[1]Surabhi Anuradha
[2*]Surabhi Sivakumar

A Hybrid Model based on Ensemble Learning from Residuals for Time Series Prediction

Journal of Electrical Systems

Abstract: - This study unveils an inventive algorithmic framework for a Hybrid Model, amalgamating the strengths of two forecasters to enhance the precision of time series predictions. The proposed model involves sequential training of two distinct models, each serving a unique purpose. The first model captures intricate patterns inherent in the data, while the second refines predictions by meticulously analyzing residuals. The synergy between these models significantly enhances the overall predictive capabilities of the system, showcasing its adaptability to the complex nature of time series data. Experimental results, precisely evaluated across various combinations of regression models, provide unequivocal evidence of the hybrid model's efficacy. Elasticnet Regressor (Enet) in combination with Random Forest Regressor (RFR), Gradient Boost Regressor (GBR), Ridge Regressor, and Decision Tree Regressor (DTR) has shown solid performance, with results falling within the 95% confidence interval. These outcomes showed that Enet can effectively capture intricate patterns that are displayed in sequential data.

Keywords: Hybrid model, Ensemble Learning, Residuals, Forecasting, Timeseries, Intricate Patterns.

I. INTRODUCTION

Time series analysis is a critical aspect of understanding temporal data, with various methodologies aiming to model and forecast these sequences effectively [1-2]. Among these, the statistical linear family of methods, including Auto-Regressive (AR) [3-4], Moving Average (MA) [5], autoregressive moving average (ARMA) [6-7] model and Auto-Regressive Integrated Moving Average (ARIMA) [8-11], stands out as traditional yet robust techniques widely used in time series modeling. However, these methods, rooted in linear correlation structures, may exhibit limitations when confronted with the complexities of real-world time series data, potentially leading to reduced forecasting accuracy [12-14].

Machine learning (ML) techniques present a viable remedy to these problems. Modeling a variety of data types can be made more flexible by using techniques like Artificial Neural Networks (ANNs) [15-20] and Support Vector Regression (SVR) [21-23], which can capture nonlinear temporal trends. Yet, because of parameter misspecification, machine learning models may find it difficult to balance the learning of both linear and nonlinear patterns, which frequently leads to underfitting or overfitting [24].

Researchers have developed hybrid systems that integrate statistical linear and machine learning models, using their respective strengths, realizing their benefits. In order to overcome the shortcomings of individual models and improve forecasting accuracy and generalization capability, these hybrid systems describe linear and nonlinear patterns separately [25-28].

One such hybrid approach involves combining a linear statistical model with an ensemble of ML models, particularly focusing on modeling the residuals. This ensemble-based method aims to improve the system's generalization capacity while effectively addressing the challenges posed by real-world time series data [29-31]. Domingos S. de O. Santos Júnior et. al [32]. handled the heteroscedasticity challenge of the residuals by combining linear statistical model with an ensemble of ML model. He employed ensemble method to model the residuals, thus improved generalization capacity of the system. Paulo S. G de Mattos Neto et. al [33]. used hybrid approach that combine ML models using residuals to enhance the performance of sea surface temperature forecasting. They did the experimentation with two types of ML models: SVR and long short-term memory (LSTM) and attained higher accuracy than individual statistical and ML models and found that the nonlinear combination of the ML models obtained the best performance.

Although we can create hybrids that work, we still need to learn more about how time series are put together. Trends, seasons, and cycles constitute the three primary dependencies observed in time series data [34]. Trends

[1] Department of CSE(AIML), Keshav Memorial Institute of Technology, Hyderabad, Telangana, India. ORCID: 0009-0008-6091-859X
[2] Department of Chemistry, Anil Neerukonda Institute of Technology, Visakhapatnam, Andhra Pradesh, India. ORCID: 0000-0002-0173-1169
[*] Corresponding author Email: sivakumar.chemistry@anits.edu.in

delineate the long-term directional shifts or tendencies within the data. Seasonality denotes the recurrent patterns or fluctuations that transpire at specific intervals, be it daily, weekly, monthly, or annually. On the other hand, cycles signify repetitive patterns or fluctuations occurring over extended time frames, even though not strictly periodic like seasonality. Many time series can be succinctly characterized by an additive model incorporating these three components, alongside unpredictable and entirely random errors termed residuals. The residuals of a model represent the disparity between the target data used for training and the predictions generated by the model. Essentially, they signify the distinction between the observed data points and the values predicted by the model, encapsulating the discrepancy between the actual curve and the fitted curve.

This research introduced a novel framework for a hybrid model focusing on ensemble learning from residuals. The approach integrates two regression models sequentially, aiming to enhance time series forecasting accuracy by leveraging the complementary strengths of each model. Specifically, the first model is trained on input features and the target variable, generating predictions and residuals. These residuals are then utilized by the second model along with a separate set of input features, refining the forecast. This innovative strategy capitalizes on the information captured in residuals to improve the overall predictive performance of the hybrid model.

II. PROPOSED METHOD

In the domain of time series prediction, the amalgamation of different models often yields more accurate and robust results. This algorithm outlines the key steps involved in creating a Hybrid Model, harnessing the strengths of two distinct regression models to enhance forecasting precision. Linear regression demonstrates proficiency in extrapolating trends but lacks the capacity to learn interactions. Conversely, XGBoost exhibits adeptness in learning interactions but faces challenges when extrapolating trends. The proposed algorithm introduces a hybrid forecaster that merges these two complementary learning algorithms, allowing the strengths of one to compensate for the weaknesses of the other.

Fig. 1 visually represents the sequential process of the Hybrid Model algorithm, breaking down each step to offer a clear and comprehensible depiction. The detailed architecture of both model training and test stages is expounded in Fig. 2 (a) and (b), providing a thorough illustration of the intricate steps involved in these key aspects of the Hybrid Model.

Algorithm: HybridModel (model_1, model_2**):**

Step 1: Initialization:

1. Set model_1 as the first regression model.
2. Set model_2 as the second regression model.

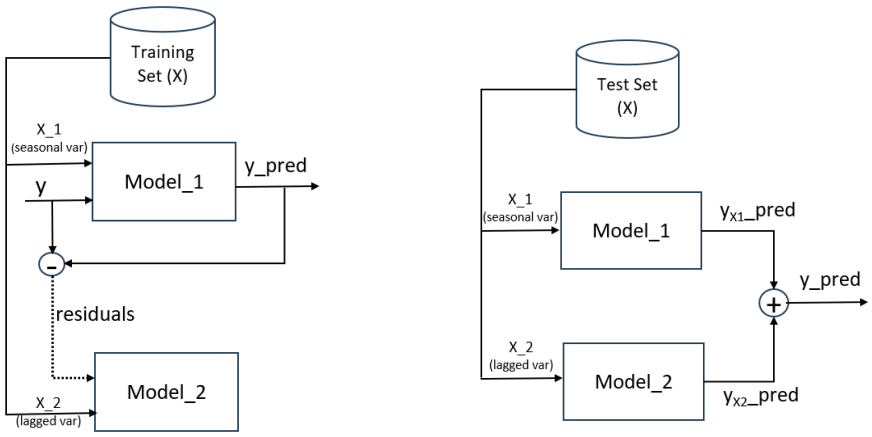
Step 2: Training the Model

1. Fit the First Model (model_1):
    a. Train model_1 using the seasonal variables (X_1) and the target variable (y).
2. Generate Predictions and Calculate Residuals:
    a. Obtain predictions (y_pred) by applying model_1 to the input features.
    b. Calculate residuals (y_diff) by subtracting the predictions from the actual target variable (y).
3. Fit the Second Model (model_2):
    a. Train model_2 using the logged variables (X_2) and residuals (y_diff) obtained in step 5b.

Step 3: Prediction

1. Generate Predictions Using the First Model:
    a. Utilize model_1 to generate predictions for the input features (X_1).
2. Calculate Residuals Using the Second Model:
    a. Apply model_2 to the second set of input features (X_2) to calculate residuals.
3. Combine Predictions and Residuals:
    a. Add the residuals obtained from model_2 to the predictions from model_1.

Fig. 1. Algorithm for Hybrid model



(a) Train (b) Test

Fig. 2. Phases of the proposed model.

III. EXPERIMENTATION

A. *Dataset Analysis*

The model underwent experimentation to ascertain the daily high price of gold in the Indian market. Prior to this experimental phase, an in-depth analysis of the dataset was undertaken to examine the inherent trends, seasonality patterns, correlation and fluctuations within the gold price data.

In an effort to examine the trends in the 'High' prices within an annual context, a 365-day rolling average was computed. This calculation aimed to accentuate yearly trends by smoothing the representation of the original data, effectively reducing short-term fluctuations or noise. The outcome is a visually clearer depiction of long-term trends or patterns in the time series data, as illustrated in Fig. 3. The graph shows the trend of the high price of gold over time. The blue line shows the actual high price of gold, and the orange line shows the rolling average of the high price of gold over the past 365 days. The analysis reveals distinctive trends over various periods. From 2014 to 2019, a gradual increase in prices is evident, suggesting a steady upward trajectory. Subsequently, there is a noticeable and rapid upswing from 2019 to 2021, indicating a phase of accelerated growth in 'High' prices. However, from 2021 to 2022, the trend appears to plateau, signifying a period of stagnation in the 'High' prices.

The periodogram plot shows the strength of different periodic components in the time series data. The x-axis shows the frequency of the periodic components, and the y-axis shows the strength of the components. The plot presented in Fig. 4 shows that there is a strong annual component in the data, with a peak at a frequency of 1. This means that there is a strong seasonal pattern in the data that repeats every year. There is also a weaker semi-annual component, with a peak at a frequency of 2. This means that there is a weaker seasonal pattern in the data that repeats every six months. Another useful tool in time series analysis is the Partial Autocorrelation Function (PACF) plot, which provides a visual depiction of the partial autocorrelations between a time series and its lag values. This analytical technique reduces the impact of shorter lags while focusing on the correlation at a particular lag. The Fig. 5 shows that the high price of gold is only significantly correlated with its first lag. This means that the price of gold today is only significantly correlated with the price of gold yesterday. The other lags in the plot fall outside the box, which means that they are not significantly correlated with the high price of gold.

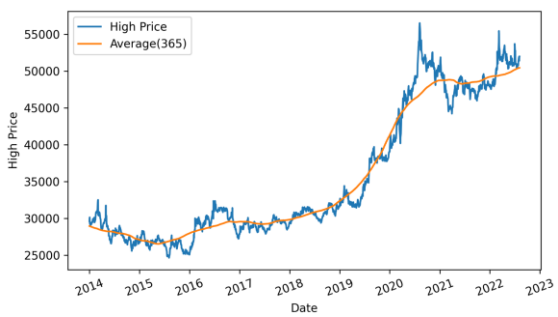


Fig. 3. Trend analysis of high price of gold over time

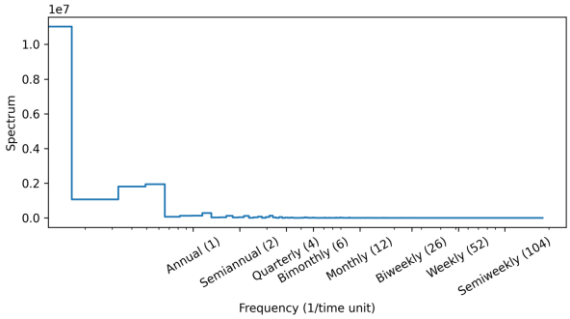


Fig. 4. Seasonality study using Periodogram plot

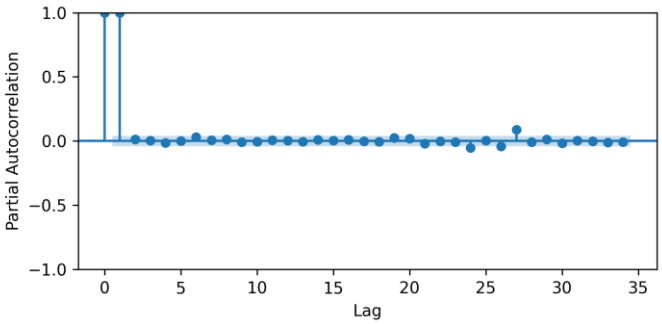


Fig. 5. Partial Autocorrelation between current and previous lags

B. Data Prepration

A comprehensive and well-structured approach was used to prepare data for training and testing the hybrid model. Two distinct sets of features were generated and the first set X_1 contained time-based features that captured the annual seasonality of the dataset. Meanwhile, the second set X_2 features were primarily based on lagged values of various columns from the input dataset, including 'Price', 'High', 'Low', 'Volume', and 'Chg%'. Lagging involved shifting these values by one time step (in this case, one day), creating a time series of historical values.

IV. Results and Discussion

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

The proposed model underwent testing and evaluation with various combinations of regression algorithms, incorporating ElasticNet (Enet), Random Forest Regressor (RFR), Decision Tree Regressor (DTR), Ridge, and Gradient Boosting Regressor (GBR) techniques. Table 1 presented the outcomes of the hybrid model across different regressor combinations, accompanied by their respective Root Mean Square (RMSE) values. A naive method is employed to assess the confidence of hybrid model predictions for time series dataset. Initially, the residuals, which represent the differences between consecutive observations in the test dataset, are calculated. These residuals capture the deviation of actual values from their preceding values, serving as a proxy for the uncertainty or variability in the data. The standard deviation of these residuals is then computed as a measure of dispersion or spread in the dataset. Subsequently, a confidence testing procedure is applied, where the standard deviation is multiplied by 1.96, corresponding to the critical value for a 95% confidence interval in a standard normal distribution. If this product exceeds the mean absolute error between the actual and predicted values generated by the hybrid model, it suggests that the model's predictions fall within the 95% confidence interval established by the data's variability. Essentially, this procedure aims to evaluate whether the model's predictions are statistically consistent with the variability observed in the test dataset, providing an indication of the reliability and robustness of the model's performance. The last column of the table showcased the hybrid model with regressor combinations falling within the 95% confidence.

Among the various combinations, Enet paired with RFR demonstrates promising results, achieving a lower train and test RMSE values, indicating better generalization and predictions fall within the 95% confidence interval, indicating reliability in forecasting. Enet paired with GBR and Ridge regressors exhibits predictions that consistently fall within the 95% confidence interval, affirming the reliability of the forecasting outcomes. Similarly, the Enet-Enet combination also performs well, falling within the 95% confidence interval. It is noteworthy that certain combinations, such as Enet-DTR and RFR-DTR exhibit a training RMSE of 0, which could signify overfitting. However, they achieve higher test RMSE values, indicating poor generalization. Despite this, predictions fall within the 95% confidence, suggesting consistency in forecasting.

The results underscore the importance of selecting appropriate regressor combinations, as different algorithms exhibit varying strengths and weaknesses in capturing the nuances of the gold price dataset. The promising performance of Enet-RFR combination suggests the potential for refining and optimizing the hybrid model for improved accuracy in predicting daily high gold prices in the Indian market. Further fine-tuning and exploration of

additional regressor combinations may enhance the model's robustness and broaden its applicability. Overall, these findings lay the groundwork for continued refinement and optimization of the hybrid model, ensuring its effectiveness in real-world forecasting scenarios.

The RMSE values for the training and test datasets of the hybrid model concerning the Enet_RFR and Enet_Ridge configurations were depicted in Figures (a), (b), and (c) of 6 and 7.

Table 1. Hybrid Models within 95% confidence interval

Model_1	Model_2	Train_RMSE	Test_RMSE	95%CI_Margin	MAE	Prediction within 95% confidence
RFR	RFR	36.923	555.156	363.447	500.904	No
Enet	RFR	88.976	245.657	363.447	230.008	Yes
Enet	Enet	221.102	234.64	363.447	205.121	Yes
Enet	GBR	40.665	294.996	363.447	271.597	Yes
Enet	Ridge	220.049	228.029	363.447	197.967	Yes
Enet	DTR	0	261.602	363.447	231.966	Yes
RFR	DTR	0	528.176	363.447	475.859	No

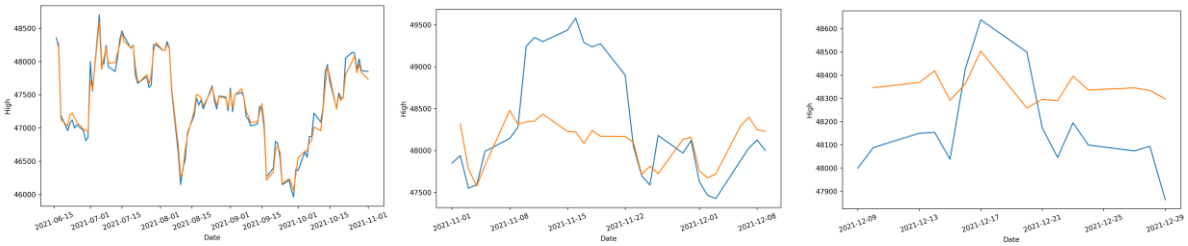


Fig. 6. (a) Train; (b) Validation; (c)Test RMSE of the ‘Enet_RFR’ Hybrid Model.

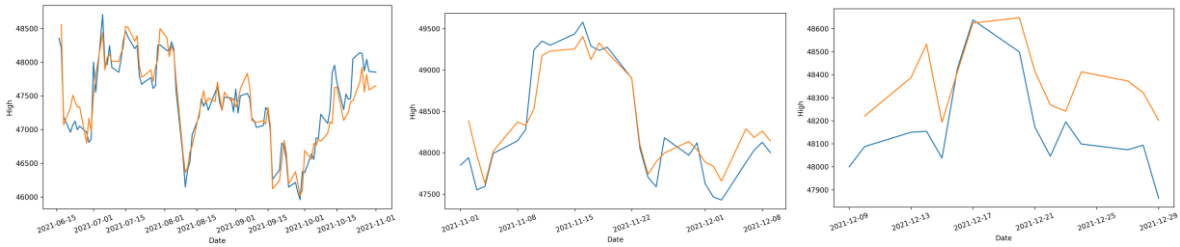


Fig. 7. (a) Train; (b) Validation; (c)Test RMSE of the ‘Enet_Ridge’ Hybrid Model.

V. Conclusion and Future Scope

The algorithmic framework proposed in this paper for the Hybrid Model entailed the sequential training of two models. The first model captured patterns in the data, and the second model refined predictions through the analysis of residuals. The synergy between these models markedly improved the overall predictive capabilities of the system, showcasing its adaptability to the inherent complexities of time series data. The experimental results provided unequivocal evidence of the efficacy of the hybrid model, demonstrating its robust performance across a range of regression model combinations, with the majority falling within the 95% confidence interval.

Looking ahead, there is a promising avenue for future research to explore more sophisticated hybrid models that extend beyond the combination of two models. A proposal for an advanced hybrid model could involve integrating three models in a synergistic manner to harness a broader range of information and patterns within the data. This multi-model approach could potentially provide even more accurate and versatile predictions, offering enhanced

insights into complex time series datasets. Further investigations into the development and evaluation of such advanced hybrid models could pave the way for improved forecasting capabilities and contribute to the evolving landscape of predictive modelling in time series analysis.

REFERENCES

[1] Bichescu, B., & Polak, G. G. (2023). Time series modeling and forecasting by mathematical programming. Computers & Operations Research, 151, 106079.

[2] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 12, pp. 11106-11115).

[3] Dansana, D., Kumar, R., Das Adhikari, J., Mohapatra, M., Sharma, R., Priyadarshini, I., & Le, D. N. (2020). Global forecasting confirmed and fatal cases of COVID-19 outbreak using autoregressive integrated moving average model. Frontiers in public health, 8, 580327.

[4] Xu, W., Peng, H., Zeng, X., Zhou, F., Tian, X., & Peng, X. (2019). Deep belief network-based AR model for nonlinear time series forecasting. Applied Soft Computing, 77, 605-621.

[5] Ivanovski, Z., Milenkovski, A., & Narasanov, Z. (2018). Time series forecasting using a moving average model for extrapolation of number of tourist. UTMS Journal of Economics, 9(2).

[6] Singh, B., & Pozo, D. (2019). A guide to solar power forecasting using ARMA models. In 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe) (pp. 1-4). IEEE.

[7] Zaman, U., Teimourzadeh, H., Sangani, E. H., Liang, X., & Chung, C. Y. (2021). Wind speed forecasting using ARMA and neural network models. In 2021 IEEE Electrical Power and Energy Conference (EPEC) (pp. 243-248). IEEE.

[8] Abonazel, M. R., & Abd-Elftah, A. I. (2019). Forecasting Egyptian GDP using ARIMA models. Reports on Economics and Finance, 5(1), 35-47.

[9] Alghamdi, T., Elgazzar, K., Bayoumi, M., Sharaf, T., & Shah, S. (2019). Forecasting traffic congestion using ARIMA modeling. In 2019 15th international wireless communications & mobile computing conference (IWCMC) (pp. 1227-1232). IEEE.

[10] Gourav, Rekhi, J. K., Nagrath, P., & Jain, R. (2020). Forecasting air quality of Delhi using ARIMA model. In Advances in Data Sciences, Security and Applications: Proceedings of ICDSSA 2019 (pp. 315-325). Springer Singapore.

[11] Tarmanini, C., Sarma, N., Gezegin, C., & Ozgonenel, O. (2023). Short term load forecasting based on ARIMA and ANN approaches. Energy Reports, 9, 550-557.

[12] Hossain, I., Rasel, H. M., Imteaz, M. A., & Mekanik, F. (2020). Long-term seasonal rainfall forecasting using linear and non-linear modelling approaches: a case study for Western Australia. Meteorology and Atmospheric Physics, 132, 131-141.

[13] Xu, W., Peng, H., Zeng, X., Zhou, F., Tian, X., & Peng, X. (2022). A hybrid modeling method based on linear AR and nonlinear DBN-AR model for time series forecasting. Neural Processing Letters, 1-20.

[14] Hossain, I., Rasel, H. M., Imteaz, M. A., & Mekanik, F. (2018). Long-term seasonal rainfall forecasting: efficiency of linear modelling technique. Environmental Earth Sciences, 77, 1-10.

[15] Borghi, P. H., Zakordonets, O., & Teixeira, J. P. (2021). A COVID-19 time series forecasting model based on MLP ANN. Procedia Computer Science, 181, 940-947.

[16] Namasudra, S., Dhamodharavadhani, S., & Rathipriya, R. (2021). Nonlinear neural network based forecasting model for predicting COVID-19 cases. Neural processing letters, 1-21.

[17] Bata, M. T. H., Carriveau, R., & Ting, D. S. K. (2020). Short-term water demand forecasting using nonlinear autoregressive artificial neural networks. Journal of Water Resources Planning and Management, 146(3), 04020008.

[18] Slater, L., Arnal, L., Boucher, M. A., Chang, A. Y. Y., Moulds, S., Murphy, C., ... & Zappa, M. (2022). Hybrid forecasting: using statistics and machine learning to integrate predictions from dynamical models. Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi. org/10.5194/hess-2022-334, in review.

[19] Tealab, A. (2018). Time series forecasting using artificial neural networks methodologies: A systematic review. Future Computing and Informatics Journal, 3(2), 334-340.

[20] Cirstea, R. G., Micu, D. V., Muresan, G. M., Guo, C., & Yang, B. (2018). Correlated time series forecasting using multi-task deep neural networks. In Proceedings of the 27th acm international conference on information and knowledge management (pp. 1527-1530).

[21] Abbasimehr, H., & Baghery, F. S. (2022). A novel time series clustering method with fine-tuned support vector regression for customer behavior analysis. Expert Systems with Applications, 204, 117584.

[22] Luo, X., Yuan, X., Zhu, S., Xu, Z., Meng, L., & Peng, J. (2019). A hybrid support vector regression framework for streamflow forecast. Journal of Hydrology, 568, 184-193.

[23] Aghelpour, P., Mohammadi, B., & Biazar, S. M. (2019). Long-term monthly average temperature forecasting in some climate types of Iran, using the models SARIMA, SVR, and SVR-FA. Theoretical and Applied Climatology, 138(3-4), 1471-1480.

[24] Júnior, D. S. D. O. S., de Mattos Neto, P. S., de Oliveira, J. F., & Cavalcanti, G. D. (2023). A hybrid system based on ensemble learning to model residuals for time series forecasting. Information Sciences, 649, 119614.
[25] Belmahdi, B., Louzazni, M., & Bouardi, A. E. (2020). A hybrid ARIMA–ANN method to forecast daily global solar radiation in three different cities in Morocco. The European Physical Journal Plus, 135, 1-23.
[26] Júnior, D. S. D. O. S., de Oliveira, J. F., & de Mattos Neto, P. S. (2019). An intelligent hybridization of ARIMA with machine learning models for time series forecasting. Knowledge-Based Systems, 175, 72-86.
[27] Shoeibi Omrani, P., Vecchia, A. L., Dobrovolschi, I., Van Baalen, T., Poort, J., Octaviano, R., ... & Muñoz, E. (2019). Deep learning and hybrid approaches applied to production forecasting. In Abu Dhabi International Petroleum Exhibition and Conference (p. D031S103R001). SPE.
[28] Lee, J., & Cho, Y. (2022). National-scale electricity peak load forecasting: Traditional, machine learning, or hybrid model?. Energy, 239, 122366.
[29] Wu, H., & Levinson, D. (2021). The ensemble approach to forecasting: a review and synthesis. Transportation Research Part C: Emerging Technologies, 132, 103357.
[30] Livieris, I. E., Pintelas, E., Stavroyiannis, S., & Pintelas, P. (2020). Ensemble deep learning models for forecasting cryptocurrency time-series. Algorithms, 13(5), 121.
[31] Wu, W., Emerton, R., Duan, Q., Wood, A. W., Wetterhall, F., & Robertson, D. E. (2020). Ensemble flood forecasting: Current status and future opportunities. Wiley Interdisciplinary Reviews: Water, 7(3), e1432.
[32] Júnior, D. S. D. O. S., de Mattos Neto, P. S., de Oliveira, J. F., & Cavalcanti, G. D. (2023). A hybrid system based on ensemble learning to model residuals for time series forecasting. Information Sciences, 649, 119614.
[33] de Mattos Neto, P. S., Cavalcanti, G. D., de O. Santos Júnior, D. S., & Silva, E. G. (2022). Hybrid systems using residual modeling for sea surface temperature forecasting. Scientific Reports, 12(1), 487.
[34] Zhang, M. (2018). Time series: Autoregressive models ar, ma, arma, arima. University of Pittsburgh.