sustainability

MDPI

Article

Wind Power Forecasting Based on WaveNet and Multitask Learning

Hao Wang [1], Chen Peng [2,*], Bolin Liao [2], Xinwei Cao [3] and Shuai Li [4]

[1] School of Communication and Electronic Engineering, Jishou University, Jishou 416000, China; 2021700516@stu.jsu.edu.cn
[2] School of Computer Science and Engineering, Jishou University, Jishou 416000, China; bolinliao@jsu.edu.cn
[3] School of Business, Jiangnan University, Wuxi 214122, China; xwcao@jiangnan.edu.cn
[4] Faculty of Information Technology and Electrical Engineering, University of Oulu, 90307 Oulu, Finland; shuai.li@oulu.fi
* Correspondence: chen.peng@jsu.edu.cn; Tel.: +86-137-2406-2118

Abstract: Accurately predicting the power output of wind turbines is crucial for ensuring the reliable and efficient operation of large-scale power systems. To address the inherent limitations of physical models, statistical models, and machine learning algorithms, we propose a novel framework for wind turbine power prediction. This framework combines a special type of convolutional neural network, WaveNet, with a multigate mixture-of-experts (MMoE) architecture. The integration aims to overcome the inherent limitations by effectively capturing and utilizing complex patterns and trends in the time series data. First, the maximum information coefficient (MIC) method is applied to handle data features, and the wavelet transform technique is employed to remove noise from the data. Subsequently, WaveNet utilizes its scalable convolutional network to extract representations of wind power data and effectively capture long-range temporal information. These representations are then fed into the MMoE architecture, which treats multistep time series prediction as a set of independent yet interrelated tasks, allowing for information sharing among different tasks to prevent error accumulation and improve prediction accuracy. We conducted predictions for various forecasting horizons and compared the performance of the proposed model against several benchmark models. The experimental results confirm the strong predictive capability of the WaveNet–MMoE framework.

Keywords: wind turbine power forecasting; WaveNet; multitask learning; multigate mixture-of-experts; multistep time series forecasting; maximum information coefficient; wavelet transform

Citation: Wang, H.; Peng, C.; Liao, B.; Cao, X.; Li, S. Wind Power Forecasting Based on WaveNet and Multitask Learning. *Sustainability* **2023**, *15*, 10816. https://doi.org/10.3390/su151410816

Academic Editors: Hua Li and Francisco Haces-Fernandez

Received: 25 May 2023
Revised: 30 June 2023
Accepted: 5 July 2023
Published: 10 July 2023

1. Introduction

Wind power forecasting is a crucial mean to ensure the reliable and efficient operation of large-scale power systems. Accurate wind turbine power forecasting can help grid operators to better manage the integration of wind energy into the grid, reduce costs, and improve system stability. The commonly used methods for time series forecasting include physics models, statistical models, and machine learning methods.

Physics models are derived based on fundamental laws and principles with the aim of capturing the underlying mechanisms and dynamic characteristics of the predicted system [1]. However, physics models often rely on simplifying assumptions and may not fully capture the complexity of real-world systems [2]. They may require precise knowledge of system parameters, which is often difficult to obtain. Additionally, physics models are highly sensitive to errors and uncertainties, especially when faced with sudden changes, unexpected events, or system variations. In such situations, these models may struggle to adapt and accurately forecast future trends [3].

On the other hand, statistical models are data driven and rely on historical patterns and statistical techniques for forecasting. These models analyze past observations and identify patterns and correlations to make future predictions [4]. While statistical models

offer flexibility and the ability to capture complex relationships within the data, they also have limitations. One of the main drawbacks of statistical models is their assumption of data stationarity, where the statistical properties remain constant over time [5]. However, in real-world scenarios, data may exhibit nonstationary behavior, such as trends, seasonality, or abrupt changes, which statistical models may struggle to capture. Additionally, statistical models may face issues of overfitting or underfitting, leading to poor generalization and unreliable predictions [6].

With the rapid development of machine learning (ML) technology, its application in the field of forecasting has become increasingly widespread [7–10]. Machine learning methods leverage large amounts of data and powerful algorithmic capabilities to automatically learn patterns and regularities from the data, which can be used to predict future trends and outcomes. In [11], the authors compared the predictive performance of traditional time series models, such as autoregressive integrated moving average (ARIMA) and exponential smoothing (ETS), and machine learning models, such as eXtreme gradient boosting (XGBoost) and neural networks, on different datasets. The results show that machine learning models have improved accuracy and flexibility compared with traditional time series models; however, there are still challenges in terms of interpretability and stability.

Machine learning models heavily rely on feature selection and feature engineering. Feature selection involves selecting the most relevant features from the raw data, while feature engineering involves transforming and combining the raw data to extract more valuable features. These processes require domain knowledge and expertise and can have a significant impact on the performance of the model [12]. Additionally, the presence of noise in the data is another important issue that prevents the accurate prediction of time series, which asks for an effective noise-suppressing technique to enhance the prediction performance.

A challenge in forecasting is the task of multistep forecasting, which refers to predicting multiple future time steps in a time series prediction problem [13–16]. It involves forecasting the values or states of a time series for several consecutive time steps ahead. In contrast to the single-step forecasting task, which focuses on predicting only the next time step, the multistep forecasting task aims to provide predictions for an extended range. Generally speaking, for multistep forecasting tasks, an iterative prediction procedure [17] is commonly used, which utilizes the result of each prediction as an input to predict the next time step. However, this can lead to error accumulation as the length of the time series increases. In contrast to the iterative prediction procedure, multistep time series forecasting (MSTF) treats multistep forecasting as a multi-output problem and predicts all outputs in one forward pass, which effectively avoids error accumulation. The downside is that this requires a more complicated model, which increases both space and time complexity.

In machine learning, MTL is a subfield that simultaneously solves multiple related tasks and improves learning efficiency and prediction accuracy by sharing model parameters. MTL has been successfully applied in various machine learning applications. For example, Carlos Busso [18] utilized the MTL method to jointly learn emotional attributes by utilizing their mutual dependencies. To overcome the limitation of multiple machine learning models for predicting multiple metrics in a single indoor space leading to contradictory predictions, Betty et al. proposed a deep-neural-network-based multitask learning model called DeepComfort in 2022 [19]. DeepComfort simultaneously predicts multiple TC output metrics, namely, TSV, TPV, and TCV, using a single model. It was validated on the ASHRAE-II database and a dataset created in this study for elementary school students. Despite facing challenges of illogical responses and data imbalance, it exhibited high F1 scores, accuracy (approximately 90%), and generalization ability. However, the application of MTL in MSTF tasks has not been explored in the literature yet. Further research is needed to demonstrate the advantages of the multitask learning approach.

In summary, the challenges addressed in this paper are listed as follows:

- In machine learning, appropriate feature selection methods are crucial for model performance. Additionally, the presence of noise in the data can also impact the performance of the model.
- Although MSTF mitigates the issue of error accumulation commonly encountered in time series prediction tasks, it necessitates the careful selection of appropriate models to effectively capture and exploit the intricate patterns and trends present in time series data, thereby facilitating improved forecasting accuracy.
- The combination of MSTF and MTL, as well as their impact on prediction results, has not been investigated in the field of wind turbine power prediction.

To address these challenges, this paper proposes a new framework, WaveNet–MMoE, that combines a special convolutional neural network, WaveNet, with a multitask learning architecture, multigate mixture-of-experts (MMoE). First, in feature engineering, the maximum information coefficient (MIC) method is employed for feature selection, while the wavelet transform technique is utilized to reduce the impact of noise. Second, WaveNet is employed to extract the representation of wind power data, leveraging its unique dilated convolutional network to capture distant time information and enhance model performance. Finally, these representations are used as inputs to MMoE, treating the multistep prediction task as a collection of different but interconnected tasks, thereby reducing error accumulation and improving prediction accuracy.

The contributions of this paper are as follows:

- The MIC method is adopted to rank the correlation of features, effectively eliminating redundant information while retaining important features. Additionally, the wavelet transform technique is utilized to remove noise present in the data.
- By leveraging the MTL framework MMoE, the prediction of multiple temporally correlated information is treated as a set of related yet mutually independent tasks, enabling these tasks to be executed in parallel, effectively avoiding error accumulation. Moreover, it facilitates the information sharing among different tasks, thereby improving prediction accuracy.
- This paper investigates the integration of MSTF and MTL in the field of wind turbine power prediction, along with the challenges arising from increased complexity in the process.

The remainder of this paper is structured as follows: Section 2 provides a literature review. Section 3 provides a detailed explanation of the data normalization process and feature selection based on the maximum information coefficient (MIC). Additionally, it discusses the utilization of wavelet transformation to remove noise from the data. Section 4 presents the new framework and its components. Section 5 showcases the experimental results. Section 6 discusses future work and concludes this paper.

2. Literature Review

Current wind turbine power forecasting methods in the literature can be divided into three categories: physical methods, statistical methods, and ML algorithms [20–22]. Physical methods solve complex weather pattern considerations by converting predicted meteorological parameters into wind speed curves to infer trends in wind power series [23]. However, due to the complex nature of calculation processes and atmospheric conditions, physical models have certain limitations.

Statistical methods generally use nonlinear and linear relationships between wind speed, wind direction and temperature, and power generation, etc., for time series forecasting. Autoregressive (AR) models [24], autoregressive moving average (ARMA) models [25], and autoregressive integrated moving average (ARIMA) models [26] perform well in processing inference problems by studying statistical regularities in wind power data. However, due to the randomness and intermittency of wind power series, statistical methods still have room for improvement.

In recent years, ML technology [27–34] has developed rapidly and has been widely applied in various fields, such as mechanics [35–38], medicine [39–42], and energy [43–46]. The advantage of ML technology in prediction lies in its ability to handle nonlinear relationships and complex data patterns between input variables and output variables [47]. Furthermore, ML technology can automatically learn features from data without the need for manual feature extraction and selection. This makes ML technology more adaptable and flexible when dealing with large-scale, high-dimensional, and complex data [48,49].

The current mainstream of ML forecasting methods primarily uses neural networks [50,51] technology. Different neural network models, including convolutional neural network (CNN) [52,53], recurrent neural networks (RNN) [54], and later Transformers [55], have had a huge impact in the field of wind power forecasting applications. When dealing with time series, traditional ML techniques can be impacted, resulting in network performance issues that affect the accuracy and stability of the model [56]. Due to the defectiveness of causal convolution in extracting distant information, there is an urgent need for better models to solve these problems. The Tensorial Encoder Transformer (TENT) model proposed in [57] has tensorial attention, so by processing weather data in tensorial format, the spatiotemporal structure of weather data can be obtained. Compared with 3D CNNs, TENT models can better simulate complex weather data patterns that may occur in temperature forecasting tasks. In 2021, Qi et al. proposed a novel asynchronous dilated graph convolutional network (ADGCN) for traffic flow prediction. ADGCN successfully extends the dilated one-dimensional causal convolution to graph convolution. With the increase in network depth, the receptive field of the model grows exponentially. Experimental results on three public transportation datasets demonstrate that ADGCN outperforms existing corresponding methods in terms of prediction performance, particularly in long-term prediction tasks [58].

Despite the increasing utilization of extended temporal convolutional networks as forecasting models, their application in predicting wind turbine power remains largely unexplored. In 2020, Zhu et al. proposed a novel network called temporal convolutional network (TCN). The proposed method addresses the issues of long-term dependencies and performance degradation in sequence prediction tasks by leveraging dilated causal convolutions and residual connections in deep convolutional models. Simulation results demonstrate that TCN exhibits stable training and strong generalization capability. Furthermore, TCN achieves higher prediction accuracy compared with existing predictors, such as support vector machines, multilayer perceptron, long short-term memory networks, and gated recurrent unit networks [59]. In 2022, He et al. proposed a novel self-calibrating temporal convolutional network (SCTCN) for remaining useful life (RUL) prediction of wind turbine gearbox bearings. This is an improved network based on TCN, which inherits TCN's dilated causal convolutions for capturing long-term historical information and introduces self-calibration modules to focus on local information within the time series. As a result, SCTCN can learn more comprehensive historical information, leading to improved accuracy in RUL prediction. Experimental evaluations were conducted on a test rig and wind turbine gearbox for bearing RUL prediction, validating the effectiveness of the proposed approach. The experimental results demonstrate that SCTCN achieves higher prediction accuracy compared with other state-of-the-art methods [60].

Feature selection is a technique within feature engineering that plays a crucial role in the field of time series prediction, directly impacting the performance and accuracy of prediction models [61,62]. The goal of feature selection is to identify and choose the most relevant and informative features from the available set of input variables. By selecting appropriate features, it is possible to improve the accuracy of predictions and reduce the computational complexity of the model [63,64]. In 2013, Yang conducted the first study on feature selection for traffic congestion prediction. By applying feature ranking and selection techniques, only the most relevant features were retained, reducing the data dimensionality and improving prediction performance. Experimental results demonstrated that using the optimally selected features for prediction outperformed using all features [65]. In [66],

Liu et al. proposed an improved high-accuracy transient stability prediction model for power systems based on the minimum redundancy maximum relevance (mRMR) feature selection and winner-take-all (WTA) ensemble learning. Two of the most informative input features were obtained through the mRMR feature selection method. Subsequently, the WTA ensemble learning method was employed to combine the predicted results of generator electromagnetic power and bus voltage magnitude, resulting in an improved transient stability prediction model that exhibited higher prediction accuracy for unstable samples. In 2019, Naik et al. proposed an artificial neural network (ANN) regression prediction model based on the Boruta feature selection technique. After considering 33 different combinations of technical indicators for stock prediction, the Boruta feature selection technique was employed to identify relevant technical indicators. Experimental results demonstrated that using the indicators identified by the Boruta feature selection technique for stock prediction reduced the prediction error rate to 12% [67]. In 2021, Bagherzadeh et al. conducted a study on the impact of various feature selection methods on the performance of machine learning algorithms. The experiments involved seven feature selection methods: variance threshold, analysis of variance (ANOVA), mutual information (MI), Pearson correlation (PC), backward elimination (BE), random forest (RF), and least absolute shrinkage and selection operator (LASSO). The results demonstrated that machine learning algorithms based on MI exhibited the strongest capability, indicating a significant dependence of model performance on feature selection [68].

In the field of machine learning and data analysis, the interference or random fluctuations caused by noise are often regarded as challenges for models, as they can lead to inaccurate and unstable predictions [69]. Yan et al. (2018) proposed a prediction model that combines wavelet analysis with a long short-term memory (LSTM) neural network to capture complex features, such as nonlinearity, nonstationarity, and sequential correlations in financial time series. The results indicated that LSTM demonstrated superior predictive performance compared with other machine learning models, such as multilayer perceptron (MLP), support vector machine (SVM), and K-nearest neighbors (KNN). This highlights the applicability and effectiveness of LSTM in financial time series forecasting [70]. In 2020, Kim et al. proposed a deep learning model combining a denoising autoencoder and convolutional long short-term memory (LSTM) for predicting global ocean weather. The proposed model aimed to forecast ocean weather one week ahead with an average error of 6.7%. The results demonstrated that the denoising autoencoder effectively removed noise that hindered the training of the deep learning model. The proposed model showed a certain level of applicability in predicting ocean weather [71]. In 2021, Samal et al. developed a time convolutional denoising autoencoder (TCDA) network, which is a combination of a time convolutional network (TCN) and a denoising autoencoder (DAE) network. The experiment utilized the DAE network to reconstruct errors and handle missing values. The results demonstrated that compared with baseline models, such as SARIMA, FbPROPHET, ANN, SVR, CNN, LSTM, GRU, BILSTM, and BIGRU, TCDA exhibited superior predictive performance [72].

Wind turbine power forecasting is an MSTF task that involves predicting wind turbine power data at multiple time points, which is different from simple single-step time series forecasting. However, in MSTF tasks, inevitably, issues of high time complexity and space complexity arise. Increasing the complexity of a model may require more computational resources and time to perform prediction tasks, leading to longer execution times and increased computational resource demands [73]. Therefore, there is an urgent need to seek methods to reduce the complexity of models and minimize the consumption of resources and time.

Multitask learning refers to a machine learning approach that involves simultaneously learning and optimizing multiple related tasks [74]. In traditional single-task learning, models are designed to solve a specific individual task, whereas multitask learning aims to handle multiple tasks concurrently and improve model performance through task interrelationships and information sharing. Multitask learning offers several benefits, including

enhanced model generalization, accelerated training process, and reduced model complexity [75]. It is applicable in various domains, such as natural language processing, computer vision, and speech recognition, where different tasks can share underlying feature representations and collectively improve overall performance. Zhang et al. [76] proposed an MTL model with three parallel LSTM layers to jointly forecast taxi pick-up and drop-off demands, which outperforms HA, ARIMA, and single-task LSTM on real-world data. In 2017, Crichton et al. proposed a multitask model based on named entity recognition (NER). The experiment treated each dataset as a separate task and employed the multitask model for joint training. The results revealed statistically significant performance differences between the multi-output model and the single-task model for six datasets. Specifically, five datasets demonstrated significantly better performance, while one dataset showed noticeably poorer performance [77]. In 2020, Yang et al. proposed a prediction model based on multitask learning for forecasting individuals' daily activities. By combining a convolutional neural network (CNN) with bidirectional long short-term memory (Bi-LSTM) units, a parallel multitask learning model was established as the prediction model. The experimental results demonstrated that the proposed model achieved an accuracy improvement of at least 2.22% compared with the single-task learning Bi-LSTM and CNN+Bi-LSTM models. Additionally, the NMAE, NRMSE, and R^2 indicators were enhanced by at least 1.542%, 7.79%, and 1.69%, respectively [78].

3. Data Processing

3.1. Correlation Analysis Based on Maximum Information Coefficient

The data analyzed in this study were obtained from a wind farm in China. The data were recorded from a wind turbine operated for a period of 24 months from 1 September 2018 to 1 September 2020. Recordings were taken at a frequency of every 10 min. Feature selection has a significant impact on wind power forecasting. By excluding irrelevant or redundant features, the selected features can better capture the underlying dynamic characteristics of the wind power system, thereby improving the model's generalization ability and enhancing prediction accuracy. Additionally, feature selection helps reduce the dimensionality of the input space, alleviating the curse of dimensionality and improving the computational efficiency of the prediction model. Therefore, we investigate the correlation between the output power and other features in SCADA data using MIC, and carefully select the features that are most relevant to the forecasting target.

The statistical measure known as MIC quantifies the degree of linear or nonlinear correlation between two variables. It is based on the concept of mutual information (MI), which can be mathematically expressed using the following formula:

$$I(x;y) = \int p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} dxdy \tag{1}$$

$$I[x;y] \approx I[X;Y] = \sum_{X,Y} p(X,Y) \log_2 \frac{p(X,Y)}{p(X)p(Y)} \tag{2}$$

where x and y, respectively, refer to the number of intervals into which the scatter plot is partitioned, and the approach involves examining the distribution of data points within each cell to resolve the challenge of calculating joint probabilities in mutual information (MI). Specifically, a and b represent the number of grid cells used to partition the scatter plot in the x and y directions, respectively. The formula for MIC is as follows:

$$mic(x;y) = \max_{a \times b < B} \frac{I(x;y)}{\log_2 \min(a,b)} \tag{3}$$

$$MIC[x;y] = \max_{|X||Y|<B} \frac{I[X;Y]}{\log_2(\min(|X|,|Y|))} \tag{4}$$

3.2. Min–Max Normalization

Max-Min normalization, also known as data normalization or feature scaling, is a commonly used data preprocessing method. It scales the data to a specific range through linear transformation, typically mapping the data to the range [0, 1]. This method applies a linear transformation to the original data, mapping the minimum value to 0 and the maximum value to 1, while scaling the other values proportionally between 0 and 1. Before inputting the input variables and output variables into our model, it is necessary to normalize them; otherwise, the loss function of the model may not converge. In our experiments, we employ the max–min normalization method, which is formulated as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{5}$$

where x and x' represent the data before and after normalization, respectively. x_{max} and x_{min} represent the maximum and minimum of the data, respectively.

This section investigated the parameter values of the MIC between the output wind turbine power data and other features in the SCADA data, as well as performed data normalization. Here, the output wind power in SCADA data is considered as the target feature. The correlation histograms between output power and other features in the SCADA dataset are shown in Figure 1, from which it can be seen that the first five features have the strongest correlation. Figure 2 shows a scatter plot of the correlation between wind turbine power and gearbox input shaft temperature. Figure 3 displays a scatter plot of the correlation between wind turbine power and wind speed.

After applying feature selection with MIC on the original dataset, the selected features were utilized as inputs, while the output power was set as the forecasting target. The size of the sliding window in time series forecasting can be chosen according to the requirements of the task. A smaller window size allows for a more sensitive capture of short-term patterns and fluctuations, while a larger window size enables the capture of longer-term trends and periodicity. Let m denote the number of time steps predicted at a time with MSTF. An experiment was conducted using the original WaveNet model to forecast the entire test set with a sliding window size of 250 and m = 3.

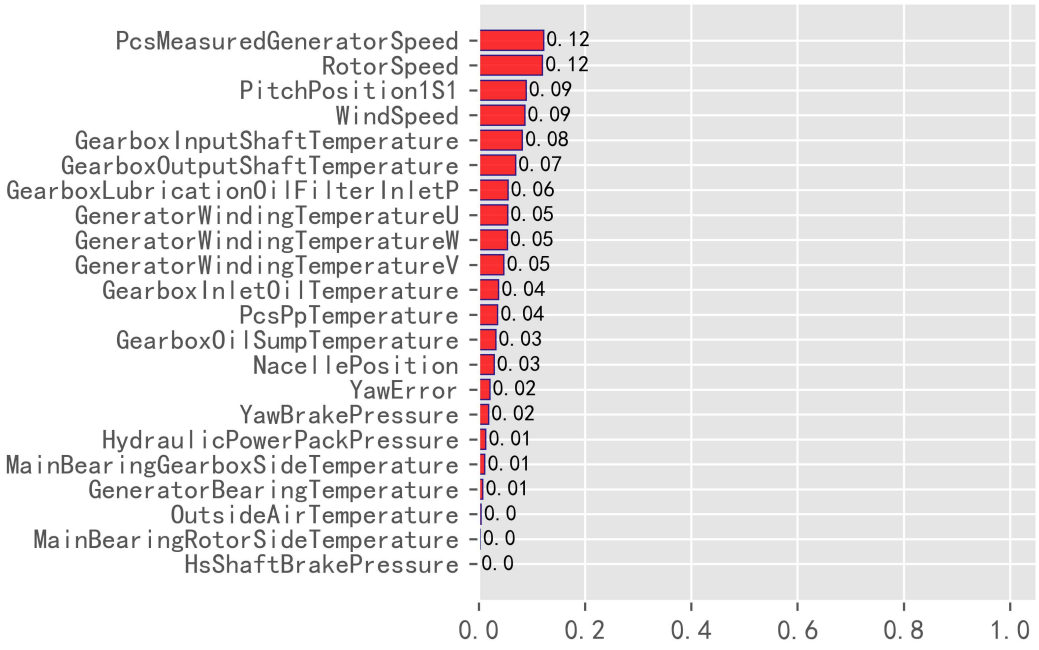


Figure 1. Maximum information factor correlation diagram for wind turbine power data.

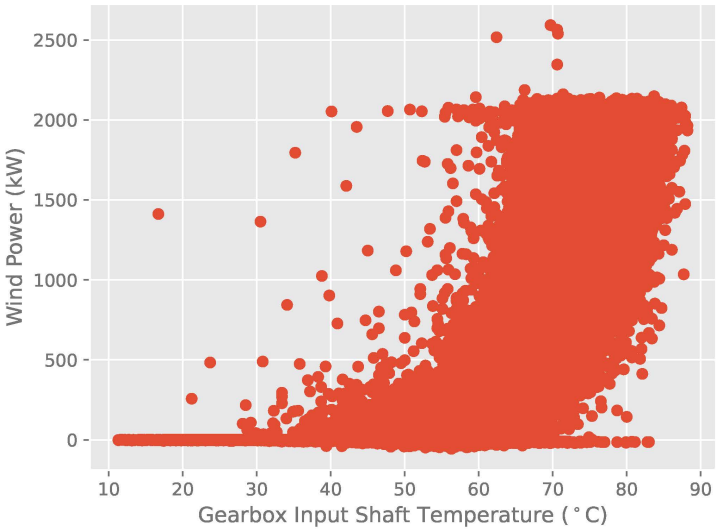


Figure 2. Scatter diagram of gearbox input shaft temperature and wind power.

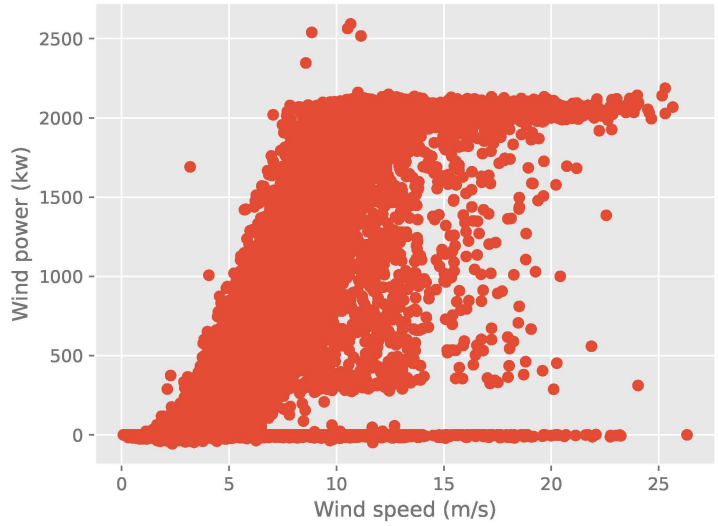


Figure 3. Scatter diagram of wind speed and wind power.

3.3. Wavelet Transform

In time series data, various types of noise often exist, such as high-frequency noise, low-frequency noise, or transient noise. Typically, this can affect the performance of prediction models. Wavelet transform can adaptively select suitable wavelet basis functions and decomposition levels based on the characteristics of the signal, thereby better adapting to different types and complexities of signals and effectively suppressing noise.

In this stage of the experiment, a wavelet transform technique was employed to remove noise from the data, aiming to enhance the predictive performance of WaveNet–MMoE. In the experiment, a wavelet level of 3 was used, and the db4 wavelet basis was chosen. Figures 4–7 illustrate the effects of noise reduction on selected feature data. The results showed that wavelet transform, when used to remove noise, resulted in data that were smoother and clearer.

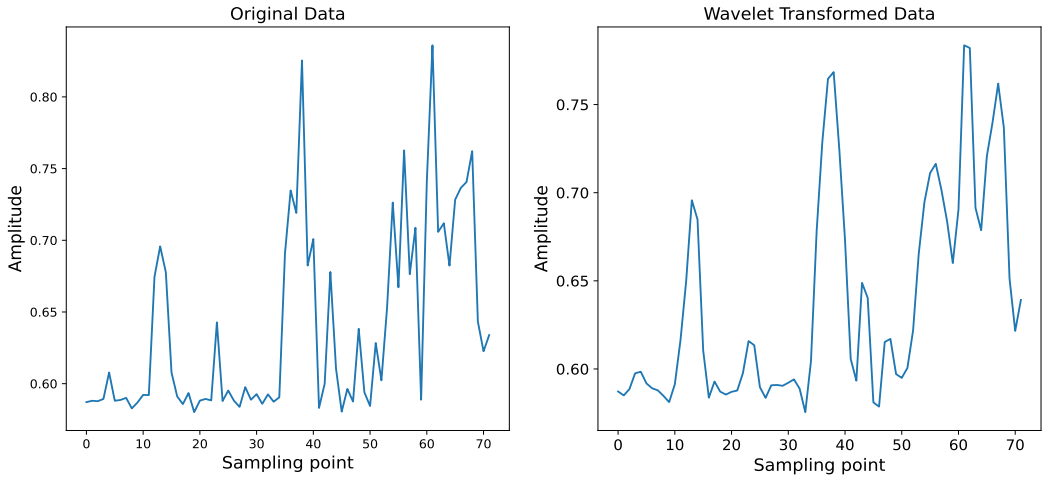


Figure 4. Denoising of pcs measured generator speed.

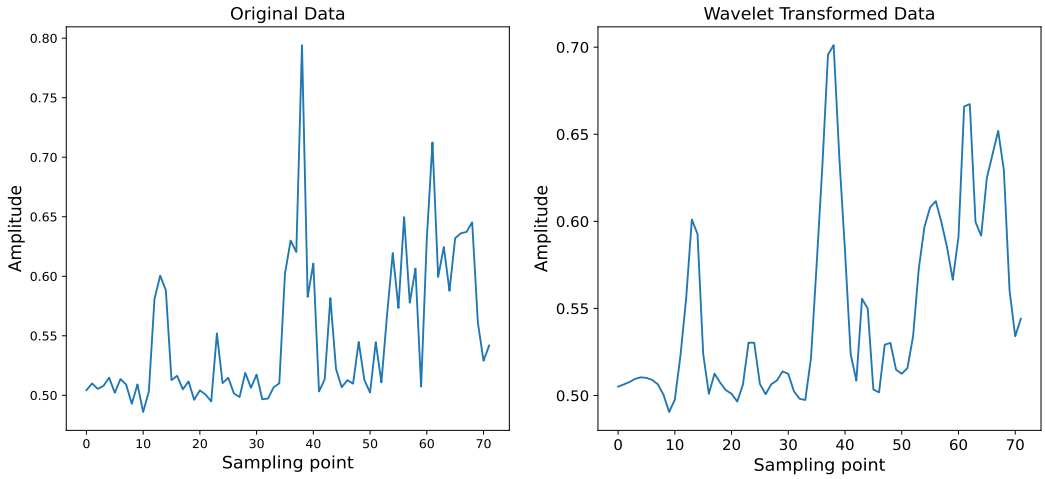


Figure 5. Denoising of rotor speed.

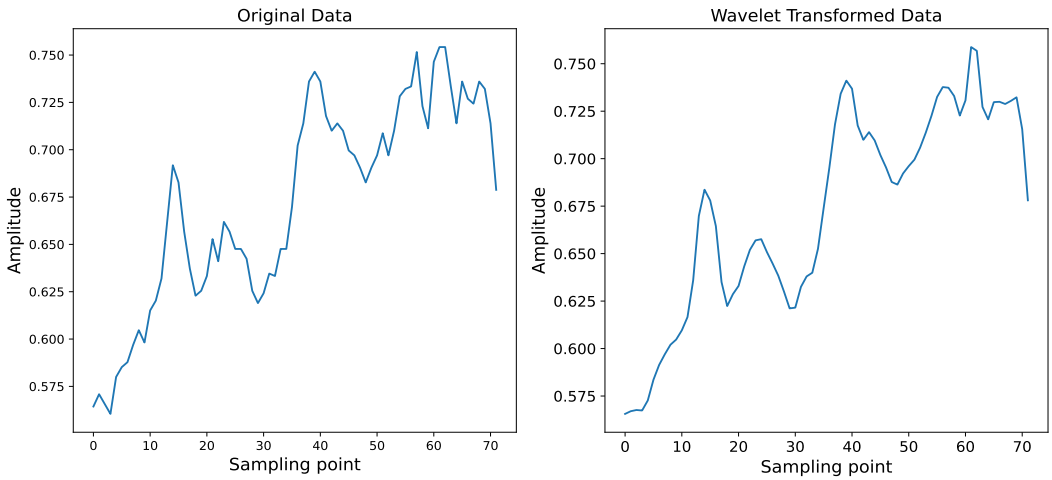


Figure 6. Denoising of gearbox input shaft temperature.

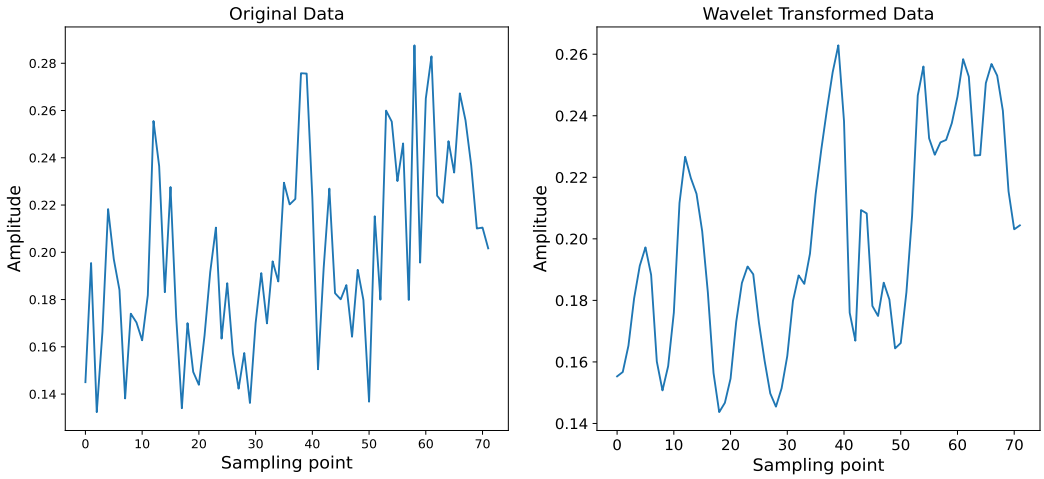


Figure 7. Denoising of wind speed.

4. The WaveNet–MMoE Architecture Based on WaveNet and MMoE

4.1. WaveNet

Many neural network models encounter difficulties when applied to real-world problems. One of the problems is the long-term dependency problem, where the model loses the ability to connect distant information as each time interval increases. WaveNet is an autoregressive probabilistic model that uses its unique dilated causal convolution. Dilated causal convolution skips some input values with a certain stride and applies the convolution kernel to an area larger than its own size, so that even with fewer layers, it has a relatively large receptive field, overcomes the drawback of causal convolution in capturing local information, and obtains more temporal information. Its modeling method is as follows:

$$p(x) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}) \tag{6}$$

where $x_1, \ldots, x_{t-1}$ denotes the sound wave, and t stands for time. Figure 8 shows the basic structure of WaveNet.

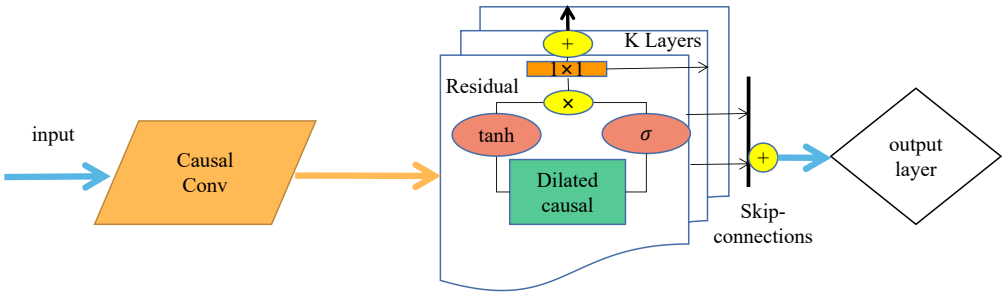


Figure 8. The structure of WaveNet.

The gated convolution can be formulated as

$$output = tanh(W_{f,k} * input) \bigodot \text{œ}(W_{g,k} * input). \tag{7}$$

where $*$ denotes the convolution operation and $\odot$ is the corresponding position multiplication operator. The short-circuit structure of residual is added for better training. The final result is obtained after superposition based on the intermediate results of the output of each layer.

The fundamental constituent of WaveNet is the causal convolution. By using causal convolution, it is ensured that the model does not violate the order of the data when modeling them. The prediction $p(x_{t+1} \mid x_1, \ldots, x_t)$ output by the model at moment t does not depend on any of the data at future moments $x_{t+1}, x_{t+2}, \ldots, x_T$. The drawback of causal convolution is that it can only capture local information. To expand the receptive field, one needs to increase the size of the convolution kernel or add more layers to the model. However, these methods are not effective for longer time series. Therefore, WaveNet uses dilated causal convolution. Dilated convolution skips input values with a certain stride and applies the convolution kernel to an area larger than its own size, so that it can have a large receptive field even with fewer layers. A visualization of a stack of dilated causal convolutional layers with causal convolutions is shown in Figure 9. With the increase in dilation, the receptive field of the dilated causal convolution in the figure expands from the original four points to eight points.

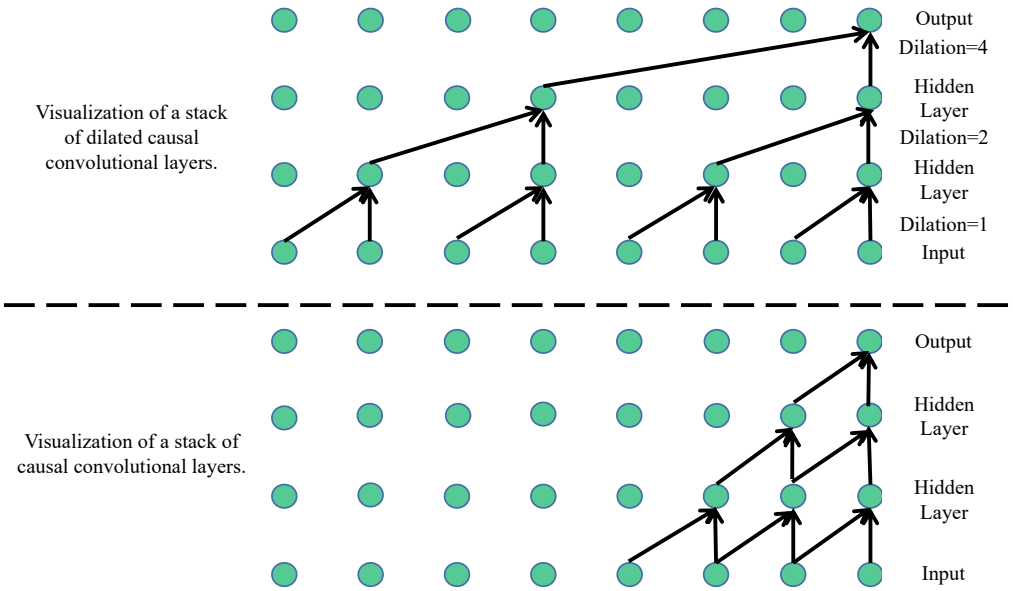


Figure 9. A stack of dilated causal convolutional layers and causal convolution.

4.2. MMoE

Single-task learning (STL) is a common approach in machine learning that aims to train models to solve specific individual tasks. In STL, the model is designed to optimize and adapt to a specific objective function, and there exists a clear one-to-one mapping between input data and output labels. However, STL also has some drawbacks. It is limited in its ability to address multiple related tasks effectively. When faced with multiple tasks, each task typically requires training a separate model, resulting in lengthy training processes and significant resource consumption. Additionally, STL fails to fully exploit the interdependencies and shared information among different tasks, which can adversely affect the predictive performance of the model. To overcome these limitations, multitask learning (MTL) has been proposed as an approach in which a single model is trained to address multiple related tasks simultaneously, aiming to leverage the shared information and knowledge among tasks to enhance the model's performance and generalization capability. Additionally, MTL aims to reduce the model's time complexity, shorten the training time, and minimize resource consumption. Figure 10 shows the STL and MTL.

Multigate mixture-of-experts (MMoE) is a model architecture designed to address the problem of multitask learning. It incorporates multiple expert networks and a gating network to enable sharing and interaction among tasks. Each expert network is responsible for learning the feature representation for a specific task, while the gating network dynamically determines how to allocate the outputs of the expert networks. Through this approach, MMoE is able to capture the relationships between tasks and achieve improved

performance through the combination of shared feature representation and expert networks. Figure 11 illustrates the structure of MMoE, and its representation equation is as follows:

$$y^k = h^k(f^k(x)) \tag{8}$$

$$f^k(x) = \sum_{i=1}^{n} g^k(x)_i f_i(x) \tag{9}$$

where x denotes the input, $W_{gk} \in \mathbb{R}^{n*d}$ represents a trainable matrix where n is the number of experts, and d is the dimensionality of the feature. k denotes the independent output unit of the kth task, h^k represents the tower network, and g^k represents the control gate of the kth task. The function $f^k(x)$ in Equation (4) is expressed as a combination of expert outputs using Equation (5), where $g^k(x) = softmax(W_{gk}x)$ is an input–output mapping function that outputs weights on all experts.

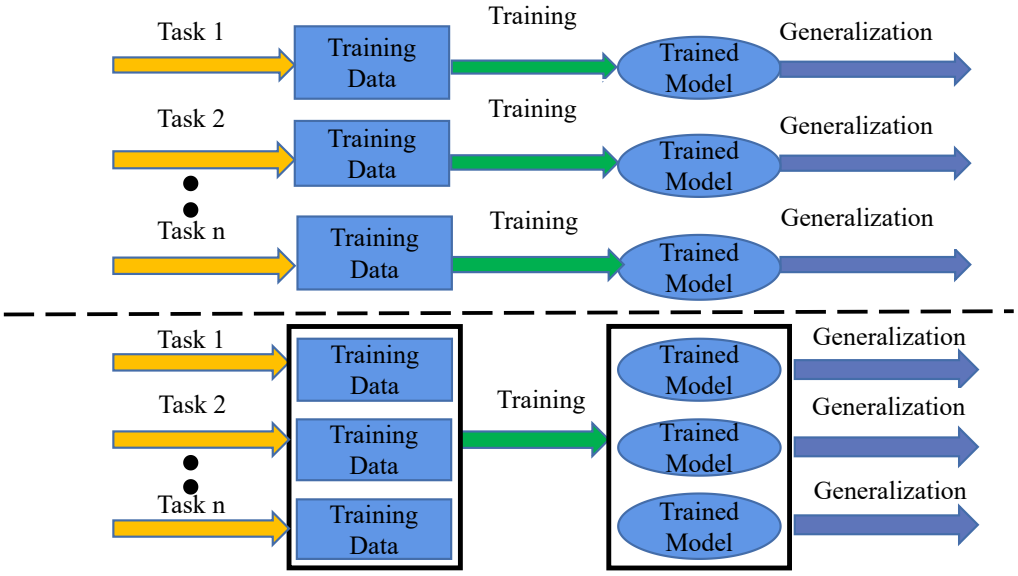


Figure 10. The operational mechanism of STL and MTL.

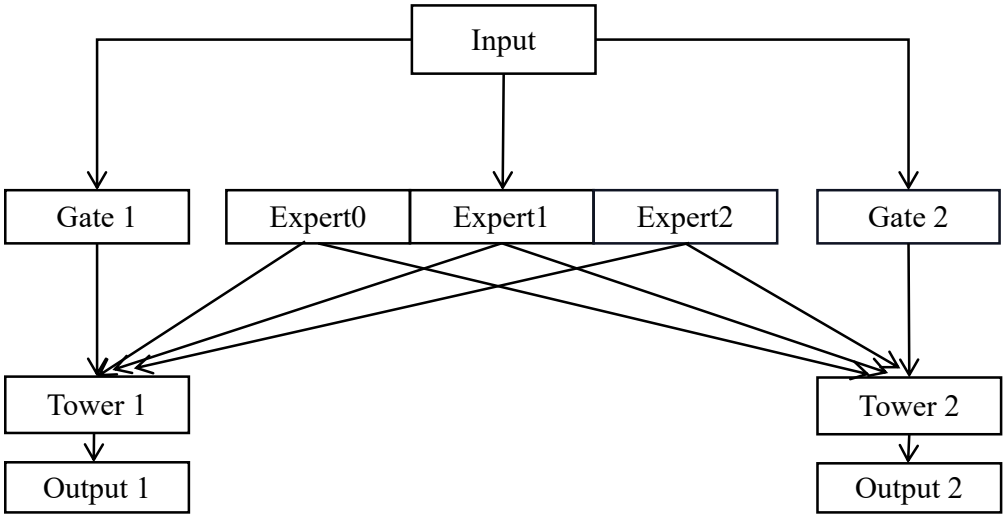


Figure 11. The structure of MMoE.

4.3. WaveNet–MMoE

In the combined model, the feature data selected by MIC is fed to WaveNet. After a simple causal convolution operation, the signal sequence passes through two layers of convolution with varying dilation rates, while creating residual connections and skip connections at each layer, to accelerate the convergence of the model. The output of each skip connection is used for subsequent calculations and added together as the final output to MMoE. After training by expert networks, the gate network performs a weighted sum of the outputs and provides predictions to multiple towers. In this experiment, conventional DNN structures are used for towers.

Figure 12 illustrates the structure of WaveNet–MMoE. When m = 2, WaveNet extracts deep-level feature information and inputs it to MMoE with the same number of gates for prediction at different time steps.

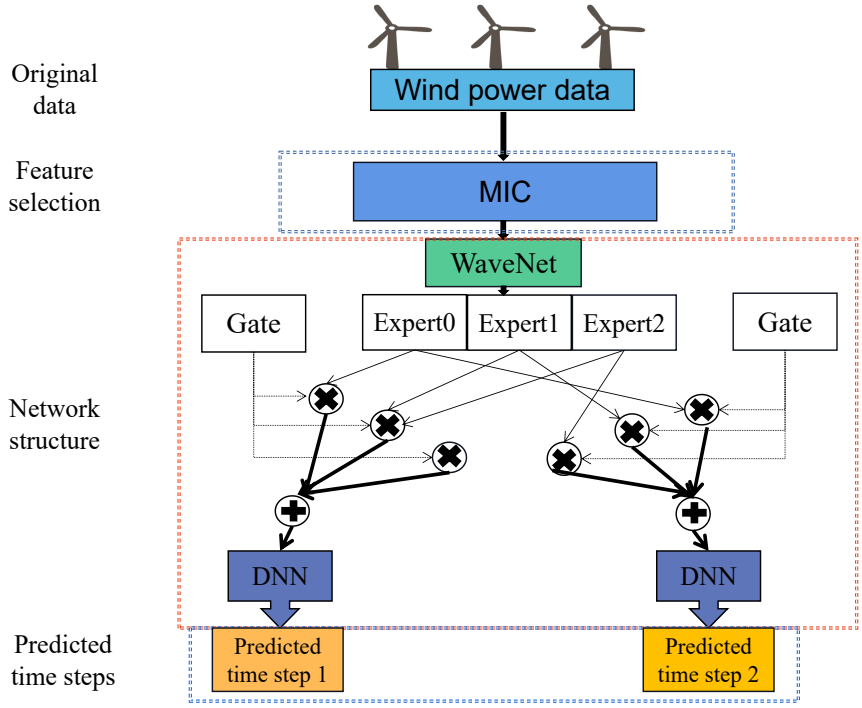


Figure 12. The overall structure of WaveNet–MMoE.

5. Experimental Analysis

5.1. Analysis of Forecast Results

In this experiment, 90% of the dataset was allocated for training purposes, whereas the remaining 10% was reserved for model evaluation through testing. In the experiment, the same network architecture was used for WaveNet and WaveNet–MMoE. The dilation rates were set to 8 and 12, the kernel size was set to 3, and the number of filters was set to 16 and 32. This was used to stack different dilated convolutions. The output layer consisted of a fully connected layer with three neurons, which was used for the final prediction task. In MMoE, the number of experts was set to 8, the input dimension was set to 72, and each expert had an output dimension of 3. The Adam optimizer was chosen, the batch size was set to 512, and a total of 100 iterations were performed. Table 1 presents a performance comparison between WaveNet–MMoE and eight other models. In order to comprehensively evaluate the performance of the model, we chose mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) as evaluation metrics. MAE measures the mean

absolute difference between the actual and predicted values, regardless of the direction of errors. Its formula can be expressed as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{10}$$

where n is the number of samples, $\hat{y}_i$ is the prediction, and y_i is the observation.

Table 1. The evaluation results of WaveNet–MMoE and other models for $m = 3$ and $m = 6$.

Prediction Window Size	Method	Metrics				
		MAE	**MAPE**	**MSE**	**RMSE**	**R^2**
3	CNN	0.079	1.340	0.014	0.118	0.474
	FNN	0.073	1.293	0.012	0.110	0.544
	CNN-LSTM	0.070	1.278	0.011	0.105	0.566
	LSTM	0.070	1.275	0.011	0.105	0.567
	Transformer	0.066	0.689	0.013	0.113	0.515
	Decision Tree	0.077	1.010	0.017	0.130	0.335
	CNN-Tree	0.108	1.690	0.029	0.170	−0.109
	WaveNet	0.056	0.712	0.009	0.095	0.645
	WaveNet–MMoE	0.052	0.556	0.009	0.095	0.656
6	CNN	0.084	1.454	0.016	0.126	0.406
	FNN	0.072	1.064	0.012	0.110	0.528
	CNN-LSTM	0.076	1.307	0.014	0.118	0.482
	LSTM	0.075	1.357	0.013	0.114	0.501
	Transformer	0.075	0.782	0.017	0.130	0.360
	Decision Tree	0.090	1.278	0.023	0.152	0.137
	CNN-Tree	0.113	1.794	0.032	0.179	−0.219
	WaveNet	0.063	0.803	0.012	0.110	0.557
	WaveNet–MMoE	0.060	0.736	0.011	0.105	0.572

The MAPE is an evaluation metric used to measure the accuracy of forecasting models. It is a relative measure that scales the MAD as a percentage instead of a unit of measurement. This approach uses absolute values to avoid the positive and negative errors from offsetting each other, providing a measure of relative error. The formula for the MAPE can be expressed as

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{11}$$

where n is the number of samples, $\hat{y}_i$ is the prediction, and y_i is the observation.

Mean squared error (MSE) measures the average squared difference between predicted values and actual values. It quantifies the average magnitude of errors to assess the overall performance of a model. A lower MSE indicates a better fit of the model to the data. The formula is as follows:

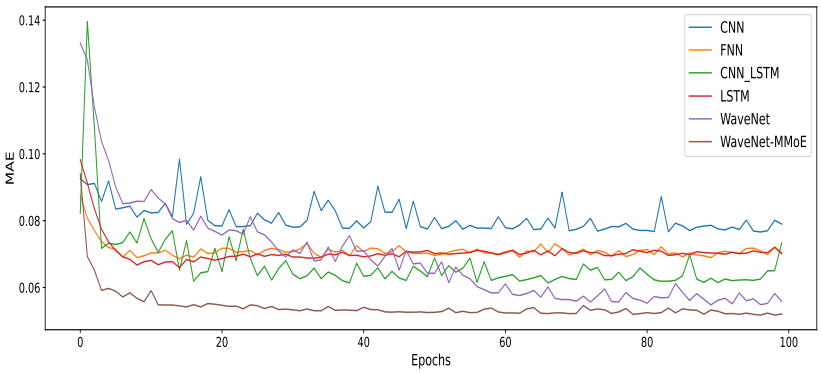
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12}$$

Root mean squared error (RMSE) is the square root of MSE and is used to interpret the error in the same units as the target variable. It is a popular evaluation metric as it provides a more interpretable measure of average error. Similar to MSE, a lower RMSE indicates a better fit of the model. The formula is as follows:

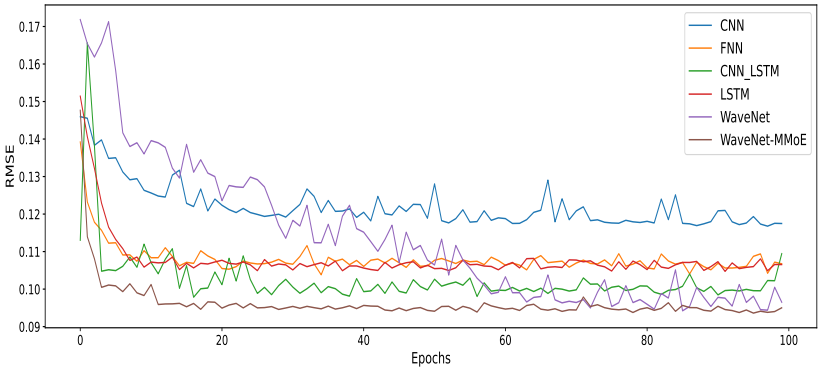
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

The R^2 coefficient, also known as the coefficient of determination, is a statistical measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variables in a regression model. It ranges between 0 and 1, where a higher value indicates a better fit of the model to the data. In other words, R^2 measures the explanatory and predictive power of the regression model with respect to the variation in the dependent variable. The formula is as follows:

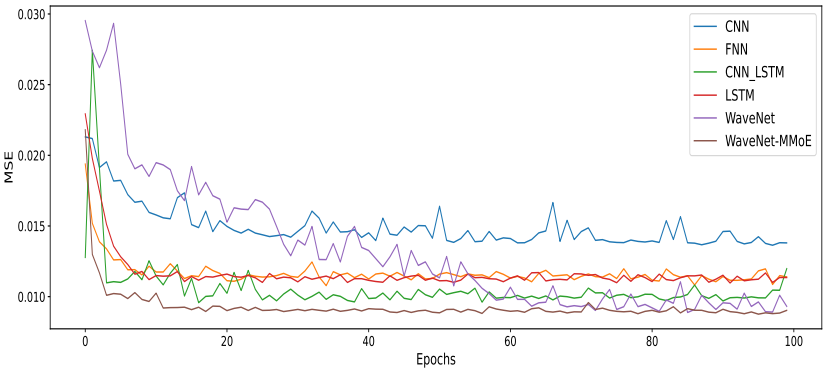
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2} \tag{14}$$



(**a**) MAE in different ways

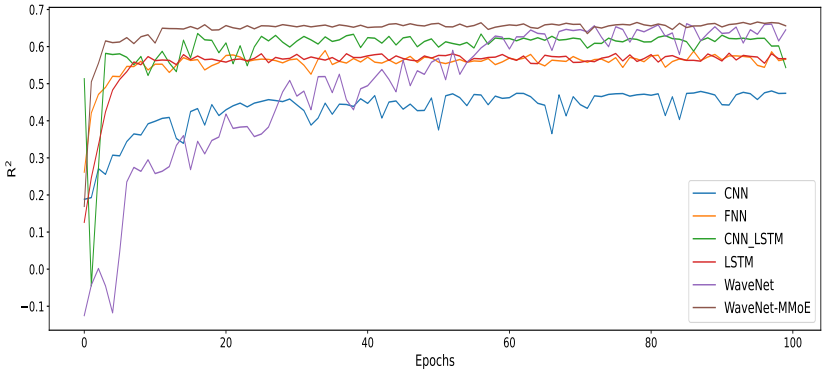


(**b**) RMSE in different ways



(**c**) MSE in different ways

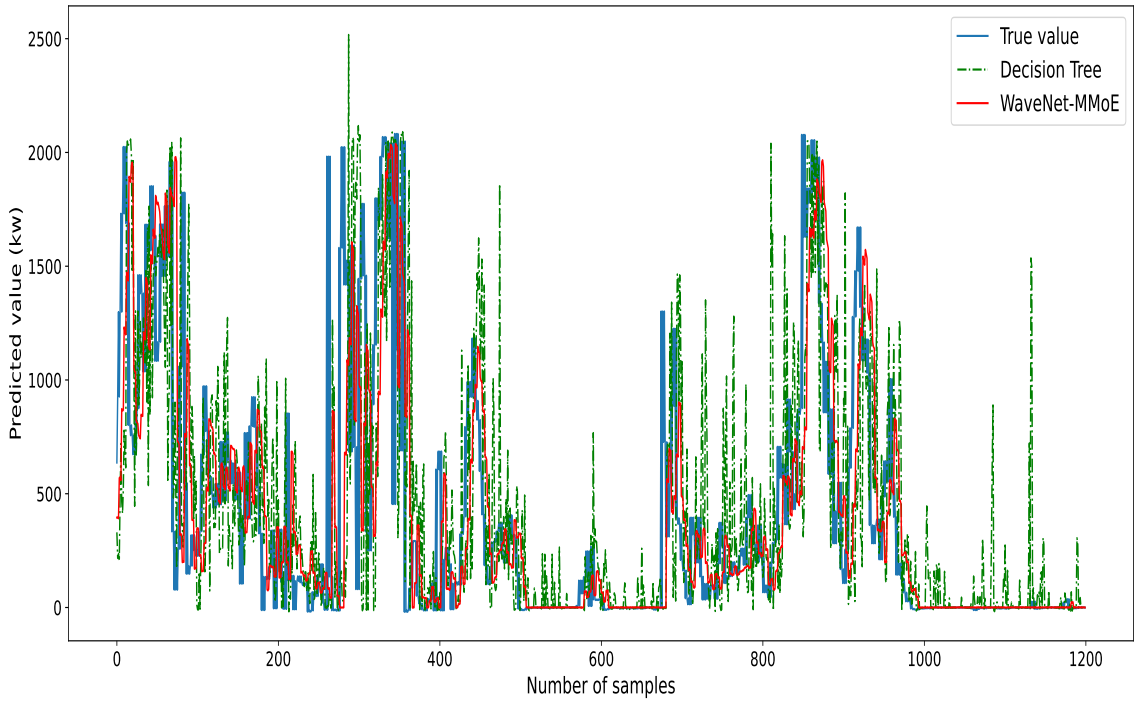
Figure 13. *Cont.*



(**d**) R^2 in different ways

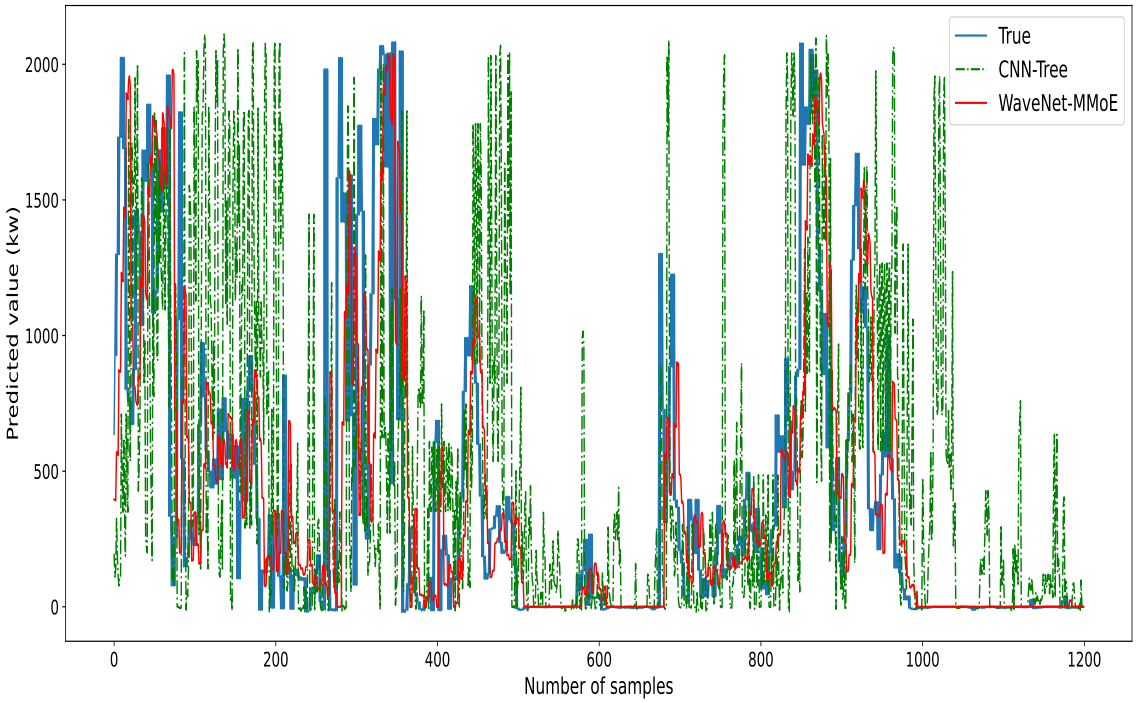
Figure 13. Scores of different methods at $m = 3$.

As shown in Table 1, the proposed multitask model outperforms several single-task models in predicting wind turbine power at different time steps. When $m = 3$, compared with the best-performing single-task model, the MAE and MAPE metrics decrease by at least 7% and 22%, respectively, while R^2 increases by 2%. When $m = 6$, compared with the best-performing model, the MAE, MAPE, MSE, and RMSE decrease by 5%, 8%, 8%, and 5%, respectively, while R^2 increases by 3%. This indicates that WaveNet–MMoE exhibits good predictive performance. Figure 13 shows the performance of various metrics on the test set at different epochs, while Figure 14 compares the predictions of WaveNet–MMoE with decision tree and CNN-Tree models, and Figure 15 depicts the prediction results of WaveNet–MMoE for $m = 3$.



(**a**) Decision Tree vs. WaveNet–MMoE

Figure 14. *Cont.*



(**b**) CNN-Tree vs. WaveNet–MMoE

Figure 14. Comparison of prediction results for $m = 3$.

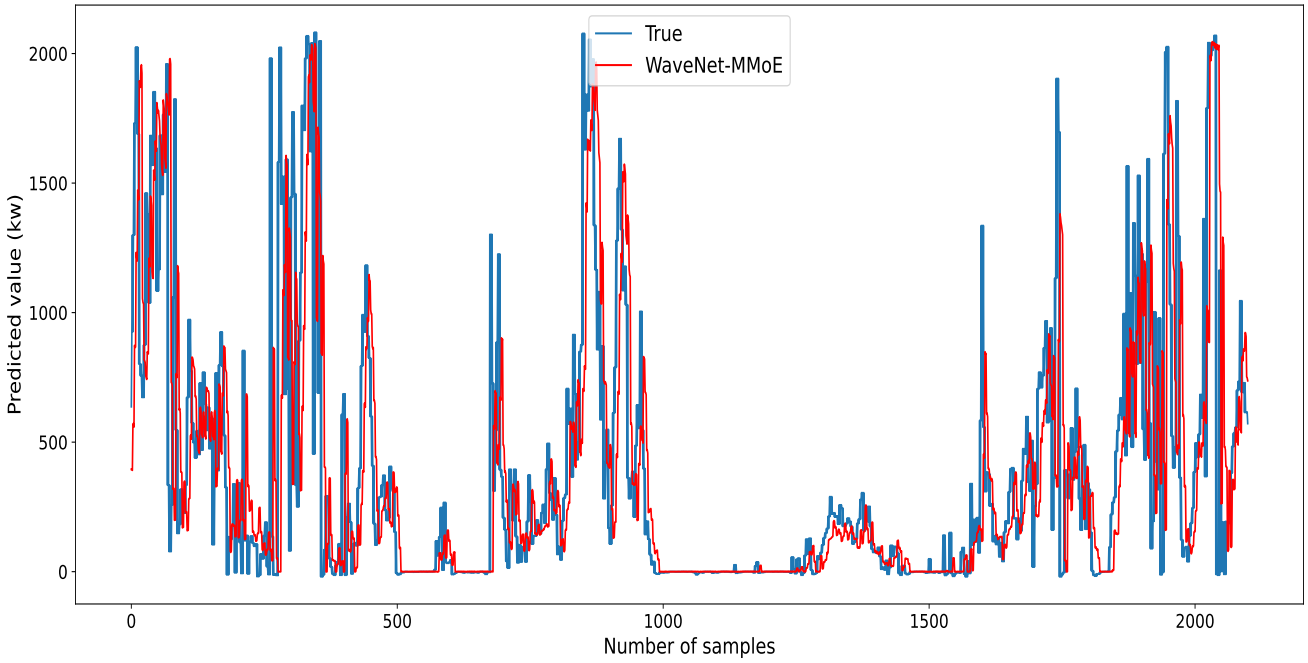


Figure 15. Prediction results of WaveNet–MMoE for $m = 3$.

5.2. Analysis of Forecast Results

Data preprocessing is a crucial step in data analysis and machine learning. It involves cleaning, transforming, and organizing raw data to provide high-quality data for subsequent analysis and modeling. The goal of data preprocessing is to eliminate noise, missing values, and outliers in the data; adjust the distribution and scale of the data; and select and

extract the most relevant features. Table 2 presents the results of selected methods after feature selection using MIC and denoising through wavelet transform.

Table 2. Data processing score.

Type of Datasets	Model	Metrics		
		MAE	MAPE	RMSE
Original	Decision Tree	0.0933	1.3522	0.1493
	Transformer	0.1259	2.4574	0.1762
	WaveNet	0.1064	0.8091	0.1571
	WaveNet–MMoE	0.0765	0.6420	0.1184
MIC	Decision Tree	0.0782	1.0515	0.1344
	Transformer	0.0664	0.7162	0.1143
	WaveNet	0.0604	0.7621	0.1016
	WaveNet–MMoE	0.0529	0.5628	0.0956
MIC + Wavelet	Decision Tree	0.0773	1.0103	0.1305
	Transformer	0.0659	0.6894	0.1128
	WaveNet	0.0559	0.7117	0.0965
	WaveNet–MMoE	0.0520	0.5561	0.0950

From Table 2, it can be observed that when the data undergo feature selection using the mutual information coefficient (MIC), some methods exhibit lower values of MAE, MAPE, and RMSE compared with predictions using the original data. Subsequently, after applying denoising through wavelet transform based on the MIC feature selection, the MAE, MAPE, and RMSE values of some methods are also lower than those of the methods without wavelet denoising. This indicates the necessity of data processing.

6. Conclusions

This paper presents a novel framework for wind turbine power forecasting, integrating the MIC method, wavelet transform, WaveNet model, and MTL architecture MMoE. Experimental results demonstrate significant improvements in multiple performance metrics of the proposed multitask model compared with several single-task models when predicting wind turbine power at different time steps. The framework provides a new perspective for multistep time series prediction compared with conventional single-task prediction models. Additionally, the necessity of using MIC and wavelet transform techniques in data preprocessing is validated, as proper data preprocessing enhances the prediction accuracy of this approach. Overall, this study offers an effective method to address the issue of error accumulation in wind turbine power forecasting. Future research can explore the application of this framework to other renewable energy sources and investigate its potential for real-time prediction.

Author Contributions: Conceptualization, H.W. and C.P.; methodology, C.P. and H.W.; software, H.W.; validation, H.W. and C.P.; formal analysis, H.W. and C.P.; investigation, B.L. and X.C.; data curation, B.L.; writing—original draft preparation, H.W.; writing—review and editing, H.W. and C.P.; visualization, H.W.; supervision, X.C. and S.L.; project administration, H.W.; funding acquisition, C.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Natural Science Foundation of China under Grant 62006095, by the Natural Science Foundation of Hunan Province, China, under Grant 2021JJ40441, and by the Jishou University Graduate Research and Innovation Project TXJD202303.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MMoE	multigate mixture-of-experts
ML	machine learning
MI	mutual information
MIC	maximum information coefficient
MSTF	multistep time series forecasting
MTL	multitask learning
STL	single-task learning
AR	autoregressive
ARMA	autoregressive moving average
ARIMA	autoregressive integrated moving average
ETS	exponential smoothing
XGBoost	eXtreme gradient boosting
CNN	convolutional neural networks
RNN	recurrent neural networks
TENT	Tensorial Encoder Transformer
TCN	temporal convolutional networks
DCCCN	dilated causal convolutional networks
ADGCN	asynchronous dilated graph convolutional network
SCTCN	self-calibrating temporal convolutional network
RUL	remaining useful life
WTA	winner-take-all
ANN	artificial neural network
ANOVA	analysis of variance
PC	pearson correlation
BE	backward elimination
RF	random forest
LASSO	least absolute shrinkage and selection operator
LSTM	long short-term memory
ConvLSTM	convolutional LSTM
MTL-TCNN	multitask learning temporal convolutional neural network
HA	historical average

References

1. Wu, Y.K.; Hong, J.S. A literature review of wind forecasting technology in the world. In Proceedings of the 2007 IEEE Lausanne Power Tech, Lausanne, Switzerland, 1–5 July 2007; pp. 504–509.
2. Lei, M.; Shiyan, L.; Chuanwen, J.; Hongling, L.; Yan, Z. A review on the forecasting of wind speed and generated power. *Renew. Sustain. Energy Rev.* **2009**, *13*, 915–920. [CrossRef]
3. Oh, K.; Kim, E.J.; Park, C.Y. A Physical Model-Based Data-Driven Approach to Overcome Data Scarcity and Predict Building Energy Consumption. *Sustainability* **2022**, *14*, 9464. [CrossRef]
4. Cox, D.R.; Gudmundsson, G.; Lindgren, G.; Bondesson, L.; Harsaae, E.; Laake, P.; Juselius, K.; Lauritzen, S.L. Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scand. J. Stat.* **1981**, *8*, 93–115.
5. Gomes, P.; Castro, R. Wind speed and wind power forecasting using statistical models: autoregressive moving average (ARMA) and artificial neural networks (ANN). *Int. J. Sustain. Energy Dev.* **2012**, *1*, 41–50. [CrossRef]
6. Khandelwal, I.; Adhikari, R.; Verma, G. Time series forecasting using hybrid ARIMA and ANN models based on DWT decomposition. *Procedia Comput. Sci.* **2015**, *48*, 173–179. [CrossRef]
7. Sapitang, M.; M. Ridwan, W.; Faizal Kushiar, K.; Najah Ahmed, A.; El-Shafie, A. Machine learning application in reservoir water level forecasting for sustainable hydropower generation strategy. *Sustainability* **2020**, *12*, 6121. [CrossRef]
8. Solyali, D. A comparative analysis of machine learning approaches for short-/long-term electricity load forecasting in Cyprus. *Sustainability* **2020**, *12*, 3612. [CrossRef]
9. Musarat, M.A.; Alaloul, W.S.; Rabbani, M.B.A.; Ali, M.; Altaf, M.; Fediuk, R.; Vatin, N.; Klyuev, S.; Bukhari, H.; Sadiq, A.; et al. Kabul river flow prediction using automated ARIMA forecasting: A machine learning approach. *Sustainability* **2021**, *13*, 10720. [CrossRef]
10. Yousaf, A.; Asif, R.M.; Shakir, M.; Rehman, A.U.; S. Adrees, M. An improved residential electricity load forecasting using a machine-learning-based feature selection approach and a proposed integration strategy. *Sustainability* **2021**, *13*, 6199. [CrossRef]
11. Pavlyshenko, B.M. Machine-learning models for sales time series forecasting. *Data* **2019**, *4*, 15. [CrossRef]

12. Fan, C.; Sun, Y.; Zhao, Y.; Song, M.; Wang, J. Deep learning-based feature engineering methods for improved building energy prediction. *Appl. Energy* **2019**, *240*, 35–45. [CrossRef]
13. Chandra, R.; Ong, Y.S.; Goh, C.K. Co-evolutionary multi-task learning with predictive recurrence for multi-step chaotic time series prediction. *Neurocomputing* **2017**, *243*, 21–34. [CrossRef]
14. Dong, G.; Fataliyev, K.; Wang, L. One-step and multi-step ahead stock prediction using backpropagation neural networks. In Proceedings of the 2013 9th International Conference on Information, Communications & Signal Processing, Beijing, China, 16–18 October 2013; pp. 1–5.
15. Du, P.; Jin, Y.; Zhang, K. A hybrid multi-step rolling forecasting model based on ssa and simulated annealing—Adaptive particle swarm optimization for wind speed. *Sustainability* **2016**, *8*, 754. [CrossRef]
16. Chandra, R.; Goyal, S.; Gupta, R. Evaluation of deep learning models for multi-step ahead time series prediction. *IEEE Access* **2021**, *9*, 83105–83123. [CrossRef]
17. Girard, A.; Rasmussen, C.; Candela, J.Q.; Murray-Smith, R. Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. *Adv. Neural Inf. Process. Syst.* **2002**, *15*, 545–552.
18. Parthasarathy, S.; Busso, C. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; Volume 2017, pp. 1103–1107.
19. Lala, B.; Rizk, H.; Kala, S.M.; Hagishima, A. Multi-task learning for concurrent prediction of thermal comfort, sensation and preference in winters. *Buildings* **2022**, *12*, 750. [CrossRef]
20. Ahmed, A.; Khalid, M. A review on the selected applications of forecasting models in renewable power systems. *Renew. Sustain. Energy Rev.* **2019**, *100*, 9–21. [CrossRef]
21. Ssekulima, E.B.; Anwar, M.B.; Al Hinai, A.; El Moursi, M.S. Wind speed and solar irradiance forecasting techniques for enhanced renewable energy integration with the grid: A review. *IET Renew. Power Gener.* **2016**, *10*, 885–989. [CrossRef]
22. Wang, Y.; Yu, Y.; Cao, S.; Zhang, X.; Gao, S. A review of applications of artificial intelligent algorithms in wind farms. *Artif. Intell. Rev.* **2020**, *53*, 3447–3500. [CrossRef]
23. Zhang, K.; Liu, Z.; Zheng, L. Short-term prediction of passenger demand in multi-zone level: Temporal convolutional neural network with multi-task learning. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1480–1490. [CrossRef]
24. Kavasseri, R.G.; Seetharaman, K. Day-ahead wind speed forecasting using f-ARIMA models. *Renew. Energy* **2009**, *34*, 1388–1393. [CrossRef]
25. Erdem, E.; Shi, J. ARMA based approaches for forecasting the tuple of wind speed and direction. *Appl. Energy* **2011**, *88*, 1405–1414. [CrossRef]
26. Chen, P.; Pedersen, T.; Bak-Jensen, B.; Chen, Z. ARIMA-based time series model of stochastic wind power generation. *IEEE Trans. Power Syst.* **2009**, *25*, 667–676. [CrossRef]
27. Dai, J.; Chen, Y.; Xiao, L.; Jia, L.; He, Y. Design and analysis of a hybrid GNN-ZNN model with a fuzzy adaptive factor for matrix inversion. *IEEE Trans. Ind. Inform.* **2021**, *18*, 2434–2442. [CrossRef]
28. Zhang, Y.; Qiu, B.; Liao, B.; Yang, Z. Control of pendulum tracking (including swinging up) of IPC system using zeroing-gradient method. *Nonlinear Dyn.* **2017**, *89*, 1–25. [CrossRef]
29. Jia, L.; Xiao, L.; Dai, J.; Qi, Z.; Zhang, Z.; Zhang, Y. Design and application of an adaptive fuzzy control strategy to zeroing neural network for solving time-variant QP problem. *IEEE Trans. Fuzzy Syst.* **2020**, *29*, 1544–1555. [CrossRef]
30. Lei, Y.; Liao, B.; Chen, J. Comprehensive analysis of ZNN models for computing complex-valued time-dependent matrix inverse. *IEEE Access* **2020**, *8*, 91989–91998. [CrossRef]
31. Lu, H.; Jin, L.; Luo, X.; Liao, B.; Guo, D.; Xiao, L. RNN for solving perturbed time-varying underdetermined linear system with double bound limits on residual errors and state variables. *IEEE Trans. Ind. Inform.* **2019**, *15*, 5931–5942. [CrossRef]
32. Cioffi, R.; Travaglioni, M.; Piscitelli, G.; Petrillo, A.; De Felice, F. Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions. *Sustainability* **2020**, *12*, 492. [CrossRef]
33. Alshboul, O.; Shehadeh, A.; Almasabha, G.; Almuflih, A.S. Extreme gradient boosting-based machine learning approach for green building cost prediction. *Sustainability* **2022**, *14*, 6651. [CrossRef]
34. Li, W.; Han, L.; Xiao, X.; Liao, B.; Peng, C. A gradient-based neural network accelerated for vision-based control of an RCM-constrained surgical endoscope robot. *Neural Comput. Appl.* **2022**, *34*, 1329–1343. [CrossRef]
35. Xiao, L.; Liao, B.; Li, S.; Zhang, Z.; Ding, L.; Jin, L. Design and analysis of FTZNN applied to the real-time solution of a nonstationary Lyapunov equation and tracking control of a wheeled mobile manipulator. *IEEE Trans. Ind. Informatics* **2017**, *14*, 98–105. [CrossRef]
36. Xiao, L.; Zhang, Y. Solving time-varying inverse kinematics problem of wheeled mobile manipulators using Zhang neural network with exponential convergence. *Nonlinear Dyn.* **2014**, *76*, 1543–1559. [CrossRef]
37. Khan, A.T.; Cao, X.; Liao, B.; Francis, A. Bio-Inspired Machine Learning for Distributed Confidential Multi-Portfolio Selection Problem. *Biomimetics* **2022**, *7*, 124. [CrossRef]
38. Liao, B.; Wang, Y.; Li, W.; Peng, C.; Xiang, Q. Prescribed-time convergent and noise-tolerant Z-type neural dynamics for calculating time-dependent quadratic programming. *Neural Comput. Appl.* **2021**, *33*, 5327–5337. [CrossRef]
39. Rajula, H.S.R.; Verlato, G.; Manchia, M.; Antonucci, N.; Fanos, V. Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. *Medicina* **2020**, *56*, 455. [CrossRef] [PubMed]

40. Dhiman, G.; Juneja, S.; Viriyasitavat, W.; Mohafez, H.; Hadizadeh, M.; Islam, M.A.; El Bayoumy, I.; Gulati, K. A novel machine-learning-based hybrid CNN model for tumor identification in medical image processing. *Sustainability* **2022**, *14*, 1447. [CrossRef]
41. Kumar, M.; Singhal, S.; Shekhar, S.; Sharma, B.; Srivastava, G. Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning. *Sustainability* **2022**, *14*, 13998. [CrossRef]
42. Khosravi Kazazi, A.; Amiri, F.; Rahmani, Y.; Samouei, R.; Rabiei-Dastjerdi, H. A New Hybrid Model for Mapping Spatial Accessibility to Healthcare Services Using Machine Learning Methods. *Sustainability* **2022**, *14*, 14106. [CrossRef]
43. Robinson, C.; Dilkina, B.; Hubbs, J.; Zhang, W.; Guhathakurta, S.; Brown, M.A.; Pendyala, R.M. Machine learning approaches for estimating commercial building energy consumption. *Appl. Energy* **2017**, *208*, 889–904. [CrossRef]
44. Mosavi, A.; Salimi, M.; Faizollahzadeh Ardabili, S.; Rabczuk, T.; Shamshirband, S.; Varkonyi-Koczy, A.R. State of the art of machine learning models in energy systems, a systematic review. *Energies* **2019**, *12*, 1301. [CrossRef]
45. Jung, H.; Jeon, J.; Choi, D.; Park, J.Y. Application of machine learning techniques in injection molding quality prediction: Implications on sustainable manufacturing industry. *Sustainability* **2021**, *13*, 4120. [CrossRef]
46. Yousaf, A.; Asif, R.M.; Shakir, M.; Rehman, A.U.; Alassery, F.; Hamam, H.; Cheikhrouhou, O. A novel machine learning-based price forecasting for energy management systems. *Sustainability* **2021**, *13*, 12693. [CrossRef]
47. Ding, L.; She, J.; Peng, S. An integrated prediction model for network traffic based on wavelet transformation. *Elektron. Ir Elektrotechnika* **2013**, *19*, 73–76. [CrossRef]
48. Liu, H.; Mi, X.; Li, Y. An experimental investigation of three new hybrid wind speed forecasting models using multi-decomposing strategy and ELM algorithm. *Renew. Energy* **2018**, *123*, 694–705. [CrossRef]
49. Tascikaraoglu, A.; Uzunoglu, M. A review of combined approaches for prediction of short-term wind speed and power. *Renew. Sustain. Energy Rev.* **2014**, *34*, 243–254. [CrossRef]
50. Li, Y.; Wu, H.; Liu, H. Multi-step wind speed forecasting using EWT decomposition, LSTM principal computing, RELM subordinate computing and IEWT reconstruction. *Energy Convers. Manag.* **2018**, *167*, 203–219. [CrossRef]
51. Ruiz, L.G.B.; Cuéllar, M.P.; Calvo-Flores, M.D.; Jiménez, M.D.C.P. An application of non-linear autoregressive neural networks to predict energy consumption in public buildings. *Energies* **2016**, *9*, 684. [CrossRef]
52. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6.
53. Liu, J.; Shi, Q.; Han, R.; Yang, J. A hybrid GA–PSO–CNN model for ultra-short-term wind power forecasting. *Energies* **2021**, *14*, 6500. [CrossRef]
54. Elsaraiti, M.; Merabet, A. A comparative analysis of the arima and lstm predictive models and their effectiveness for predicting wind speed. *Energies* **2021**, *14*, 6782. [CrossRef]
55. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv* **2019**, arXiv:1910.03771.
56. Gao, X.; Li, X.; Zhao, B.; Ji, W.; Jing, X.; He, Y. Short-term electricity load forecasting model based on EMD-GRU with feature selection. *Energies* **2019**, *12*, 1140. [CrossRef]
57. Bilgin, O.; Mąka, P.; Vergutz, T.; Mehrkanoon, S. TENT: Tensorized Encoder Transformer for temperature forecasting. *arXiv* **2021**, arXiv:2106.14742.
58. Qi, T.; Li, G.; Chen, L.; Xue, Y. Adgcn: An asynchronous dilation graph convolutional network for traffic flow prediction. *IEEE Internet Things J.* **2021**, *9*, 4001–4014. [CrossRef]
59. Zhu, R.; Liao, W.; Wang, Y. Short-term prediction for wind power based on temporal convolutional network. *Energy Rep.* **2020**, *6*, 424–429. [CrossRef]
60. He, K.; Su, Z.; Tian, X.; Yu, H.; Luo, M. RUL prediction of wind turbine gearbox bearings based on self-calibration temporal convolutional network. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [CrossRef]
61. Tyralis, H.; Papacharalampous, G. Variable selection in time series forecasting using random forests. *Algorithms* **2017**, *10*, 114. [CrossRef]
62. Bouktif, S.; Fiaz, A.; Ouni, A.; Serhani, M.A. Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies* **2018**, *11*, 1636. [CrossRef]
63. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
64. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
65. Yang, S. On feature selection for traffic congestion prediction. *Transp. Res. Part C Emerg. Technol.* **2013**, *26*, 160–169. [CrossRef]
66. Liu, J.; Sun, H.; Li, Y.; Fang, W.; Niu, S. An improved power system transient stability prediction model based on mRMR feature selection and WTA ensemble learning. *Appl. Sci.* **2020**, *10*, 2255. [CrossRef]
67. Naik, N.; Mohan, B.R. Optimal feature selection of technical indicator and stock prediction using machine learning technique. In Proceedings of the Emerging Technologies in Computer Engineering: Microservices in Big Data Analytics: Second International Conference, ICETCE 2019, Jaipur, India, 1–2 February 2019; Springer: Singapore, 2019; pp. 261–268.
68. Bagherzadeh, F.; Mehrani, M.J.; Basirifard, M.; Roostaei, J. Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. *J. Water Process Eng.* **2021**, *41*, 102033. [CrossRef]

69. Nettleton, D.F.; Orriols-Puig, A.; Fornells, A. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artif. Intell. Rev.* **2010**, *33*, 275–306. [CrossRef]
70. Yan, H.; Ouyang, H. Financial time series prediction based on deep learning. *Wirel. Pers. Commun.* **2018**, *102*, 683–700. [CrossRef]
71. Kim, K.S.; Lee, J.B.; Roh, M.I.; Han, K.M.; Lee, G.H. Prediction of ocean weather based on denoising autoencoder and convolutional LSTM. *J. Mar. Sci. Eng.* **2020**, *8*, 805. [CrossRef]
72. Samal, K.K.R.; Babu, K.S.; Das, S.K. Temporal convolutional denoising autoencoder network for air pollution prediction with missing values. *Urban Clim.* **2021**, *38*, 100872. [CrossRef]
73. Syfert, M.M.; Smith, M.J.; Coomes, D.A. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS ONE* **2013**, *8*, e55158. [CrossRef]
74. Zhang, Y.; Yang, Q. An overview of multi-task learning. *Natl. Sci. Rev.* **2018**, *5*, 30–43. [CrossRef]
75. Zhang, Y.; Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5586–5609. [CrossRef]
76. Zhang, C.; Zhu, F.; Wang, X.; Sun, L.; Tang, H.; Lv, Y. Taxi demand prediction using parallel multi-task learning model. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 794–803. [CrossRef]
77. Crichton, G.; Pyysalo, S.; Chiu, B.; Korhonen, A. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinform.* **2017**, *18*, 368. [CrossRef] [PubMed]
78. Yang, H.; Gong, S.; Liu, Y.; Lin, Z.; Qu, Y. A multi-task learning model for daily activity forecast in smart home. *Sensors* **2020**, *20*, 1933. [CrossRef] [PubMed]