Very-Short Term Wind Speed Forecasting Via Distance Algorithm in Machine Learning

Alireza Shaterzadeh Yazdi[1] , Cavit Fatih Küçüktezcan[1]

***Abstract*— This paper proposes distance matrices, Euclidean, and offset translation methods in machine learning prediction of wind speed. The primary aim for this research is to design forecasting models for very short-term and short-term wind speed prediction based on these two methods by using historical data on wind speed. The test data is collected at a wind power station at 10 minutes intervals. Furthermore, we evaluate the output in different time horizons in comparison to the benchmark method (persistence). To ensure the output results, comparing this method with the persistence method is essential. The proposed method performance was evaluated and compared with the conventional persistence method performance in terms of mean absolute error.**

***Index Terms*— Very short-term prediction, Wind speed prediction, Distance matrices, Machine learning**

I. INTRODUCTION

The approximate of 70 to 80 percent of the pollution in the world is triggered by the use of traditional energy sources [1]. These sources are limited so renewable energy has a more significant role in providing future global power as an inexhaustible resource. Greenhouse emissions reduction is the other advantage of using renewable energy sources.

By 2030 wind power could approach 2.11 GW which could supply up to 20 percent of electricity in the world and create 2.4 million new jobs which redact 3.2 billion tons of CO2 emission per year, this attracts 200 billion EUR annual investment [2]. Wind energy has significant economic benefits, including job creation, and cheap and clean fuel sources. Critical to authorities and investors is the cost-effective extension of the grid, which can be provided through the use of wind power in a variety of applications.

Wind energy is intermittent and impacts the sustainability of system operations. System reliability is jeopardized by these unexpected fluctuations. In this case, the system will cost more to compensate for the increased primary load. As a result, markets that cannot predict future wind speeds must increase speeds to prevent loss of wind speed variability. To reduce these costs, the system needs accurate wind speed prediction. In other words, grid operators' survival depends on accurate wind forecasts [3].

(1) Nowadays, wind speed prediction provides effective reference information for the development of the environment and energy industry [4][5]. These models can be divided into two directions: causality analysis and time series analysis [7]. The Historical relationship is the base of the causality analysis among the explanatory valuable, which is triggered by multiple collinearity negative effects [8].

(2) So far, time series analysis has four main methods which are: (1) statistical models; (2) Physical methods; (3) Spatial measurement algorithms; (4) Artificial intelligence algorithms [9][10]. These algorithms are summarized below:

(3) The statistical model has good performance in linear time series and needs to consider the distribution of data before making a prediction. However, temperature, pressure, terrain, latitude, and altitude are the most nonlinear characteristics which make the wind speed prediction weak.

(4) Physical methods often are poor in accuracy and effectiveness due to the big data calculation, and difficulty of measurement implementation, and need to be modeled separately in different regions. The cost of method research is high as a consequence of the mentioned reasons.

(5) Spatial measurement algorithms' approach to acceptance results in difficulty. Collection of lots of information from different stations, measurement, accuracy, and time delay leads to a complex of forecasting.

(6) Unlike the mentioned methods artificial intelligence can achieve the result for the nonlinear part of the original data. However, it has some shortcomings, such as low convergence, overfitting, and failing to locally optimal solutions easily [10-14].

Short-term wind speed forecasting (STWSF) models cover forecasts minutes or a few hours ahead. In STWSF, load forecasting by monitoring the system's balance requirements could improve the overall uncertainty of the system. STWSF helps system operators to estimate the supply side of load balancing [15]. Our new method is empirically being discussed with the most powerful method in short-term wind speed forecasting which is the persistence method.

The new algorithm according to the length of similarity takes candidate vectors. Euclidean distance and Offset transmission are used to optimize the estimation. Euclidean distance matrices are enforced to know the input data sample to make data decisions so, can be used to compute the distance between two

1: Department of Electrical Engineering, Bahcesehir University, Istanbul, Turkey

Corresponding author: alireza.yazdi@bahcesehir.edu.tr

data points in a plane. This structure has the advantage of getting vectors that are as similar as possible in terms of distance from the target vector. A distance matrix helps algorithms to detect similarities between content. A distance function returns the distance between the elements of a set. If the distance is zero, the elements are equivalent, otherwise, they are different. Furthermore, we use different time horizons on our input data wind speed to fairly approach the most promising results.

Given knowledge the next part of this paper has a definition for a better understanding of the system structure. Some papers have used ANN, Fuzzy Logic to predict wind speed. They train a big part of the data to be able to forecast wind speed. These models are very time-demanding in terms of time and computation.

The main contribution of this study is to evaluate our algorithm forecasting with the persistence method for a very short-term scale in a detailed manner. wind speed dataset in the whole study has been obtained from Afyon.

This research focuses on time range, similarity length (representing the length of data selected for processing), and thresholds for making short-term predictions. The implicit value of forecasts may compile disproportionate charges and penalties, real-time competitive knowledge advantages, and day-ahead energy market trading. Therefore, the proposed method triggers more efficient project construction, operations, and maintenance. Also, the proposed method performance was evaluated and compared with the conventional persistence method performance in terms of mean absolute error.

This article consists of five sections: Introduction (section I), Methodology (section II), Results (section III), Conclusion (section IV), References (section V).

II. METHODOLOGY

One of the most important parts of the study is to turn the information into knowledge. For this purpose, we use machine learning which is undeniably one of the most influential and powerful technologies in today's world. They considered data coming from Afyon in Fig. 2 and Data logger is recorded every 10 minutes. The total dataset used for our method has 35064 recorded numbers. During our study, we either use 10 minutes or bigger interval data to find out better results. Intermittent wind power affects the sustainability of power system operation which has to affect the reliability of the system and the market has to increase its rate in case of any loss. Accurate wind speed prediction is necessary and could reduce this cost. Data preparation to have a vector format and getting rid of the broken data is done in the next step. To select the training sample, we implement the Euclidean distance method coupled with offset transmission. Euclidean distance aid us to find out the most similar data to the target vector and adding them to our archive. To learn the best length of similarity in the learning process, we find out how well they perform. we define similarity length for our samples. Similarity length selected according to our strategy in learning algorithm in case the length is too short, prediction accuracy will decrease, and if it chooses too long, not only the speed will slow down but also the prediction accuracy will decrease. So, the model input has some influence on the prediction performance [16].

For reference in comparison, we define the target vector. The target vector is the vector whose next value is estimated and it takes this value directly from the next value in our dataset. After that vector is getting normalized to use the same scale, without distorting differences in the ranges of values or losing information to model our data correctly.

The main purpose of this paper is to extract the active learning algorithm as a solution to optimize the training sample sets for short-term wind speed prediction. In terms of our purpose, after training samples, it filters out noisy training samples according to their distance. Before considering the length of samples to select candidate one, we calculate the weight of each sample. Weighting is the process by which data is adjusted to reflect the known population profile. This helps compensate for large deviations between the actual and target profiles. Now you can evaluate the data based on the weights and all selected samples. In particular, the active learning approach is based on the idea of adding samples that do not violate the distance constraint, which is defined in our work as the threshold. The transmission offset applied to our sample data can act as a simple correction to the response. This helps the model to predict the offset of response.

The flowchart of our study is shown in Fig.1 which can be summarized as follows:

Step (1) Define the initial parameter and reduce distorted information.

Step (2) Set the target vector as the next value to estimate X(i+1) (i = 1, 2…, length of similarity(L)). Then, compute Euclidean distance and find all candidate vectors according to the target vector.

Step (3) Check all created vector distances with the target vector and eliminate vectors whose distance is more than the threshold.

Step (4) Take all nominated vectors that are similar to the target vector and add them to our samples.

Step (5) Start the learning algorithm to predict the next value of the sample's vectors. Transmission offset is applied to act as a simple correction to the response and then, calculate the error by MAPE (mean absolute percentage error) for both persistence and our methodology.

Step (6) Reestablish the model to predict the next learning model, to reach the preset number of learning.

A. Materials and Experiment

Considered data are coming from Afyon in Turkey and Data logger is recorded every 10 minutes. The total dataset used for our method has 35064 recorded numbers. During our study, we either use 10 minutes or bigger interval data to find out better results.

The wind speed potential in Afyon City is good in various wind parameters such as wind speed, wind direction, and air density temperature, and the wind farm considers a hub height. Fig. 3 displays the time series of wind speed and in Table I it is describe statistically.

Wind speed experiment data are from June 1, 2011, to July 30, 2013. Time horizon has defined in our data which fixes a point of time in the future and evaluates a certain process to end. Also, enough data is provided for training otherwise, the resulting model may perform poorly. Depending on the time horizon data included at least 1000 rows.

The hourly wind speed illustrated in Fig.3 was employed to test the performance of the proposed.

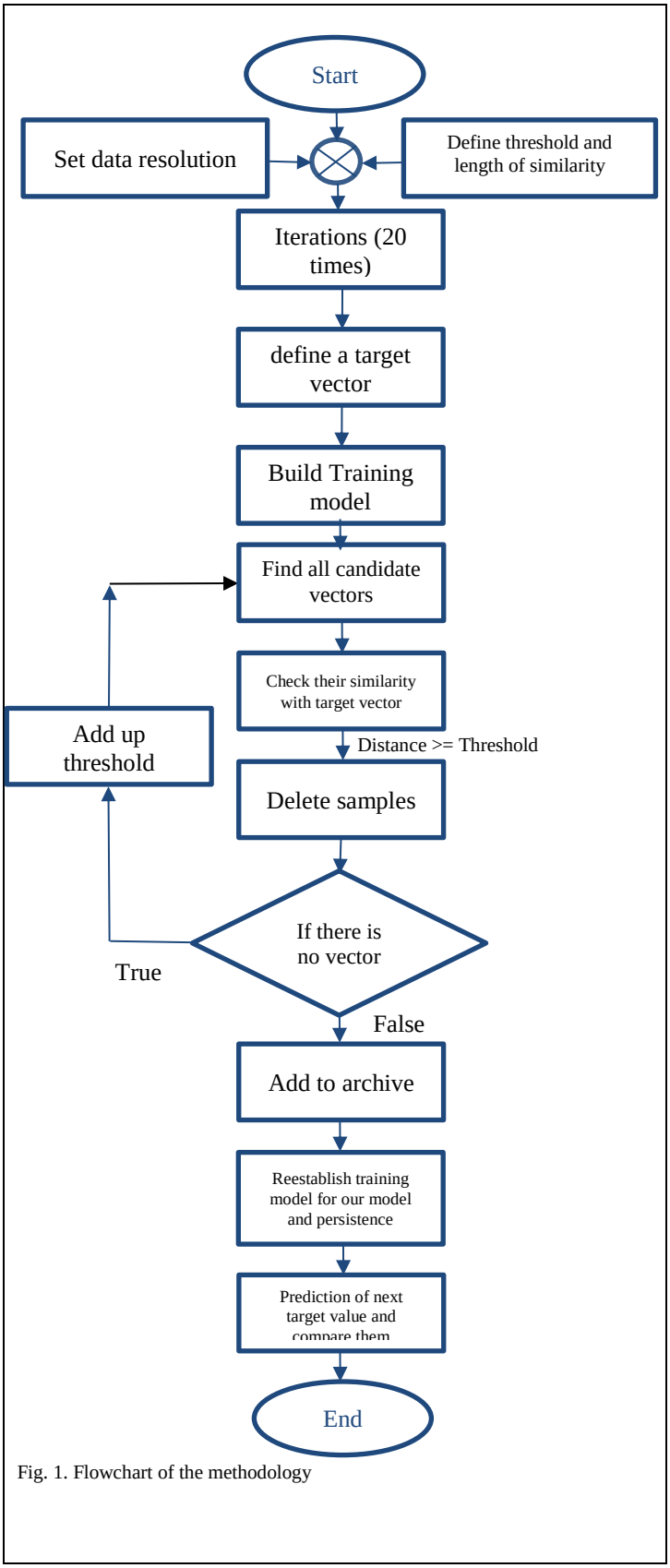

Fig. 1. Flowchart of the methodology

Fig. 2. Location of Afyon on the map

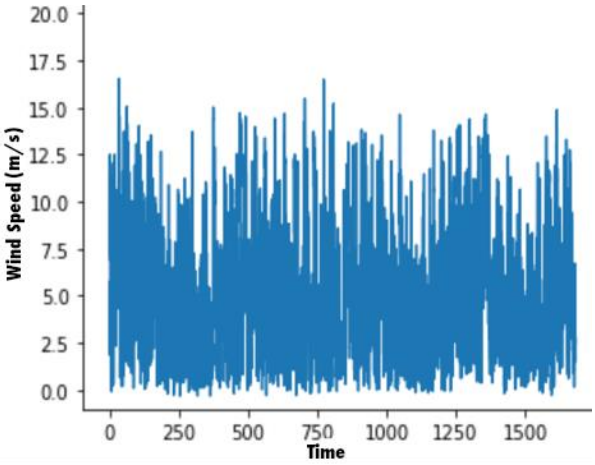

Fig.3. The wind speed data

TABLE I
Descriptive Statistics of Wind Speed

	Min	Mean	Max
All	0.237	5.471	17.430

B. Model Input Selection

Model input selection influences the prediction of the methodology therefore, getting the appropriate input is important. At first special numbers which represent null values has eliminated. we define similarity length for our samples. The similarity length is chosen according to the strategy of the learning algorithm. A length that is too short will result in poor prediction accuracy, and a length that is too long will not only slow down but will also result in poor prediction accuracy. Therefore, model inputs have some impact on prediction performance.

Time-lag method has been implied in our system to reconstruct the space from observed scalar data [17]. Basic to single-variable time series { x_t,t=1,2,…, N, } is observed, this method converts scalar time series {x_t} to vector in m-dimensional Euclidean space. Time lag responsibility is to delay vector x_t from the equation:

$$x_t = \{x_t, x_{t-\tau}, \dots, x_{t-(m-1)\tau}\} \quad (1)$$

Where x_t represents the point in the m-dimensional reconstructed phase-space τ is delay time and m is the embedding dimension. Time horizon has defined in our data which fixes a point of time in the future and evaluates a certain process to end.

C. Training Sampling Set

Preparing enough training data guarantees a suitable result. For best performance, data training iteration is set to twenty times in our model. To select the training samples, we implement the Euclidean distance method to find the data most similar to the target vector and add them to the archive. The similarity is usually expressed as a number. when the data are more alike, it gets higher and normally specifies between one and zero. Zero means data objects are dissimilar and one shows high similarity for data objects. We use the Euclidean distance method to compute the distance between the candidate archive and target archive to design the reference code as A_i, the other point as B_i:

$$D(A, B) = ||A - B||_0 = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \quad (2)$$

Where:

$$A = (A_1, A_2, \dots, A_n), \text{ and } B = (B_1, B_2, \dots, B) \quad (3)$$

The distance calculates in this formula represents the smallest distance between each pair of points. Fig. 4 shows the Euclidean distance metric.

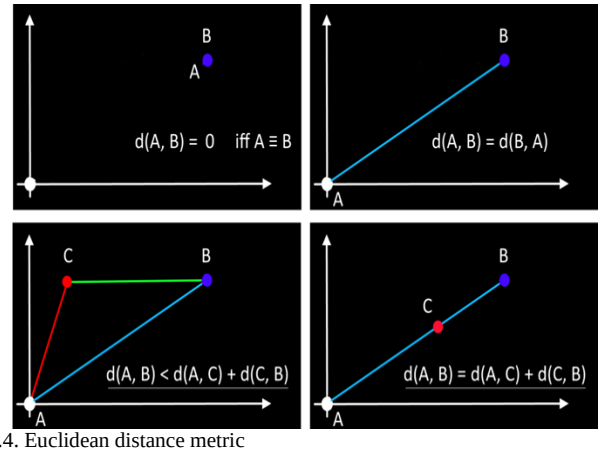

Fig.4. Euclidean distance metric

It is a useful formula to compute the distance between two data points in absence of obstacles on the pathways. In the next step, the vector is getting normalized to use the same scale, without distorting differences in the ranges of values or losing information to model our data correctly. In this case one over calculated distance gives normalization results which are weight:

$$W = 1 / \text{Distance} \quad (4)$$

Weighting is the process by which data is adjusted to reflect the known population profile. It is used to balance out any significant variance between the actual and target profiles. Now our data are ready to evaluate by their weights and all the selected samples.

In the next step, archive elements have been provided. According to the threshold (e), measurement of Euclidean distance results are considered, Therefore elements that do not violate the threshold sum up to archives, and the others have been ignored. If our methodology does not get any members in the archive then, it will multiply our threshold to 1.01 and check the similarity again. In this case, we will manage to find some members for our archive.

$$e_{new} = e * 1.01 \quad (5)$$

Along with finding archives, the persistence method is being fulfilled. Persistence is our competitor's method to see how our methodology approaches the result.

D. Development and Prediction Model

Transmission offset applied to our data during model training can act as a simple correction to the response. This helps the model to predict the offset of response. For this purpose, the transmission offset of K_i is calculated:

$$\text{Offset (K)} = \text{mean K / K New} \quad (6)$$

The mean of another point over the next value of that point gives us the transmission offset. In the following formula, K is specified as one similarity length. Calculate the model transmission by:

$$\text{Offset}\,(K_1, K_2, \dots, K_n) = \frac{\text{offset}(K_1)*\text{weight}(K_1)+\text{offset}(K_2)*\text{weight}(K_2)+\cdots \text{offset}(K_n)*\text{weight}(K_n)}{weight(K_1,K_{2,\dots,}K_n)} \quad (7)$$

By multiplying the result by the mean of the target vector, new data will be approached. This proceeds to lead us to the forecasted wind speed result. The discussed process was reset twenty times to get the best-predicted performance.

III. RESULTS

The Propose approach was programmed based on the Python platform (Jupyter Notebook). Performance is evaluated in terms of MAPE (mean absolute percentage error).

$$MAPE = \sqrt{\frac{1}{N_x}\sum_{t=1}^{N}\frac{(y(t)-y(t_1))}{y(t)}100} \quad (8)$$

Different time horizon has been applied both persistence and a new method to check the performance of our method in short-term wind speed prediction.

TABLE II
Result of New Method

Index	Threshold	Length of similarity	Time horizon	MAPE Result
1	0.1	7	10	3.11
2	0.1	7	30	18.53
3	0.1	7	60	27.37
4	0.1	7	90	58.25
5	0.1	7	120	48.86
6	0.1	7	240	105.83

TABLE III
Result of the Persistence Method

Index	Threshold	Length of similarity	Time horizon	MAPE Result
1	0.1	7	10	4.76
2	0.1	7	30	23.17
3	0.1	7	60	29.48
4	0.1	7	90	30.56
5	0.1	7	120	41.29
6	0.1	7	240	96.29

According to Table II and Table III, the proposed method has significantly better performance in very short-term prediction (10, 30, and 60 minutes). For clericity, Fig. 5 depicts prediction results and the real wind data. As noticed wind speed is predicted very well in the proposed method.

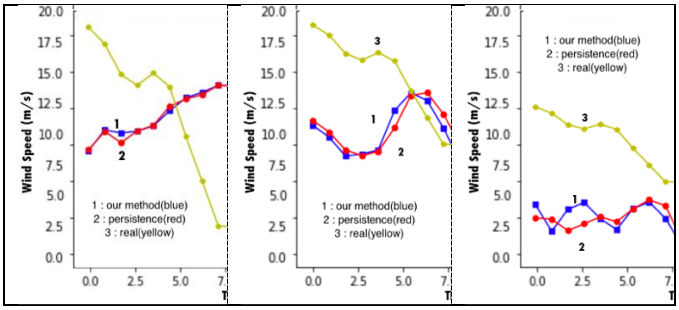

Fig.5. Comparison between the prediction results and the real one

From this point of view beating persistence in a short-term wind, speed prediction is difficult because persistence is performing well in this time scale. The new method has improved the prediction accuracy in this time scale.

IV. CONCLUSIONS

Reductant the potential risk of unexpected wind speed variation in the system is the most important objective of this paper. Furthermore, reducing the primer charge and maintenance cost will give a more reliable system. To reach this goal, wind speed data have been investigated and provided for the system to be examined and trained. This process is used to figure out a prediction for wind speed in short-term forecasting to have a sustainable power system.

The outcome is compared with the persistence method which is one of the most reliable methods in short-term wind speed prediction. Results prove that the new method was successful in the first hour. This approach combined the Euclidean distance and Offset transmission and shows we have a reliable output result in comparison to the persistence method. The proposed approach greatly improves prediction accuracy.

V. REFERENCES

[1] K. Kiranvishnu, K. Sireesha, J. Ramprabhakar, A Comparative study of wind power forecasting Techniques, March 16, 2016.
[2] Wind could supply Fifth of World Electricity by 2030, October 17, 2016.
[3] Suarabh S.Soman, Hamidreza Zareipour, Om Malik, Paras Mandal, A Review of Wind power and Wind Speed Forecasting Methods With Different Time Horizons , September 26, 2010.
[4] J.Wang, S.wang, W.yang, A novel non-linear combination for short -term wind speed forecast, Renew. Energy 143 (2019).
[5] Q.Zhou, wang, G.zhang, A combined forecasting system based on modified multi-objective optimization and submodel selection strategy for shor-term wind speed, Appl. Soft comput. J.94 (2020).
[6] P.giang, Z.Lui, J.Wang, L.Zhang, Decomposition-selection-ensemble forecasting system for energy futures price forecasting based on multi-objective-version of chaos game optimization algorithm(2021).
[7] L.liu, Q.Wang, J.wang, M.liu, A rolling gray optimization in economic prediction, comput. Intel. 32 (2016)
[8] P.Zhiang, Z.Liu, J.Wang, L.zhang, Decomposition-selection-ensemble forecasting system for energy futures price forecasting based on muli-objective version of chaos game optimization algorithm, Resour, Policy 73(2021).
[9] P.Jiang, Z.Lui, X.Niu, L.Zhang, A combined forecasting system based on statistical method, artificial neural network, and deep

learning methods for short-term wind speed forecasting, Energy(2021).
[10] E.erdem, J.sgi, ARMA based approaches for forecasting the tuple of wind speed and direction, appl. Energy 88 (2011).
[11] Y.wang, J.wang, G.zhao, Y.dong, application of residual modification approach in seasonal ARIMA for electricity demand forecasting: a case study of china energy policy(2012).
[12] D.C.Kiplangat, K.Asokan, Ks.kumar, improved week-ahead prediction of wind speed using simple linear model with wavelet decomposition, Renew energy(2016).
[13] R.G. kavasery, K.seetharaman day-ahead wind speed forecasting using F-ARIMA models, renew. Energy 43(2009).
[14] N.M, Zhi, C.Q. Yuan, Y.J. Yang, forecasting china energy demand and self-sufficiently by gray forecasting model and Markov model, Int.J.Electr. power enery syst 66 (2015).
[15] Edward Beleke Sekulima, M.Anvar, Wind speed and solar irradiance forecasting techniques for enhanced renewable energy integration with the grid(2016)
[16] H. De and G.Aquah, comparison of Akaike information criterion (AIC) Bayesian information criterion (BIC) in selection of an asymmetric price relationship, jurnal of development & Agricultural economics, Vol.2, No.1, 1-6, 2010.
[17] Dong Lie, Gao Shuang “Chaos characteristic analysis on the time series of wind power generation capacity”, Acta Energiae Solaries Sinica, vol. 28, pp. 1290-1294, Nov. 2007.