**
	Test	**32**	**4**	**2**	**0.2**	**16**	**123.38**	**8.78**	**0.17%**
Model_10	Train	16	4	3	0.3	64	18771.26	131.25	2.72%
	Test	16	4	3	0.3	64	19756.16	133.07	2.70%
Model_11	Train	32	8	3	0.2	32	6038.78	66.00	1.62%
	Test	32	8	3	0.2	32	6637.64	68.56	1.63%
Model_12	Train	16	2	2	0.1	16	132.47	6.87	0.18%
	Test	16	2	2	0.1	16	145.50	7.35	0.20%
Model_13	Train	32	8	2	0.3	32	686.52	24.17	0.49%
	Test	32	8	2	0.3	32	687.50	23.98	0.49%
Model_14	Train	16	4	2	0.1	64	9894.99	85.12	2.35%
	Test	16	4	2	0.1	64	10328.33	87.74	2.39%
Model_15	Train	32	2	2	0.1	16	445.74	19.98	0.54%
	Test	32	2	2	0.1	16	443.69	19.75	0.55%

Table 4.4: Train and Test performance for Transformer models for 36h sequence input

Based on the comprehensive evaluation of Transformer architectures using different input sequence lengths and the comparison between the best model for the architectures, it can be appreciated in Table 4.5 that the 24-hour configuration, particularly Model_3, demonstrated the most favorable balance between predictive accuracy and model efficiency. Its superior performance across all key metrics underscores the suitability of this configuration for short-term WPF. To further validate its effectiveness, Model_3 (24-hour input) will now be benchmarked against traditional RNN models, specifically GRU and LSTM architectures. This comparative analysis aims to determine the most effective DL approach for capturing the temporal dynamics of wind power generation in short-term forecasting scenarios.

Sequence Length - Model	Set	Dim	Heads	Layers	Dropout	Batch	MSE	MAE	MAPE
12h - Model_1	Train	32	4	3	0.1	16	137.97	7.50	0.24%
	Test	32	4	3	0.1	16	147.09	8.06	0.25%
24h - Model_3	Train	16	2	2	0.2	16	35.43	3.94	0.08%
	Test	16	2	2	0.2	16	40.76	4.24	0.09%
36h - Model_9	Train	32	4	2	0.2	16	117.60	8.55	0.17%
	Test	32	4	2	0.2	16	123.38	8.78	0.17%

Table 4.5: Train and Test performance for the best Transformer model for each time sequence input

4.3 Comparative Analysis with Recurrent Neural Networks (GRU and LSTM)

For the evaluation, the LSTM and GRU models were implemented following a comparable structure. Each model consisted of recurrent layers (LSTM or GRU, respectively), followed by a dense output layer to generate the forecast. Hyperparameters such as the number of model dimension (16 or 32), hidden units (1 or 2), number of layers (2 or 3), dropout rate (0.1 or 0.2), batch size (16 or 32) and sequence length (12, 24 or 36) were optimized through a grid search procedure. Multiple combinations were tested to identify the most suitable configuration for each architecture. Additionally, early stopping was applied during training to prevent overfitting and ensure generalization on unseen data.

For both models, the data was normalized and split into training and testing sets (70% and 30%, respectively) and sequences of 12h, 24h or 36h, same sequence lengths that were tested the Transformer models. Also, both models used the same evaluation metrics as the Transformer: MSE, MAE and MAPE.

In the table below, Table 4.6, are the results for the best combination of hyperparameter for each type of model. It can be observed that the Transformer model outperformed both LSTM and GRU models in all evaluated metrics. Overall, the Transformer consistently provided the most accurate and generalizable predictions for short-term wind power generation, highlighting its superior performance over the other models.

Model	Set	Seq Length	Dim	Head/Hidden	Layers	Dropout	Batch	MSE	MAE	MAPE
Trans	Train	24	16	2	2	0.2	16	35.43	3.94	0.08%
former	Test	24	16	2	2	0.2	16	40.76	4.24	0.09%
LSTM	Train	36	32	2	2	0.1	16	116,529.32	247.48	4.76%
	Test	36	32	2	2	0.1	16	143,538.88	267.08	4.88%
GRU	Train	36	32	1	3	0.1	16	118,938.93	254.02	5.17%
	Test	36	32	1	3	0.1	16	144,631.42	272.47	5.28%

Table 4.6: Train and Test performance for Transformer Model, LSTM and GRU

To complement the tabular results and provide a more intuitive understanding of model performance, Figure 4.6, Figure 4.7 and Figure 4.8 compare the actual wind power generation, which was the test set, with the predictions made by the best-performing configurations of each model: Transformer, LSTM and GRU. These visualizations offer valuable insight into how closely each model follows the real data over time.

The visual analysis of model predictions reveals clear differences in performance across the GRU, LSTM and Transformer architectures. The GRU model tends to underpredict certain peaks and exhibits a noticeable lag when adjusting to abrupt increases or drops in wind generation. Although it successfully captures general trends, its forecasts appear overly smoothed, likely due to its simplified gating mechanism, which limits its responsiveness to

sharp variations. In contrast, the LSTM model shows greater reactivity to sudden spikes in generation, particularly in the mid-range values where it demonstrates strong accuracy. However, it occasionally overshoots during periods of high volatility, indicating a certain sensitivity to fluctuations in the input sequence. The Transformer model offers the most accurate visual alignment with actual generation values, closely tracking both gradual shifts and abrupt changes. This consistency highlights its superior capacity to model long-range temporal dependencies and complex dynamics, supporting the low test MAPE (0.09%) previously reported and reinforcing its status as the best-performing architecture in this study.

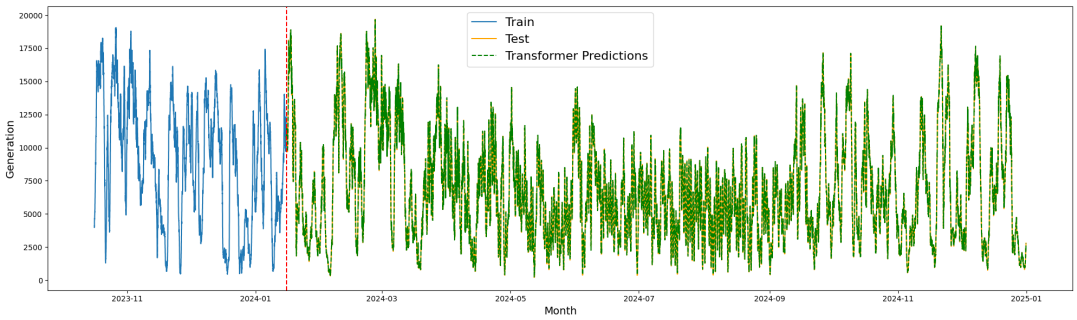


Figure 4.6: Wind Power Generation Forecast for Transformer Model 3, 24h sequence length Self-Elaboration

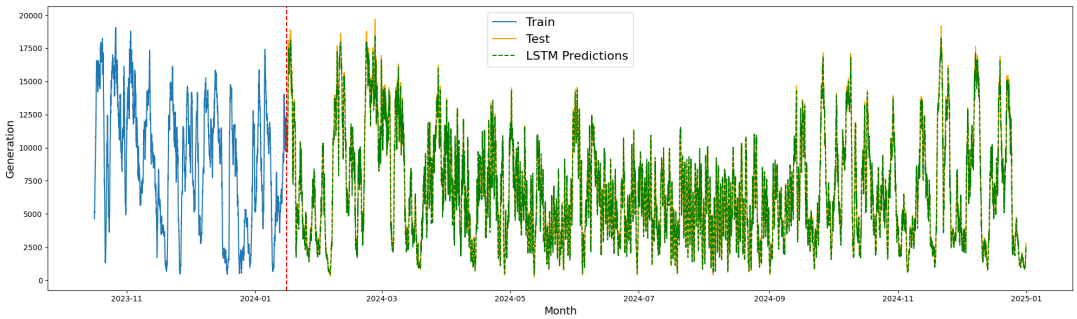


Figure 4.7: Wind Power Generation Forecast for LSTM Best Model, 36h sequence length Self-Elaboration

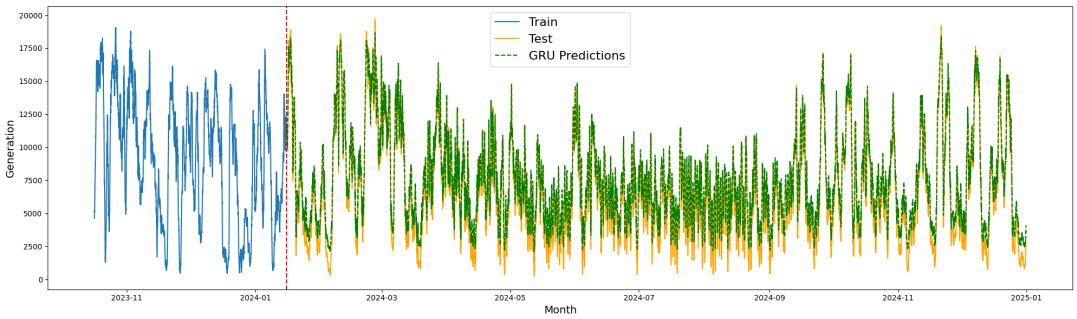


Figure 4.8: Wind Power Generation Forecast for GRU Best Model, 36h sequence length Self-Elaboration

In conclusion, the visual inspection of predicted vs. actual generation curves confirms the superiority of the Transformer model in accurately tracking wind power dynamics. Although LSTM and GRU exhibit very decent forecasting capabilities, their relative performance lags behind, particularly during abrupt changes. These findings support the quantitative metrics and highlight the importance of model selection based on both predictive performance and operational constraints.

Chapter 5

Conclusions

Accurate forecasting of wind power generation constitutes a fundamental component in the effective integration of renewable energy sources into modern power systems. As the share of wind energy increases globally, so does the complexity of managing its inherent variability and unpredictability. In particular, short-term wind forecasting is essential for maintaining grid stability, optimizing energy dispatch strategies, reducing operating costs and enabling more efficient participation in electricity markets. Within this context, the present Bachelor's thesis provides a meaningful contribution by demonstrating the applicability and effectiveness of Transformer-based architectures for short-term wind power prediction using historical generation data.

The work is situated within a broader methodological evolution observed in the literature, where traditional forecasting techniques such as ARIMA and persistence models have been progressively replaced by more advanced ML and DL models. Among these, Transformer architectures have gained increasing attention due to their ability to model long-range temporal dependencies through self-attention mechanisms. Unlike RNNs or LSTMs, which rely on sequential processing, Transformers leverage parallel computation and a self-attention mechanism to dynamically weight the relevance of different time steps. These capabilities make them particularly effective for time series forecasting tasks with complex temporal dynamics, such as wind power generation.

To contribute to this evolving research landscape, the study follows a rigorous and structured methodology that spans data acquisition, preprocessing, model design, training and evaluation. The dataset used comprises wind power generation data for Spain from 2020 to 2024. This time period captures a representative sample of operational conditions and variability in wind production, providing a strong empirical foundation for model development.

A key pillar of the project was the ensurance of data quality through preprocessing, which is crucial for time series forecasting models to function correctly and reliably. The dataset underwent a detailed cleaning process, beginning with the detection and interpolation of missing values, which were primarily caused by daylight saving time shifts. These were addressed

using linear interpolation to maintain continuity in the series without introducing artificial fluctuations. In parallel, outliers were identified using the IQR method, but were retained to preserve the dataset's variability and realism. Furthermore, normalization and sliding window segmentation were applied to prepare the time series for input into Transformer models, ensuring both temporal consistency and efficient learning dynamics.

One of the most relevant contributions of this work is the decision to implement and compare multiple Transformer configurations, rather than relying on a single model instance. This strategy was driven by the recognition that Transformer performance can vary significantly depending on hyperparameter choices, particularly in terms of model dimension, attention head count, number of encoder layers, sequence length, dropout rate and batch size. Consequently, a comprehensive grid search was performed across these parameters, training around fifty Transformer models with varied configurations. This exhaustive approach allowed for a robust evaluation of how model complexity, regularization and sequence granularity affect forecasting accuracy and generalization capability.

The results showed that the best performance was achieved using a 24-hour input window, which struck an optimal balance between information richness and noise. Specifically, Model_3, trained with a dimensionality of 16, 2 attention heads, 2 encoder layers, dropout of 0.2 and batch size of 16, achieved the lowest error rates across all metrics (MSE = 40.76, MAE = 3.94, MAPE = 0.09%), outperforming the configurations trained with the other sequence lengths (12h and 36h).

To further contextualize the performance of Transformer models, a comparative analysis was conducted using LSTM and GRU architectures, both of which have traditionally been used for time series prediction tasks. These models were carefully implemented with equivalent input formats and underwent their own hyperparameter optimization through grid search. The same evaluation metrics (MSE, MAE and MAPE) were used for all models, ensuring a consistent and fair comparison. Despite these efforts, the best LSTM and GRU configurations were unable to outperform the top Transformer model. Their forecasts exhibited slightly higher error values. These findings support the previously mentioned hypothesis that the self-attention mechanism in Transformer models offers a structural advantage in capturing relevant temporal features.

Nonetheless, it is important to acknowledge the computational demands associated with Transformer architectures, which represent a trade-off for their superior performance. Training the models required significant processing time and memory, particularly when larger dimensions or deeper encoder stacks were tested. This limitation suggests that future applications in operational environments should consider performance-cost trade-offs carefully and explore avenues to improve model efficiency without compromising accuracy.

Building upon the findings of this study, several future lines of research can be proposed to enhance forecasting robustness and broaden the applicability of Transformer-based models. First, longer input sequences (e.g., 48, 72 or 96 hours) could be evaluated to deter-

mine whether additional historical context further improves performance, though such models would require higher computational resources. Second, deeper architectures with more attention heads and encoder layers could be tested using more powerful GPUs or cloud-based solutions to assess the limits of model capacity for time series tasks. Third, spatio-temporal modeling could be incorporated by extending the dataset to include wind farms in other regions or countries. This would allow the model to learn geographic correlations and improve its generalization across diverse meteorological contexts.

In conclusion, this Bachelor's thesis provides compelling empirical evidence for the superiority of Transformer-based models in short-term wind power forecasting. Through a detailed methodology that prioritizes data quality, and hyperparameter exploration, the study validates the model's ability to outperform traditional DL alternatives. At the same time, it opens up multiple research avenues that could further advance forecasting capabilities in renewable energy systems. As energy systems become increasingly data-driven, models like the Transformer will play a key role in enabling more reliable, sustainable and intelligent grid operations.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que *ChatGPT* u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Teresa Oriol Guerra, estudiante de Doble Grado en Business Analytics y Relaciones Internacionales (E6 Analytics) de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Wind Power Generation Forecasting Using Transformer-based Time Series Models", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa *ChatGPT* u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Brainstorming de ideas de investigación: Utilizado para idear y esbozar posibles áreas de investigación.
2. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
3. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
4. Interpretador de código: Para realizar análisis de datos preliminares.
5. Constructor de plantillas: Para diseñar formatos específicos para secciones del trabajo.
6. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.
7. Generador previo de diagramas de flujo y contenido: Para esbozar diagramas iniciales.
8. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.

9. Generador de datos sintéticos de prueba: Para la creación de conjuntos de datos ficticios.

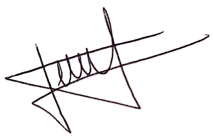
10. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

11. Traductor: Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado *ChatGPT* u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 17 de junio de 2025

Firma:



Teresa Oriol Guerra

References

Ahmed, U., Muhammad, R., Abbas, S. S., Aziz, I., & Mahmood, A. (2024). Short-term wind power forecasting using integrated boosting approach. *Frontiers in Energy Research*, *12*. (https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2024.1401978/full) doi: 10.3389/fenrg.2024.1401978

Brownlee, J. (2020). *Deep learning for time series forecasting*. Machine Learning Mastery. ((Accessed: September 5, 2024))

Bu, S.-J., & Cho, S.-B. (2020). Time series forecasting with multi-headed attention-based deep learning for residential energy consumption. *Energies*, *13*(18), 4722. (https://www.mdpi.com/1996-1073/13/18/4722)

Chen, P., Pedersen, T., Bak-Jensen, B., & Chen, Z. (2009). Arima-based time series model of stochastic wind power generation. *IEEE transactions on power systems*, *25*(2), 667–676. (https://ieeexplore.ieee.org/document/5340622)

Commission, E. (2020a). *Emissions trading system (ets).* ((Accessed: September 22, 2024) https://ec.europa.eu/clima/eu-action/eu-emissions-trading-system-eu-ets_en)

Commission, E. (2020b). *Fit for 55 package.* ((Accessed: September 22, 2024) https://ec.europa.eu/clima/eu-action/european-green-deal/delivering-european-green-deal_en)

Commission, E. (2020c). *Renewable energy directive (red ii).* ((Accessed: September 22, 2024) https://energy.ec.europa.eu/topics/renewable-energy/renewable-energy-directive-targets-and-rules/renewable-energy-directive_en)

Commission, E. (2021). *European green deal.* ((Accessed: September 22, 2024) https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en)

ENTSO-E. (2025). *Actual generation per production type.* ((Accessed: January 23, 2025) https://transparency.entsoe.eu/generation/r2/actualGenerationPerProductionType/show)

Hanifi, S., Liu, X., Lin, Z., & Lotfian, S. (2020). A critical review of wind power forecasting methods—past, present and future. *Energies*, *13*(15), 3764. ((Accessed: September 24, 2024) https://www.mdpi.com/1996-1073/13/15/3764)

Hassan, Q., Algburi, S., Sameen, A. Z., Salman, H. M., & Jaszczur, M. (2023). A review of hybrid renewable energy systems: Solar and wind-powered solutions: Challenges, opportunities, and policy implications. *Results in Engineering*, *20*, 101621. ((Accessed: October 8, 2024) https://www.sciencedirect.com/science/article/pii/S259012302300748X) doi: 10.1016/j.rineng.2023.101621

IRENA. (2023). *Wind energy.* ((Accessed: September 3, 2024) https://www.irena.org/Energy-Transition/Technology/Wind-energy)

Kavasseri, R. G., & Seetharaman, K. (2009). Day-ahead wind speed forecasting using f-arima models. *Renewable Energy*, *34*(5), 1388–1393. ((Accessed: October 3, 2024) https://www.sciencedirect.com/science/article/pii/S0960148108003327)

Lange, M., & Focken, U. (2006). *Physical approach to short-term wind power prediction*. Springer-Verlag Berlin Heidelberg. Retrieved from https://link.springer.com/book/10.1007/3-540-31106-8

Li, W., & Law, K. E. (2024). Deep learning models for time series forecasting: a review. *IEEE Access*. (https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10583885)

Lima, J., Guetter, A., Freitas, S., Panetta, J., & de Mattos, J. G. (2017). A meteorological–statistic model for short-term wind power forecasting. *Journal of Control, Automation and Electrical Systems*, *28*, 679–691. (https://www.researchgate.net/publication/318291467_A_Meteorological-Statistic_Model_for_Short-Term_Wind_Power_Forecasting) doi: 10.1007/s40313-017-0329-8

Lin, Z., & Liu, X. (2020). Wind power prediction based on high-frequency scada data along with isolation forest and deep learning. *International Journal of Electrical Power Energy Systems*, *118*, 105835. (https://www.sciencedirect.com/science/article/abs/pii/S0142061519332491#!) doi: 10.1016/j.ijepes.2020.105835

Liu, H., & Zhang, Z. (2024). Development and trending of deep learning methods. *Artificial Intelligence Review*, *57*, 112. ((Accessed: September 24, 2024) https://link.springer.com/article/10.1007/s10462-024-10728-z#citeas) doi: 10.1007/s10462-024-10728-z

Oriol, T. (2025). *Python code.* (https://github.com/toriolg/TFG_Business_Analytics)

Our World in Data. (2024). *Cumulative installed wind energy capacity (gigawatts).* ((Accessed: September 18, 2024) https://ourworldindata.org/grapher/cumulative-installed-wind-energy-capacity-gigawatts?time=earliest..2022)

Pelletier, F., Masson, C., & Tahan, A. (2016). Wind turbine power curve modeling using artificial neural network. *Renewable Energy*, *89*, 207–214.

(https://www.researchgate.net/publication/291012629_Wind_turbine_power_curve_modelling_using_artificial_neural_network) doi: 10.1016/j.renene.2015.11.065

Shan, R., Niu, J., Chai, X., & Gu, Q. (2024). Mid-to-long term wind power forecasting based on arima-bp combined model. *Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering)*, *17*(4), 401–407. doi: 10.2174/2352096516666230818145947

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*. (https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)

Wang, J., Botterud, A., Bessa, R., Keko, H., Carvalho, L., Issicaba, D., … Miranda, V. (2011). Wind power forecasting uncertainty and unit commitment. *Applied Energy*, *88*(11), 4014-4023. ((Accessed: September 24, 2024) https://www.sciencedirect.com/science/article/abs/pii/S0306261911002339) doi: 10.1016/j.apenergy.2011.04.011

Wu, H., Meng, K., Fan, D., Zhang, Z., & Liu, Q. (2022). Multistep short-term wind speed forecasting using transformer. *Energy*, *261*, 125231. (https://www.sciencedirect.com/science/article/pii/S0360544222021193) doi: 10.1016/j.energy.2022.125231

Zhang, J., Yan, J., Infield, D., Liu, Y., & Lien, F. S. (2019). Short-term forecasting and uncertainty analysis of wind turbine power based on long short-term memory network and gaussian mixture model. *Applied Energy*, *241*, 229–244. (https://strathprints.strath.ac.uk/67694/5/Zhang_etal_AE_2019_forecasting_and_uncertainty_analysis_of_wind_turbine_power_based_on_long_short_term_memory_network.pdf) doi: 10.1016/j.apenergy.2019.03.026