

Data-augmented sequential deep learning for wind power forecasting

Hao Chen^{a,*}, Yngve Birkelund^b, Qixia Zhang^c

^a Department of Technology and Safety, UiT The Arctic University of Norway, Tromsø 9019, Norway

^b Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø 9019, Norway

^c Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Keywords:

Renewable energy
Wind power forecasting
Data augmentation
Deep learning
Encoder-decoder networks
Big data

ABSTRACT

Accurate wind power forecasting plays a critical role in the operation of wind parks and the dispatch of wind energy into the power grid. With excellent automatic pattern recognition and nonlinear mapping ability for big data, deep learning is increasingly employed in wind power forecasting. However, salient realities are that in-situ measured wind data are relatively expensive and inaccessible and correlation between steps is omitted in most multistep wind power forecasts. This paper is the first time that data augmentation is applied to wind power forecasting by systematically summarizing and proposing both physics-oriented and data-oriented time-series wind data augmentation approaches to considerably enlarge primary datasets, and develops deep encoder-decoder long short-term memory networks that enable sequential input and sequential output for wind power forecasting. The proposed augmentation techniques and forecasting algorithm are deployed on five turbines with diverse topographies in an Arctic wind park, and the outcomes are evaluated against benchmark models and different augmentations. The main findings reveal that on one side, the average improvement in RMSE of the proposed forecasting model over the benchmarks is 33.89%, 10.60%, 7.12%, and 4.27% before data augmentations, and increases to 40.63%, 17.67%, 11.74%, and 7.06%, respectively, after augmentations. The other side unveils that the effect of data augmentations on prediction is intricately varying, but for the proposed model with and without augmentations, all augmentation approaches boost the model outperformance from 7.87% to 13.36% in RMSE, 5.24% to 8.97% in MAE, and similarly over 12% in QR90. Finally, data-oriented augmentations, in general, are slightly better than physics-driven ones.

1. Introduction

Wind is a renewable, sustainable, and environmentally friendly energy resource. As wind technology has developed in recent years, wind energy has received attention from a growing number of countries for its low-cost operation and maintenance, small turbine footprint, flexibility in development scale, and rapidly decreasing electricity generation costs. [1]

Meanwhile, massive electricity generated by wind energy is volatile, intermittent, and with low power density. These features influence the

power production of generation companies, the balance of the grid and may profoundly jeopardize its security. [2] In a large-scale grid-connected system involving wind power, an unplanned load increase or an unscheduled wind power decrease will cause a supply-demand imbalance when thermal power or hydropower ceases generation or is insufficient. [3] Hence, the uncertainty in wind power production enlarges the required reserve capacity of the system. An accurate wind power forecast minimizes the spare capacity and enables optimal dispatch of power in systems with wind power generation. Furthermore, an effective prediction serves as a basis for wind parks to engage in

Abbreviations: \hat{P}_{t+n} , n timestep ahead predicted wind power; P_t , Measured wind power; v_t , Measured wind speed; u_{t+n} , n , timestep ahead wind speed calculated from weather model; m , Sample number of the testing set; Cap , Designed capacity of the wind turbine; T , Statistic of paired T-test; F , Statistic of paired Friedman test; BA, Bionic optimized neural networks constructed Adaboost; DA#, Physics-oriented data augmentation strategy – number #; ED, Encoder-Decoder; EDLSTM, Proposed Encoder-Decoder Long Short-Term Memory neural networks; LSTM, Long Short-Term Memory; MAE, Mean Absolute Error; MSE, Mean Square Error; NLP, Natural Language Processing; NN, Three-layer backpropagation Neural Networks; NWP, Numerical Weather Prediction; PA#, Data-oriented data augmentation strategy – strategy number #; PR, Persistence model; QR90, Qualification Rate at the 90% threshold; RMSE, Root Mean Square Error; RNN, Recurrent Neural Networks; seq2seq, Sequence-to-Sequence; STD, Standard deviation; T#, Wind turbines with different terrain – turbine number.

* Corresponding author.

E-mail address: hao.chen@uit.no (H. Chen).

<https://doi.org/10.1016/j.enconman.2021.114790>

Received 7 July 2021; Accepted 18 September 2021

Available online 1 October 2021

0196-8904/© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

generation bidding, determines a reasonable charging and discharging strategy for energy storage, and lowers the occurrence and duration of wind curtailments.

Wind power forecasting methodology is generally divided into physical, statistical, and hybrid approaches. [4] The first predicts wind power by extensive numerical computation of physical equations. It is based on fluid dynamics and uses Numerical Weather Prediction (NWP) data such as wind speed and pressure, and geoinformation like ground roughness and altitude. The method performs best in medium or long-term forecasting and applies to the wind resource assessment of new wind parks that lack historical observations. The statistical approach aims to establish linear or nonlinear patterns within wind data sequences that can be utilized in forecasting. In particular, machine learning-based wind power forecasting methods developed in recent years are widely applied. The hybrid approach is a combination of the former categories and has shown its edge profoundly. [5]

In 2006, Hinton et al. successfully trained deep neural networks (i.e., artificial neural networks with several hidden layers) and achieved excellent performance on multiple datasets, [6] which signified the birth of deep learning. Since then, deep learning techniques based on neural networks of different designs have flourished and solved long-standing challenges, such as voice and image recognition and generation, preliminary implementation of autonomous driving, etc. [7] Recently, the application of deep learning to energy science has also become popular because of its powerful auto-pattern recognition and nonlinear mapping capabilities. [8] The two major drivers of deep learning evolution are progressive computational capabilities and the influx of big data. It is generally agreed that larger datasets yield better deep learning models. [9]

The effectiveness of deep supervised learning relies on the volume and quality of labeled training data as well as the topology and parameters tuning of deep networks. [10] Notably, an effective solution to establish large sets of training data is data augmentation, since the training set typically lacks a sufficient number of manually labeled samples. Especially in wind energy, it is generally challenging to acquire high-quality and long-duration meteorological and power production data.

Data augmentation is a technique to make supervised machine learning, especially deep networks, more efficient. It extends the amount of available training data by adding modified versions of existing data or new data generated based on existing data. Technically, data augmentation imposes a sort of perturbation or noise on the datasets, both of which are viewed as unfavorable factors in signal processing and statistical modeling and need to be removed by implementing filters. [11, 12] However, the technique effect in deep learning is to regularize the model and assist in mitigating overfitting during deep training, thereby improving the generalizability and ubiquity of the learned models. Overfitting is a phenomenon that occurs as a learner learns a function with extraordinarily large variance, such as perfectly fitting the training data. Generalizability defines the difference in performance when a model is assessed in relation to data in the training set previously seen compared to previously unseen data in the testing set. [13]

Essentially, using multi-inputs to make multistep wind power forecasting can be regarded as a Sequence-to-Sequence (seq2seq) prediction that is framed as a mapping of multiple inputs to multiple time-series outputs. It was demonstrated that the seq2seq model “*approaches or surpasses all currently published results*” [14] in Natural Language Processing (NLP), like Google Translate, and recently it has also shown its promise in renewable energy forecasting. [15, 16] The Encoder-Decoder (ED) Recurrent Neural Networks (RNN) has successfully handled seq2seq problems [17] and exhibits state-of-the-art performance in the area of text translation that is fundamentally a time-series problem.

1.1. Previous work review

In computer science research, there are several developed

methodologies in data augmentation. [18] Shorten and Khoshgoftaar [13] systematically presented current imagery data augmentation methods, their promising advances, and methodologies used to implement them to boost the performance of imagining deep learning tasks. Cubuk et al. [19] investigated several commonly used image recognition datasets and designed an augmentation strategy that learns from the datasets. The strategy consists of many sub-strategies and is automatically selected in the model training process and helps gain 0.4% to 0.6% imagine classification accuracy on different datasets. But the data augmentation technique is mainly in the field of image recognition and has received little attention to transfer the technique to the sequence domain. However, both image and sequence deep learning tasks intrinsically focus on automatically exploiting data features while avoiding overfitting. So, researchers should concentrate more on data augmentation applied to sequential deep learning. DeVries and Taylor [20] summarized and utilized interpolation and extrapolation, etc., and domain-agnostic approach to reach the predictions with deep learning for time-series datasets, and tentatively proved the techniques are timely and effective in some supervised learning problems. Park et al. [21] presented a speech recognition augmentation approach named Spe-Augment consisting of masking features, frequency channels, and time steps to reach leading capabilities on two speech recognition mission sets.

Deep learning techniques have gotten much attention from researchers in renewable energy forecasting. [8] With its distinctive automatic nonlinear recognition capabilities, deep learning has gradually emerged as an important approach to the challenge of forecasting sharply volatile wind power. [5, 22] Yildiz et al. [23] extracted wind datasets with features with variational mode decomposition and converted these features into images. Then the images were handled by an improved residual-based deep convolutional neural network to forecast wind power for a wind park in Turkey. The edge of the proposed process was proved by a comparison between some existing well-used large networks. Kisvari et al. [24] constructed a framework consisting of data preprocessing, anomaly detection, feature engineering, and gated recurrent deep learning models for wind power prediction and demonstrated that the framework offered more effective predictions than traditional recurrent neural networks. Shahid et al. [25] piled up Long Short-Term Memory (LSTM) units into a large network and tunes the network by using the genetic algorithm to forecast wind power validated the statistical advantage of the network over a single unit by the Wilcoxon Signed-Rank test. Memarzadeh et al. [26] applied a bionic algorithm, wavelet transform, feature selection, and LSTM networks to forecast wind power of two wind parks in Spain and Iran, and showed the effectiveness of the proposed method by comparison with benchmark neural networks.

While numerous wind power models based on a hybrid of traditional data methodologies and deep learning have been developed and advanced in forecasting for many sites, nevertheless, further sophistication of forecasting models may render the results specific, i.e., wind power forecasts are restricted to a certain category of terrain and weather features and difficult to be generalized and not be easily employed because their consisting techniques such as signal processing, feature engineering, etc. require a prolonged and special training to master. Lipu et al. [27] also summarized the most recent progress of wind power forecasting using artificial intelligence and pointed out the issues and challenges in the field. The challenges include many various data preprocessing techniques for diverse wind data, model structure, and optimization, etc. In particular, Reichstein et al. [28] recommended that more attention should be given to Earth system science problems to the coupled data approaches with physical phenomena and deep learning methods themselves, rather than building more complex traditional methods-based models.

In the present study, in the contrast, we return to the physical process of wind power generation, the statistical characteristics of wind data, and the nature of deep learning to approach the forecasting problem.

After synthesizing numerous data augmentation methodologies and drawing on multiple state-of-the-art advances in sequential data prediction, the robust and efficacious encoder-decoder deep neural networks with stacking LSTM units are proposed for wind turbine power forecasting in the Arctic.

1.2. Contributions

Leveraging the aforementioned literature review, attention is paid to a wind park, inside the Arctic, in complex terrain. The principal contributions of the present study paper are as follows:

1. This paper systematically applies data augmentation to wind power forecasting for the first time. Specifically, eight time-series data augmentation approaches are proposed according to physical characteristics of wind energy and statistical properties of data in wind engineering. The approaches are implemented in four benchmarking models and proposed advanced deep learning models. The methodology is particularly suitable for new wind parks that have a short period of operation and therefore a limited amount of accumulated data. It enables to fully and automatically deepen the information and value of these limited data.
2. We exhaustively develop a seq2seq deep learning predictive end-to-end model with inputs of historical wind speed and power data and wind speed from NWP as well as simultaneously interrelated outputs of multistep, futuristic wind power. The model is based on an encoder-decoder constructed with LSTM and shows its superiority in forecasting power.
3. It is demonstrated that the impact of various augmentation approaches is different in each forecasting algorithm. Augmentations somewhat increase linear, like persistence, model errors. Nonetheless, augmentations improve the performance, most notably the proposed deep learning model, of neural networks-based algorithms, where data-oriented augmentations generally contribute greater than physics-oriented ones.
4. The data augmentations combined with the proposed and benchmark forecasting models are utilized to predict power generated by five turbines in various landscapes. The results are analyzed by rigorous statistical methods and indicate that the augmentations and the proposed forecasting model have wind engineering values and potentially extensive applicability in other energy sectors.

The architecture avenue opens the article with an introduction on wind energy forecasting and its deep learning utilization status quo as well as contributions presented in Section 1. Section 2 illustrates the principle of wind power generation and the utilized data and scheme. Section 3 delves into proposed data augmentation techniques and a novel predictive deep neural network. Section 4 provides detailed experiment procedures and model assessment metrics. In Section 5, hierarchical experimental results and discussions, from comparisons of models themselves to data augmentation approaches, are presented. Finally, the main findings, research outlooks, and derivative policy recommendations are demonstrated in Section 6.

2. Data preparation and forecast scheme

Wind power generation is a conversion from wind energy to electricity. Ideally, the output generation of a wind turbine is expressed as in (1):

$$P = \begin{cases} 0 & v < v_{min} \\ P_v(C_p, \rho, A; v) & v_{min} < v < v_n \\ P_r & v_n < v < v_{max} \\ 0 & v > v_{max} \end{cases} \quad (1)$$

where P is the output power of the wind turbine (W); $P_v(\cdot)$, typically

proportional to the cubic of the wind speed, is the wind curve function at the speed interval, C_p means wind energy utilization efficiency; ρ is the air density (kg/m^3); A is the effective area swept by turbine blades (m^2), v denotes the wind speed (m/s); v_{min} , v_{max} , and v_n respectively are cut-in, cut-off, and rated wind speed. P_r is turbine rated wind power. From (1), the output of a wind turbine is mainly influenced by the third power of wind speed, air density, and swept area.

The study centers on the wind turbine, 3.0 MW Vestas V90, electricity production of a wind park, Fakken, with an installed capacity of 54 MW with 18 turbines, average annual production is 139 GWh in the Arctic region. Wind is predominantly influenced by the terrain; wind anomalies occur when wind moves through these areas. The influence is dependent on the height and width of the barriers. The terrain of Fakken wind park is with low and flat hills and narrow valleys, and towards a fjord.

The timescale of data in this study is from 0:00 1st January 2017 to 23:50 31st December 2017. Raw wind speed and power data of each turbine, 10 mins temporal resolution and recorded by Supervisory Control And Data Acquisition SCADA, are supplied by a local wind energy operator. The NWP wind speed data, calculated by the Meteorological cooperation on operational Ensemble Prediction System (MEPS) NWP model, are with 2.5 km horizontal resolution that is taken as the mesoscale. The model, operating by the Norwegian Meteorological Institute, updates at 00, 06, 12, and 18 UTC, and its forecasts for the next 66 h are available around 1 h 15 min later. The wind speed data sequences from NWP comprise the nearest accessible weather prediction data.

To verify the generality and portability of the proposed methodology, five wind turbines separately situated in different topographic conditions in the wind park are selected as study subjects. Moreover, wind measurements are taken at the turbine nacelle, which is 80 m about the ground. Their topographic features and statistics of annual in-situ measured wind speed and power are shown in Table 1.

Statistically, wind power forecasting can be regarded as a multi-variable regression problem, in which wind power time series is autoregressed, and wind speed serves as the supplementing information to the autoregression. Updating the wind speed from NWP of the predicted time, the current information, is also the key feature in the prediction since according to an extensively cited reference by Giebel and Kariniotakis [29], forecasting wind power beyond three to six hours typically requires consideration of information on NWP wind speed at the moment of prediction. In this study, we chose measured data of the previous six hours to make multistep forecasts for the wind power from the next six to twelve hours with the assistance of wind speed from NWP.

The fundamental multistep forecasting model $f(\cdot)$ with timestep $i + n$ is described as:

$$\hat{P}_{i+n} = f(P_{i-j}; v_{i-j}; u_{i+n}) + \varepsilon_n \quad (2)$$

where i represents the base current time $i = 1, 2, \dots, 7$, and with each $i, j = 0, 1, \dots, 6$. \hat{P}_{i+n} is n timestep ahead predicted wind power, $n \in \{6, 7, 8, 9, 10, 11, 12\}$, v is the wind speed observed in the turbine, u represents the wind speed calculated from the mesoscale NWP wind model for the site. ε_n is the error of the forecasting model.

Since the scopes of wind power and speed are not the same, it is beneficial to rescale the raw data into a new set with a similar scale. Data standardization is rescaling variables with a mean of zero and a standard deviation (STD) of one. The technique can accelerate convergence speed and improve algorithms' accuracy of neural networks. [30]

3. Methodology

3.1. Wind data augmentation

In practice, testing errors need to be continuously reduced along with training errors to construct meaningful deep learning models. Data

Table 1
The terrain and statistics of wind turbines.

Wind Turbine	Terrain	Wind power				Wind speed			
		Mean [kW]	STD [kW]	Skew	Kur	Mean [m/s]	STD [m/s]	Skew	Kur
T1	Plateau	825.58	990.43	1.01	-0.35	3.98	5.15	1.18	1.31
T2	Valley	826.19	987.92	0.95	-0.43	3.91	5.06	1.10	1.12
T3	Lakeside	738.37	914.33	1.23	0.28	3.55	4.65	1.33	2.01
T4	Hilltop	804.40	971.86	1.04	-0.25	4.02	5.27	1.27	1.71
T5	Seaside	783.86	950.42	1.06	-0.14	3.93	5.10	1.19	1.49

Note: STD is standard deviation, Skew is skewness and Kur is relative kurtosis (actual kurtosis minus 3).

augmentation is a phenomenally robust approach to accomplish this aim. It embarks on overfitting from the origin, the training data themselves, of the problem, assuming that further information can be retrieved from the source dataset.

Based on know-how in wind energy technology and state-of-the-art data science, we divide the techniques for augmenting wind data for forecasting with robust and efficient deep learning into two categories: physics-oriented and data-oriented.

3.1.1. Physics-oriented approaches

Inspired by the physics of wind power engineering, we propose three strategies to augment training set data for forecasting models. The first is the explicit perturbation of the wind power curve according to Eq. (1). The second is the implicit perturbation based on the difference between the numerical weather predicted wind speed of the wind park area and the actual measured wind speed of turbines. The third considers the operational data of the other wind turbines in the vicinity of the studied wind turbines. These three physics-oriented approaches are shortened as PA1, PA2, and PA3, respectively.

PA1: Considering the wind speed as the independent variable and differentiating Eq. (1), the following Eq. (3) is obtained.

$$dP = \begin{cases} 0 & v < v_{min} \\ (P'_v(C_p, \rho, A; v)/P)dv & v_{min} < v < v_n \\ 0 & v_n < v < v_{max} \\ 0 & v > v_{max} \end{cases} \quad (3)$$

from Eq. (1), it is observed that when v is in the cut-in and rated wind speed interval, the derivative of the power curve, the ratio of tiny variations in wind turbine power and wind speed, is proportional to the quadratic of this point wind speed. Therefore, according to Eq. (3), it is possible to artificially adhere a slight random perturbation in a wind speed point in the interval and calculate the corresponding power variation in accordance with the speed.

PA2: According to Eq. (2), the input to the power forecasting model contains the wind speed from measurements and the NWP model, but they correspond to different time stamps when entering the model. Since NWP datasets also have wind speeds that correspond to the same time stamps as the measured wind speeds, and there is no significant difference in wind speed probability distribution from two wind speed resources in the wind park based on our previous study. [31] So, we resort to a random replacement strategy with a fixed probability to replace the wind speeds in the measured datasets with the correspondent NWP wind speeds.

PA3: Since the neighboring turbines to the target turbine have similar wind conditions in operation. Therefore, adopting the measured wind speed of the neighboring turbine with a specific probability to replace the target turbine could be a strategy to augment the target wind speed dataset.

3.1.2. Data-oriented approaches

The proposed taxonomy for the data-oriented methods for wind power forecasting is enlightened by the feature space expansion, signal

processing, and machine learning techniques. It consists of five approaches. DA1: Various simple interpolation and extrapolation methods are used to obtain data on larger time scales. DA2: Implements noise to the original dataset. DA3: Sequential augmentation approaches, named geometric transformations, draw on image processing, symmetry or flipping, translation, and random erasing. DA4: Methodology used for decomposition in time-series data. DA5: Scenario generation methods for the single turbine include statistical and machine learning generation.

DA1: Averaging is usually required to calculate the data in hourly units as the original measured dataset is in ten-minute increments. The new hourly data can be acquired by performing some interpolation or extrapolation modification to this averaging process. The new averaging is defined as:

$$x'_t = \sum_{j=1}^6 \omega_j x_j \quad (4)$$

where x'_t is the hourly data and x_j donates the raw 10-mins data. ω_j is the stochastic weight that fulfills: $\sum_{j=1}^6 \omega_j = 6$, $(-0.3 \leq \omega_j \leq 1.3)$, which when $\omega_j < 0$ is extrapolation while $\omega_j \geq 0$ means interpolation.

DA2: Another simple, probably the simplest, method of data augmentation is the addition of white noise, following the standard normal distribution, to data. A wind power forecasting study considered noise in data as a detrimental factor for prediction and removed it by signal processing. [32] Nonetheless, in machine learning research, applying noise to the neural network's inputs increases the generalizability of the networks. [18] The noise injection is determined with a scaling parameter δ :

$$x'_t = x_t + \delta X, \quad X \sim \mathcal{N}(0, \sigma_t) \quad (5)$$

where x'_t is the enhanced data and x_t donates the original hourly data.

DA3: Geometric transformations are among the initial data augmentation methods with excellent effectiveness in deep learning for image recognition, such as flipping, cropping, and color transformations. [13] Based on the characteristics of the measured wind speed time series and referring to image geometric augmentations, we stochastically opt for, 10% respectively, symmetry along with the average point, substitution of prior or posterior values, and stochastic erasing of some data.

DA4: Wind power forecasting is known mathematically as a special time-series problem. Ordinarily, the time series x_t can be decomposed into base α_t , trend τ_t , season s_t , and residual γ_t part as in Eq. (6).

$$x_t = \alpha_t + \tau_t + s_t + \gamma_t, \quad t = 1, 2, \dots, N \quad (6)$$

The extensively implemented approach is firstly based on the time-domain figure of the time series or its Fourier analysis to obtain its period corresponding to seasonality, and then decomposes the time series with the loess smoothing technique, [33] a locally weighted autoregression, into the above four components. The weights of these four components are subsequently and stochastically adjusted by Eq. (7) to form an augmented series.

$$x'_t = \omega_1 \alpha_t + \omega_2 \tau_t + \omega_3 s_t + \omega_4 \gamma_t, \quad \sum_{i=1}^4 \omega_i = 4, \quad 0.9 \leq \omega_i \leq 1.1 \quad (7)$$

DA5: The data augmentation methodologies described above all involve randomness, data selections, and/or weight adjustments, so they are relatively independent of the data and require considerable manual fine-tuning. Wind power scenario generation is an effective tool to resolve uncertainties in stochastic planning of the energy system with the integration of wind power. [34] Classical and advanced statistical methods and machine learning models are broadly employed [35] to predict wind power scenarios. Intrinsically, these models profile conditional distributions of time series by assuming that the current value depends on previous points: a new time series may be generated from the learned conditional distributions provided that original series values are perturbed in some way.

3.2. Encoder-decoder LSTM deep networks

RNN has achieved tremendous success and wide application in numerous sequence applications. [18] RNN is designed to process learning tasks with sequential data. ‘Recurrent’ means the current output is related to the previous output. The nodes in hidden are structurally connected to each other to reach inputs of the hidden layers includes not only outputs of the input layer but also ones of the previous-time hidden layers.

Among the RNN network structures, the most extensively used and highly successful model is the LSTM network, with a kind of unique memory unit in its hidden layers and is generally more expressive of long-short time dependencies than the other RNNs. [36] Typically, the LSTM unit consists of three gates, i.e., input gate, forget gate, and output gate. There are three primary internal phases of the unit. The first is forget phase, which retains the important information coming in from the previous node and forgets the unimportant details. The next phase is the selective memory phase, which optionally remembers inputs of this phase. Finally, an output phase determines which ones should be treated as outputs of the current state. Mathematically, the long-short memory unit can be expressed as [37]:

$$\begin{aligned} i_t &:= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \\ f_t &:= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \\ o_t &:= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \\ c_t &:= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \\ c_t &:= f_t \odot c_{t-1} + i_t \odot c_t, \\ h_t &:= o_t \odot \tanh(c_t). \end{aligned} \quad (8)$$

where x_t is the input and h_{t-1} is the hidden state of the previous time-step. i_t , f_t , and o_t are input, forget, and output gates, W denotes the corresponding weight parameter, and b is the corresponding bias parameter. c_t is the candidate memory cell, c_t is the memory cell, and c_{t-1} is its previous time step state. h_t is the hidden state. $\sigma(\cdot)$ is the sigmoid function, $\tanh(\cdot)$ is hyperbolic tangent function, and \odot represents the pointwise multiplication.

The encoder-decoder LSTM is a type of EDLSTM network designed to deal with seq2seq, and its architecture is innovative in terms of sequence embedding, i.e., the usage of a reading-in and exporting-out fixed-size sequences. The encoder-decoder LSTM includes an input layer, LSTM based encoder and decoder, and an output layer in this study. The LSTM unit achieves the extraction and utilization of important information in the sequence through its gate controls. The encoder reads input sequences and encodes them into fixed-length vectors by the weight of each time step with a context vector. The decoder decodes these fixed-length vectors and outputs predicted sequences. The fixed-length context vector introduces a mechanism called Attention, which enables highly summarize and highlight the information learned by the encoder and uses it as input to the decoder for translation. The encoder and decoder networks are mutually independent, which indicates that their LSTM units do not share parameters during the process of networks training.

3.3. Proposed deep EDLSTM for wind power forecasting

According to Eq. (2), wind power prediction involves autoregression, multiple sources of wind speed, and nonlinear functional relationships, all of which may lead to the application of EDLSTM networks. In addition, multistep wind power forecasting is appropriate to be handled as a seq2seq problem since the historical data of the inputs are linked and interactive. Therefore, a deep, stacked multilayers EDLSTM, shorten as EDLSTM, is proposed and utilized to extract the implicit features from layer to layer. The detailed deep EDLSTM employed in this article is illustrated in Fig. 1.

First, the encoder consists of a stack of three-layer LSTMs, which sequentially extracts complex time-dependent features of inputting measured and meteorological data deeply layer by layer with transferring hidden states h . And then generate a fixed-length context vector containing the extracted characteristic information. The structure and transmission of information for the decoder are basically identical to those for the encoder. Then, the context vector serves as the initial input to the decoder. Regardless of the updating from the encoder of the context vector, the vector is sent to the first layer of the decoder as its input, and its output is used as the input of the second layer. Sequentially, the third layer output is transformed through the output layer and cyclically fed back to the first layer as its next input. Eventually, the decoder generates a time series of the predicted wind power.

4. Experiments

4.1. Experimental scheme

The scheme of forecasting individual turbine wind power by employing EDLSTM with data augmentation is animatedly illustrated in Fig. 2. Firstly, the measured wind speed and power with the ten-minute resolution are averagely interpolated into, except for the DA1 augmentation measure, data with hourly resolution. All hourly data are segmented into training and testing sets, accounting for 65% and 35%, respectively. Secondly, the measured wind speed and/or wind power data in the training set are separately augmented with the approaches proposed in Section 3.1 to enlarge the data amount to five times the original training set size. i. e., the new data with the four times larger size of the original training set are generated with augmentations. Thirdly, the unexpanded and expanded training sets are individually fed into the benchmark models, i.e., Persistence (PR), simple three-layer backpropagation Neural Networks (NN), basic LSTM RNN (LSTM), Bi-omic optimized neural networks constructed Adaboost (BA) ensemble leaning (regarded as a popular and advanced hybrid forecasting model have been proven to perform well and have been extensively studied [39, 40, 41, 42], namely, ensemble learning perdition models) and the proposed deep EDLSTM network to conduct training and obtain multiple learned models. The benchmark models have been introduced in Ref. [41, 43, 44] and their parameters are briefly summarized in Table 2. Finally, the testing set data are imported into the trained models to yield the multistep predicted wind power and to assess and compare the forecasting models’ performance.

4.2. Data augmentation program

Our data augmentation strategy fine-tunes the data without altering the temporal order of the original data and ensures that the augmented training data and the previous ones maintain statistical consistency. This study augments the training samples and scales up their number to five times the original sample size. The data augmentation techniques explained above, apart from DA5, all involve stochastic perturbation of the original data. Our method is to gradually enlarge the perturbation amplitude and accordingly generate new data four times. For the DA5 method, four new datasets are produced by individually operating autoregressive models based on four machine learning models. Details

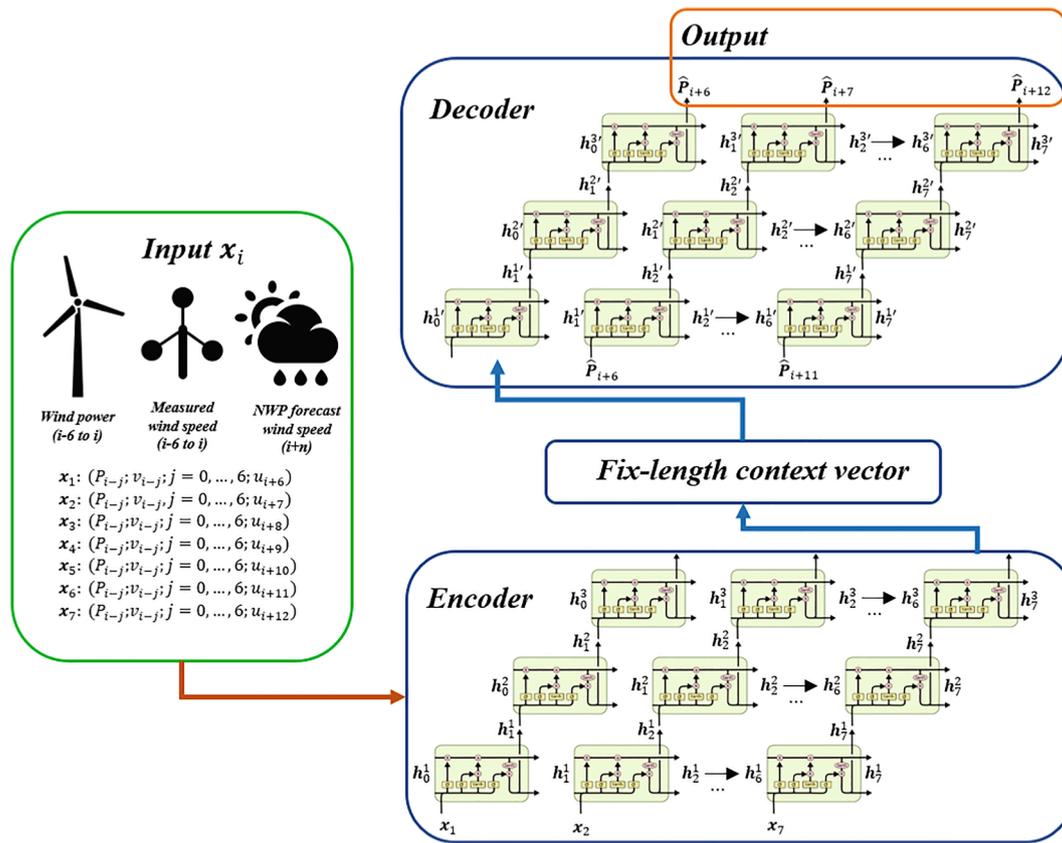


Fig. 1. The structural diagram of the proposed deep EDLSTM for wind power forecasting (The LSTM unit graph in is cited from Ref. [38]), (The three-layer stacking structure of the LSTM of the encoder and decoder is designed to mine the point information and the dependence information of sequences through the state transfer twice between the three layers.).

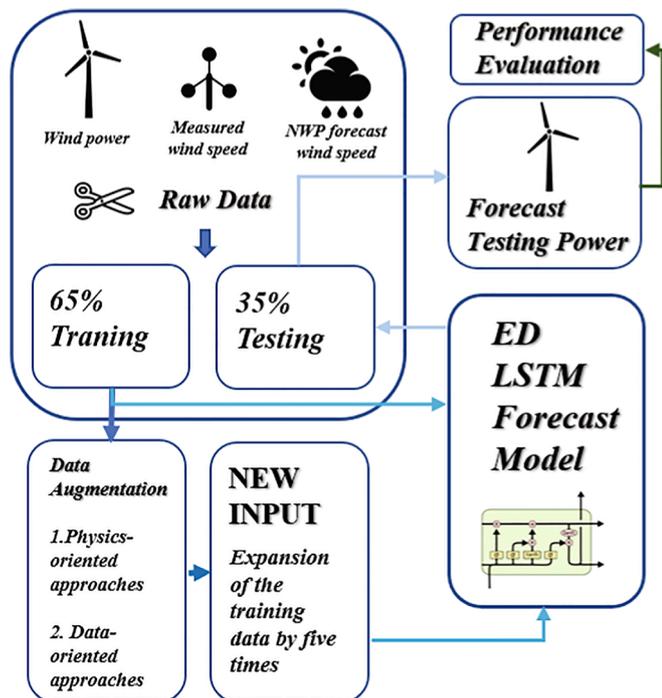


Fig. 2. The main procedure of the data augmentation based EDLSTM for predicting wind power.

Table 2

A summary of forecasting models' parameters.

Forecasting model	Main parameters
PR	The predicted value for the next moment is the current moment's value.
NN	The input, hidden, and output layers are with 15, 30, and 7 neurons, respectively; sigmoid activation function, and MSE loss function. (The number of hidden layer neurons is determined based on a grid search with density of 5 from 10 to 100.)
LSTM	One fully connected dense NN layer, Seven LSTM units (TensorFlow optimized default settings for regression problems), and one dense NN with 7 neurons as output layer; sigmoid activation function, MSE loss function, and Adam algorithm optimizer. (Dense NN layers are the same with NN model.)
BA	As the performance of a neural network is intimately linked to neuron number in the hidden layer, the genetic algorithm, [45] a bionic algorithm, is applied in training iterations to automatically search for the adaptive neuron number and constitute optimized neural networks as Adaboost's base learners. The node number searching interval is set as [10,100] and the max iteration is 50. The Adaboost emphasizes (with bigger weights) data mislearned in the previous base learner to establish an ensemble model that boosts the performance of single base learners. The number of base learners is 10 and Adaboost max iteration is 20. (BA model behaves like a deep learning model called the residual deep network, [46] which is similar to seq2seq structure in forward propagation to achieve the integration of input and output for effectively mining the features; while in backpropagation some gradients are fed directly to output, avoiding gradient vanishing.)
EDLSTM	As described in Section 3.3 and Fig. 1. (LSTM unit is TensorFlow optimized default settings for regression problems.)

Table 3

A detailed description of each data augmentation process.

Physics-oriented	PA1	The Vestas V90 3 MW wind turbine corresponds to a cut-in and rated wind speed of 4 and 15 m/s, respectively, according to its power curve. Select the measured wind speed v_i in the corresponding interval; $v_i' = v_i + X$, $X \sim U[-0.1n, 0.1n]$, $n = 1, 2, 3, 4$, where U represents the uniform distribution. Then the power variation corresponding to the wind speed variation is calculated by Eq. (3), and new power data are generated accordingly.
	PA2	The measured wind speeds are randomly substituted with 50 % probability four times with NWP wind speed data with the same timestamps, and the wind power data are added a white noise following $N(0,0.1)$.
	PA3	We select measured wind speeds of the two closest turbines to the target turbine and randomly substitute, with a probability of 15% for each and a total of 30%, the target wind speed dataset. The power data are with the same treatment in PA2.
Data-oriented	DA1	As described in DA1 introduction in Section 3.1.2.
	DA2	Two normally distributed noises, $N(0,0.1n)$ and $N(0,0.02n)$, are separately loaded into the measured wind speed and power data four times, where $n = 1, 2, 3, 4$.
	DA3	As described in DA3 introduction in Section 3.1.2.
	DA4	As described in DA4 introduction in Section 3.1.2.
	DA5	Four learning algorithms to augment measured wind data, such as; $x_i' = f_i(x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}, x_{t-6})$, $i = 1, 2, 3, 4$, where x_i' is the generating data, $f_1(\cdot)$ represents a single step ahead forecasting model established by learning algorithms. $f_1(\cdot)$ is linear regression, $f_2(\cdot)$ is support vector regression, $f_3(\cdot)$ is classification and regression tree, and $f_4(\cdot)$ is simple three-layer neural networks with 15 hidden neurons regression models, respectively. All four are well-established and widespread machine learning algorithms, and a detailed description of them can be found in Ref. [43] for space constraints.

Note: The units of wind speed and power in the table are m/s and MW, respectively.

of the various data augmentation approaches are shown in Table 3.

4.3. Performance evaluation

Collectively, data-driven wind power forecasting is inherently a matter of using advanced neural networks for regression in which Mean Square Error (MSE) serves as the loss function. So, Root Mean Square Error (RMSE) is naturally selected as the metric to measure the performance of the models. The metric is negative-oriented to the modeling performance, which means a smaller value corresponds to better performance.

$$RMSE = \sqrt{\frac{\sum_{i=1}^m (P_i - \hat{P}_i)^2}{m}} \quad (9)$$

where P_i and \hat{P}_i are normalized measured and corresponding predicted wind power, m is the sample number of the testing set.

Nevertheless, the RMSE is with a disproportionately big effect of larger errors and, sometimes, is close when comparing some different forecasting models. Therefore, in these cases, Mean Absolute Error (MAE) and Qualification Rate (QR) [47] indices are introduced as below to comprehensively assess the performance of models. MAE uniformly examines the forecasting errors while the QR emphasizes the smaller errors.

$$MAE = \frac{\sum_{i=1}^m |P_i - \hat{P}_i|}{m} \quad (10)$$

$$QR = \frac{1}{m} \sum_{i=1}^m \begin{cases} 1, & \left(1 - \frac{|P_i - \hat{P}_i|}{Cap}\right) \geq Q \\ 0, & \left(1 - \frac{|P_i - \hat{P}_i|}{Cap}\right) < Q \end{cases} \quad (11)$$

where Cap is the designed capacity of the turbine. Q is the quantile percentage for qualified predictions, chosen as 90% in this study.

Two statistical tests are employed to check whether there are statistically significant differences exist in the performance of forecasting models. And both of their confidence values are set as 0.05. The first is paired T-test for the two comparisons. The null hypothesis H_0 : The averages of these samples are equivalent; H_a : The averages are not equivalent. And its test statistic T is:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{STD(\bar{Y}_1 - \bar{Y}_2)} \sim t_{2l-2} \quad (12)$$

where \bar{Y} is the average and l is the number of samples.

The second is the Friedman test, for multiple comparisons, is harnessed to examine across multiple trials and checks column effects after statistically eliminating potential row effects. [48]

H_0 : The column data do not have a significant difference.

H_a : They have a significant difference.

The statistic F is given as:

$$F = \frac{12l}{k(k+1)} \left[\sum_{i=1}^k r_i^2 - \frac{k(k+1)^2}{4} \right] \quad (13)$$

where k is the number of columns. r_i is the average value of row i , which follows $\chi^2_{(k-1)}$ distribution under H_0 .

5. Results and discussion

This section reveals the experimental results at three levels, firstly, the superiority of the proposed forecasting model is verified by analyzing different models' performance on the original dataset. Secondly, the overall effects of data augmentations on different forecasting algorithms are illustrated by the comparison of their performance before and after data augmentations. And finally, the impacts of various augmentation approaches on the proposed model's forecasting effectiveness are statistically explored.

5.1. Benchmarks and proposed deep EDLSTM model forecasting outcomes

The standardized measured and NWP wind data of chosen five wind turbines are respectively loaded into the four benchmarks and proposed deep EDLSTM models to make six to twelve hours ahead of wind power forecasts. The RMSE is displayed in Fig. 3. In general, the RMSE of all forecasting models grows as increasing prediction steps. The PR grows faster compared to the other models. The proposed deep EDLSTM outperforms best among all models for multistep power prediction for all wind turbines in almost all cases. The RMSE of the NN, LSTM, BA, and EDLSTM all constructed on neural networks is noticeably smaller than the one of PR, suggesting that neural networks can reflect the nonlinear characteristics of wind power. Moreover, these characteristics are better retained by the forecasting models as the networks are deeper and more tailored. On the overall average, the benchmarking PR, NN, LSTM, and BA models have RMSE that is 51.46%, 11.89%, 7.67%, and 4.46% higher than EDLSTM. This demonstrates that the proposed model can efficiently and accurately predict the power generated by the five wind

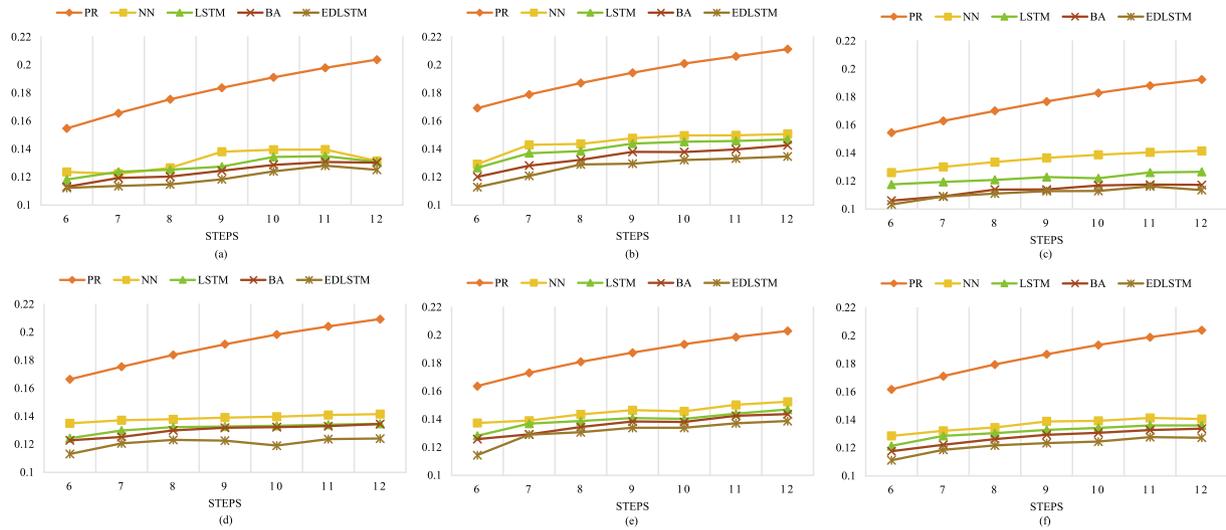


Fig. 3. The multistep performance of benchmarking and deep EDLSTM forecasting models for each turbine: (a) 1. Plateau, (b) 2. Valley, (c) 3. Lakeside, (d) 4. Hilltop, (e) 5. Seaside, (f) Average.

turbines under attention. Besides, EDLSTM’s RMSE maintains relative stability with the increasing step, indicating that the seq2seq with multiple inputs and multiple outputs reduces the cumulative error in multistep forecasting. Reasonably, the forecasting algorithms outcome relatively low RMSE of the wind turbines situated on plateau and lakeside, both of which are regarded as flat terrains. In contrast, the unique fjord topography on the Norwegian coast causes wind turbines located on hilltops, valleys, and seashores to be challenging, but handled properly by EDLSTM, to predict their electricity generation. Therefore, the proposed model allows for effective and robust power predictions of wind turbines on several different topographical conditions.

5.2. Holistic validity of data augmentations

Aiming to investigate the applicability of data augmentation in wind power prediction, the original measured data are enlarged following the eight augmentation approaches presented in Section 3.1 and are predicted by the four benchmarks and the proposed EDLSTM models. The RMSE for the six to twelve-step forecasts by the forecasting algorithms based on the eight data-augmented sets is averaged separately. The results are compared to the RMSE equally averaged of the models without augmentations. Fig. 4 shows the comparison, and Table 4 offers their performance difference with paired T-test.

As can be seen, the average effect of data augmentation is tightly linked to forecasting algorithms. The RMSE of PR with data augmentation is the same as the previous one for all wind turbines in focus. The

Table 4

The T test for average performance difference between without and with data augmentations.

Paired T	PR	NN	LSTM	BA	EDLSTM
Mean	0	0.003876	0.005813	0.004610	0.010291
p-values	/	0.002219	0.000083	0.000173	0.000001

reason is there is no learning process in the PR method and its RMSE remains the same when the used data augmentations give stochastic perturbations in data or generations of new data based on patterns of primitive data. So, it is meaningless to further discuss the augmentation in the PR approach. Within one STD, there is an apparent difference, with p-values smaller than 0.05, between RMSE of all network-based NN, LSTM, BA, and EDLSTM forecasting algorithms. It can be interpreted that these algorithms can not only respectively learn the dominant or trending patterns in the input space, but data augmentations also provide additional valuable information in these network-based models training phrases.

Most notably, a significant improvement, with a statistical average difference over 0.0102, in the performance of the EDLSTM forecasting algorithm is evident with augmented input data. On the one hand, it means that the limited original data restrict the proposed deep learning model’s potential or possibly cause overfitting. On the other hand, it demonstrates that the augmented data more adequately train the complex deep networks to yield better predictions by insight into more

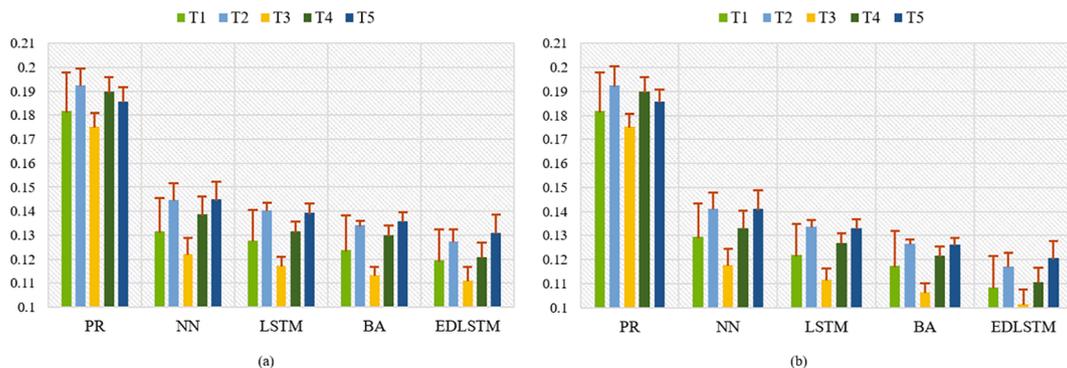


Fig. 4. The overall average RMSE of multistep forecasting models without and with data augmentations for each turbine: (a) Without augmentations, (b) With augmentations.

hidden and sophisticated patterns in the forecasting. In addition, the STD of RMSE between multiple predictions shows no significant variation before and after data augmentations, which points out that the effects of data augmentations are approximate for each step. Generally, the average RMSE of augmented models of NN, LSTM, and BA separately grows by 21.47%, 13.30%, and 7.60% compared with augmented EDLSTM.

To more explicitly show outcomes of the various data-augmented models, the RMSE of each step prediction based on the eight augmentation approaches is averaged and plotted in Fig. 5.

By comparing Fig. 5 with Fig. 3, it can be found that: first, the tendency of gradually increasing RMSE persists of data-augmented multistep predictions. Secondly, the augmented EDLSTM model outperforms its counterpart based on raw data in almost every step of prediction for all wind turbines. And thirdly, the power prediction of T3 wind turbine is the best, corresponding to the RMSE of the data augmented EDLSTM model is barely less than 0.11, and the second-best one is T1. Furthermore, the predictions for T2, T4, and T5, located in complex terrain, are also significantly improved. Thus, data augmentation improves EDLSTM for power forecasting, resulting in satisfactory reductions in model RMSE errors.

5.3. Competition between diverse data augmentation methodologies

The superiority of data augmentation approaches as a whole in wind power prediction is elaborated in Section 5.2. To further investigate which data augmentation approaches are more effective, the average and STD of RMSE for each step of prediction by algorithms based on different augmentation approaches are taken and presented in Fig. 6. As can be seen, there is no obvious regularity in the average multistep forecasting performance with different augmentation-based models. That is, the results of various augmentation approaches in different forecasting algorithms are not tendentious. The overall RMSE of distinct augmentations is comparable in NN, LSTM, and BA but the opposite is the view in EDLSTM. Nevertheless, certain patterns exist for augmentations in the prediction of different turbines. Regardless of what augmentations, the errors in predictions for turbines in flatter terrain are smaller, consistent with the predictions without augmentations.

As a further statistical examination to test the variation in different data augmentation in multistep predictions, the Friedman test to answer whether there is a difference between the RMSE averages of the five wind turbines with different augmentations in the same time step. The p -

values are demonstrated in Table 5. Among the power forecasts based on data augmentations for all turbines, The effect of different augmentation approaches for forecasting models is not statistically significant in most cases, such as in NN, LSTM, and most cases of BA. Particularly, the proposed EDLSTM models' RMSE, with a relatively complex p -value set, differs only in sixth and seventh step forecasts with varying augmentations. Additionally, in view of the EDLSTM's favorable outperformance in wind power forecasting, the decrease rate of average multistep RMSE for each augmented versus unaugmented model based on the same forecasting algorithm is computed. The rate is averaged among five turbines and illustrated in Fig. 7. The p -value for the multivariate comparison between these RMSE decrease rates is 0.00033, much less than 0.05, indicating that overall improvements in EDLSTM performance with various augmentations are statistically different. In general, based on RMSE, PA3, PA2, DA1, and PA1 provide modest improvements, from 7.87% to 9.96%, to the EDLSTM model, while DA5, DA4, DA3, and DA2 improve, sequentially from 10.80% to 11.36%, the model relatively substantially.

Despite the varying decrease degrees in RMSE for the EDLSTM models with different augmentation approaches, the difference is minimal between some approaches, like DA4 and DA5. To further compare the effects of different augmentations, the average MAE and QR90 of forecasts with the same scenario as in Fig. 7 are gained and their change rates before and after augmentations are calculated and tested in Figs. 8 and 9. The p -value of MAE decrease rate comparison is 0.0023, less than 0.05, also smaller than its counterpart of RMSE, which also means varying augmentations give statistically different boosts in EDLSTM. Similar to Fig. 7, the DAs are better than PAs, but Fig. 8 offers a clearer distinction between several DAs. DA4 and DA5 have a greater MAE decline, 8.97% and 8.82%, than DA2 and DA3, 8.49% and 7.79%, which generally indicates that the former two provide closer predictions to the real values. But DA4 and DA5 may have big deviations in some forecasting points, so these data-oriented augmentations are quite close in Fig. 7. The p -value of QR90 increase rate comparison is 0.0052, bigger than 0.05, which illustrates different augmentations have no significant different improvements, around 12% to 13%, in QR90. This phenomenon reveals that either augmentation technique can elevate the qualification rate of the EDLSTM model in a relatively similar amount, and provide satisfactory forecasts in terms of this evaluation index.

To summarize, the impact of the different data augmentation methods on the benchmark models is not significantly different. However, the improvement for the deep EDLSTM is slightly varied,

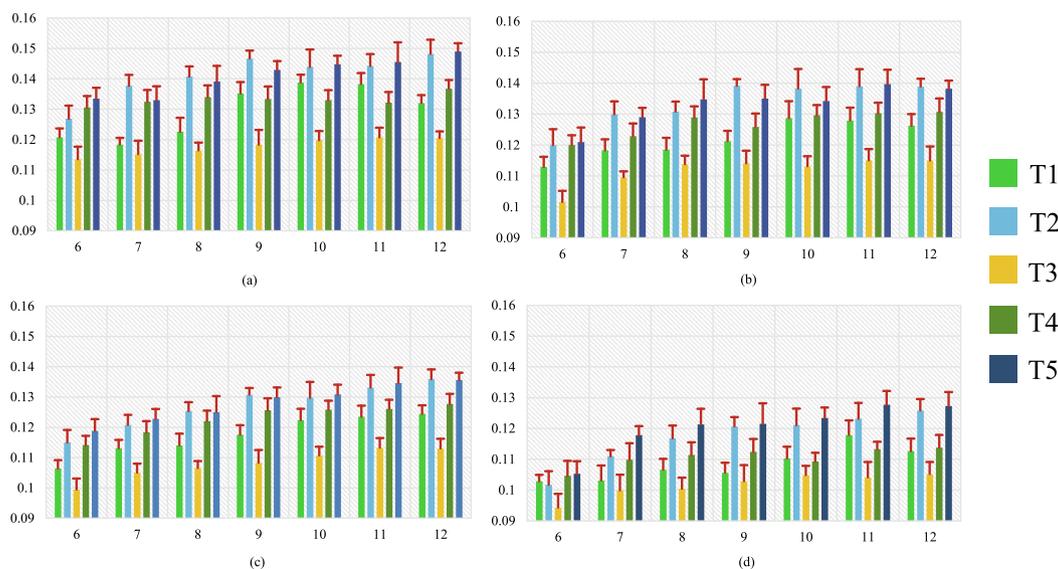


Fig. 5. The multistep average RMSE of forecasting models with data augmentations for each turbine: (a) NN, (b) LSTM, (c) BA, (d) EDLSTM.

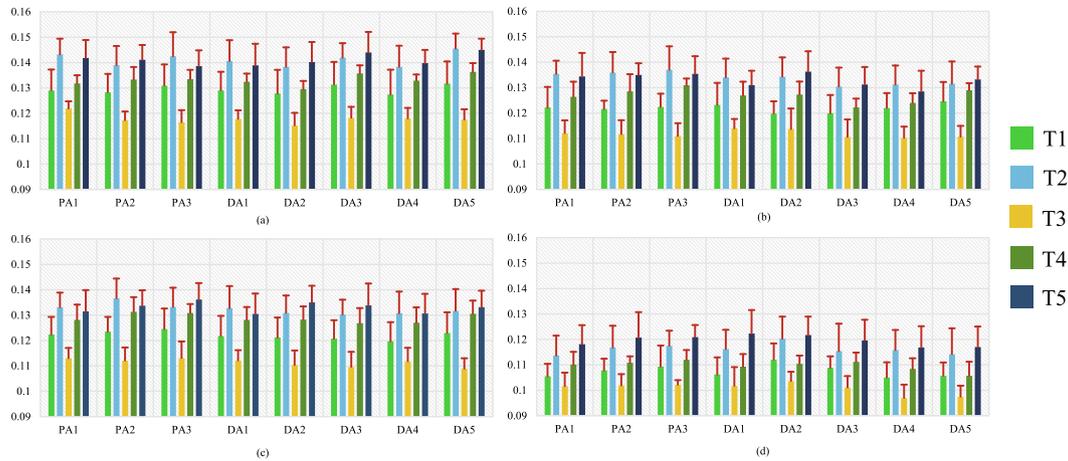


Fig. 6. The multistep average RMSE of forecasting models with various data augmentations for each turbine: (a) NN, (b) LSTM, (c) BA, (d) EDLSTM.

Table 5

The *p*-values of RMSE Friedman test within five turbines for multiple comparisons in different data-augmented approaches.

<i>P</i> -values	6	7	8	9	10	11	12
NN	0.4717	0.2772	0.1013	0.54	0.1705	0.1046	0.7608
LSTM	0.1274	0.6809	0.3268	0.6809	0.0183	0.0335	0.6113
BA	<i>0.0202</i>	0.4084	0.3445	0.0558	0.1775	0.532	0.1507
EDLSTM	<i>0.0049</i>	<i>0.0012</i>	0.0626	0.3041	0.1213	0.0901	0.1239

Note: The *p*-values less than 0.05 are marked in italics meaning H_0 is rejected.

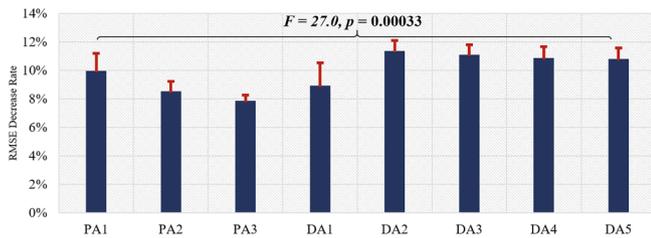


Fig. 7. The average RMSE decrease rate of multistep EDLSTM forecast with various augmentations for averaging five turbines.

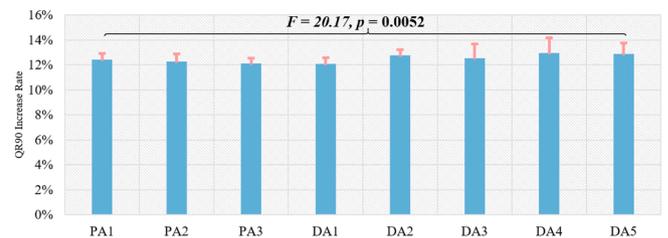


Fig. 9. The average QR90 increase rate of multistep EDLSTM forecast with various augmentations for averaging five turbines.

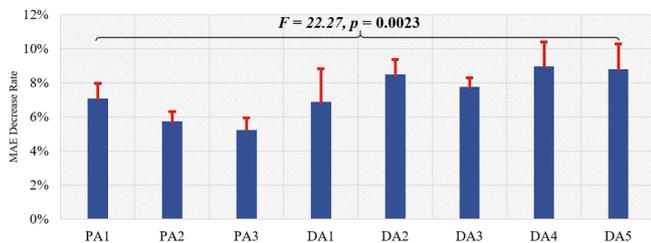


Fig. 8. The average MAE decrease rate of multistep EDLSTM forecast with various augmentations for averaging five turbines.

unremarkable in QR90 metric. DAs, on the whole, outperform PAs in RMSE and MAE, and MAE further reveals that DA4 and DA5 have edges among the DA methods.

6. Conclusions

This paper initially scrutinizes the usefulness of data augmentation approaches in wind power forecasting and proposes a multi-input and multi-output prediction algorithm with verified superiority. Inferences on the results of multistep forecasting five wind turbines with various

topologies, conclusions are given as follows.

The proposed seq2seq-based deep EDLSTM enables highly effective and robust multistep power forecasting, by highlighting the sequential dependence of the problem, for wind turbines under different terrain conditions. Also, compared with the benchmark PR, NN, LSTM, and BA algorithms, its overall RMSE is lowered by 33.89%, 10.60%, 7.12%, and 4.27%, respectively.

Since EDLSTM is a complex deep learning model, its strength requires so-called big data. It is demonstrated that five-fold expansions of the primary data with data augmentations statistically boost neural network-based NN, LSTM, BA, and EDLSTM wind power forecasting capabilities. The boost is particularly evident in EDLSTM, where, on average, the performance of the data-augmented model provides better forecasting with lower RMSE, which is 10.2% smaller than its counterpart without data augmentations. This boosting can be interpreted as expanding the training set, it is equivalent to adding a regular term to the loss function when training models, which can effectively avoid overfitting. Besides, due to the stochasticity involved in data augmentations, the learned model built on the techniques presents better robustness. Moreover, the data-augmented EDLSTM edges over the benchmarks, PR, NN, LSTM, and BA with the same expanding inputs, extending to 40.63%, 17.67%, 11.74%, and 7.06% decrease in RMSE, respectively since the proposed EDLSTM further learns deeper

information, like signal decompositions, of the wind data by mentioned augmentation techniques.

The impact of the eight data augmentation approaches employed, three physics-oriented and five data-oriented, on wind power prediction is forecasting arithmetic sensitive. For the proposed well-performing EDLSTM, various augmentations can approximately, by over 12%, boost the forecasting qualification rate at the 90% threshold. But augmentations improve the forecasting performance to slightly different degrees when evaluated by RMSE and MAE: multistep and multiturbine meanly, the improvement varies from approximately 7.87% to 11.36% of RMSE and 5.24% to 8.97% of MAE within one standard deviation, and generally, data-oriented augmentations outperform physics-oriented ones. Among data-oriented augmentations, the results illustrate that EDLSTM's forecasting RMSE is significantly decreased even by simply appending noisy and randomly perturbing, or moving data the same way as sophisticated statistical data decomposition and learning data generation, however, as per MAE, the latter two provide overall closer predictions to the real power.

Our future research, on the basis of this paper, foresees to further investigate de facto more advanced data augmentation techniques and integrate them into the proposed model to conduct in-depth point and probability predictions and attempt industrial applications in extensive comparisons with other forecasting models.

Additionally, ensuing policy recommendations may be extrapolated.

Drawing on state-of-the-art deep learning techniques and increasing computational abilities, wind power forecasting and deriving data issues in energy fields shall be approached progressively from traditional statistical and parameters-sensitive classical machine learning methods to deep learning approaches that can automatically identify complex patterns. Besides, the sophisticated deep networks are particularly reliant on data amounts. Motivated by this article, limited data of wind parks or other energy sectors could be artificially enlarged by appropriate data augmentations to serve as the stepping stone for further applications of deep learning to challenge related scientific and engineering difficulties.

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.

Acknowledgments

The open-access publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway. And thanks to the support of UiT Arctic Centre for Sustainable Energy and Dr. Fuqing Yuan and Stian Normann Anfinsen for their comments on the manuscript.

Data availability

The NWP data are publicly available from The Norwegian Meteorological Institute. The measured wind data from Fakken wind park are the property of the power company Troms Kraft AS. These data are available on reasonable request from the authors.

References

- [1] Letcher TM. *Wind Energy Engineering: A Handbook for Onshore and Offshore Wind Turbines*. Academic Press; 2017.
- [2] Rahimi E, Rabiee A, Aghaei J, Muttaqi KM, Esmaeel Nezhad A. On the management of wind power intermittency. *Renew Sustain Energy Rev* 2013;28:643–53.
- [3] Cai Y, Bréon F-M. Wind power potential and intermittency issues in the context of climate change. *Energy Convers Manage* 2021;240:114276.
- [4] Liu H, Chen C, Lv X, Wu X, Liu M. Deterministic wind energy forecasting: A review of intelligent predictors and auxiliary methods. *Energy Convers Manage* 2019;195:328–45.
- [5] Hanifi S, Liu X, Lin Z, Lotfian S. A critical review of wind power forecasting methods—past, present and future. *Energies* 2020;13(15):3764.
- [6] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006;313(5786):504–7.
- [7] Kim P. In: *MATLAB Deep Learning*. Berkeley, CA: Apress; 2017. p. 1–18. https://doi.org/10.1007/978-1-4842-2845-6_1.
- [8] Wang H, Lei Z, Zhang X, Zhou B, Peng J. A review of deep learning for renewable energy forecasting. *Energy Convers Manage* 2019;198:111799.
- [9] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 843–852.
- [10] Wang X, Shi Y, Kitani KM. Deep supervised hashing with triplet labels. In: *Asian conference on computer vision*. Springer; 2016. p. 70–84.
- [11] Solomon C, Breckon T. *Fundamentals of Digital Image Processing: A practical Approach with Examples in Matlab*. John Wiley & Sons; 2011.
- [12] Chatterjee S, Hadi AS. *Regression Analysis by Example*. John Wiley & Sons; 2013.
- [13] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *Journal of Big Data* 2019;6(1):1–48.
- [14] I. Sutskever, O. Vinyals, and Q. V. Le. "Sequence to sequence learning with neural networks," arXiv preprint arXiv:1409.3215, 2014.
- [15] Zhang Y, Li Y, Zhang G. Short-term wind power forecasting approach based on Seq2Seq model using NWP data. *Energy* 2020:118371.
- [16] Pirhooshyaran M, Scheinberg K, Snyder LV. Feature engineering and forecasting via derivative-free optimization and ensemble of sequence-to-sequence networks with applications in renewable energy. *Energy* 2020;196:117136.
- [17] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.
- [18] Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep Learning*. MIT press Cambridge; 2016.
- [19] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. "Autoaugment: Learning augmentation strategies from data," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 113–123.
- [20] T. DeVries and G. W. Taylor, "Dataset augmentation in feature space," arXiv preprint arXiv:1702.05538, 2017.
- [21] D. S. Park et al., "Specaugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.
- [22] Deng X, Shao H, Hu C, Jiang D, Jiang Y. Wind power forecasting methods based on deep learning: A survey. *Computer Modeling in Engineering & Sciences* 2020;122(1):273–302.
- [23] Yildiz C, Acikgoz H, Korkmaz D, Budak U. An improved residual-based convolutional neural network for very short-term wind power forecasting. *Energy Convers Manage* 2021;228:113731.
- [24] Kisvari A, Lin Z, Liu X. Wind power forecasting—A data-driven method along with gated recurrent neural network. *Renewable Energy* 2021;163:1895–909.
- [25] Shahid F, Zameer A, Muneeb M. A novel genetic LSTM model for wind power forecast. *Energy* 2021;223:120069.
- [26] Memarzadeh G, Keynia F. A new short-term wind speed forecasting method based on fine-tuned LSTM neural network and optimal input sets. *Energy Convers Manage* 2020;213:112824.
- [27] M. H. Lipu et al., "Artificial Intelligence Based Hybrid Forecasting Approaches for Wind Power Generation: Progress, Challenges and Prospects," IEEE Access, 2021.
- [28] Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N. Deep learning and process understanding for data-driven Earth system science. *Nature* 2019;566(7743):195–204.
- [29] Giebel G, Kariniotakis G. Wind power forecasting—A review of the state of the art. In: *Renewable Energy Forecasting*. Elsevier; 2017. p. 59–109. <https://doi.org/10.1016/B978-0-08-100504-0.00003-2>.
- [30] Ng A. *Advice for applying machine learning*. Machine learning 2011.
- [31] Chen H, Birkelund Y, Anfinsen SN, Staupe-Delgado R, Yuan F. Assessing probabilistic modelling for wind speed from numerical weather prediction model and observation in the Arctic. *Sci Rep* 2021;11(1):1–11.
- [32] Dong Q, Sun Y, Li P. A novel forecasting model based on a hybrid processing strategy and an optimized local linear fuzzy neural network to make wind power forecasting: A case study of wind farms in China. *Renewable Energy* 2017;102:241–57.
- [33] Cleveland RB, Cleveland WS, McRae JE, Terpenning I. STL: A seasonal-trend decomposition. *Journal of official statistics* 1990;6(1):3–73.
- [34] Li J, Zhou J, Chen B. Review of wind power scenario generation methods for optimal operation of renewable energy systems. *Appl Energy* 2020;280:115992.
- [35] Chen H, Birkelund Y, Anfinsen SN, Yuan F. Comparative study of data-driven short-term wind power forecasting approaches for the Norwegian Arctic region. *J Renewable Sustainable Energy* 2021;13(2):023314. <https://doi.org/10.1063/5.0038429>.
- [36] F. A. Gers, J. Schmidhuber, F. Cummins. Learning to forget: Continual prediction with LSTM. 1999.
- [37] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM—a tutorial into Long Short-Term Memory Recurrent Neural Networks," arXiv preprint arXiv:1909.09586, 2019.
- [38] C. Olah. *Understanding lstm networks*. 2015.
- [39] Ren Ye, Suganthan PN, Srikanth N. Ensemble methods for wind and solar power forecasting—A state-of-the-art review. *Renew Sustain Energy Rev* 2015;50:82–91.
- [40] Jiajun H, Chuanjin Y, Yongle L, Huoyue X. Ultra-short term wind prediction with wavelet transform, deep belief network and ensemble learning. *Energy Convers Manage* 2020;205:112418.
- [41] Lee J, Wang W, Harrou F, Sun Y. Wind power prediction using ensemble learning-based models. *IEEE Access* 2020;8:61517–27.

- [42] da Silva RG, Ribeiro MHDM, Moreno SR, Mariani VC, dos Santos Coelho L. A novel decomposition-ensemble learning framework for multi-step ahead wind energy forecasting. *Energy* 2021;216:119174.
- [43] Murphy KP. *Machine learning: a probabilistic perspective*. MIT press; 2012.
- [44] Sun W, Zhang T, Tao R, Wang A. Short-Term Photovoltaic Power Prediction Modeling Based on AdaBoost Algorithm and Elman. In: 2020 10th International Conference on Power and Energy Systems (ICPES). IEEE; 2020. p. 184–8.
- [45] Whitley D. A genetic algorithm tutorial. *Statistics and computing* 1994;4(2):65–85.
- [46] Veit A, Wilber MJ, Belongie S. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems* 2016;29:550–8.
- [47] Lijuan L, Hongliang L, Jun W, Hai B. A novel model for wind power forecasting based on Markov residual correction. In: IREC2015 The Sixth International Renewable Energy Congress. IEEE; 2015. p. 1–5.
- [48] Gibbons JD, Chakraborti S. *Nonparametric Statistical Inference; Revised and Expanded*. CRC Press; 2014.