Computers and Electrical Engineering 127 (2025) 110613

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Advancing harmonic prediction for offshore wind farms using synthetic data and machine learning

Alp Karadeniz *

Department of Electrical and Electronics Engineering, Balıkesir University, Balıkesir, Turkey

ARTICLE INFO

Keywords:
Offshore wind farms
Harmonic prediction
Data augmentation
Generative adversarial networks (GAN)
Machine learning
Deep learning
Total harmonic distortion (THD)
Renewable energy forecasting
Wind power quality

ABSTRACT

This study presents a novel forecasting model for accurate harmonic prediction in offshore wind farms (OWFs) using data augmentation and machine learning techniques. A Generative Adversarial Network (GAN) is employed to generate synthetic meteorological data, enhancing the training set for improved accuracy. The model utilizes wind speed data from Bozcaada, Turkey, and simulates voltage and current waveforms to predict Total Harmonic Distortion Voltage (THDV). Machine learning (Random Forest) and deep learning (LSTM, GRU) models are compared to assess prediction performance. Results show that the GAN-based data augmentation significantly enhances prediction accuracy. This study provides a valuable methodology for harmonic forecasting in OWFs, offering insights for future renewable energy system planning and grid stability.

1. Introduction

Amid the global drive towards achieving carbon neutrality in response to climate change, renewable energy sources, particularly offshore wind power, are gaining prominence. Major oil corporations are redirecting their investments towards renewables, policymakers are increasingly endorsing emerging technologies, and offshore wind energy, abundant and consistent, is at the forefront due to its expanding capacity and reduced visual impact compared to onshore wind farms. As floating wind turbines become more cost-effective, the growth trajectory of offshore wind energy is expected to continue, with both the quantity and scale of offshore wind farms expanding rapidly [1,2].

Wind Power Plants (WPPs) are required to uphold stringent power quality standards such as stable voltage and frequency to ensure grid reliability [3–5], with harmonic distortion [6] emerging as a significant concern. Wind turbine generators (WTGs) are categorized into four types, where Type 1 and Type 2 employ soft starters to mitigate in-rush currents, while Type 3 and Type 4, equipped with power converters [7], may introduce harmonics, particularly in offshore wind farms where Type 3 (doubly-fed induction generator) generators are prevalent [8]. To gauge the extent of distortion within the original signal, metrics like Total Harmonic Distortion (THD) and Total Demand Distortion (TDD) have been established for voltage and current harmonics.

Furthermore, harmonics forecasting serves as a methodology to design effective harmonic mitigation devices, including passive and active filters, aligning with IEEE 519 standards. In this study, harmonic estimation is conducted using typical meteorological year (TMY) data provided by the European Commission's Joint Research Centre (JRC), covering a period from August 2019, spanning 1 month with hourly intervals [9]. The TMY dataset encompasses parameters such as wind speed and solar radiation. Given that this

* Corresponding author at: Department of Electrical and Electronics Engineering, Balıkesir University, Balıkesir, Turkey.
E-mail address: akaradeniz@balikesir.edu.tr.

https://doi.org/10.1016/j.compeleceng.2025.110613
Received 20 July 2024; Received in revised form 18 March 2025; Accepted 31 July 2025
Available online 5 August 2025

study focuses solely on DFIG Type 3 offshore wind farms, only wind speed data is utilized. Subsequently, the prepared model undergoes simulation, yielding voltage and current values, which are then employed to train GAN, machine and deep learning techniques. Finally, a comparative analysis is conducted between the predicted models of ED and DD. 2019 August data is selected because it was the most recent wind speed data which is provided for Bozcaada region. Furthermore, after 2019 August there is no supplied data. Therefore, for GAN model, this situation is suitable case to produce more data to supply these expanded data to ML and DL methods to train models.

According to the literature summary (Section 2), it can be mentioned that these studies are related onshore wind and PV system's harmonic estimations. But in this study, it focused offshore DFIG system with long distance submarine cable. Also, the pivot point (Bozcaada) is considered as sample data locations in Aegean Sea region. In Turkey, at this time there is no application of offshore wind turbine. But recent studies [10,11], are focused offshore wind power generation capacity of Bozcaada region. In near future, for possible application of OWF in Bozcaada, Turkey, this study shows the harmonic estimation model performance with respect to meteorological data. Also, to predict future values by short-term data (for this system one month data) Generative Adversarial Network (GAN) is used to create extra data. Moreover, after that, these expanded data (ED) is used to train Random Forest (RF), LSTM and GRU models. Furthermore, these trained models are comparatively analyzed to show the efficiency of data augmentation technique to get successful prediction on futuristic harmonic estimations. Also, there have been no studies found in the literature about Harmonic estimation with GAN-based data augmentation technique. These findings are invaluable for advancing predictive capabilities in harmonic forecasting and related fields.

Moreover, the remainder of this paper is organized as follows: Section 2 provides a comprehensive literature review on harmonic prediction and data augmentation techniques in renewable energy systems. Section 3 outlines the motivation behind this study and highlights its key contributions. Section 4 describes the data preparation process, including the details of the meteorological dataset and simulation setup. Section 5 explains the data preprocessing steps, including feature selection, normalization, and the Generative Adversarial Network (GAN)-based data augmentation approach. Section 6 introduces the machine learning and deep learning algorithms used for prediction. Section 7 presents the experimental results and comparative analysis of different models. Finally, Section 8 summarizes the key findings and conclusions, along with potential directions for future research.

2. Literature overview

Several research papers in the literature focus on harmonic forecasting and data augmentation by GAN method in power generation systems [12–19]. Firstly, a paper [12] examines the ST-DAGANs-CapNet method which is developed for diagnosing wind turbine gearbox faults extracts time-frequency features using Stockwell transformation (ST), augments training data using GANs for data augmentation, and diagnoses single and compound faults using capsule neural networks (CapsNet). Experiments demonstrate that this approach effectively addresses the shortage of training data and yields better results than existing fault diagnosis methods.

Another paper [13] demonstrates the issue of data imbalance in deep learning-based fault diagnosis, faulty datasets are augmented using an enhanced flow-based generative model (EFBG). This proposed method has increased the accuracy of fault diagnosis in wind turbine gearboxes from 83.26 % to 88.01 %.

Also, the study [14] investigates how generative Data Augmentation (DA) techniques improve the performance of Machine Learning (ML) and Deep Learning (DL) algorithms in Extreme Wind Speed (EWS) prediction, crucial for minimizing turbine damage and outage events in wind farms. Various Variational AutoEncoders (VAEs) including Conditional VAEs and Class-Informed VAEs are proposed and analyzed. The results demonstrate that these techniques outperform traditional DA algorithms in EWS prediction in Spain, with the Class-Informed VAE achieving the best results using a Convolutional Neural Network approach, resulting in up to a 4 % improvement in Accuracy, Recall, and F1 score.

Moreover, the study [15] has increased the accuracy of day-ahead wind power predictions by approximately 5 % by creating datasets based on mutation rates and augmenting them using generative adversarial networks (GANs). The utilization of wind speed data and GAN-based techniques significantly enhances the accuracy and reliability of the prediction.

Additionally, the study [16] aims to develop a hybrid prediction model for accurate and reliable harmonic forecasts in renewable energy systems. Six different hybrid prediction models were constructed using combinations of multilayered Artificial Neural Networks (ANN) and Adaptive Neuro Fuzzy Inference System (ANFIS) and employed to perform harmonic predictions. Model-3 and model-6 were identified as the best and most consistent performers, demonstrating the effectiveness of the developed prediction methodology.

Also, the study [17] emphasizes the critical role of harmonic prediction in the development of devices aimed at minimizing harmonic distortions. To provide accurate and reliable predictions for harmonics in renewable energy systems, an Adaptive Neuro Fuzzy Inference System (ANFIS) was combined with a Long Short-Term Memory Network (LSTM) in two different structured models. The results demonstrate that the model utilizing ANFIS in the initial stage and LSTM in the second stage (referred to as the ANFIS-LSTM model) outperforms all other models, showing a significant improvement compared to previous methods in the literature.

Additionally, in [18], the paper illustrates modeling and predictive results for the total harmonic distortion (THD) of both current and voltage in a nonlinear high-power load. This is accomplished by employing sophisticated techniques such as neural networks and fuzzy inference. By utilizing data from a steel charging operation, a neuro-fuzzy adaptive system is trained and tested with diverse architectures, providing valuable insights for designing harmonic current filters to mitigate productivity loss and address power quality issues in industrial settings.

Additionally, a study [19], a predictive model is introduced that utilizes Long Short-Term Memory (LSTM) networks to forecast the generation of voltage harmonics in wind power systems, which is vital for ensuring grid stability. By utilizing data from the Jeffreys

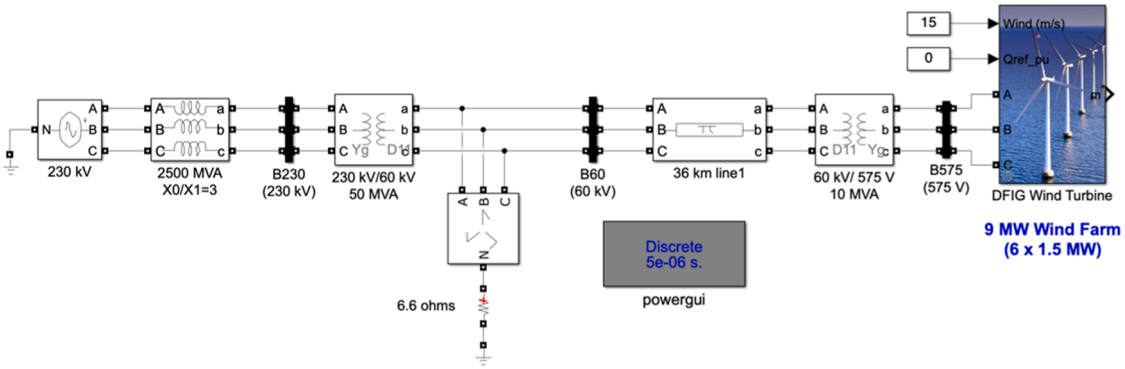


Fig. 1. The studied system Simulink model.

Bay Wind Farm, this model achieves accurate predictions of voltage harmonics with a low root mean square error (RMSE), thereby enabling effective grid management.

3. Motivation

From the literature summary presented here, it can be mentioned that studies on harmonic forecasting for wind energy systems have gained importance in the recent literature. However, currently, there is no application of offshore wind turbine in Turkey. In addition, recent studies [20–23] have focused on offshore wind power generation capacity of Turkey. In parallel with these studies, harmonic estimation for Bozcaada region and also, short-term data usage with GAN to create expanded data to train selected models are not studied before. In this regard, this study has novelty.

3.1. Contributions to the knowledge

This study makes several significant contributions to the field of harmonic forecasting in offshore wind farms (OWFs), particularly in the context of Turkey. The key contributions are summarized as follows:

- In this study, the forecasting model specifically designed for accurate and reliable prediction of harmonics in OWFs. This model leverages GAN-based data augmentation combined with multiple machine learning (ML) and deep learning (DL) models, including Random Forest (RF), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU).
- Also, the application of Double-Feed Induction Generator (DFIG) configurations to enhance the accuracy of harmonic predictions in wind turbine systems. This aspect of the model ensures robust simulation of output power and waveform characteristics.
- Moreover, integration of actual meteorological data from Bozcaada, Aegean Sea, Turkey, into the model. By using wind speed data as input parameters, the study provides realistic and region-specific predictions.
- Additionally, the use of Generative Adversarial Networks (GANs) for data augmentation represents a significant methodological advancement. The GAN model is trained on wind speed data from August 2019 to generate expanded data for September 2019, thereby overcoming limitations posed by short-term datasets.
- A thorough comparative analysis of the effectiveness of different ML and DL techniques (RF, LSTM, GRU) on forecasting Total Harmonics Distortion Voltage (THDV). This comparison offers valuable insights into the relative performance and suitability of these algorithms for harmonic prediction tasks.
- Also, by focusing on the Bozcaada region, this study addresses a gap in the literature regarding harmonic forecasting for offshore wind energy systems in Turkey. It provides novel insights that are directly applicable to the region's specific meteorological and wind conditions.
- In addition to that, the research showcases the effectiveness of GAN-based data augmentation techniques in improving THDV forecasting accuracy. The expanded data (ED) created by the GAN model, when used alongside daily data (DD), demonstrates superior predictive performance, thereby validating the approach.
- Lastly, the findings and methodologies presented in this study contribute to the broader body of knowledge in renewable energy system forecasting. They offer a foundation for future research aimed at enhancing harmonic prediction models, particularly in the context of renewable energy sources like wind power.

Overall, this study not only introduces a novel approach to harmonic forecasting using advanced data augmentation and machine learning techniques but also fills a critical gap in the literature regarding the application of these techniques to offshore wind farms in Turkey. The insights gained from this research are expected to inform and guide future studies in this domain.

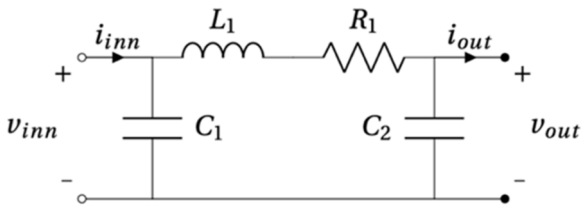


Fig. 2. Subsea cable π- model.

Table 1
Cable data from ABB data sheet.

Cable	R [mΩ/*km*]	L [*mH/km*]	C [*μF/km*]
500 mm^2 Cu cable	33.6	0.41	0.24

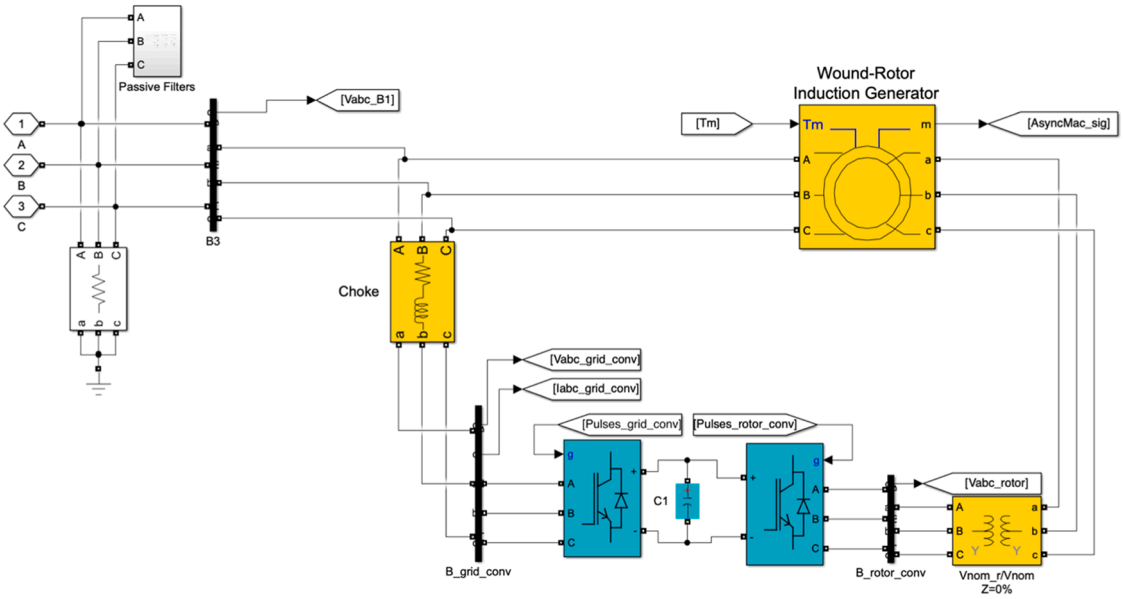


Fig. 3. Under mask illustration of DFIG wind turbine block.

Table 2
Parameters for 6 × 1.5 MW wind turbine.

Rated Power	1.5 MW x 6	Nominal DC voltage	1150 V
Rated Speed	12 m/s	Line filter capacitor ($Q = 50$)	50e3 F
Nominal Voltage	575V	Grid-side converter nominal voltage	575 V
Nominal Frequency	60 Hz	DC bus capacitor	10000e-6 F
Nominal DC bus voltage	1150 V	Grid-side coupling inductor [L, R]	0.1 H, 0.01 Ohm
Stator, rotor leakage inductance	0.18 pu. 0.16 pu.	Shaft spring constant	1.1
Stator, rotor resistance	0.023 pu. 0.016 pu.	Wind turbine inertia constant H (s)	0.432s
Switching Frequency	2700 Hz	Shaft base speed	125.66 rad/s

4. Data preparation

In this section, the modeling of the examined system and the wind farm is presented. The diagram of the analyzed system containing the offshore wind farm with type 3 DFIG turbines is shown in Fig. 1. It has a 230 kV grid network, a 230/60 kV transformer γ/Δ connected 60 MVA, a 60kV/575 V transformer Δ/γ connected 10 MVA, and a 36 km submarine cable for the offshore wind power system. Additionally, in the setup, the π-model parameters of the 36-km-long submarine cable (Fig. 2), which is used for the experimental offshore wind farm, are sourced from ABB data sheets [24,25]. They are provided in Table 1 and Fig. 3.

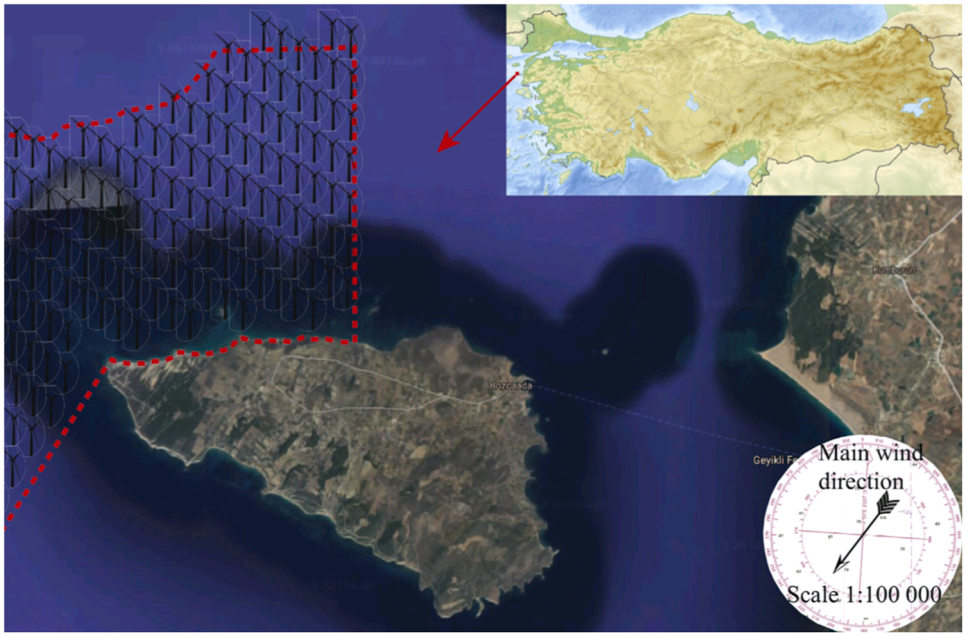


Fig. 4. The geological location Bozcaada, Aegean Sea, Turkey [10].

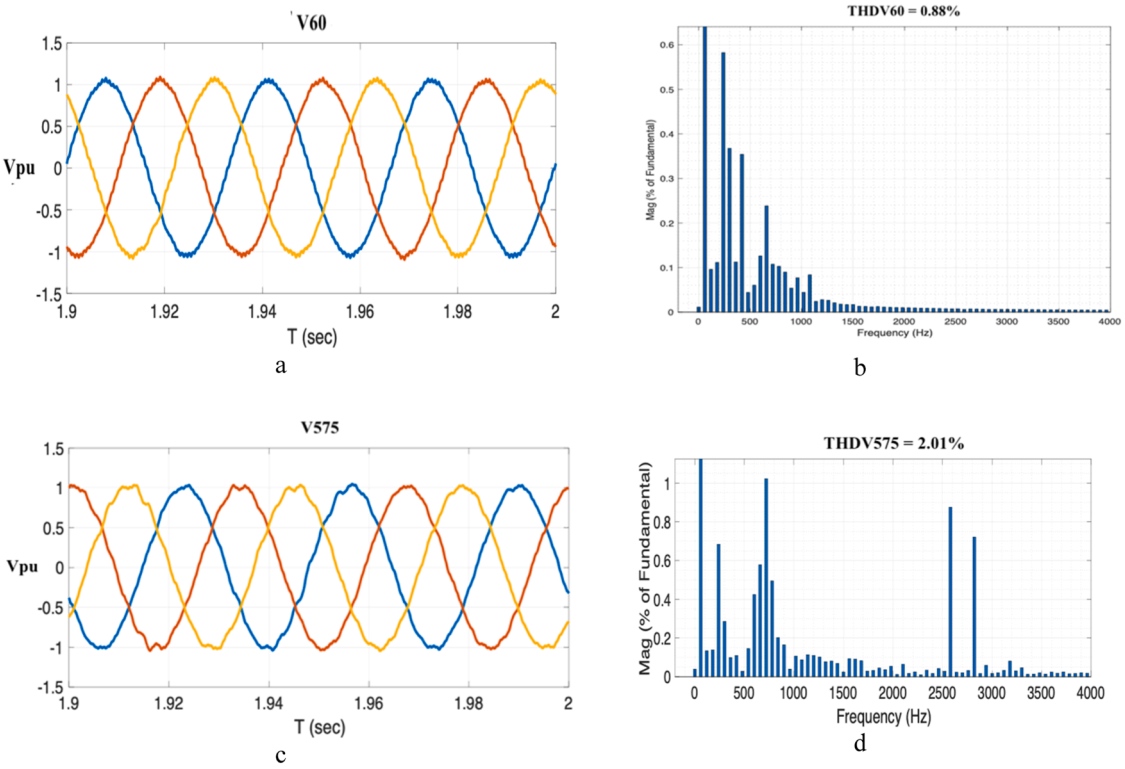


Fig. 5. The (a) voltage waveform of 60 kV Bus, (b) voltage harmonic spectrum at the 60 kV, (c) voltage waveform of 575 V Bus and (d) voltage harmonic spectrums at the 575 V bus for Bozcaada area.

Furthermore, the simulated setup comprises two transformers. The initial transformer is rated at 230/60 kV and is connected in a wye/delta configuration with a power rating of 50 MVA. The resistances and inductances of its windings and magnetization inductance are 2.66 × 10–3 and 0.08 per unit (p.u.), respectively. The second transformer, with a rating of 60 kV/575 V, is connected in a delta/

wye configuration from the primary side and has a power rating of 10 MVA. The resistances and inductances of its windings and magnetization inductance are 26.66 × 10–4 pu. and 0.008 pu., respectively. Additionally, the parameters of the induction generator are outlined in Table 2. Harmonic estimation is conducted using the standard meteorological year (TMY) dataset obtained from the European Commission, Joint Research Centre (JRC) [9]. The TMY dataset (time (UTC), WS10m) spans August 2019, encompassing approximately one month, with hourly data available. It includes wind speed data (m/s) and time data which has year/month/-day/hour format. Also, the modelled system is utilized to generate voltage values for training prediction techniques. Moreover, the reference point (Bozcaada) is regarded as the sample data location in the Aegean Sea region. The geological coordinates of the reference point are latitude 39.837 and longitude 25.967 for the Bozcaada area. Fig. 4 illustrates the geological locations of the reference point in the Aegean Sea region.

Moreover, Bozcaada is selected due to its optimal wind speed values, making it one of the most suitable locations in Turkey for offshore wind farm (OWF) installation [10]. The electrical grid is represented as a standard distribution network. The model undergoes simulation over a span of 100 s, starting from September 2019. A total of 720 data is proportionally scaled for this interval, with each point corresponding to 0.138 s. To replicate real-world wind speed inputs and produce the associated voltage harmonics, an OWF model is integrated. Harmonic distortions are noticeable in the voltage waveforms. One of the three-phase voltage waveforms is specifically analyzed due to the balanced system. Additionally, Fourier Transform (FT) analysis is utilized to extract harmonics from the data gathered from the scope. The FT window covers 2 cycles, capturing samples from the voltage waveforms over the 100-second period, yielding 720 samples.

Total Harmonic Distortion (THDV) values for voltage waveforms are extracted from the simulated signals. Following the simulation process, the distorted waveforms of the 575 V and 60 kV bus voltages and THDV values for Bozcaada are presented, along with the harmonic spectra in Fig. 5. It should be noted that the system includes its own passive filter for harmonic distortion. Without these filters, the Doubly Fed Induction Generator (DFIG) produces higher harmonic distortions that exceed IEEE 519 standards. Consequently, in this system, all harmonics for voltage waveforms are maintained within IEEE 519 standards.

From the data in Fig. 5, for the Bozcaada region, $THDV_{575V}$ and $THDV_{60kV}$ values are recorded at 2.01 % and 0.88 %, respectively. Notably, 575 V Bus THD figure reveal elevated harmonic values at frequencies of 2560 Hz and 2820 Hz, attributable to the IGBT-based PWM converter switching frequency of the OWF, which is 2700 Hz. These findings are treated as empirical data in the research and are subjected to comparison following the application of Machine/Deep Learning techniques.

5. Data preprocessing

In the initial phase of data processing, the dataset was retrieved. Subsequently, a subset of columns deemed essential for analysis was selected, including 'time(UTC)', 'WS10m', 'V60′, 'I60′, 'V575′, 'I575′, 'THDI575′, 'THDV60′, 'THDI60′, and 'THDV575′. Following column selection, the dataset underwent normalization to ensure uniformity across features. This involved rescaling both the input and output variables to a predefined range, utilizing min-max scaling. Specifically, the input features were normalized to a range of [0, 1], while the output variable underwent similar normalization to maintain consistency. Further data augmentation, employing a Generative Adversarial Network (GAN) model, was deferred to subsequent processing stages for enhanced dataset diversity and model robustness.

Moreover, to facilitate robust model training and evaluation, the dataset was partitioned into distinct training and testing subsets. This division was crucial to assess the performance of the developed models accurately. The splitting process was conducted using a standard practice, where a portion of the dataset typically around 20 % [26] was reserved for testing purposes, while the remaining data was allocated for model training. The random sampling technique ensured that both the training and testing sets were representative of the overall dataset, thus preventing bias and ensuring the generalization capability of the models. This approach enabled the models to learn from the training data and subsequently validate their performance on unseen test samples, providing insights into their effectiveness in real-world scenarios.

Additionally, the detailed explanation of the ***flowchart*** steps can be given as:

1. Read Data File: The dataset is read from a data file.
2. Select Required Columns: Relevant columns are selected from the dataset.
3. Data Augmentation: The dataset is augmented using a Generative Adversarial Network (GAN) model.
4. Rescale Data: The data is rescaled to a specific range.
5. Train and Test Sets: The dataset is split into training and testing sets.
6. Model Training (LSTM, GRU and Random Forest Regression): LSTM, GRU and RF models are trained on the training set.
7. Performance Evaluation: The performance of each model is evaluated.
8. Visualization of Results: Model predictions and actual values are visualized on a graph.

This sequential process outlines the steps involved in data preprocessing, model training, performance evaluation, and result visualization for the offshore wind farm data analysis.

5.1. Data augmentation model based on GAN

In this paper, a data augmentation model using Generative Adversarial Networks (GANs) was implemented [27] to enhance the robustness and performance of machine learning models. The GAN architecture consists of two neural networks: a generator and a

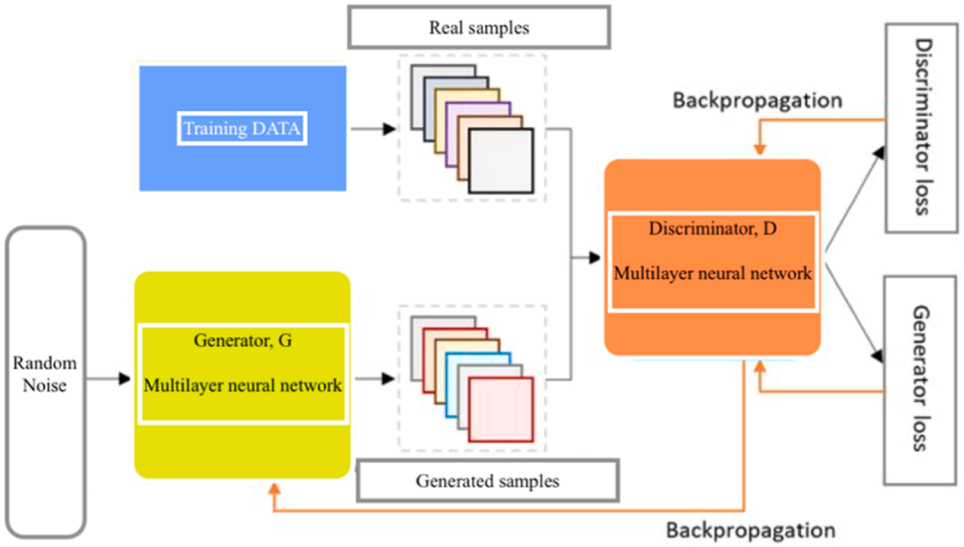


Fig. 6. An example of GAN architecture.

discriminator, trained simultaneously through adversarial processes. Synthetic data samples are generated by the generator to mimic real-world meteorological wind speed data, while the authenticity of these generated samples is evaluated by the discriminator. An example of GAN architecture is given at Fig. 6.

The primary objective of employing GANs in the data augmentation strategy is to produce a diverse and realistic set of wind speed data points, crucial for training and validating models. By expanding the dataset with synthetic yet realistic samples, the issues of data scarcity are addressed, and the model's ability to generalize to unseen data is enhanced. The generated data closely mirrors the statistical properties of the empirical data obtained from the EU Science HUB, ensuring that the augmented dataset is comprehensive and representative of real-world conditions.

The incorporation of GAN-based data augmentation has shown significant improvements in the predictive accuracy and stability of machine learning and deep learning models. This approach enriches the training dataset and mitigates the risk of overfitting, leading to more reliable and robust performance in OWF (Offshore Wind Farm) simulations and analyses.

Moreover, the selected models (Random Forest (RF), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU)) were chosen for their effectiveness in managing nonlinear relationships and temporal dependencies. RF was selected for its robustness in feature selection, whereas LSTM and GRU are well-established architectures for time-series forecasting due to their capability to model long-term dependencies [28–30].

GRU-MODEL

1. Build GAN

- Input Layer: This layer initializes the input dimension of the model. It consists of a dense layer with 100 neurons and a ReLU activation function [31].
- Hidden Layers: Two dense layers follow the input layer, each comprising 100 neurons and utilizing the ReLU activation function.
- Output Layer: This layer determines the output dimension of the model. It is a dense layer with a ReLU activation function. In total 100/100 Layers are used in this model.

2. Generate Synthetic Data with GAN model

- Input Layer: The input and output dimensions of the GAN model are determined based on the input and output data provided.
- Hidden Layers: The hidden layers are constructed during the compilation and training phases of the GAN model, and their details are not explicitly defined in this function.
- Output Layer: The GAN model outputs synthetic data points, which correspond to the augmented dataset.

These layers collectively constitute the architecture of the GAN model employed for data augmentation. Through the training process, the model learns to generate synthetic data points that closely resemble the original dataset, thereby expanding the dataset and enhancing the model's generalization capabilities. In this paper, the original data was August 2019. The data is small with respect to train a artificial intelligent model but GAN is used to create next month data as synthetic data. ED is created by combining August and the synthetic data to simulate for September 2019.

It should be noted that, not only wind speed data, the all V575pu, THDV575, THDV60, THDI60, THDI575, I575pu, Time, V60pu and I60pu data is generated by GAN model.

6. Machine learning algorithm

In this paper, to predict harmonic values in the particular area using both ED and DD, various machine learning methods were

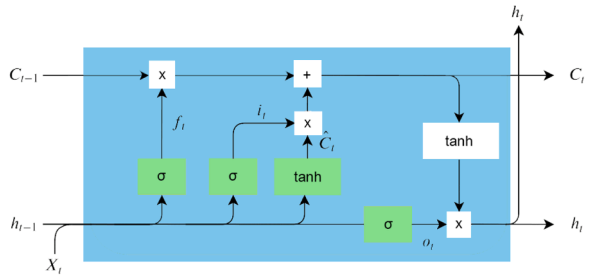


Fig. 7. Long short-term memory architecture.

explored, including random forest (RF). Below is an overview of the RF technique is explained.

6.1. Random forest regressor (RF)

In this study, as ML method, Random Forest technique is selected. RF constructs multiple decision trees during training and averages their predictions for regression or selects the mode prediction for classification. The "random" aspect encompasses two main facets: randomly sampling data for training each tree and considering only a random subset of features at each node for splitting. The construction process involves several crucial steps [32]. Bootstrap sampling randomly selects subsets of training data with replacements. Decision trees are then built on these subsets, taking into account only random feature subsets at each node. Predictions are aggregated, typically by averaging, to derive the final output. Mathematically, as expressed in Eq. (1), predictions for unseen samples x' can be made by averaging the predictions from all individual trees on x':

$$\widehat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{1}$$

To estimate uncertainty in predictions in Eq. (2), compute the standard deviation of predictions from each regression tree for a given input x′:

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \widehat{f})^2}{B-1}} \tag{2}$$

The number of samples/trees, B, in a model is adjustable, depending on the dataset's size and characteristics. Optimal B can be determined through cross-validation or by evaluating the out-of-bag error.

7. Deep learning architectures

In this paper, utilizing machine learning techniques, two renowned deep learning (DL) methods, namely long short-term memory (LSTM) and gated recurrent units (GRU), were employed to predict harmonic values in the Bozcaada region with both ED and DD. Below, we introduce these methods.

7.1. Long short-term memory (LSTM)

Moreover, a specialized variant of recurrent neural network (RNN) architecture, Long Short-Term Memory (LSTM), was developed to mitigate the vanishing gradient issue and adeptly capture long-range dependencies in sequential data. LSTMs are extensively utilized across various domains [26], including natural language processing, speech recognition, and time-series forecasting. Fig. 7 offers a schematic illustration of the LSTM architecture.

Also, at the heart of LSTMs lies the concept of maintaining a cell state that can store information across time steps, empowering the network to selectively retain or discard information at each step. This capability is achieved through a series of gates, which are small neural networks responsible for controlling the flow of information. The three key components of an LSTM cell are:

Forget Gate: It determines which information from the preceding cell state should be discarded and which information should be preserved. It accepts the current input and the output from the previous layer as input and generates a forget gate vector ft (where t denotes the current time step) with values ranging from 0 to 1 for each element. Subsequently, this vector is element-wise multiplied with the previous cell state Ct-1 to ascertain which information should be forgotten.

Input Gate: It discerns which new information should be incorporated into the cell state. Also, taking the current input and the output from the previous layer as input, it generates an input gate vector i_t. Additionally, it produces a candidate cell state $\widetilde{C}_t$, which serves as a contender for updating the cell state. The input gate vector regulates the extent to which the candidate cell state should be added to the current cell state.

Output Gate: It determines the output of the LSTM cell at the current time step. It takes as input the current input and the output

Table 3
Regression results of $THDV_{575V}$ with ML and DL methods.

MODEL Algorithms	THDV values of 575 V Bus			
	With ED		With DD	
	RMSE (%)	MAE (%)	RMSE (%)	MAE (%)
Random Forest (RF)	6.372	4.799	5.618	4.030
Gated Recurrent Units (GRU)	7.595	5.825	11.773	9.299
Long Short-Term Memory (LSTM)	7.266	5.556	11.121	8.654

from the previous layer and produces an output gate vector. The current cell state C_t is passed through a *tanh* activation function to squash the values between −1 and 1, and then multiplied element-wise with the output gate vector to produce the output of the LSTM cell h_t. Sequentially as given in Eq. (3):

$$\left.\begin{aligned} f_t &= \sigma\big(W_f[h_{t-1}, x_t] + b_f\big) \\ i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \\ \widetilde{C}_t &= \tanh(W_C[h_{t-1}, x_t] + b_C) \\ C_t &= f_t C_{t-1} + i_t \widetilde{C}_t \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ ht &= o_t \cdot tanh(C_t) \end{aligned}\right\} \tag{3}$$

In these equations, W_f, W_i, W_c, W_o are weight matrices, b_f, b_i, b_c, o_o are bias vectors, σ is the sigmoid activation function and [h_{t-1}, x_t] represents the concatenation of the previous hidden state h_{t-1} and the current input x_t.

By controlling the flow of information through the forget, input, and output gates, LSTMs can learn long-range dependencies in sequential data and mitigate the vanishing gradient problem, making them effective for tasks that require modeling temporal dynamics.

7.2. Gated recurrent units (GRU)

Also, another DL method, GRU is modelled in this paper. In detail, at the heart of GRUs lies the notion of hidden states, acting as memory units that retain information from previous time steps in a sequence. Yet, what distinguishes GRUs is their incorporation of gating mechanisms, which govern the flow of information within the network. These mechanisms comprise an update gate (Eq. (4)) and a reset gate (Eq. (5)).

$$z_t = \sigma(W_z[h_{t-1}, x_t]) \tag{4}$$

The update gate, represented by z_t, evaluates how much information from the previous hidden state should be preserved or adjusted in light of the current input. This gate aids the GRU in gauging the importance of past context for the present time step. Conversely, the reset gate, labeled as r_t, controls the extent to which the previous hidden state should be reset or disregarded, considering the current input.

$$r_t = \sigma(W_r[h_{t-1}, x_t]) \tag{5}$$

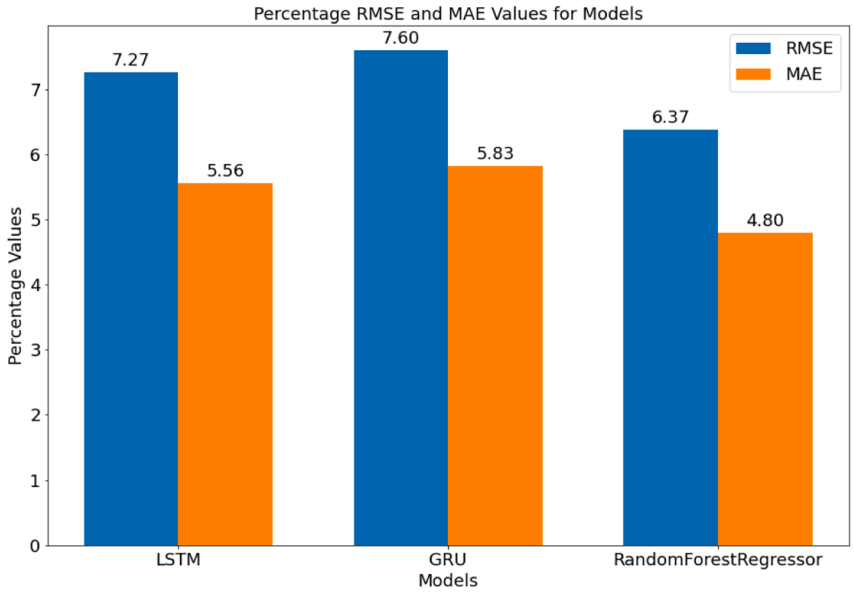
Upon computing the update and reset gates, GRUs generate a candidate hidden state ($\widetilde{h}_t$), which incorporates both the current input and the modified previous hidden state (Eq. (6)). This candidate state reflects the network's updated understanding of the current timestep's significance in the context of the sequence.

$$\widetilde{h}_t = \tanh(W[r_t \odot h_{t-1}, x_t] \tag{6}$$

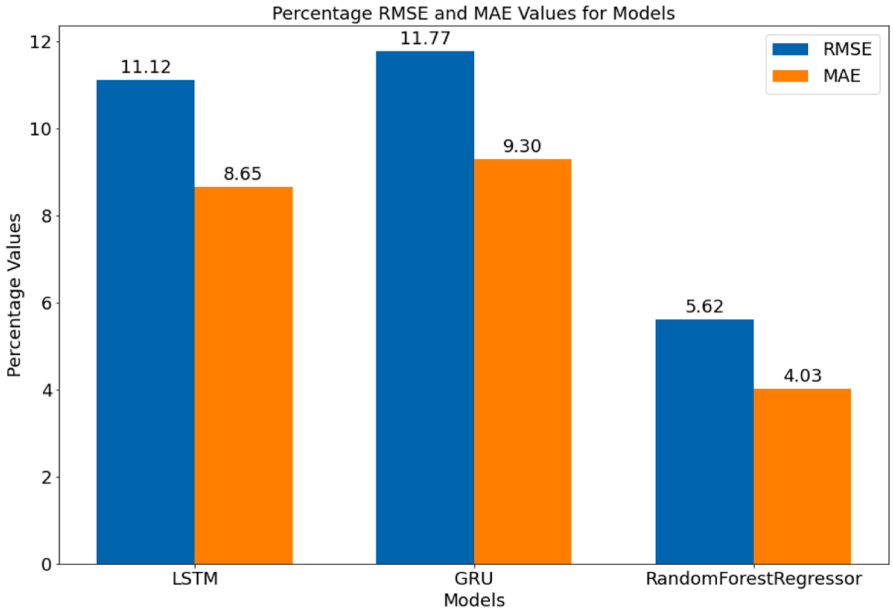
Ultimately, the update gate governs the fusion of the candidate hidden state with the previous hidden state to generate the final hidden state (h_t) for the current time step. This weighted amalgamation, as expressed in Eq. (7), guarantees the preservation of pertinent information while discarding irrelevant or obsolete data. Consequently, the network can dynamically adjust to the input sequence, ensuring adaptability and responsiveness.

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t \tag{7}$$

The mathematical formulation of GRUs involves a series of operations governed by trainable parameters, including weight matrices (W) and activation functions (σ). By learning these parameters from training data, GRUs optimize their ability to capture long-term dependencies and patterns within sequential data [33].



a

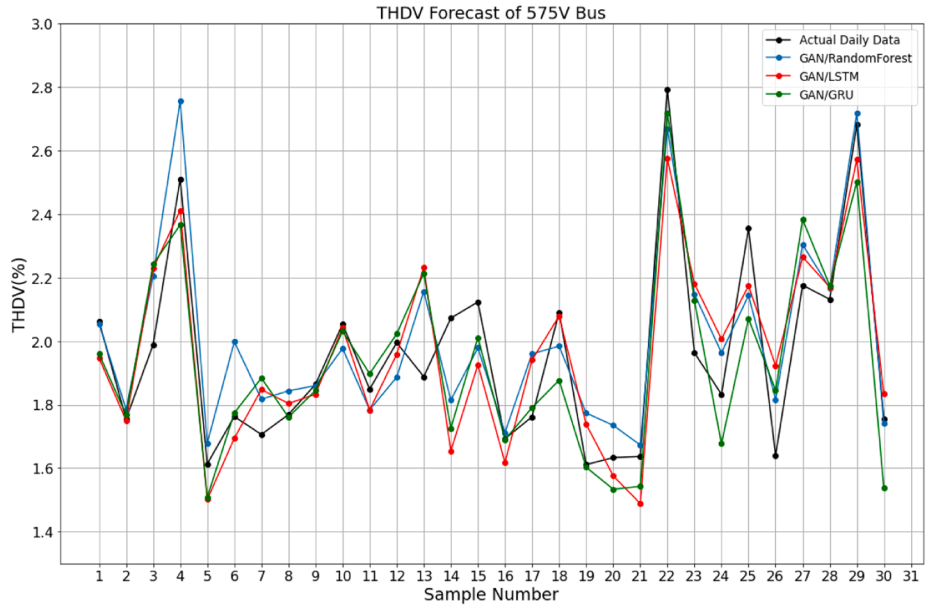


b

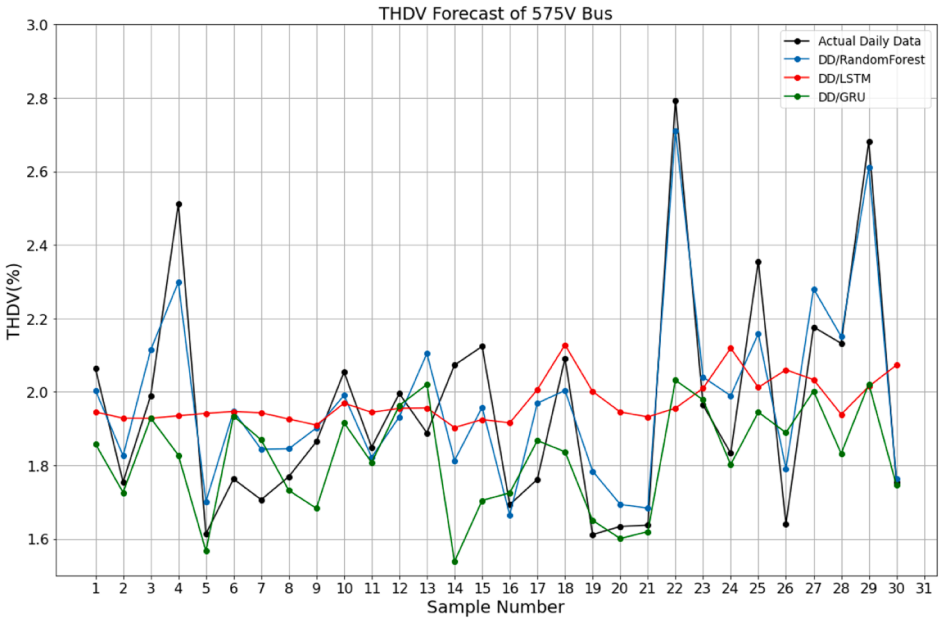
Fig. 8. The RMSE and. MAE values of a) with ED and b) with DD.

8. Experiments and results

In this research, initially, experiments were conducted with RF, LSTM and GRU techniques. The models were trained using data from the Bozcaada region as ED and DD. The performance analysis of these techniques for predicting THDV values of 575 V and 60 kV buses of the Bozcaada region provides detailed insights into their performance characteristics. The buses 575 V and 60 kV are selected because 575 V bus is the nearest source of harmonic distortion and 60 kV bus is point of common coupling (PCC) that connects wind power and grid power.



a



b

Fig. 9. Actual vs. predicted a) with GAN/ED and b) with DD graphs for Bozcaada.

8.1. $THDV_{575V}$ results

The regression results of $THDV_{575V}$ for DD and ED sets are presented in Table 3. Also, the illustrations of RMSE and MAE values of RF, LSTM and GRU models are given in Fig. 8.

The performance metrics presented in Table 3 offer valuable insights into the predictive capabilities of different model algorithms for forecasting total harmonic distortion voltage (THDV) values on the 575 V Bus.

From Table 3, the performance metrics of different model algorithms in predicting the total harmonic distortion voltage (THDV) values for the 575 V Bus under both ED and DD conditions are presented. The results are measured in terms of root mean square error

Table 4
Regression results of $THDV_{60kV}$ with ML and DL methods.

MODEL Algorithms	THDV values of 60 kV Bus			
	With ED		With DD	
	RMSE (%)	MAE (%)	RMSE (%)	MAE (%)
Random Forest (RF)	6.266	4.682	6.738	5.720
Gated Recurrent Units (GRU)	7.305	5.728	12.779	10.612
Long Short-Term Memory (LSTM)	7.363	5.796	12.126	9.818

(RMSE) and mean absolute error (MAE) percentages. From these terms, Random Forest (RF) model demonstrates a RMSE of 6.372 % and MAE of 4.799 % in the ED scenario, and a RMSE of 5.618 % and MAE of 4.030 % in the DD scenario. Conversely, the Gated Recurrent Units (GRU) model yields higher errors, with an RMSE of 7.595 % and MAE of 5.825 % in the ED scenario, and an RMSE of 11.773 % and MAE of 9.299 % in the DD scenario. Similarly, the Long Short-Term Memory (LSTM) model also exhibits elevated errors, with an RMSE of 7.266 % and MAE of 5.556 % in the ED scenario, and an RMSE of 11.121 % and MAE of 8.654 % in the DD scenario. These findings highlight the superior performance of the RF model compared to GRU and LSTM models, particularly evident in both ED and DD scenarios.

Fig. 9 show the graph of the actual and predicted daily values for 9/2019 (Bozcaada) data sets. It has been observed that the predicted values converge to actual ones at *a* part. But as seen from b part of Fig. 9, without GAN model, the trained models that are trained by DD show poor performance and the error between test values and predicted values are bigger. From Fig. 9 it can be said that, the expanded dataset (by GAN model) is superior that daily dataset and it is a proof that GAN method is applied successfully for data generation from short-term data.

8.2. $THDV_{60kV}$ results

Additionally, Table 4 presents the regression results of total harmonic distortion voltage (THDV) at the point of common coupling (PCC) for both dataset DD and ED, aiming to demonstrate the performance analysis of ED at the 60 kV Bus. Furthermore, in Fig. 10, the graphical representations of the root mean square error (RMSE) and mean absolute error (MAE) values for the Random Forest (RF), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) models are depicted.

The provided Table 4 outlines the performance metrics of various model algorithms in predicting the total harmonic distortion voltage (THDV) values for the 60 kV Bus under both ED and DD conditions. In the case of the Random Forest (RF) model, it achieves an RMSE of 6.266 % and MAE of 4.682 % in the ED scenario, while in the DD scenario, the RMSE increases to 6.738 % and MAE to 5.720 %. Conversely, both the Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) models exhibit higher errors in both ED and DD scenarios. For the GRU model, the RMSE is 7.305 % and MAE is 5.728 % in the ED scenario, while in the DD scenario, the RMSE spikes to 12.779 % and MAE to 10.612 %. Similarly, for the LSTM model, the RMSE is 7.363 % and MAE is 5.796 % in the ED scenario, and increases to 12.126 % RMSE and 9.818 % MAE in the DD scenario. These results indicate that the RF model outperforms the GRU and LSTM models in predicting THDV values for the 60 kV Bus under both ED and DD conditions.

These findings underscore the consistent superiority of the RF model in predicting THDV values across different bus voltages and distortion conditions, suggesting its robustness and effectiveness in such predictive tasks. However, it is essential to consider the specific requirements and characteristics of the application context when selecting the most appropriate predictive model.

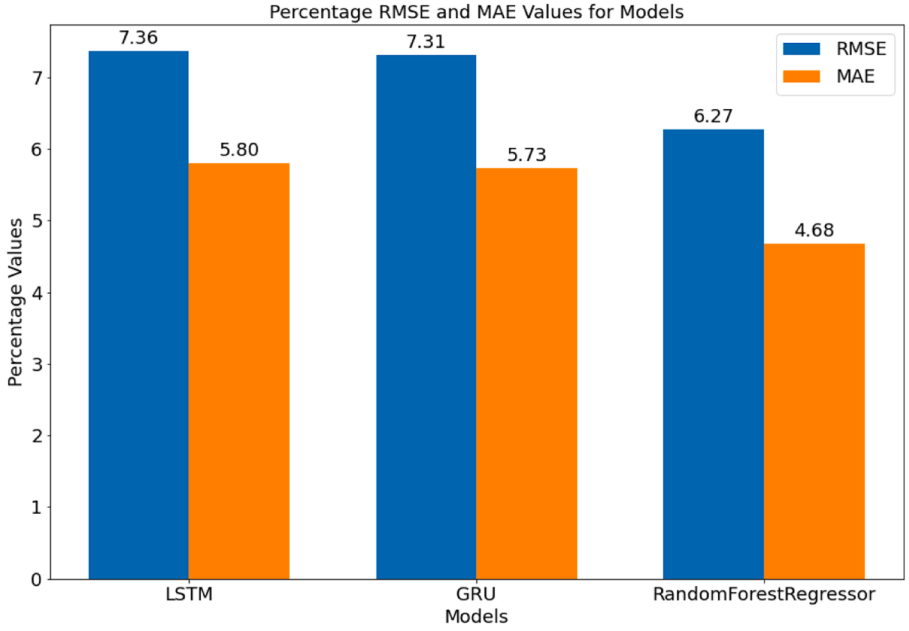
Moreover, Fig. 11 depicts the graphical representation of the observed and forecasted daily values for the Bozcaada datasets in September 2019. It is evident that the predicted values align closely with the actual data for a portion of the dataset. However, as illustrated in section b of Fig. 11, the absence of the Generative Adversarial Network (GAN) model results in inferior performance, particularly noticeable in the discrepancies between the test values and predicted values. This observation suggests that the models trained solely on DD data exhibit poorer predictive capabilities. This highlights the critical role of GAN methods in addressing the challenges associated with data scarcity and improving the robustness of predictive models in real-world applications.

9. Conclusıon

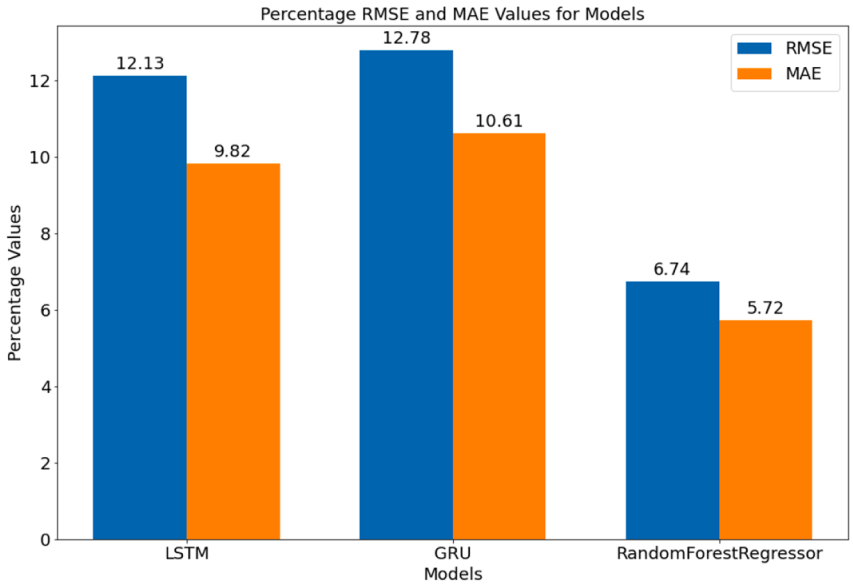
This study presents an advanced methodology for harmonic prediction in offshore wind farms by integrating data augmentation techniques with machine learning and deep learning models. The proposed approach employs Generative Adversarial Networks (GANs) to generate synthetic wind speed data, addressing challenges associated with limited datasets. The augmented data enhances the predictive capability of Random Forest (RF), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) models in forecasting Total Harmonic Distortion Voltage (THDV).

Experimental results demonstrate that GAN-based data augmentation significantly improves prediction accuracy compared to using only daily data. Specifically, for the 575 V bus, the RF model achieved an RMSE of 6.372 % and MAE of 4.799 % with expanded data (ED), compared to 5.618 % RMSE and 4.030 % MAE with daily data (DD). Similarly, for the 60 kV bus, the RF model yielded 6.266 % RMSE and 4.682 % MAE with ED, while the DD-trained model resulted in 6.738 % RMSE and 5.720 % MAE. These results highlight that the GAN-augmented dataset enhances model performance, reducing errors and improving prediction reliability.

Among the evaluated models, RF consistently outperforms LSTM and GRU, which exhibited higher RMSE values, particularly in DD



a



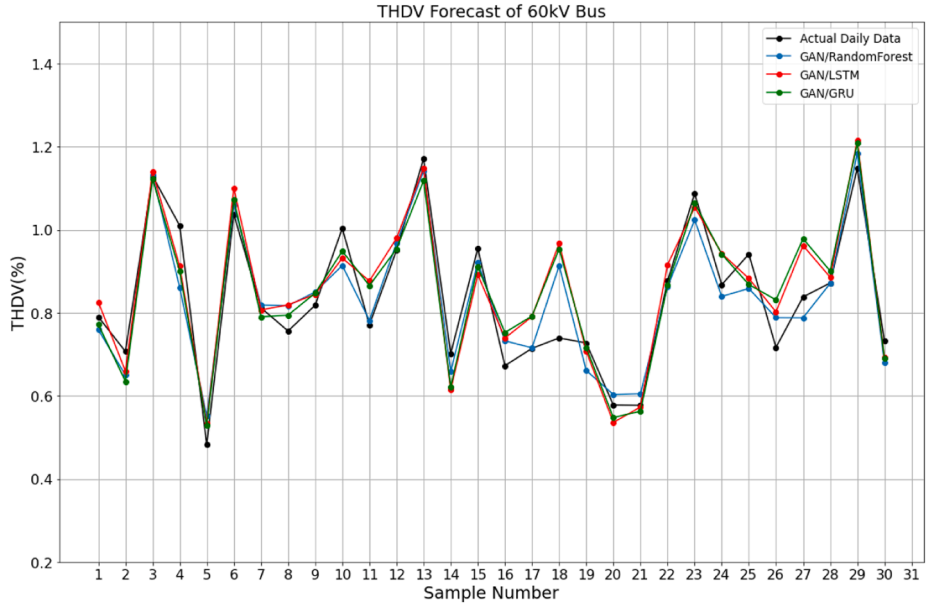
b

Fig. 10. The RMSE and. MAE values of a) with ED and b) with DD.

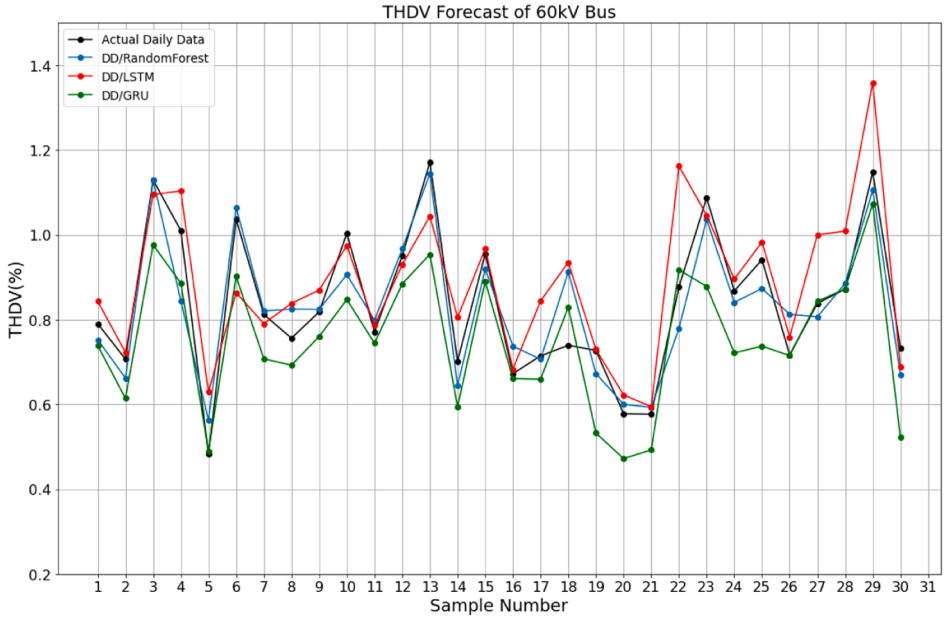
scenarios. The results confirm that RF is more robust in harmonic forecasting applications, while deep learning models still contribute to capturing time-dependent variations.

By focusing on the Bozcaada region, this research provides insights into offshore wind farm feasibility in Turkey, where offshore wind energy deployment is still in its early stages. The study underscores the importance of data augmentation in improving harmonic forecasting accuracy, contributing to enhanced power quality and grid stability.

Future work can explore different GAN architectures to further refine synthetic data generation and investigate hybrid modeling approaches to optimize prediction performance. Extending the study to real-world offshore wind farm datasets would further validate the proposed methodology. These findings contribute to the broader field of renewable energy forecasting, offering practical solutions



a



b

Fig. 11. Actual vs. predicted a) with GAN and b) with DD graphs for Bozcaada.

for offshore wind power integration.

Author agreement statement

I the undersigned declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by

all of us.

We understand that the Corresponding Author is the sole contact for the Editorial process. He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs

CRediT authorship contribution statement

Alp Karadeniz: Conceptualization, Methodology, Validation, Supervision, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] G. Van Kuik, B. Ummels, R. Hendriks, Perspectives on wind energy, 2008.
[2] C. Shan, Harmonic analysis of collection grid in offshore wind installations, 2017.
[3] PWC, unlocking Europe's offshore wind potential moving towards a subsidy free industry, PWC,Tech.Rep. MAY. (2017).
[4] Faiz J, Ebrahimpour H, Pillay P. Influence of unbalanced voltage on the steady-state performance of a three-phase squirrel-cage induction motor. IEEE Trans Energy Convers 2004;19:657–62. https://doi.org/10.1109/TEC.2004.837283.
[5] Ebrahimzadeh E, Blaabjerg F, Wang X, Bak CL. Harmonic stability and resonance analysis in large PMSG-based wind power plants. IEEE Trans Sustain Energy 2018;9:12–23. https://doi.org/10.1109/TSTE.2017.2712098.
[6] Hasan KNBM, Rauma K, Luna A, Candela JI, Rodríguez P. Harmonic compensation analysis in offshore wind power plants using hybrid filters. IEEE Trans Ind Appl 2014;50:2050–60. https://doi.org/10.1109/TIA.2013.2286216.
[7] Radhakrishnan K. Passive filter design and optimisation for harmonic mitigation in wind power plants. Institutt for Elkraftteknikk Master Thesis; 2016.
[8] Gautam D, Vittal V, Harbour T. Impact of increased penetration of DFIG-based wind turbine generators on transient and small signal stability of power systems. IEEE Trans Power Syst 2009;24:1426–34. https://doi.org/10.1109/TPWRS.2009.2021234.
[9] EU Science HUB. https://re.jrc.ec.europa.eu/pvg_tools/en/#TMY; 2024.
[10] Argin M, Yerci V, Erdogan N, Kucuksari S, Cali U. Exploring the offshore wind energy potential of Turkey based on multi-criteria site selection. Energy Strategy Rev 2019;23:33–46. https://doi.org/10.1016/j.esr.2018.12.005.
[11] Argın M, Yerci V. The assessment of offshore wind power potential of Turkey. 2015 9th International Conference on Electrical and Electronics Engineering (ELECO). Bursa, Turkey; 2015. p. 966–70. https://doi.org/10.1109/ELECO.2015.7394519.
[12] Liang P, Deng C, Yuan X, Zhang L. A deep capsule neural network with data augmentation generative adversarial networks for single and simultaneous fault diagnosis of wind turbine gearbox. ISA Trans 2023;135:462–75. https://doi.org/10.1016/j.isatra.2022.10.008.
[13] Du W, Zhu P, Pu Z, Gong X, Li C. Data augmentation on fault diagnosis of wind turbine gearboxes with an enhanced flow-based generative model. Measurement (Lond) 2024;225. https://doi.org/10.1016/j.measurement.2023.113985.
[14] Vega-Bayo M, Pérez-Aracil J, Prieto-Godino L, Salcedo-Sanz S. Improving the prediction of extreme wind speed events with generative data augmentation techniques. Renew Energy 2024;221. https://doi.org/10.1016/j.renene.2023.119769.
[15] Liu R, Song Y, Yuan C, Wang D, Xu P, Li Y. GAN-based abrupt weather data augmentation for wind turbine power day-ahead predictions. Energies (Basel) 2023;16. https://doi.org/10.3390/en16217250.
[16] Hadi FMA, Aly HH, Little T. Harmonics forecasting of wind and solar hybrid model based on deep machine learning. IEEE Access 2023;11:100438–57. https://doi.org/10.1109/ACCESS.2023.3314742.
[17] Hadi FMA, Aly HH, Little T. Harmonics forecasting of wind and solar hybrid model driven by DFIG and PMSG using ANN and ANFIS. IEEE Access 2023;11:55413–24. https://doi.org/10.1109/ACCESS.2023.3253047.
[18] Manuela Panoiu CPLG. Neuro-fuzzy modeling and prediction of current total harmonic distortion for high power nonlinear loads. In: 2018 Innovations in Intelligent Systems and Applications (INISTA); 2018.
[19] Kuyunani EM, Hasan AN, Shongwe T. Improving voltage harmonics forecasting at a wind farm using deep learning techniques. In: IEEE International Symposium on Industrial Electronics. Institute of Electrical and Electronics Engineers Inc.; 2021. https://doi.org/10.1109/ISIE45552.2021.9576357.
[20] Ganea D, Mereuta E, Rusu L. Estimation of the near futurewind power potential in the black sea. Energies (Basel) 2018;11. https://doi.org/10.3390/en11113198.
[21] Argin M, Yerci V. Offshore wind power potential of the Black Sea region in Turkey. Int J Green Energy 2017;14:811–8. https://doi.org/10.1080/15435075.2017.1331443.
[22] Koletsis I, Kotroni V, Lagouvardos K, Soukissian T. Assessment of offshore wind speed and power potential over the Mediterranean and the Black Seas under future climate changes. Renew Sustain Energy Rev 2016;60:234–45. https://doi.org/10.1016/j.rser.2016.01.080.
[23] Diaconita A, Andrei G, Rusu L. New insights into the wind energy potential of the west Black Sea area based on the North Sea wind farms model. Energy Reports 2021;7:112–8. https://doi.org/10.1016/j.egyr.2021.06.018.
[24] ABB, XLPELandCableSystems-User'sGuide, vol.Rev5 (2010).
[25] ABB, XLPE Submarine Cable Systems attachment to XLPE Land Cable Systems - user 's guide. Rev 2010;5.
[26] Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: lstm cells and network architectures. Neural Comput 2019;31:1235–70. https://doi.org/10.1162/neco_a_01199.
[27] Tran NT, Tran VH, Nguyen NB, Nguyen TK, Cheung NM. On data augmentation for GAN training. IEEE Trans Image Process 2021;30:1882–97. https://doi.org/10.1109/TIP.2021.3049346.
[28] Velarde G, Brañez P, Bueno A, Heredia R, Lopez-Ledezma M. An open source and reproducible implementation of LSTM and GRU networks for time series forecasting. In: Proceedings of the 8th International Conference on Time Series and Forecasting. 18; June 2022. p. 30. https://doi.org/10.3390/engproc2022018030 [Online]. Available:.
[29] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, https://arxiv.org/abs/1412.3555; 2014.
[30] Breiman L. Random forests. Mach Learn 2001;45(1):5–32 [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324.

[31] J. Schmidt-Hieber, Nonparametric regression using deep neural networks with ReLU activation function, (2017). https://doi.org/10.1214/19-AOS1875.
[32] Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl 2019; 134:93–101. https://doi.org/10.1016/j.eswa.2019.05.028.
[33] Tufts University, IEEE Circuits and Systems Society. In: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS); 2017. August 6-9n.d.

Alp Karadeniz received a Ph.D in Electrical & Electronics Engineering from Balikesir University, 2019. He is currently involved in research on renewable energy systems, with a focus on offshore wind farms and harmonic prediction. His interests include machine learning, deep learning applications, and power quality analysis, particularly in integrating AI-driven forecasting techniques for sustainable energy solutions.