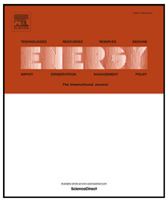
Contents lists available at ScienceDirect

Energy

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



A transformer-based deep neural network with wavelet transform for forecasting wind speed and wind energy

Erick Giovani Sperandio Nascimento [a,b,c,*], Talison A.C. de Melo [d], Davidson M. Moreira [c]

[a] Surrey Institute for People-Centred Artificial Intelligence, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, UK
[b] Global Centre for Clean Air Research (GCARE), Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, UK
[c] Stricto Sensu Department, SENAI CIMATEC, Salvador, Bahia, Brazil
[d] Department of Electrical Engineering, UFBA, Salvador, Bahia, Brazil

ARTICLE INFO

Keywords:
Transformer
Wavelet
Wind speed forecasting
Wind power forecasting
Deep learning
Renewable energy
Multivariate time series forecasting

ABSTRACT

This work presents a novel transformer-based deep neural network architecture integrated with wavelet transform for forecasting wind speed and wind energy (power) generation for the next 6 h ahead, using multiple meteorological variables as input for multivariate time series forecasting. To evaluate the performance of the proposed model, different case studies were investigated, using data collected from anemometers installed in three different regions in Bahia, Brazil. The performance of the proposed transformer-based model with wavelet transform was compared with an LSTM (Long Short Term Memory) model as a baseline, since it has been successfully used for time series processing in deep learning, as well as with previous state-of-the-art (SOTA) similar works. Results of the forecasting performance were evaluated using statistical metrics, along with the time for training and performing inferences, both using quantitative and qualitative analysis. They showed that the proposed method is effective for forecasting wind speed and power generation, with superior performance than the baseline model and comparable performance to previous similar SOTA works, presenting potential suitability for being extended for the general purpose of multivariate time series forecasting. Furthermore, results demonstrated that the integration of the transformer model with wavelet decomposition improved the forecast accuracy.

1. Introduction

Electric energy is a fundamental part of modern society's life. It is used for various applications from simple daily household activities to complex industrial tasks. According to the International Energy Agency (IEA), a significant portion of the world's electricity is provided by non renewable energy sources, being essentially coal, oil and natural gas, as illustrated in Fig. 1.

Non renewable energy sources are already in short supply and harmful for the environment. Some disadvantages of using fossil fuels as an energy source are: generation of greenhouse gases, depletion of ozone layer, non-biodegradable waste generation and soil, water and air pollution. Alternative renewable energy sources, for instance wind and solar energy, are a plausible solution to the exhaustion of non renewable sources. They are clean energy sources and available in nearly unlimited supply.

In favour of sustainable development, efforts should be aligned with the reduction of greenhouse gases emissions, with more investments in renewable and sustainable energy generation, among other actions. In this context, wind energy is increasingly gaining ground due to its benefits such as its renewable and sustainable nature, along with its low cost, low negative impact on the environment and abundance [2]. Because of this, several methods are found in the literature to predict the wind speed and wind energy for different forecasting horizons. For instance, some studies approached stochastic methods such as autoregressive (AR) model [3] and AutoRegressive Integrated Moving Average (ARIMA) [4]. Other works used deep learning architectures, for example Multilayer Perceptron (MLP) combined with wavelet transforms for nowcasting wind power [5] and wind ramp [6], Long Short Term Memory (LSTM) [7], Gated Recurrent Unit (GRU) [8], Convolutional Neural Network (CNN) [9], and hybrid models (e.g. based on a combination of CNN and LSTM) [10]. Still in the field of artificial intelligence, we observe the application of neuro-fuzzy [11], that refers to combinations of artificial neural networks and fuzzy logic for wind prediction.

[*] Corresponding author at: Surrey Institute for People-Centred Artificial Intelligence, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, UK.
E-mail address: erick.sperandio@surrey.ac.uk (E.G.S. Nascimento).

https://doi.org/10.1016/j.energy.2023.127678
Received 26 December 2022; Received in revised form 24 March 2023; Accepted 27 April 2023
Available online 2 May 2023

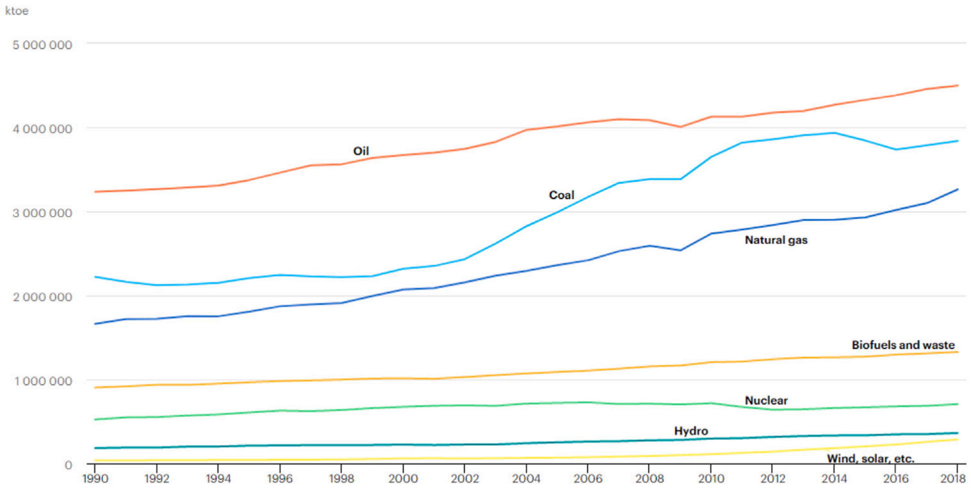


Fig. 1. Total energy supply by source, World 1990–2018.
Source: World Energy Balances 2020 - International Energy Agency (IEA) [1].

In the area of sequential data modelling, the transformer model with self-attention mechanism [12] is currently one of the most used neural network architectures in the field of Natural Language Processing (NLP), and was initially proposed for machine translation applications [13]. The revolution that the transformer architecture promoted in the NLP field motivated the development of new approaches based on the transformers in other areas such as computer vision and time series. For instance, in the computer vision area, [14] developed a Multiscale Vision Transformers (MViT) for video and image recognition, by connecting the seminal idea of multiscale feature hierarchies with transformer models, while [15] introduced the concept of Convolutional vision Transformer (CvT), that seeks to improve Vision Transformer (ViT) in performance and efficiency by introducing convolutions into ViT to yield the best of both designs. The work of [16] added locality to ViT by introducing depth-wise convolution into the feed-forward network. Authors in [17] propose a conditional positional encoding (CPE) scheme for ViT, while [18] empirically showed that a spatial dimension reduction is beneficial to a transformer architecture due to the application of CNNs, proposing a novel Pooling-based Vision Transformer (PiT) upon the original ViT model, and [19] introduced the idea of dense prediction transformers by assembling tokens from various stages of the ViT into image representations at different resolutions and further combine them into full resolution using a convolutional decoder.

More recently, new studies have demonstrated the feasibility of the application of the transformer architecture in the task of time series forecasting. Some works attempted to develop a general transformer-based neural network for time series forecasting, e.g. [20] developed a transformer-based model in two real-world datasets from different domains, [21] applied it to a case study of influenza-like illness, [22] used transformer neural networks to predict Dogecoin price, and [23] developed a transformer-based model to forecast time series of aero-engine gas-path performance. The work of [24] studied the application of the transformer architecture for short-term wind power forecasting, applying it to data of wind power output of multiple wind farms, and comparing against the LSTM model as a baseline. Based on the Mean Absolute Error (MAE) and the Mean Average Percentage Error (MAPE) metrics, they found that the transformer outperformed the baseline model. However, few details are available about the implementations, case study information (such as the height of the measurements), hyperparameters optimisation etc. Authors in [25] developed a model that implements a masked interpretable multi-head attention layer on top of LSTM encoder/decoder layers to predict wind speed using temporal diffusion transformers technique [26], which is a general approach for multi-horizon time series forecasting based on the attention mechanism. The work presented in [27] implements a multi-modal multi-task transformer architecture for ultra-short-term wind power forecasting based on multi-source heterogeneous data.

Therefore, we notice that there are very few works that explore the transformers architecture for wind speed and power forecasting. Actually, none of them focused on the systematic and detailed development, exploration and evaluation of a specially designed transformer-based deep neural network model integrated with wavelet decomposition for wind speed and wind energy forecasting, using quantitative and qualitative analysis based on statistical metrics, and the assessment of its computational cost for training and for inference, presenting clear knowledge gaps in the literature.

In order to contribute to the current state of the art and fulfilling the identified knowledge gaps, the purpose of this research is to propose, develop and investigate the implementation of a new deep neural network based on the transformer architecture, but with improvements specially designed for wind speed and wind energy forecasting for the next 6 h. The new proposed deep learning architecture is powered by the integration with wavelet transform for feature augmentation, that basically refers to techniques used to increase the number of features in the data by adding more relevant information from the existing data, aiming at adding and supporting more learning capabilities. By integrating with the proposed transformer model for the task of wind speed and wind energy forecasting, we developed and studied the influence of the feature augmentation using wavelet decomposition on the predictions, both for the proposed transformer architecture and for the baseline LSTM model, for comparison purposes.

As a case study, we used data from three different cities in Bahia state, Brazil. For the analysis, the time series of wind and other meteorological variables were collected from towers equipped with anemometers at heights of 100.0, 120.0, and 150.0 m. The investigations applied to this case study also contribute to the operation of wind energy plants in Brazil, justified by:

- **Wind Energy Potential**: In Brazil, according to the Brazilian Association of Wind Energy (ABEEólica), in 2021 wind energy reached 18 GW of installed capacity with more than 8300 wind turbines in 695 wind farms. Thus, wind energy occupies the second place (10.3%) in the national electricity matrix, staying only behind of the hydraulic energy (58.7%) [28].
- **Hydraulic Crisis**: In the last hydrological year, between August 2020 and September 2021, rainfall recorded in Brazil pointed to a historical shortage (worse in 91 years), as reported by the

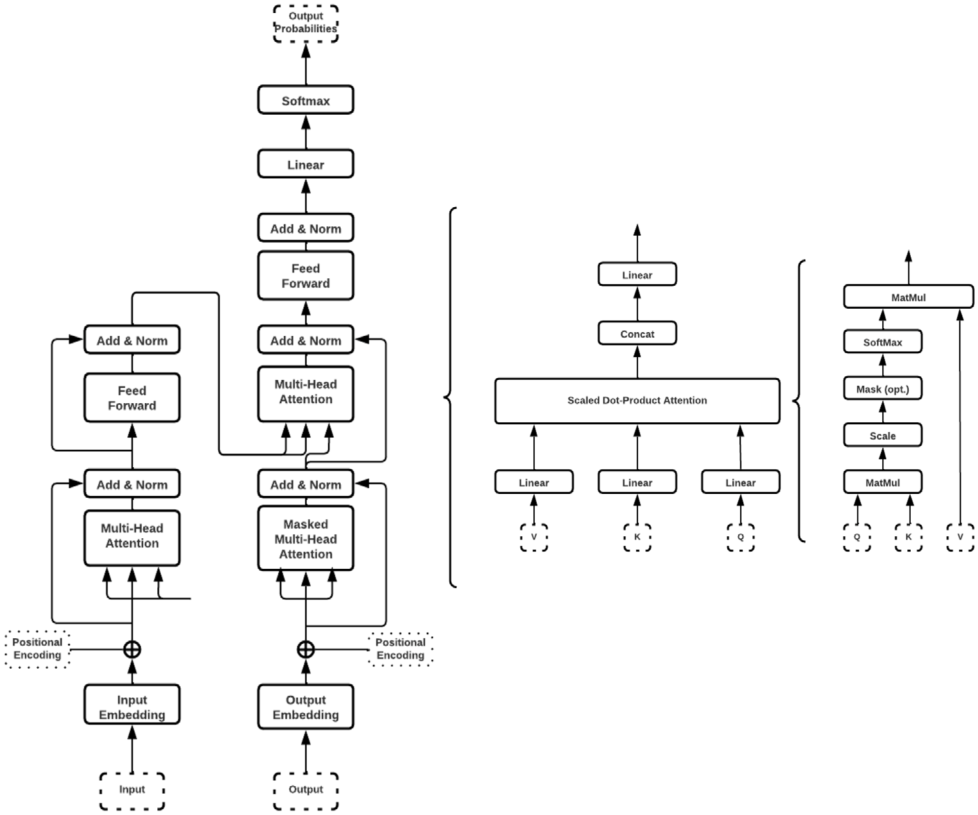


Fig. 2. (left) The original transformer architecture; (centre) Multi-Head Attention; and (right) Scaled Dot-Product Attention.
Source: Adapted from Vaswani et al. [12].

National Electric System Operator (ONS), which makes energy production more difficult and expensive [29].

Therefore, the main novelties and contributions of this work to the current state of the art are summarised as follows:

1. The proposal, development and investigation of a new deep learning methodology using a specially designed transformer-based architecture integrated with wavelet transform for predicting wind speed and wind energy for the next 6 h ahead using multiple meteorological variables as input;
2. The integration of the proposed model with wavelet decomposition for feature augmentation, and a study of the influence of its usage for improving the predictions;
3. A systematic full assessment and comparison of the performance of the proposed model with a traditional deep learning architecture for time series processing, the LSTM, with hyperparameters optimisation in order to guarantee a fair comparison approach, evaluating both using quantitative and qualitative analysis, assessing and comparing their computational costs for training and for inference;
4. The application of the proposed approach and the baseline model to a case study comprised of data collected from anemometric towers at different heights, namely 100.0, 120.0 and 150 m, which are very representative of usual wind turbine's hub heights, in three different regions in Bahia State, Brazil, which is one of the most important regions in Brazil for wind power generation;
5. The analysis of feature augmentation with wavelet transform for both models, in order to assess the real impact of its application for improving the forecasting performance of the proposed approach.

The remainder of the work is organised as follows. Section 2 presents the methodology employed to achieve the aforementioned objectives. Section 3 presents and discusses the obtained results, and Section 4 contains the conclusions of the research.

2. Methodology

2.1. The transformers architecture

The original transformer architecture, shown on the left of Fig. 2, consists on an encoder–decoder layers structure using stacked self-attention and fully connected layers for both the encoder and decoder [12]. The encoder is in the left side of the image, and decoder layer at the right.

The encoding component is a stack of encoders and each one is composed of two main sub-layers: multi-head self-attention mechanism and feed forward neural network. Similar to the encoder, the decoding component is a stack of decoders. In addition to the two sub-layers of the encoder, the decoder has a third sub-layer between them that performs multi-head attention over the output of the encoder stack.

The multi-head attention and scaled dot-product attention shown on the centre and right of Fig. 2, respectively, use tree vectors: Query vector (Q), a Key vector (K), and a Value vector (V). Transformers use a "Scaled Dot-Product Attention" to obtain the context vector and calculate the attention as:

$$\text{Attention(Q,K,V)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where $Q = W^Q x$, $K = W^K x$, $V = W^V x$ on input $x = \{x_1, \ldots, x_n\}$. W^Q, W^K and W^V are weight matrices to generate Q, K and V via linear transformations on x. Multi-head Attention is a module for attention

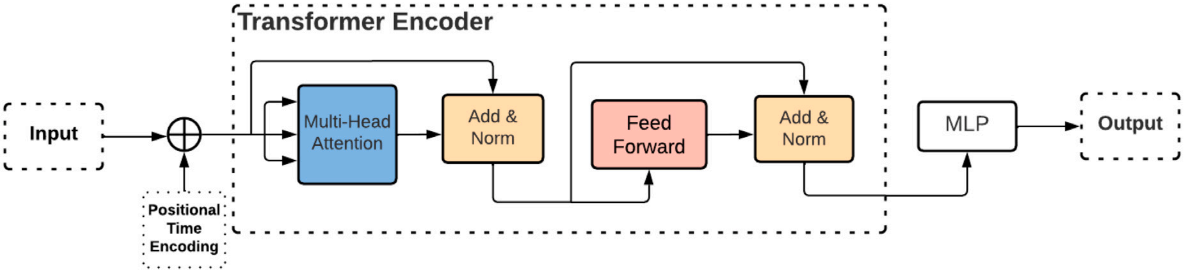


Fig. 3. The proposed transformer architecture for wind speed and power generation forecast. The illustration of the transformer encoder was inspired by [12].

mechanisms which allows the model to jointly attend to information from different representation subspaces in parallel, that is:

$$\text{Multi-Head(Q,K,V)} = \text{Concat}\left(head_1, \dots, head_n\right) W^O \tag{2}$$

where:

$$head_i = \text{Attention}(WQ_i^Q, WK_i^K, WV_i^V) \tag{3}$$

2.2. The proposed transformer-based architecture

The transformer architecture needs to be changed in order to be able to deal with the task of time series forecasting, because it was originally designed for machine translation systems. For this end, the proposed transformer-based model for wind speed and wind energy forecasting consists of, in general terms:

- the implementation of an encoder-only transformer followed by a multilayer perceptron (MLP), as seen in Fig. 3, aiming at capturing relevant temporal information by the encoder layer to latter enable the prediction by the last MLP layer;
- the implementation of a new positional encoding mechanism, better suited for handling time series forecasting tasks that present cyclical/seasonal behaviour, as in the case of wind and other weather time series data;
- the removal of the embedding layers of the model's input, so that temporal patterns can be better recognised from the input time series;
- the replacement of the softmax activation function in the output layer by a more specialised one for time series forecasting (as depicted below), once this activation function is better suited for classification tasks but not for time series forecasting tasks.

For the proposed model, we restrict the use of the Adam [30] optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-7}$. The output layer of the proposed transformer architecture was set to use the Rectified Linear Unit (ReLU) and Leaky Rectified Linear Unit (LeakyReLU) activation functions. ReLU has output 0 if the input is less than 0, and raw output otherwise. LeakyReLU has a small slope for negative values instead of a flat slope. We tested both activation functions, and the one that performed better was considered.

In order to enable this proposed model to deal with time series, the idea was to pre-process the timestamp associated to each sample as new features using the trigonometric functions sine and cosine. For this end, the positional encoding was implemented in order to represent the timestamp associated with a sample as three new features: hour of the day, day of the month and month of the year. Each element was decomposed into a sine and cosine component, such that hour, day and month are represented cyclically, according to the following equations:

$$\sin\left(t_a, f\right) = \sin\left(\frac{2\pi t_a}{f}\right) \tag{4}$$

$$\cos\left(t_a, f\right) = \cos\left(\frac{2\pi t_a}{f}\right) \tag{5}$$

where t_a is the value of the timestamp attribute associated to a sample, i.e. month of the year, day of the month, or hour of the day, and f is the number of months, days or hours in each corresponding time scale of year, month or day, respectively. This approach can be easily expanded to other time frequencies, e.g. seconds, minutes etc.

Table 1
The different types of discrete Wavelet families.

Families	Short name	Order
Haar	haar	–
Daubechies	db	1–38
Symlets	sym	2–20
Coiflets	coif	1–17
Biorthogonal	bior	1.1–6.8
Reverse biorthogonal	rbio	1.1–6.8
Discrete Meyer	dmey	–

2.3. Integration with wavelet decomposition for feature augmentation

Signal processing techniques, like the Fourier transform, Wavelet transform, and Wavelet packet decomposition, have been successfully applied together with machine learning techniques aiming to improve the forecast accuracy of time series [5,6,31]. In this study, the feature augmentation was applied using wavelet decomposition. Since there are many families of Wavelets (see Table 1) for the discrete Wavelet transform (DWT), finding the optimal mother wavelet is essential for the best performance of the machine learning methods. To determine which type of wavelet presents more correlation in the reconstruction of meteorological signals, as proposed by [5], a comparison and a statistical evaluation (through RMSE) were made between the approximate level 1 reconstructed signal by wavelet decomposition and the original signal. The wavelet type that has the lowest RMSE between the approximate reconstructed signal at level 1 and the original signal has been selected. Fig. 4 summarises the adopted strategy, as presented by [5].

2.4. The case study

For this project, we used hourly wind speed time series collected at 100.0 m, 120.0 m and 150.0 m height of anemometric towers installed in the cities of Esplanada, Mucugê and Mucuri, located in Bahia, Brazil, totalling nine different time series with 744 h each. The State of Bahia has a size comparable to countries like Spain and France, presenting a huge potential for wind power generation [5]. The respective geographic locations are shown in Fig. 5. Table 2 presents the geographic information and reference period of data captured by anemometers in Esplanada, Mucugê and Mucuri.

The database consists of the following attributes: hour, day, month, and important meteorological parameters: wind speed [m/s], wind direction [°] (oriented northwards), air temperature [C], air humidity [%], and air pressure [Bar].

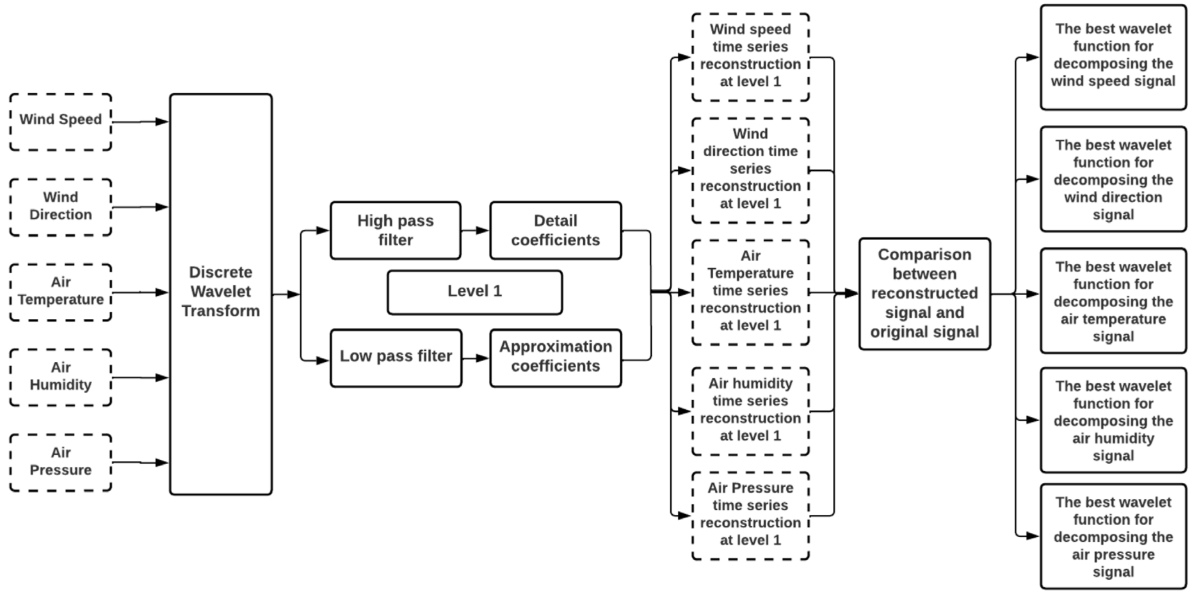

Wind Speed
Wind Direction
Air Temperature
Air Humidity
Air Pressure
Discrete Wavelet Transform
High pass filter
Detail coefficients
Level 1
Low pass filter
Approximation coefficients
Wind speed time series reconstruction at level 1
Wind direction time series reconstruction at level 1
Air Temperature time series reconstruction at level 1
Air humidity time series reconstruction at level 1
Air Pressure time series reconstruction at level 1
Comparison between reconstructed signal and original signal
The best wavelet function for decomposing the wind speed signal
The best wavelet function for decomposing the wind direction signal
The best wavelet function for decomposing the air temperature signal
The best wavelet function for decomposing the air humidity signal
The best wavelet function for decomposing the air pressure signal

Fig. 4. Summary of the wavelet decomposition strategy.

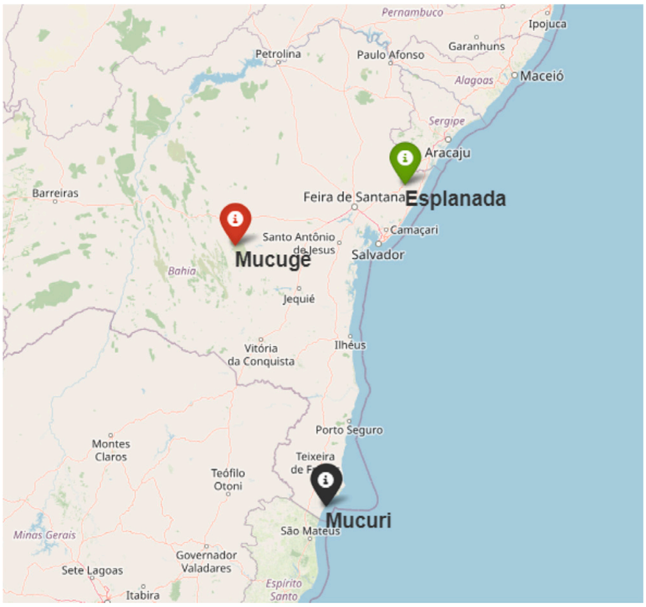


Fig. 5. Locations of the three towers in Esplanada, Mucugê and Mucuri.

Table 2
Geographic location, reference period and recorded hours of data collected by anemometers in Mucugê, Esplanada and Mucuri.

Site	Latitude	Longitude	Reference period
Mucugê	13° 21’01.92”S	41° 31’53.76”W	09/07/2016 at 00:00 a.m. to 10/07/2016 at 11:00 p.m.
Esplanada	11° 50’55.22”S	37° 55’44.31”W	04/28/2016 at 09:00 a.m. to 05/29/2016 at 08:00 a.m.
Mucuri	18° 1’31.52”S	39° 30’51.69”W	11/30/2015 at 02:00 p.m. to 12/31/2015 at 01:00 p.m.

2.5. Experimental design

The training set with 550 recorded samples was divided into 70% as training data, and the remaining 30% as validation data. The test set was formed with 194 recorded hours, never seen by the models during the training and validation. In all experiments, we normalised the features by scaling them ranging from 0 to 1 through the minimum–maximum scaler procedure.

Early stopping was used to stop the training once the model's performance stops improving on a hold out validation dataset. The Keras library supports the early stopping feature via a callback function called $EarlyStopping$. The early stopping method was applied in all simulations with patience = 30.

Several preliminary tests were carried out to verify the best settings for Multi Head Attention. We verified that d_{model} values (model dimension) less than 100 presented better results. We varied the dimensions of keys (d_k) and values (d_v) from 4 to 128, and the model converged more efficiently with 8 and 16. We tested dropout rate (P_{drop}) of 0.0, 0.1, 0.2, and 0.3, where the value of 0.2 best fitted to our application. In all cases, we obeyed the Multi Head Attention settings by following the equality $d_k = d_v = \frac{d_{model}}{h}$, where h is the number of attention heads. Based on all these aspects, Table 3 presents the configurations tested in the Multi-Head Attention layer.

In this study, we used LSTM models as a baseline. The hyperparameters for the baseline models were tuned by hyperparameter optimiser called $Hyperband$ [32]. It is an optimised version of random search and is one of the tuners available in the Keras Tuner library. For further theoretical details, please see [32]. According to Fig. 6, the hyperparameters tuned are the number of units in a dense and LSTM layers, the dropout rate and the activation function from the hidden and output layers. For all simulations, we set the optimiser's learning rate to 0.001. We also compared both LSTM and the proposed transformer-based model with a simple persistence model, aiming at evaluating whether these models would perform better or worse than a persistence approach.

For each configuration of the predictive model, tests were performed with and without the wavelet decomposition strategy, that is, there were 6 simulations for each configuration proposed, as follows:

1. Simulation 1: dataset formed only from original signals;
2. Simulation 2: original signals + signals reconstructed by wavelet decomposition (approximation and detail signals) at level 1;

Table 3
Variations on the Multi-Head Attention.

Configuration	h	d_{model}	d_k	d_v	P_{drop}
1	2	32	16	16	0.2
2	3	48	16	16	0.2
3	4	64	16	16	0.2
4	5	80	16	16	0.2
5	6	96	16	16	0.2
6	7	56	8	8	0.2
7	8	64	8	8	0.2
8	9	72	8	8	0.2
9	10	80	8	8	0.2

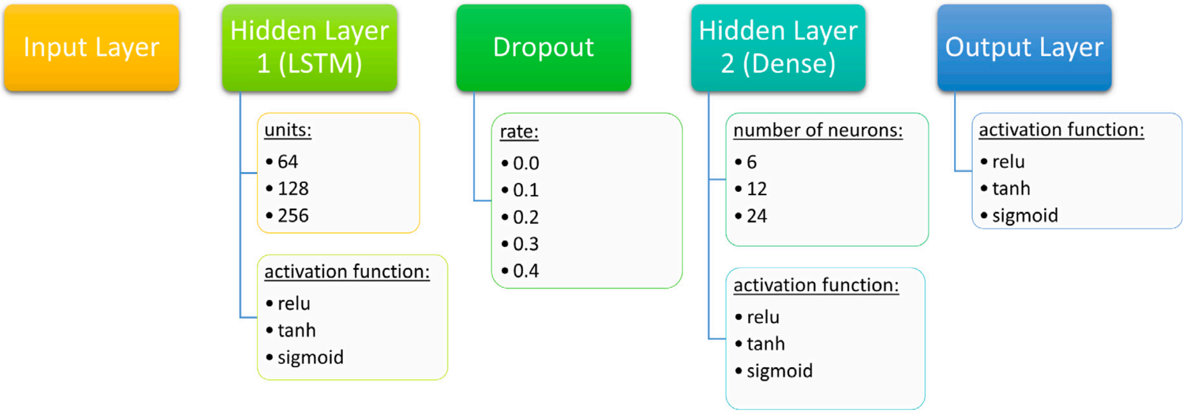


Fig. 6. The hyperparameters tuned through Hyperband.

3. Simulation 3: original signals + signals reconstructed by wavelet decomposition (approximation and detail signals) at levels 1 and 2;
4. Simulation 4: original signals + signals reconstructed by wavelet decomposition (approximation and detail signals) at levels 1, 2 and 3;
5. Simulation 5: original signals + signals reconstructed by wavelet decomposition (approximation and detail signals) at levels 1, 2, 3 and 4;
6. Simulation 6: original signals + signals reconstructed by wavelet decomposition (approximation and detail signals) at levels 1, 2, 3, 4 and 5.

2.6. Evaluation metrics

Initially, we used the following metrics to assess the performance of all the evaluated configurations:

$$\text{NMSE} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - x'_i)^2}{\text{var}(x)} \tag{6}$$

$$\rho(r) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(x'_i - \overline{x'})}{\sqrt{\text{var}(x)\,\text{var}(x')}} \tag{7}$$

where n is the number of samples, NMSE is the normalised mean square error, $\rho(r)$ is the Pearson's correlation coefficient, x and x' are the observed and prediction sets, respectively, x_i and x'_i are the ith observed sample and predicted value, respectively, and $var()$ is the variance of an input set of variables.

For the best configuration, i.e. the one with the smallest NMSE and the highest $\rho(r)$, we evaluated the following additional metrics:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x'_i| \tag{8}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x'_i)^2 \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x'_i)^2} \tag{10}$$

$$\text{Fac2} = \frac{\sum_{i=1}^{n}[0.5 \leq \frac{x_i}{x'_i} \leq 2.0]}{N} \tag{11}$$

where MAE is the mean absolute error, MSE is the mean square error, RMSE is the root mean square error, Fac2 is the fraction of data that lies within a factor of two, according to the notation of the Iverson bracket [33], where it returns 1 (one) if the result of the calculation between the brackets evaluates to true, and 0 (zero) otherwise. For these statistical metrics, values close to 0.0 are adequate for the NMSE, MAE, MSE, and RMSE. Values close to 1.0 are adequate for the $\rho(r)$ and Fac2.

Moreover, we applied quantile regression to estimate the errors and the prediction intervals of the baseline and the proposed models. We used the validation dataset to calculate the errors and the prediction intervals, and employed a quantile of $q = 0.95$. This means that the predicted future values would fall within a certain prediction interval with a probability of 95% [34]. Then, the test dataset was used to finally calculate the prediction intervals for each model.

To assess whether each model is statistically different from each other, we employed the Wilcoxon signed-rank test [35], which is a nonparametric statistical test for determining whether two paired or related samples have equal or different distributions. Given α as the statistical significance, $H0$ the null hypothesis, and p the calculated p-value from the test, if $p \geq \alpha$ then $H0$ can be rejected, meaning that the samples are drawn from different distributions; but, if $p>\alpha$, $H0$ cannot be rejected, the samples are drawn from the same distribution. We chose $\alpha = 0.05$ in this work.

The baseline and proposed models were also evaluated with respect to the time needed to train and to perform the inferences, or predictions.

Table 4
Technical specifications of the wind turbines.

Characteristics	Wind turbines		
Anemometric height [m]	100.0	120.0	150.0
Rated power [kW]	2000.0	4500.0	5000.0
Hub height [m]	100.0	120.0	140.0
Rotor diameter [m]	80.0	128.0	128.0
Swept area [m^2]	5027.0	12,868.0	12,868.0
Number of blades	3	3	3
Cut-in wind speed (v_{ci}) [m/s]	3.5	1.0	2.0
Rated wind speed (v_R) [m/s]	12.0	12.0	14.0
Cut-out wind speed (v_{co}) [m/s]	25.0	27.0	27.0

2.7. Calculation of the wind power generation

Wind power is considered the flux of wind energy through an area of interest, as defined in [36]. In this way, schematic of air flow at velocity U through area A can be considered. Thus, Eq. (12)[6,37] was used to predict the wind energy output from a wind turbine:

$$P_i(v) = \begin{cases} 0, & \text{if } v < v_{ci} \\ (\phi \pi r^2 v^3 C_p \eta)/2, & \text{if } v_{ci} \le v < v_R \\ P_R & \text{if } v_R \le v < v_{co} \\ 0, & \text{if } v \ge v_{co} \end{cases} \tag{12}$$

where v is the wind speed [m/s]; v_{ci} is the cut-in speed [m/s]; v_R is the rated speed [m/s]; v_{co} is the cut-out speed [m/s]; P_R is the rated power [W]; $P_i(v)$ is the power generated in the related wind speed [W]; ϕ is the density of the fluid [kg/m^3]; r is the radius [m]; C_p is the coefficient of power; η is the efficiency of the generator set — mechanical and electrical power transmission. In this study, we considered the density of the fluid, the coefficient of power and the efficiency of the generator set equal to 1.225 kg/m^3, 0.50 and 0.95, respectively.

Therefore, to simulate the final wind power output, it is necessary to specify the parameters of a wind turbine. For this study, we used the technical specifications of the investigated wind turbines from [6], as shown in Table 4.

All simulations were implemented in Python and performed on a high-performance computer with NVIDIA GPU Tesla V100-SXM3-32-GB.

3. Results and discussion

In this section, we present a detailed description of the results, followed by a comprehensive discussion. Tables 5–7 show the selected wavelet functions (with lower RMSE values) for the decomposition of the wind direction and speed signals collected in Mucugê, Esplanada and Mucuri (anemometers at 100, 120 and 150 m), respectively. In the case of air temperature, humidity and pressure signals, the defined wavelet functions are the same for the three anemometric heights (100, 120 and 150 m), as follows:

- Mucugê:
  - temperature — rbio3.5 with RMSE value of 1.220;
  - humidity — db4 with RMSE value of 2.108;
  - pressure — rbio3.3 with RMSE value of 0.246.
- Esplanada:
  - temperature — sym18 with RMSE value of 0.198;
  - humidity — db34 with RMSE value of 0.988;
  - pressure — rbio3.5 with RMSE value of 0.285.
- Mucuri:
  - temperature — dmey with RMSE value of 0.222;
  - humidity — sym12 with RMSE value of 1.078;
  - pressure — rbio1.3 with RMSE value of 1.617.

Figs. 7, 8, and 9 demonstrate the results of applying the Hyperband optimiser with Keras Tuner in Mucugê, Esplanada and Mucuri, respectively, for anemometers at heights of 100.0, 120.0, and 150.0 m.

Based on the hyperparameters set for the LSTM model in Figs. 7, 8 and 9 for Mucuge, Esplanada and Mucuri, respectively, and based on the settings described in Table 3 for the transformer model, Tables S.1, S.2 and S.3 show the evaluation metrics (MSE, RMSE, MAE, Pearson's r correlation, Fac2 and NMSE) of the mean value of the six time steps predictions for each model and anemometer height in Mucugê, Esplanada and Mucuri, respectively, based on the best values of NMSE and Pearson's r correlation metrics. The best results are highlighted in bold. For the Mucuri dataset (anemometers at height 100 m, 120 m and 150 m), we observed that the inclusion of data beyond the second level of the wavelet transform made the network unable to learn, generating a constant prediction for any input. Thus, these results were disregarded.

The training time and the average inference time per sample of the simulations presented for each model and anemometer height in Mucugê, Esplanada and Mucuri are plotted in Figures S.1 and S.2, respectively. Figures S.3, S.4 and S.5 show the evolution of the loss during the training and validation processes of the best LSTM and

Table 5
The best wavelet function for decomposing the wind direction and speed signals collected in Mucugê (anemometer at 100, 120 and 150 m).

Anemometric height 100 m			
Wind speed [m/s]	RMSE [m/s]	wind direction [°]	RMSE [°]
db33	0.383	db10	9.425
Anemometric height 150 m			
Wind speed [m/s]	RMSE [m/s]	wind direction [°]	RMSE [°]
db33	0.368	bior1.1	7.314
Anemometric height 120 m			
Wind speed [m/s]	RMSE [m/s]	wind direction [°]	RMSE [°]
db33	0.375	db9	3.341

Table 6
The best wavelet function for decomposing the wind direction and speed signals collected in Esplanada (anemometer at 100, 120 and 150 m).

Anemometric height 100 m			
Wind speed [m/s]	RMSE [m/s]	wind direction [°]	RMSE [°]
db8	0.325	db7	6.073
Anemometric height 120 m			
Wind speed [m/s]	RMSE [m/s]	wind direction [°]	RMSE [°]
sym5	0.323	rbio1.3	7.347
Anemometric height 150 m			
Wind speed [m/s]	RMSE [m/s]	wind direction [°]	RMSE [°]
sym5	0.328	rbio3.5	9.792

Table 7
The best wavelet function for decomposing the wind direction and speed signals collected in Mucuri (anemometer at 100, 120 and 150 m).

Anemometric height 100 m			
Wind speed [m/s]	RMSE [m/s]	wind direction [°]	RMSE [°]
db18	0.423	db3	27.623
Anemometric height 120 m			
Wind speed [m/s]	RMSE [m/s]	wind direction [°]	RMSE [°]
db18	0.417	sym5	23.858
Anemometric height 150 m			
Wind speed [m/s]	RMSE [m/s]	wind direction [°]	RMSE [°]
db18	0.402	db38	25.286

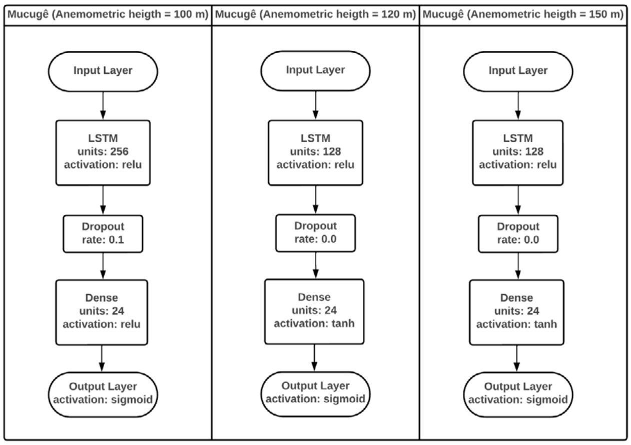


Fig. 7. The best parameters for fitting the LSTM model with Hyperband for Mucugê for anemometers at heights of 100.0, 120.0, and 150.0 m.

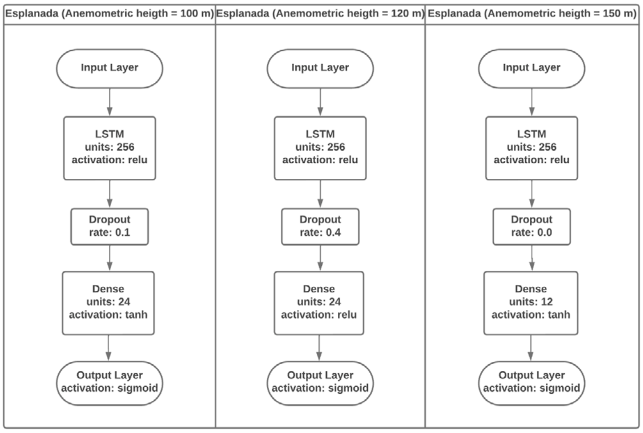


Fig. 8. The best parameters for fitting the LSTM model with Hyperband for Esplanada for anemometers at heights of 100.0, 120.0, and 150.0 m.

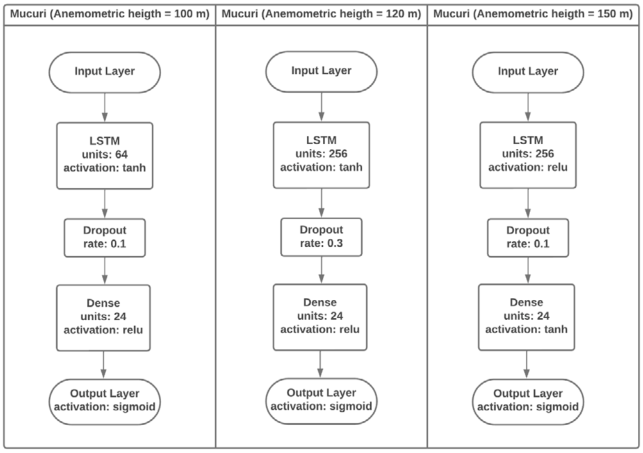


Fig. 9. The best parameters for fitting the LSTM model with Hyperband for Mucuri for anemometers at heights of 100.0, 120.0, and 150.0 m.

transformer models for each anemometer height in Mucugê, Esplanada and Mucuri.

Tables S.1, S.2 and S.3 present the quantitative analysis of the comparison between the persistence, LSTM and transformer models for Mucuge, Esplanada e Mucuri anemometric towers and each height, respectively. Figures S.6, S.7 and S.8 illustrate the results of the predictions of wind speed made by the persistence model for Mucugê at 150 m, Esplanada at 120 m and Mucuri at 100 m, respectively.

Table 8 present the prediction intervals for all forecasting horizons for LSTM and transformer models measured in the test dataset for each anemometric tower at selected heights. In addition, Figures S.9, S.10 and S.11 present a graphical representation of the prediction intervals made by the LSTM model for Esplanada, Mucuge and Mucuri, respectively. The same analysis is presented for the transformer model in Figures S.12, S.13 and S.14 for the same anemometric towers.

Figures S.15 and S.16 show, at left, the wind speed predicted by the LSTM and transformer models in Mucugê's anemometer at height of 150.0 m, respectively, and, at right, the corresponding calculated wind power. We performed the same analysis for Esplanada's anemometer at 120.0 m considering both LSTM and transformer models in Figures S.17 and S.18, respectively, and for Mucuri's anemometer at 100.0 m for LSTM and transformer models in Figures S.19 and S.20, respectively.

Tables 9–11 present a comparative overview of the results of the statistical analyses (MSE, RMSE, MAE, Pearson's r correlation, Fac2 and NMSE) for the best results of the mean value and standard deviation of all time steps predictions, for the LSTM and transformer models in Mucugê, Esplanada and Mucuri, respectively, for the anemometers at 100.0 m, 120.0 m and 150.0 m. For these tables, *sd* is the standard deviation associated to each calculated metric. Tables S.4, S.5 and S.6 present these results in deeper details including the metrics for wind speed forecasting for each forecasting horizon. Table 12 presents a quantitative comparison of the results of the LSTM and transformer models with previous similar state-of-the-art works [5,6] applied to the same dataset for wind speed and wind power forecasting. Finally, Table 13 present the result of the Wilcoxon signed-rank test to assess whether the LSTM and the proposed transformer-based model are statistically different from each other.

3.1. Discussion

Following the described procedure for selecting the best wavelet family for each feature, the reconstruction of the original meteorological signals (wind speed, wind direction, air temperature, air humidity and air pressure) using wavelets was mostly carried out by the Daubechies family, demonstrating greater precision between the original and the reconstructed signals at level 1. Among the 48 wavelet functions, in addition to Daubechies (or db), other families showed significant results for the reconstruction of the signals, which was the discrete Meyer (dmey), Symlet (sym), Biorthogonal (bior) and Reverse Biorthogonal (rbio). The application of feature augmentation using wavelet decomposition in the weather signals has significantly improved the forecasting performance of the models. The best results found in practically all simulations with exception of two cases involved the addition of signals reconstructed in one or more levels through the wavelet transform.

We fine-tuned the proposed and the baseline models in order to guarantee that each model would perform with an improved set of hyperparameters. In almost all cases studied in this work, the proposed transformer architecture presented the best results, in relation to the metrics compared to the baseline and persistence models, with the exception of one case that it performed slightly worse than the LSTM model. In all cases, the transformer-based models presented more stable results when looking forward into the future, for all forecasting horizons, than the baseline model. This is an empirical evidence of the greater ability of the multi-head attention mechanism, along with the proposed transformer architecture for wind speed forecasting, which expands the model's ability to focus on information from distinct positions in time. For the case studies presented in this work, in general, values greater than or equal to five heads showed better results. We also observed that, in the configurations with the best results, the transformer model, in most cases, performed much better than the

Table 8
Average prediction error intervals for all forecasting horizons for LSTM and transformer models in the test dataset.

Site	Height (m)	Model	Prediction Interval (+/−)	Mean prediction Interval lower bound	Mean Prediction	Mean prediction Interval upper bound
Mucuri	100.0	LSTM	2.7603	4.3069	7.0672	9.8276
Mucuri	100.0	Transformer	2.8923	5.4738	8.3661	11.2584
Esplanada	120.0	LSTM	2.0031	2.5304	4.5336	6.5368
Esplanada	120.0	Transformer	2.3842	3.1319	5.5162	7.9004
Mucugê	150.0	LSTM	2.7674	5.4886	8.2561	11.0236
Mucugê	150.0	Transformer	3.4241	5.2223	8.6464	12.0705

Table 9
Results of the statistical analyses (MSE, RMSE, MAE, Pearson's r correlation, Fac2 and NMSE) for the best results of the wind speed forecasting by the LSTM and the transformer models in Mucugê for anemometers at heights of 100.0, 120.0 and 150.0 m. Best mean values for each metric are highlighted in bold.

	MSE		RMSE		MAE		Pearson R		Fac2		NMSE	
	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.
Mucugê for anemometer at height of 100.0 m												
Mean	4.846	**3.473**	2.167	**1.844**	1.703	**1.469**	0.674	**0.780**	0.960	**0.968**	0.554	**0.396**
sd	1.634	1.113	0.421	0.293	0.370	0.235	0.139	0.094	0.017	0.009	0.190	0.130
Mucugê for anemometer at height of 120.0 m												
Mean	**4.709**	5.526	**2.133**	2.325	**1.722**	1.837	**0.701**	0.678	**0.950**	0.922	**0.547**	0.642
sd	1.742	1.759	0.437	0.379	0.378	0.323	0.112	0.119	0.014	0.030	0.206	0.210
Mucugê for anemometer at height of 150.0 m												
Mean	5.876	**4.568**	2.369	**2.133**	1.912	**1.721**	0.642	**0.713**	0.920	**0.952**	0.653	**0.506**
sd	2.363	0.620	0.561	0.141	0.472	0.126	0.149	0.055	0.041	0.005	0.267	0.072

Table 10
Results of the statistical analyses (MSE, RMSE, MAE, Pearson's r correlation, Fac2 and NMSE) for the best results of the wind speed forecasting by the LSTM and the transformer models in Esplanada for anemometers at heights of 100.0, 120.0 and 150.0 m. Best mean values for each metric are highlighted in bold.

	MSE		RMSE		MAE		Pearson R		Fac2		NMSE	
	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.
Esplanada for anemometer at height of 100.0 m												
Mean	2.775	**1.429**	1.641	**1.192**	1.317	**0.902**	0.421	**0.618**	0.937	**0.983**	1.271	**0.652**
sd	0.939	0.222	0.311	0.096	0.269	0.079	0.166	0.075	0.050	0.006	0.446	0.113
Esplanada for anemometer at height of 120.0 m												
Mean	2.999	**1.538**	1.705	**1.236**	1.367	**0.961**	0.462	**0.598**	0.949	**0.991**	1.250	**0.638**
sd	1.043	0.261	0.330	0.107	0.283	0.083	0.157	0.100	0.045	0.002	0.457	0.123
Esplanada for anemometer at height of 150.0 m												
Mean	3.067	**1.909**	1.717	**1.376**	1.352	**1.071**	0.459	**0.569**	0.977	**0.991**	1.086	**0.672**
sd	1.238	0.360	0.372	0.132	0.324	0.096	0.197	0.128	0.014	0.005	0.462	0.146

Table 11
Results of the statistical analyses (MSE, RMSE, MAE, Pearson's r correlation, Fac2 and NMSE) for the best results of the wind speed forecasting by the LSTM and the transformer models in Mucuri for anemometers at heights of 100.0, 120.0 and 150.0 m. Best mean values for each metric are highlighted in bold.

	MSE		RMSE		MAE		Pearson R		Fac2		NMSE	
	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.
Mucuri for anemometer at height of 100.0 m												
Mean	2.737	**2.094**	1.619	**1.437**	1.318	**1.148**	**0.904**	0.889	0.982	**0.995**	0.324	**0.248**
sd	1.096	0.521	0.368	0.180	0.309	0.154	0.027	0.028	0.017	0.004	0.128	0.058
Mucuri for anemometer at height of 120.0 m												
Mean	2.869	**2.048**	1.647	**1.423**	1.343	**1.144**	**0.885**	0.879	0.991	**0.997**	0.367	**0.262**
sd	1.296	0.458	0.430	0.159	0.358	0.125	0.032	0.036	0.012	0.002	0.163	0.055
Mucuri for anemometer at height of 150.0 m												
Mean	3.285	**2.109**	1.771	**1.438**	1.418	**1.153**	0.848	**0.864**	0.985	**0.997**	0.454	**0.291**
sd	1.362	0.605	0.422	0.216	0.349	0.194	0.040	0.032	0.009	0.004	0.184	0.080

LSTM model for longer forecasting windows, from 3 to 6 h ahead. As expected, the persistence model showed worse results than both LSTM and transformer-based models.

In some cases, we saw a noticeable difference between the computational time required to training the models. The transformer-based model presented faster training time, since the computations made in its training are executed concurrently due to its parallelised architecture [12]. Indeed, this is one of the main advantages of the transformers, which enabled them to training large foundation models in the research field of natural language processing. Consequently, the model converges in fewer epochs. In some situations, the training time for the transformer model was longer just enough to reach the early stopping condition, but with clear convergence at earlier epochs. And, regarding the inference time, i.e. the time to make predictions after the models were trained, it is evident that there is no significant difference, as the models perform similarly for all cases. Despite of this, it is noticeable that the average inference time per sample is no greater than 10 ms, which is one of the major advantages of using deep learning models for time series forecasting, because, after their training and if

Table 12
Comparison of statistical analysis results between AI models from previous state-of-the-art similar works.

Paper	Mean RMSE	Mean Pearson's r	Mean Fac2
MLP [5]	1.7897	0.6502	0.9656
RNN + Wavelets [5]	1.3809	0.7928	0.9855
RNN + Wavelets [6]	–	0.7689	–
Fine-tuned LSTM [this work]	1.7864	0.6876	0.9653
Proposed transformer-based model [this work]	1.5871	0.7393	0.9772

Table 13
Results of the Wilcoxon signed-rank test comparing the predictions made by the LSTM model with the proposed transformer-based model.

Anemometric tower (all heights)	*p*-value	Test result
Esplanada	1.848376085273096e−13	reject H0: different distribution
Mucugê	8.808823247154469e−11	reject H0: different distribution
Mucuri	4.5027702670859653e−35	reject H0: different distribution

they are suitable for the task they were designed, the time to perform an inference, i.e. a forecast, is negligible.

When analysing the loss curves of each model, for each case, it is clear that the transformer-based architecture achieves superior performance for generalisation purposes, presenting no under/overfitting, with convergence state being reached in fewer epochs. On the contrary, the baseline model suffered of overfitting in all the anemometric heights of Esplanada and Mucugê, taking even more epochs to train successfully. This highlights the ability of the proposed transformer-based architecture to learn more general temporal patterns, enabling it to better generalise and thus avoiding under/overfitting issues.

The quantitative and qualitative analysis of the prediction error intervals for both LSTM and transformer models showed that they present similar behaviour for all forecasting horizons and anemometric towers, considering a confidence of $q = 0.95$ for the quantile regression approach. Furthermore, the proposed transformer-based architecture was capable of consistently predict the seasonal variation and behaviour of the wind speed and wind power throughout different periods of the year for each anemometer and height, which is an important feature needed for forecasting models that rely on meteorological data to perform their predictions.

The quantitative statistical analysis comparing the final performance of the LSTM and transformer-based models showed that, in general, the proposed transformer architecture was superior than the fine-tuned LSTM model, with the exception of Mucugê at 120 m (all metrics) and for Mucuri at 100 m and 120 m (only Pearson's r metric), where LSTM presented better results. This presents an empirical evidence of the superiority of the proposed transformer architecture for multivariate wind speed and power forecasting over the baseline fine-tuned LSTM model. Moreover, when comparing their results with similar previously published works, it can be noticed that they achieve comparable performance with other state-of-the-art models. Finally, the assessment of the statistical difference between the predictions of both the LSTM and proposed transformer model using the Wilcoxon signed-rank test showed that the null hypothesis $H0$ was rejected for all anemometric towers, which means that their predictions are statistically different from each other.

It is noticeable, therefore, that the proposed transformer model presented superior performance than the fine-tuned LSTM model, with both achieving performance comparable to similar state-of-the-art works for this same problem. Notwithstanding, despite of its overall superiority for wind speed and power forecasting, there are some limitations that must be highlighted, based on the results. Firstly, the proposed transformer model was not able to beat the fine-tuned LSTM model for Mucugê at 120 m for all metrics and for Mucuri at 100 m and 120 m for metric Pearson's r. Additionally, as the prediction intervals were equivalent for both, it presents an opportunity to further increase the proposed transformer model's performance by minimising its prediction error intervals, consequently improving its prediction's reliability.

4. Conclusions

In this paper, we proposed a new transformer-based architecture to develop a deep neural network model integrated with wavelet transform for wind speed and energy forecasting for 6 h ahead, comparing the results with the LSTM architecture as a baseline model, using a hyperparameter optimisation technique to guarantee a fair comparison between the models. The model is able to deal with multiple meteorological variables as input.

In this work, we applied the proposed modelling framework to a case study using data collected from anemometric towers at heights of 100.0, 120.0 and 150.0 m, in three different locations in the State of Bahia, Brazil, namely Mucugê, Esplanada and Mucuri. The performance of the models was assessed by the statistical metrics MAE, MSE, RMSE, NMSE, Fac2 and Pearson's r, along with the evaluation of the time required for training and for making inferences (or forecasts) by the proposed and the baseline models. We performed qualitative evaluations of the forecasts performed using graphical analysis of the predictions of wind speed, as well as wind power generation based on hypothetical wind turbines previously studied in the literature.

Based on the proposed methodology and presented results, we can summarise the main conclusions drawn from this work, as follows:

1. The proposed transformer-based architecture is effective for multivariate wind speed and wind power forecasting using multiple meteorological variables as input;
2. The integration of the proposed transformer model with feature augmentation using wavelet decomposition has successfully improved the forecast accuracy;
3. The employment of a systematic procedure to choose the best wavelet family for decomposing the time series of each input meteorological variable proved to be successful;
4. The proposed transformer architecture has better performance than the LSTM model for this application, even after performing hyperparameter optimisation, guaranteeing a fair comparison between the approaches;
5. The transformer model also proved to generalise well during the training and validation phases, presenting no under/overfitting, in contrast with the baseline model;
6. The proposed transformer architecture trains faster than the LSTM model, with both presenting negligible inference time per sample, leading to a more energy-efficient and sustainable solution, thus contributing to reduce carbon footprint and to achieve net zero carbon emissions targets;
7. The transformer model performed better than the LSTM model, with a more stable performance for longer forecasting horizons, noticeably from the 3^{rd} to the 6th hours ahead, when the LSTM's performance considerably drops;
8. Both LSTM and proposed transformer models achieved performance comparable with state-of-the-art similar works addressing the same problem;
9. The proposed transformer model, integrated with wavelet transform, has the potential to be applied and adapted to other multivariate time series forecasting tasks.

Based on the proposed methodology, investigations and on the findings, we believe that this research can contribute to foster the adoption and operationalisation of wind farms around the world, with a more robust, faster and reliable technique for wind power forecasting, especially in a time when climate change issues due to air pollution caused by non-renewable energy generation are on the agenda, and sustainable development is increasingly necessary.

CRediT authorship contribution statement

Erick Giovani Sperandio Nascimento: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Talison A.C. de Melo:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Davidson M. Moreira:** Formal analysis, Funding acquisition, Investigation, Resources.

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

The authors do not have permission to share data.

Acknowledgements

We thank the Reference Center on Artificial Intelligence (CRIA) and the Supercomputing Center for Industrial Innovation (CS2i), both from SENAI CIMATEC, for providing all the technical and infrastructure support. We also thank the NVIDIA/CIMATEC AI Joint Lab for the technical support, and the Surrey Institute for People-Centred AI and GCARE, both from the University of Surrey (UK), for their institutional support. We thank the National Council for Scientific and Technological Development (CNPq, Brazil), for partially fund this work. Erick G. Sperandio Nascimento is a CNPq technological development fellow (Proc. 308963/2022-9).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.energy.2023.127678.

References

[1] Agency IE. World energy balances 2020. 2021, URL https://bityli.com/txFAS.

[2] Oğulata RT. Energy sector and wind energy potential in Turkey. Renew Sustain Energy Rev 2003;7(6):469–84.

[3] Brown BG, Katz RW, Murphy AH. Time series models to simulate and forecast wind speed and wind power. J Appl Meteorol Climatol 1984;23(8):1184–95.

[4] Torres JL, Garcia A, De Blas M, De Francisco A. Forecast of hourly average wind speed with ARMA models in Navarre (Spain). Sol Energy 2005;79(1):65–77.

[5] Zucatelli PJ, Nascimento EGS, Santos AÁB, Moreira DM. Nowcasting prediction of wind speed using computational intelligence and wavelet in Brazil. Int J Comput Methods Eng Sci Mech 2020;21(6):343–69.

[6] Zucatelli PJ, Nascimento EGS, Santos A, Arce A, Moreira D. An investigation on deep learning and wavelet transform to nowcast wind power and wind power ramp: A case study in Brazil and Uruguay. Energy 2021;230:120842.

[7] Chen G, Tang B, Zeng X, Zhou P, Kang P, Long H. Short-term wind speed forecasting based on long short-term memory and improved BP neural network. Int J Electr Power Energy Syst 2022;134:107365.

[8] Niu Z, Yu Z, Tang W, Wu Q, Reformat M. Wind power forecasting using attention-based gated recurrent unit network. Energy 2020;196:117081.

[9] Zhao X, Jiang N, Liu J, Yu D, Chang J. Short-term average wind speed and turbulent standard deviation forecasts based on one-dimensional convolutional neural network and the integrate method for probabilistic framework. Energy Convers Manage 2020;203:112239.

[10] Chen Y, Zhang S, Zhang W, Peng J, Cai Y. Multifactor spatio-temporal correlation model based on a combination of convolutional neural network and long short-term memory neural network for wind speed forecasting. Energy Convers Manage 2019;185:783–99.

[11] Sideratos G, Hatziargyriou ND. An advanced statistical method for wind power forecasting. IEEE Trans Power Syst 2007;22(1):258–65.

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems. 2017, p. 5998–6008.

[13] Xiong R, Yang Y, He D, Zheng K, Zheng S, Xing C, Zhang H, Lan Y, Wang L, Liu T. On layer normalization in the transformer architecture. In: International conference on machine learning. PMLR; 2020, p. 10524–33.

[14] Fan H, Xiong B, Mangalam K, Li Y, Yan Z, Malik J, Feichtenhofer C. Multiscale vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 6824–35.

[15] Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L. Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 22–31.

[16] Li Y, Zhang K, Cao J, Timofte R, Van Gool L. Localvit: Bringing locality to vision transformers. 2021, arXiv preprint arXiv:2104.05707.

[17] Chu X, Tian Z, Zhang B, Wang X, Wei X, Xia H, Shen C. Conditional positional encodings for vision transformers. 2021, arXiv preprint arXiv:2102.10882.

[18] Heo B, Yun S, Han D, Chun S, Choe J, Oh SJ. Rethinking spatial dimensions of vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 11936–45.

[19] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 12179–88.

[20] Mohammdi Farsani R, Pazouki E. A transformer self-attention model for time series forecasting. J Electr Comput Eng Innov (JECEI) 2021;9(1):1–10.

[21] Wu N, Green B, Ben X, O'Banion S. Deep transformer models for time series forecasting: The influenza prevalence case. 2020, arXiv preprint arXiv:2001.08317.

[22] Sridhar S, Sanagavarapu S. Multi-head self-attention transformer for dogecoin price prediction. In: 2021 14th international conference on human system interaction. IEEE; 2021, p. 1–6.

[23] Zhou H, Zhang S, Zhao X. Condition time series prediction of aero-engine gas-path performance based on self-attention mechanism. In: 2021 40th Chinese control conference. CCC, IEEE; 2021, p. 6219–24.

[24] Qu K, Si G, Shan Z, Kong X, Yang X. Short-term forecasting for multiple wind farms based on transformer model. Energy Rep 2022;8:483–90. http://dx.doi.org/10.1016/j.egyr.2022.02.184, ICPE 2021 - The 2nd international conference on power engineering, URL https://www.sciencedirect.com/science/article/pii/S2352484722004310.

[25] Wu B, Wang L, Zeng Y-R. Interpretable wind speed prediction with multivariate time series and temporal fusion transformers. Energy 2022;252:123990. http://dx.doi.org/10.1016/j.energy.2022.123990, URL https://www.sciencedirect.com/science/article/pii/S0360544222008933.

[26] Lim B, Arik SÖ, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int J Forecast 2021;37(4):1748–64. http://dx.doi.org/10.1016/j.ijforecast.2021.03.012, URL https://www.sciencedirect.com/science/article/pii/S0169207021000637.

[27] Wang L, He Y, Liu X, Li L, Shao K. M2TNet: Multi-modal multi-task transformer network for ultra-short-term wind power multi-step forecasting. Energy Rep 2022;8:7628–42. http://dx.doi.org/10.1016/j.egyr.2022.05.290, URL https://www.sciencedirect.com/science/article/pii/S2352484722011374.

[28] Godoi M. Energia eólica chega a 18 GW de capacidade instalada no Brasil (in portuguese). 2021, URL https://bityli.com/DjTeIAT.

[29] Cherem C. Brasil experimenta a maior crise hídrica em 91 anos (in portuguese). 2021, URL https://bityli.com/sCFYnWy.

[30] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proceedings of the 3rd international conference on learning representations. 2015.

[31] Galvão SLJ, Matos JCO, Kitagawa YKL, Conterato FS, Moreira DM, Kumar P, Nascimento EGS. Particulate matter forecasting using different deep neural network topologies and wavelets for feature augmentation. Atmosphere 2022;13(9):1451.

[32] Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. J Mach Learn Res 2017;18(1):6765–816.

[33] Iverson KE. A programming language. In: Barnard III G, editor. Proceedings of the 1962 spring joint computer conference, AFIPS 1962 (Spring), San Francisco, California, USA. ACM; 1962, p. 345–51. http://dx.doi.org/10.1145/1460833.1460872.

[34] Hoshmand AR. Business forecasting: A practical approach. Routledge; 2009.

[35] Corder GW, Foreman DI. Nonparametric statistics for non-statisticians. John Wiley & Sons, Inc.; 2011.

[36] Kalmikov A. Wind power fundamentals. In: Wind energy engineering. Elsevier; 2017, p. 17–24.

[37] Gölçek M, Erdem HH, Bayülken A. A techno-economical evaluation for installation of suitable wind energy plants in Western Marmara, Turkey. Energy Explor Exploit 2007;25(6):407–27.