Journal of Advanced Zoology
ISSN: 0253-7214
Volume 44 Issue S-6 Year 2023 Page 2418:2425

New Coupled Wavelet-Random Forest Method for Wind Speed Prediction

Mohammad Hossein Kazemi[a*], Sepideh Karimi[b], Jalal Shiri[a]

[a*]*Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Iran.*
kazemi_m95@yahoo.com; j_shiri2005@yahoo.com
[b]*Water Engineering and Science Research Institute, University of Tabriz, Tabriz, Iran.*
karimi_sepide@yahoo.com

***Corresponding Author:** Mohammad Hossein Kazemi,*
Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Iran.
Email: kazemi_m95@yahoo.com.

Abstract

Accurate prediction of wind speed records is an important task in various disciplines including agriculture, meteorology, climatology, navy, energy studies, wind power, etc. Although some traditional models have been suggested and applied for wind forecast, machine learning (ML) approaches can be suitable alternatives for such models due to their successful performance inn wide range of subjects and phenomena. On the other hand, using ML techniques alone might not be suitable/successful in all cases especially when the studied data series has strong time dependency and the series show clear periodicity. So, applying wavelet transform to resolve the issue might be a good choice to increase the generalization ability of the ML techniques. The present study aimed at assessing the performance of the random forest (RF) method for predicting daily wind speed records at four sites in Iran. The wavelet transform was used for producing new sub series of data and make the new wavelet- random forest (WR) models. Both the RF and WR models were fed with the previously recorded wind speed values with different lag times. The obtained results revealed that the WRF has improved the performance of the RF model inn all studied locations, considerably.

Keywords: Wind speed, Iran, Random forest, Wavelet-random forest.

Introduction

Accurate prediction of wind speed is very important in agricultural disciplines, for industrial sites and hydrodynamic, coastal, wind wave, and ocean studies. Electrical system dependencies is one of the mostly applied techniques for wind speed prediction (1) Mohandas (2022) (2), while semi-empirical correlation (3) Ingle (2023) (4) and the stochastic time series analysis (5) and Verma (2022) have been applied dofar (6). Another alternative for those techniques might be using machine learning (ML) methods that capture the auto correlation nature of the wind speed to generate future predictions in different time scales and prediction intervals. Considerable studies focusing on using ML techniques have been reported by literature (e.g. Mohandes *et al*., 1998 (7); More and Deo, 2003 (8); Li and Shi, 2010 (9); Fadulemulla, 2023 (10); Wang *et al*. 2021 (11); Alkabbani *et al*. 2023 (12); Kumar Saini *et al*. 2023 (13)).

Although outstanding and powerful, ML methods can't map the nonlinear relations between the inputs (predictors) and target parameters when the data carry considerable noise and needs a set of pre-processing procedures. Among different pre-processing techniques, the wavelet transform is one of the commonly used ones that have been applied in various disciplines alike wind speed modeling. It generate a set of details and approximates for an available input patterns, so the most influential ones are selected as predictors. Owing to those advantages, the present study aimed at assessing the prediction ability of random forest (RF) methodology in predicting daily wind speed values at four locations of Iran. A further coupled wavelet-RF (WRF) methodology was also suggested for improving the prediction accuracy (14). For increasing the models validity, k-fold cross validation strategy was used to make a complete scan of the available patterns.

Materials and Methods

Random Forest (RF)

Employing decision trees to perform non-parametric classification, Random Forest (RF) is an ensemble machine learning technique, which improves the model's performance by mitigating inter-tree correlations through two key adjustments: training each tree with a subset of the training data and randomly selecting a subset of predictor variables for each decision node. RF assigns a class label to each pixel in an image based on the majority vote of the decision trees, and provides the option to measure uncertainty. Among various advantages of this method, one may mention its resilience when dealing with high-dimensional, correlated, or small datasets, as well as its capability to rank the importance of input variables.

Discrete Wavelet Transform (DWT)

The mother wavelet function $\psi(t)$ defined as $\int_{-\infty}^{+\infty}\psi(t)\mathrm{d}t = 0.\psi(t)$ can be obtained by compressing and expanding the following term $\psi(t)$:

$$\psi_{a.b}(t) = |a|^{-\frac{1}{2}}\psi\left(\frac{t-b}{a}\right) \qquad b\,\varepsilon\,R.\ a\,\varepsilon\,R.a \neq 0 \tag{1}$$

where $\psi a,b(t)$ = the successive wavelet, a = scale parameter, b = time parameter; R = the domain of real numbers. If $\psi a,b(t)$ satisfies Eq. 1, for the time series $f(t)$ ε $L^2(R)$ or finite energy signal, successive wavelet transform of $f(t)$ reads:

$$W_\psi f(a.b) = |a|^{-\frac{1}{2}}\int_R f(t)\,\bar{\psi}\left(\frac{t-b}{a}\right)dt \tag{2}$$

where$\bar{\psi}(t)$ = complex conjugate functions of $\psi(t)$. The wavelet transform is the decomposition of $f(t)$ under different resolution level (scale) as can be observed in Eq. 2. Now, let $a = a_0^j$, $b = kb_0a_0^j$, $a_0 > 1$, b_0 ϵ R, k, j are integer numbers. So, discrete wavelet transform of $f(t)$ can be written as:

$$W_\psi f(j.k) = a_0{}^{-j/2}\int_R f(t)\,\bar{\psi}(a_0^{-j}t - kb_0)dt \tag{3}$$

The most common (and simplest) choice for the parameters *a0* and *b0* is 2 and 1 time steps, respectively. This power of two logarithmic scaling of the time and scale is known as dyadic grid arrangement (15). Eq. 3 becomes binary wavelet transform when *a0* = 2, *b0* = 1:

$$W_\psi f(j.k) = 2^{-j/2}\int_R f(t)\,\bar{\psi}(2^{-j}t - k)dt \tag{4}$$

For a discrete time series *f(t)*, where occurs at different time t (i.e., here integer time steps are used), the DWT can be defined as

$$W_\psi f(j.k) = 2^{-j/2}\sum_{t=0}^{N-1} f(t)\,\bar{\psi}(2^{-j}t - k) \tag{5}$$

where $W\psi\, f(j,k)$ is wavelet coefficient for the discrete wavelet of scale a = 2j, b = 2jk.

Hybrid Models

In order to increase the prediction ability of the applied RF models, a further coupled wavelet-random forest (WRF) model was designed and implemented. Hence, the original time series of the studied variable (wind speed) were decomposed with a certain decomposition level into some sub-series using the Mallat's algorithm (15). Then, the generated sub-series were introduced as RF inputs.

Used Data

Daily wind speed records from four locations of I.R. Iran was used in the present study. **Figure 1** presents the locations of the sites. Data period covered a period of 5 years. At each location, a k-fold cross validation mode was applied to train and test the adopted methods. So, entire data were divided by 2 parts (based on the available data set) and each time the patterns of one complete year were used as testing data set, while the rest of patterns were used in developing/training the models. Repetition of this procedure allowed a complete sacn of the data during the study period. Once the best architecture of RF was identified, the model was applied and the 75% of the whole data were used for training and the last 25% was reserved for testing.

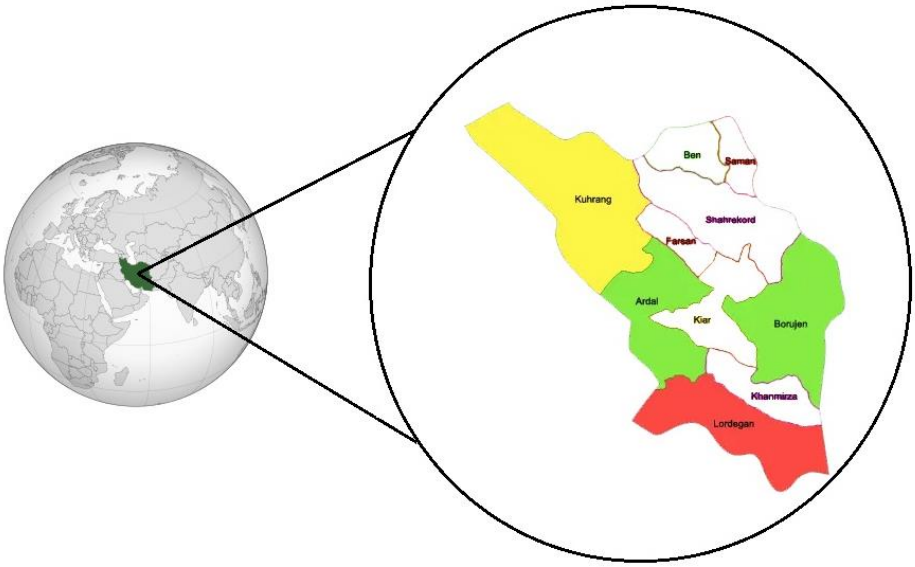


Figure 1. Geographical position of the study areas

Some statistical indices of the wind speed records at the studied sites have been listed in **Table 1**. In the table, Xmean, Xmax, Xmin, SD, CV and CSX show, respectively, the mean, maximum, minimum, standard deviation, coefficient of variation and skewness coefficient of wind records. As can be seen from the tables, the skewness coefficient is higher for the test data in all the locations. This may encounter the models with some difficulties during the temporal interpolation for testing. Nevertheless, the coefficient of variation (CV) values are higher in Kuhrang that can cause some difficulties with wind speed prediction there.

Performance Evaluation Measures

Three statistical indices as follows, were applied for evaluation of the RF and WRF models: the correlation coefficient (*R*), the root mean squared error (*RMSE*) and the scatter index (*SI*).

$$R = \frac{\sum_{i=0}^{n}(WS_0 - \overline{WS}_0)(WS_M - WS_M)}{\sqrt{\sum_{i=0}^{n}(WS_0 - \overline{WS}_0)^2 \sum_{i=0}^{n}(WS_M - \overline{WS}_M)^2}} \quad (6)$$

$$SI = \frac{RMSE}{\overline{WS}_0} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(WS_r - WS_0)^2}}{\overline{WS}_0} \quad (7)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(WS_r - WS_0)^2} \quad (8)$$

where WS_o is the observed wind speed value at the ith time step, WS_r is the simulated wind speed value at the ith time step, and n shows the patterns number.

Table 1. Statistical parameters of used hourly and daily wind speed data

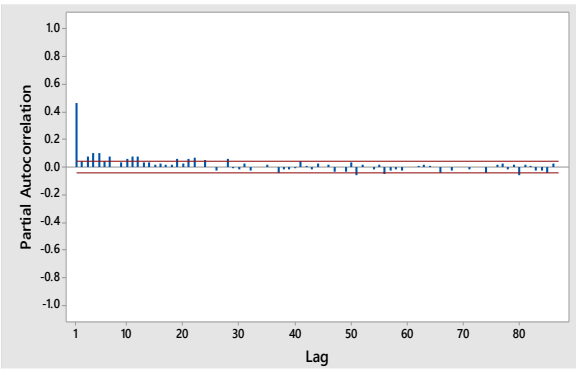
	X_{mean}	X_{max}	X_{min}	SD	C_V	C_{SX}
Ardal						
Training set	1.655	4.862	0.5	0.646	0.390	0.480
Testing set	1.680	3.740	0.561	0.537	0.320	0.709
Whole data	1.662	4.862	0.5	0.620	0.373	0.513
Borujen						
Training set	2.431	7	0.5	0.879	0.362	0.964
Testing set	2.323	6.358	0.654	0.866	0.373	1.370
Whole data	2.404	7	0.5	0.877	0.365	1.058
Kuhrang						
Training set	1.085	4.114	0.5	0.644	0.594	1.554
Testing set	0.911	3.459	0.5	0.456	0.501	1.605
Whole data	1.042	4.114	0.5	0.607	0.583	1.648
Lordegan						
Training set	1.681	6.919	0.5	0.759	0.452	0.978
Testing set	1.332	4.488	0.5	0.705	0.529	1.274
Whole data	1.594	6.919	0.5	0.761	0.477	1.007

Decomposing Wind Speed Records by Wavelet Transform

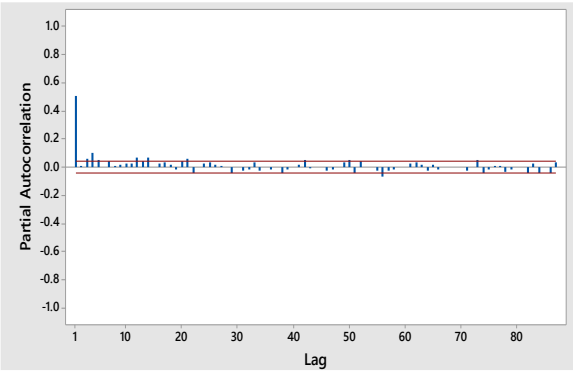
After decomposing the original time series of the wind speed records, the correlation values between each decomposition (D sub series) and original wind speed values were analyzed and those with the highest correlation values were selected to be included in the input matrix. Once the most effective signals were selected, they were used as inputs of RF model (instead of original wind speed records) to built the coupled WRF model.

Results and Discussion

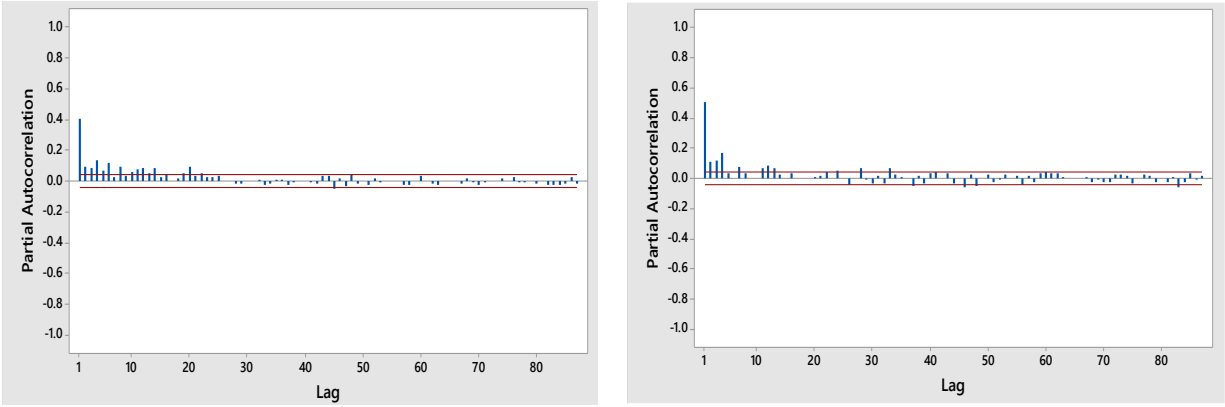
As the first step, the input parameters of the models should be determined. Here, based on the partial auto correlation function (PACF), the models were feed by different time lags of wind speed records. **Figure 2** illustrates the P diagrams of the wind speed records at the studies locations. Based on the PACF information, four time lags were selected as predictors of RF model to predict daily wind speed 1-day ahead. These lags were included in the input matrix step-by-step, so each time one predictor was additionally involved in the matrix so that the effect of each newly included variable can be distinguished, too. The same mode was adopted for the coupled WRF models to feed the input space.



Ardal



Borujen



Lordegan Kuhrang

Figure 2. Partial auto-correlation function (PACF) of daily wind speed at the studied lications.

Table 2. Statistics of wind speed records during the study period

	RF			WRF		
	SI	***NS***	***R***	***SI***	***NS***	***R***
			Ardal			
WS_t	0.337	0.218	0.469	0.111	0.659	0.771
WS_{t-1} , WS_t	0.334	0.154	0.421	0.118	0.649	0.764
WS_{t-2} , WS_{t-1} , WS_t	0.339	0.219	0.457	0.109	0.721	0.807
WS_{t-3} , WS_{t-2} , WS_{t-1} , WS_t	0.327	0.201	0.445	0.108	0.689	0.788
			Borujen			
WS_t	0.330	0.221	0.472	0.109	0.663	0.775
WS_{t-1} , WS_t	0.318	0.207	0.470	0.108	0.722	0.808
WS_{t-2} , WS_{t-1} , WS_t	0.330	0.229	0.473	0.108	0.683	0.785
WS_{t-3} , WS_{t-2} , WS_{t-1} , WS_t	0.315	0.222	0.475	0.109	0.607	0.742
			Lordegan			
WS_t	0.402	0.199	0.447	0.137	0.707	0.801
WS_{t-1} , WS_t	0.414	0.168	0.421	0.134	0.723	0.821
WS_{t-2} , WS_{t-1} , WS_t	0.403	0.201	0.440	0.143	0.682	0.789
WS_{t-3} , WS_{t-2} , WS_{t-1} , WS_t	0.405	0.213	0.462	0.142	0.750	0.833
			Kuhrang			
WS_t	0.537	0.264	0.525	0.173	0.699	0.799
WS_{t-1} , WS_t	0.481	0.301	0.556	0.162	0.834	0.868
WS_{t-2} , WS_{t-1} , WS_t	0.498	0.254	0.507	0.172	0.786	0.850
WS_{t-3} , WS_{t-2} , WS_{t-1} , WS_t	0.473	0.315	0.566	0.165	0.800	0.851

Table 2 sums up the statistical indices of the applied models during the testing period. Attending to the RF models, the results are generally far from the observed values as can be seen from higher *SI* and lower *NS* quantities. Among the studied locations, the models presented higher accuracy in Borujen, although the error values are still high. On the other hand, the quadruple-input RF model that uses the wind records of 4 successive days gave better results than the rest of input sets that may be due to the inclusion of more time steps (and consequently, more information on time series characteristics of wind records) in the input matrix. Overall, the obtained results by RF can't be applied for theoretical/practical issues due to the higher error values shown by *SI* and *NS* indices. Therefore, the necessity of coupling the model with wavelet transform for improving the outcomes of the RF is emphasized again.

o the coupled WRF models, the same four input combinations were constructed by including each time the details and approximations of a new wind speed series. Analyzing the error statistics of the WRF models in **Table 2** shows that, again, the models presented the most accurate results in Borujen. The average differences between the models performance metrics among the studied locations is low and it may be stated that the coupled WRF has been successful in predicting daily wind speed records of the studied locations.

urther, although the quadruple-input RF provided better predictions in all locations, the differences between the WRF models fed with different combinations of predictors are not more obvious and monotonously fluctuate among them at every four sites (16).
Comparing the results obtained by RF and WRF models revealed the performance improvement of the RF when the wavelet transform was applied for decomposition of the original time series.

Table 3. Performance improvement gained by WRF for two sample input set

	WS_{t-1} , WS_t			WS_{t-3} , WS_{t-2} , WS_{t-1} , WS_t		
	SI	**NS**	**R**	**SI**	**NS**	**R**
Ardal						
RF	0.334	0.154	0.421	0.327	0.201	0.445
WRF	0.118	0.649	0.764	0.108	0.689	0.788
Accuracy increment	35 %	24.0 %	55 %	33	29 %	56 %
Borujen						
RF	0.318	0.207	0.470	0.315	0.222	0.475
WRF	0.108	0.722	0.808	0.109	0.607	0.742
Accuracy increment	29%	29%	58%	29%	37%	64%
Lordegan						
RF	0.414	0.168	0.421	0.405	0.213	0.462
WRF	0.134	0.723	0.821	0.142	0.750	0.833
Accuracy increment	31%	23%	51%	29%	28%	55%
Kuhrang						
RF	0.481	0.301	0.556	0.473	0.315	0.566
WRF	0.162	0.834	0.868	0.165	0.800	0.851
Accuracy increment	30%	36%	64%	29%	39%	67%

Table 3 provides an illustration of performance improvement obtained through developing the WRF for two sample input set (double-input and quadruple-input models). In Ardal, WRF has improved the *SI* and *NS* indices, respectively, by 35% and 24% for double-input model and by 33% and 29% for quadruple-input model (17, 18). In Borujen, the performance improvement for SI and NS are, respectively, 29% and 29% for double-input model and 29% and 37% for quadruple-input model. In Lordegan, the WRF model improved the SI and NS values for duuble input model with 31% and 23% and for quadruple-input model with 29% and 28%, respectively (19). Finally, in Kuhrang, SI and NS improvement by WRF model were, respectively, 30% and 36% for double-input model and 29% and 39% for quadruple-input model. SI improvements were higher for double-input models in all locations, while the NS improvements of quadruple-input models were higher than double-input models (20). This may be due to the effect of model error reduction and variance simulation, as SI presents the error magnitude while NS includes the variance similarity, too.
Overall, it is seen that the WRF model is more powerful than the single RF model inn prediction of daily wind speed values in all locations. The resolution power of wavelet transform on the original wind speed signals that separate them into daily, monthly, etc periods may cause such performance improvement as discussed by Kisi *et al*. (2011) and Ahmed (2022) (21, 22). Since the wind speed records time series is periodic, clear presentation of such periods/cycles would be essential to better mapping of the nonlinear relations between different time events of a certain series.

Conclusion

A modeling process using random forest (RF) and wavelet random forest (WRF) methodologies were adopted here for predicting daily wind speed values at four stations of Iran. Daily data were introduced with various input combinations (based on partial auto correlation functions) to RF model and the performances of the models were assessed using statistical indices. Based on the overall evaluation, RF couldn't simulate daily wind speed variations well. So, wavelet decomposition was applied to the original wind speed time series and the decomposed components were coupled with the RF model with the same input combinations to generate the wind speed values through WRF model. Analysis confirmed that the WRF improved the RF performance with considerable reduction inn error terms in all locations and adopted input combinations. The results suggests that using wavelet decomposition for periodic data when the task is simulating the phenomena by machine learning techniques, would be essential to overcome some difficulties encountered

due to periodicity that detract the models accuracy level. Further studies might revise these outcomes using other machine learning techniques and different time scales to generate global general conclusions.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Njau EC. An electronic system for predicting air temperature and wind speed patterns. Renew Energ. 1994a;4(7):793-805.
2. Mohandas R, Ramani P, Mohapatra S. Corono-Condylar Distance: A Novel Indicator of Chronological Age–A Digital Radiographic Study. Ann Dent Spec. 2022;10(2):73-5. doi:10.51847/mPFYio61oR
3. Njau EC. Predictability of wind speed patterns. Renew Energ. 1994b;4(2):261-3.
4. Ingle NA, Algwaiz NK, Almurshad AA, AlAmoudi RS, Tariq A. Oral Health Utilization and Factors Affecting Oral Health Access Among Adults in Riyadh, KSA. Ann Dent Spec. 2023;11(1):65-9. doi:10.51847/9dlEqelquE
5. Rehman S, Halawani TO. Statistical characteristics of wind in Saudi Arabia. Renew Energ. 1994;4(8):949-56.
6. Verma P, Pandian SM. Prevalence of endodontically treated posteriors in patients undergoing orthodontic treatment cross-sectional radiographic evaluation. Ann Dent Spec. 2022;10:1-6. doi:10.51847/VtxY3JqaJ5
7. Mohandes MA, Rehman S, Halawani TO. A neural networks approach for wind speed prediction. Renew Energ. 1998;13(3):345-54.
8. More A, Deo MC. Forecasting wind with neural networks. Mar Struct. 2003;16(1):35-49.
9. Li G, Shi J. On comparing three artificial neural networks for wind speed forecasting. Appl Energ. 2010;87(7):2313-20.
10. Fadulemulla IA, AlShammari AD, ElHussein N, Seifeldin SA, AlShammari QT. Evaluation of the Anterior Cruciate Ligament Injury of Knee Joint Using Magnetic Resonances Imaging. Arch Pharm Pract. 2023;14(1):56-61. doi:10.51847/LxagvnoXIS
11. Wang Y, Zou R, Liu F, Zhang L, Liu Q. A review of wind speed and wind power forecasting with deep neural networks. Appl Energ. 2021;304:117766.
12. Alkabbani H, Hourfar F, Ahmadian A, Zhu Q, Almansoori A, Elkamel A. Machine Learning-based Time Series Modelling for Large-Scale Regional Wind Power Forecasting: a Case Study in Ontario, Canada. Clean Energ Syst. 2023:100068.
13. Saini VK, Kumar R, Al-Sumaiti AS, Sujil A, Heydarian-Forushani E. Learning based short term wind speed forecasting models for smart grid applications: An extensive review and case study. Electr Power Syst Res. 2023;222:109502.
14. Low LF, Islahudin F, Saffian SM. Development of Written Counseling Tool for Subcutaneous Anticoagulant Use in COVID-19 Patients. Arch Pharm Pract. 2023;14(2):19-24. doi:10.51847/RguC2DClhY
15. Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell. 1989;11(7):674-93.
16. Mahmoud Muddathir AR, Abdallah EI, Osman Elradi WE, Elbasheir ME, Abdelgadir RE, Waggiallah HA. Prevalence of HDNF due to ABO, Rh (D) and Other Blood Groups among Newborns, Sudan. J Biochem Technol. 2022;13(1):25-8. doi:10.51847/qvdQ4XmLif
17. Osadchuk MA, Osadchuk AM, Vasilieva IN, Trushin MV. The State Biology Museum Named after Kliment Arkadyevich Timiryazev as a Scientific and Educational Center. J Biochem Technol. 2023;14(1):7-12. doi:10.51847/OLKERwxo55
18. An TB, Linh DH, Anh NP, An TT, Tri N. Immobilization and Performance of Cellulase on Recyclable Magnetic Hydrotalcites. J Biochem Technol. 2022;13(1):13-9. doi:10.51847/APmQMAcejg
19. Agrawal M, Shrivastava S, Khare RL, Jaiswal S, Singh P, Hishikar R. Nephrotoxıcıty in Patıents on Tenofovır vs Non-Tenofovır Contaınıng Art Regımen: An Observatıonal Study. Pharmacophore. 2022;13(4):23-31. doi:10.51847/kNev4sPshf

20. Virtucio IL, Punzalan JM, Billones JB. Virtual Screening for SARS-COV-2 Entry Inhibitors by Dual Targeting of TMPRSS2 and CTSL. Pharmacophore. 2023;14(1):9-18. doi:10.51847/6IMWqjwVPa
21. Kisi O, Shiri J, Makarynskyy O. Wind speed prediction by using different wavelet conjunction models. Int J Ocean Clim Syst. 2011;2(3):189-208.
22. Ahmed NF, Albalawi AH, Albalawi AZ, Alanaz TA, Alanazi SN. Primary Immune Deficiency Disease in Saudi Children: Systematic Review. Pharmacophore. 2022;13(4):119-24. doi:10.51847/isksJQNQxO