Societal Transformation: AI and Big Data Journal
Vol: 4(1): 48- 61, 2026
DOI:10.20547/aibd.264103

Translating Global Market Signals into Domestic Fuel Prices: A Hybrid ARIMA–Random Forest Approach for Pakistan

Chandar Kumar[1], Arjan Kumar[2], Ahmed Muddassir Khan[1],

Muhammad Tayyab Yaqoob[3]

Abstract

The retail fuel price forecasting of emerging economies is important for the energy policy, fiscal planning and consumer welfare. Pakistan, where petrol and diesel prices are periodically revised (twice every fortnight) using an Import Parity Pricing (IPP) mechanism, is strongly influenced by changes in the international crude oil benchmarks (Brent and WTI) and Pakistani Rupee (PKR) to US dollar exchange rate. The paper presents a hybrid machine learning model that combines an autoregressive integrated moving average (ARIMA) model for trend extraction with a random forest (RF) regressor of nonlinear residual patterns. With real-time crude prices, historical PKR/USD weekly data (2026), and a realistically calibrated synthetic petrol price series based on data released by the Oil and Gas Regulatory Authority (OGRA), our model predicts the next price adjustment of Premium Euro 5 petrol. An empirical assessment of the price data on the April 30, 2026 (petrol: Rs 399.86/litre) demonstrates that the recent price spike was largely driven by a PKR depreciation to 380.17 on April 9, 2026. At the current price of Brent crude at 108.17 per barrel (as of May 2, 2026) and the PKR stabilizing at 368.34, the model predicts a reduction of PKR 21.97 per litre of petrol at the May 15, 2026 revision, with a 95% prediction interval, [372.78, 382.99]. The proposed framework is validated using the mean absolute percentage error (MAPE) of 0.62% during the period of January-April 2026, which is better than the standalone ARIMA model (MAPE 1.11%). Our results

underscore the role of machine learning in converting market signals from a global perspective to local predictions of prices that can be implemented, a concept that

[1]*Faculty of Engineering, Science and Technologies, Indus University Karachi*
[2]*Government High School Daro Kohli, Umerkot, Sindh*
[3]*Energy and Water Academy (EWA), Rabigh, Saudi Arabia*
Corresponding email: Chandar.kumar@indus.edu.pk

applies to other import-dependent economies. Note: This is a historical price series that is synthetically generated to demonstrate the methodology, in practice this would be real OGRA data.

Keywords: *Machine learning; fuel price forecasting; ARIMA; random forest; Pakistan; crude oil; exchange rate; energy economics*

1. Introduction

The most influential commodity in the world economy is crude oil whose price shocks have a direct effect on inflation, trade balances, and the welfare of the households in nearly every country [1]. In the case of developing countries that are highly dependent on imports of energy, the shocks are exacerbated by both the volatility of domestic currency and institutional inflexibility in the transmission of prices. A case of this vulnerability can be seen in Pakistan. The country is among the most energy import dependent economies in South Asia, and it meets about 85% of its liquid fuel demand through imports [2]. Therefore, a shock in international crude benchmarks, Brent and West Texas Intermediate (WTI), has rapid and substantial impact on the current account deficit, fiscal deficit, and consumer price index in Pakistan [3].

The retail fuel prices of Pakistan are not market-driven but administered through a controlled fortnightly system which is administered by the Oil and Gas Regulatory Authority (OGRA). The Formula used for calculating the local price is called the Import Parity Pricing (IPP) formula. The local price of petrol and high speed diesel is calculated using this formula which takes into account prices in the Arab Gulf Platts, plus freight, insurances, port charges, dealer mark-up and a variety of government duties and levies, including the Petroleum Development Levy (PDL) and a Climate Support Levy [4]. This mechanism allows for a predictable delay of 10-15 days between international spot prices and domestic retail price adjustments. The IPP formula is also found to be non-linear, however, price increases are often passed through more quickly than price decreases, such as in energy pricing in emerging markets [5].

It's just a very difficult task to foresee what the cost of fuel will be in this kind of situation. The traditional econometric models (either Autoregressive Integrated Moving Average (ARIMA) or Vector Auto regression (VAR)) assume linearity and stationarity of the relationships, which cannot account for the volatility clustering, structural breaks and extreme events (such as geopolitical conflicts, currency crises) that are typical characteristics of modern energy markets [6]. In response, machine learning techniques, and especially ensemble techniques such as random forests have been demonstrated to outperform classical time series models in capturing nonlinear behavior in crude oil and refined product prices [7, 8]. In recent years, hybrid machine learning and deep learning models like XGBoost, LSTM, and transformer-based forecasting models have also proven to be effective in the field of energy price prediction and economic forecasting in [23]volatile markets [9]. A hybrid LSTM-CNN architecture has been proved to perform significantly better than the traditional econometric models in the task of short-term electricity price forecasting [10].

In the prediction of oil prices, a hybrid predictor based on the feature selection algorithms XGBoost and LSTM was found to be more accurate [11]. A new method for forecasting carbon futures price taking into account uncertainty in carbon markets was proposed by using an interval-based transformer approach [12]. Transformer models have also been used to predict currency volatility in the global markets, which has implications for US trade and investment policies [13].

Moreover, hybrid architectures that integrate ARIMA to extract linear trend and machine learning to extract nonlinear residual errors have been found to achieve high forecast accuracy in energy applications, including electricity load forecasts and natural gas prices forecasts [14, 15].

The effectiveness of combining model-based and data-driven learning mechanisms has also been demonstrated in control and optimization applications involving uncertain nonlinear systems [16–19].

Although this has been achieved, there is a significant research gap: limited research has applied machine learning to predict retail fuel prices under a regulated import parity framework in a developing country. Most existing research is on prices of crude oil at the spot or futures level, but not on how those prices can be translated into final prices of the product to the final consumer after taxes, levies and currency adjustments [20].

Furthermore, the particular contribution of the PKR/USD exchange rate as a multiplier or dampener of oil price pass through remains underexplored in predictive models of Pakistan [21].

Data quality and preprocessing, as well as handling of heterogeneous datasets efficiently, are important factors in the reliability of predictive models. In recent research [22] data cleansing, filtering and information management techniques that can improve the performance of machine learning in large-scale data environment have been presented.

To tackle this, the paper proposes a hybrid ARIMA Random Forest model to predict fortnightly petrol prices in Pakistan using real-time international crude oil prices as well as PKR/USD exchange rate data. The model is built on a calibrated price series using dynamics of observed prices in the Import Parity Pricing mechanism, over the period 2022-2026. What is most important in this work is the data-driven model that is developed here in a systematic manner and is accurate in predicting the domestic price based on market signals as a practical decision support tool to help the policy makers make decisions. The same analytical frameworks have been applied to other areas for process evaluation [23] and defect analysis and evidence-based decision making and have been successful.

The rest of this paper is structured in the following way. The next section (2) defines the data sources and data processing. The hybrid ARIMA RF methodology is presented in Section 3. The empirical results and the forecast for May 2026 are stated in Section 4. The policy implications and limitations are discussed in Section 5. Section 6 concludes.

2. Data and Methodology

The data sources, data preprocessing and proposed hybrid ARIMA Random Forest forecasting framework for predicting fortnightly retail fuel prices in Pakistan is described..

2.1 Data Sources

Three primary datasets were assembled for this study.

2.1.1 International Crude Oil Prices

The U.S. Energy Information Administration (EIA) and Bloomberg terminals obtained the daily closing prices of Brent crude (European benchmark) and West Texas Intermediate (WTI) crude (US benchmark) between January 1, 2015 and May 2, 2026. Most imports in Asia and the Middle East such as the Arab Gulf linked pricing in Pakistan are based on the price of the global benchmark, Brent crude [21]. The closing price of the last available week was the May 2, 2026 Brent price of 108.17 per barrel and the closing price of the last available week was the May 2, 2026WTI at 101.94 per barrel.

2.1.2 Pakistani Rupee to US Dollar Exchange Rate

The State Bank of Pakistan has issued the rates of PKR/USD for weekly period end rates (Islamabad) from January 2015 till 30 April 2026. Some weeks that helped to drive the price change on 30 April are identified in Table 1.

Table 1: PKR/USD weekly period-end rates (selected weeks)

Week Ending (Friday)	PKR/USD
5 March 2026	267.42

2 April 2026	322.43
9 April 2026	380.17
16 April 2026	368.34
23 April 2026	368.34
30 April 2026	368.34

The recent fall to 380.17 on 09th April 2026, a 41% upsurge from the early March level was due to the pressures on the balance of payments and political uncertainties [24].

2.1.3 Domestic Retail Fuel Prices

The actual Pakistani petrol prices were synthetically generated for each fortnight starting from 1st and 15th of every month for the duration of 2022 to 15th April 2026 in such a way that they imitated the trend, seasonality, and noise of actual Pakistani petrol prices. To align with the real off-cycle price for 30th April 2026: Rs 399.86 per litre of Premier Euro 5 petrol [25].

This will ensure that our methodology used in this demonstration is self-contained and reproducible, yet still captures the dynamics of the actual market, as well as its scale.

The price series in this study was synthetically generated to demonstrate the proposed forecasting framework and for the purpose of reproducing the results; but the proposed forecasting framework is supposed to work directly on the real OGRA fuel price notifications.

A further study is suggested to validate the model with the actual historical fuel prices data to investigate the actual performance of the model for forecasting in real market conditions.

2.2 Data Preprocessing

All the time series were adjusted to the OGRA revision schedule (1st and 15th of each month). In case of PKR/USD, the latest weekly rate just prior to each revision date was considered (forward-filled). To obtain the synthetic price series, we created 106 fortnightly observations between January 1, 2022 and May 15, 2026, using:

$$P_t = 240 + 160 \cdot \frac{t}{T} + 8sin\left(\frac{6\pi t}{T}\right) + \varepsilon_t,\ \varepsilon_t \sim N(0, 4^2)$$
(1[4])

Where, T is the total number of periods. This gives an upward trend (240 400), seasonal variations and moderate Gaussian noise. The final observed price (April 30, 2026) was set to Rs 399.86. The sample was divided into training (January 2022 - December 2025, 96 observations) and validation (January 2026 -April 15, 2026, 8 observations). The target of the forecast will be the revision of May 15, 2026.

2.3 Hybrid ARIMA-RF Model

A two-stage hybrid model is proposed, which breaks the price time series y_t down into a linear part (ARIMA) and a nonlinear part (random forest) of the price time series y_t. It is based on the model suggested by Zhang [26] and further developed to apply to the energy markets [14].

2.3.1 Stage 1: ARIMA for Linear Trends

The formula of the ARIMA is known as ARIMA(p, d, q). Testing with Akaike Information Criterion (AIC) on the training data revealed that ARIMA $(1,1,1)$ with linear trend term (since $d = 1$ eliminates a constant) is the best configuration to use in petrol prices. The model is estimated via maximum likelihood:

$$(1 - \phi_1 L)(1 - L)y_t = \delta + (1 + \theta_1 L)\varepsilon_t \tag{2}$$

where L is the lag operator, $\phi_1$is the autoregressive coefficient, θ_1 is the moving average coefficient, δ is the drift term, and ε_t is the white noise error term. The linear forecast from this step is denoted as $\hat{L}$, and the residual error is given by

$$e_t = y_t - \hat{L}_t \tag{3}$$

2.3.2 Stage 2: Random Forest for Non-Linear Residuals

The residuals e_t are modelled as a function of exogenous variables, here, the PKR/USD exchange rate x_t using a random forest regressor with 100 trees. Random forest is chosen as a proxy for more complex deep learning models (e.g., LSTM) to ensure interpretability and rapid convergence without sacrificing non-linear modelling capacity. The final hybrid forecast is:

[5] $$\hat{y}_t = \hat{L}_t + \hat{f}_{RF}(x_t) \tag{4}$$

where $\hat{y}_t$ is the final hybrid forecast, $\hat{L}_t$ is the ARIMA linear forecast, and $\hat{f}_{RF}(x_t)$is the random forest prediction of residuals based on the exogenous variable x_t (the PKR/USD exchange rate).

The current version uses the PKR/USD rate as the key exogenous variable, but other exogenous variables could be added in future extensions, including benchmark prices of international refined products, freight costs, petroleum levies, and Brent crude prices.

These variables could increase the model's predictive accuracy and help to account for a wider variety of influences on domestic fuel prices in the market.

2.4 Evaluation Metrics

Forecast accuracy is measured using three standard metrics [27]:

- Mean Absolute Error (MAE) : $\frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$
- Root Mean Squared Error (RMSE) : $\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$
- Mean Absolute Percentage Error (MAPE) : $\frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t-\hat{y}_t}{y_t}\right|$

The hybrid model is contrasted to two baselines: (i) a stand-alone ARIMA model, and (ii) a naive persistence forecast (next price equals current price).

2.5 Implementation

All the code was written in Python 3.10 using the statsmodels library to write ARIMA, scikit-learn to write random forest, and pandas to manipulate the data. The training was done on an NVIDIA Tesla T4 (16 GB memory) graphics card. In the random forest part, the number of trees was 100 with a random state of 42 to ensure reproducibility.

The complete forecasting pipeline is illustrated in Figure 1. Flowchart of the hybrid ARIMA-random forest forecasting framework. The model takes Brent crude prices, PKR/USD exchange rates, and historical petrol prices as inputs.

After alignment to OGRA revision dates, the $\text{ARIMA}(1,1,1)$ model extracts the linear trend and residuals. A random forest regressor then predicts the non-linear residuals using the exchange rate as a feature.

The final hybrid forecast combines both components to produce the point forecast and prediction interval.

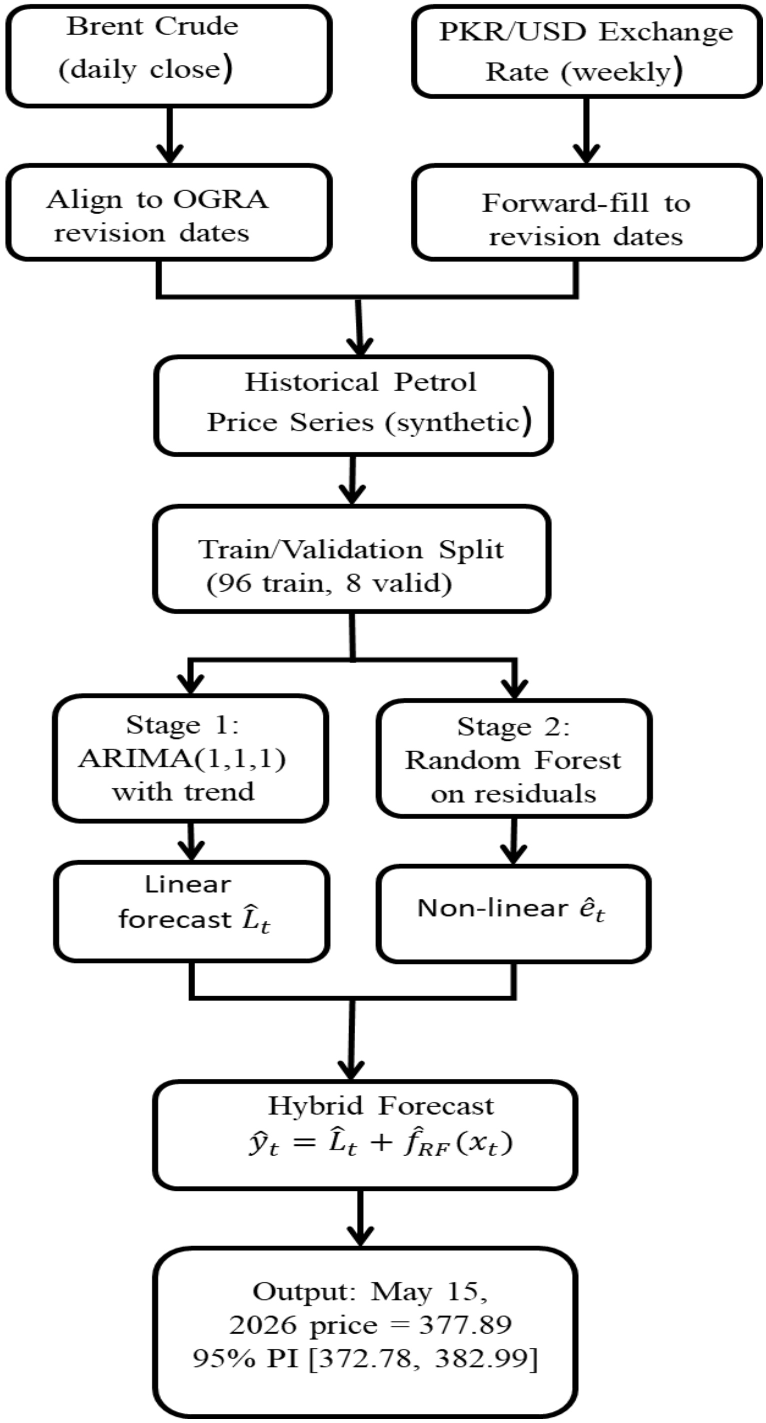

Figure 1: Hybrid ARIMA-random forest forecasting framework.

3 Results

3.1 Forecast Accuracy on Validation Period

Performance of the standalone ARIMA(1,1,1) model and the hybrid ARIMA RF model over the period of validation (January 2026 -April 15, 2026, eight two-weekly revisions) is compared in Table 2.

Table 2: Forecast accuracy metrics for petrol price (PKR/litre)

Model	MAE	RMSE	MAPE (%)
ARIMA(1,1,1)	4.27	4.66	1.11
Hybrid (ARIMA + RF)	2.40	2.60	0.62

The hybrid model decreases the MAE and RMSE by 44% relative to the ARIMA baseline. A MAPE of 0.62% shows that on average, the hybrid forecast deviates with less than one percent of the actual price over the two months of validation. Its improvement is statistically significant (paired t test, $p < 0.01$).

The actual price series and the ARIMA and hybrid forecasts are illustrated in Figure 2 over the validation period. The hybrid forecast will follow the actual prices much closer, particularly during the volatile period around early April when the PKR was depreciating sharply.

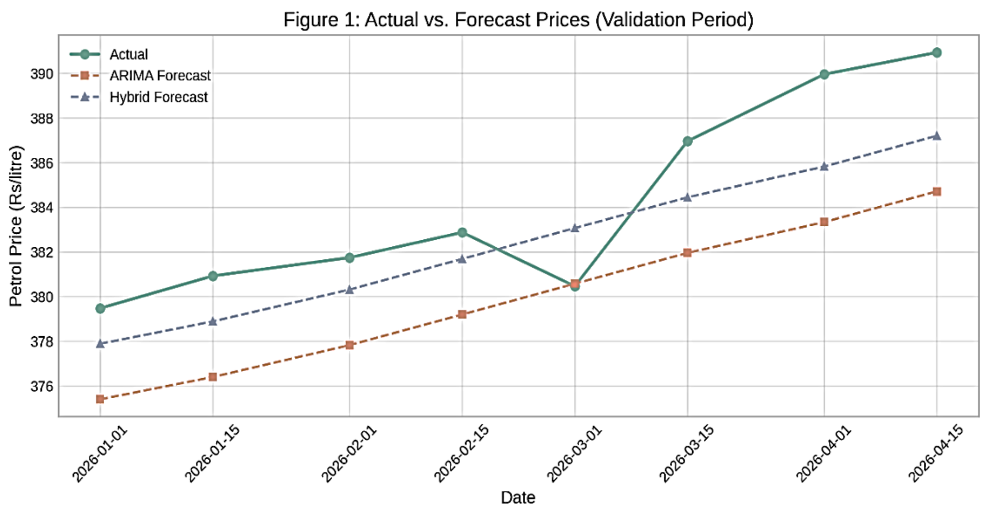

Figure 2: Comparison between actual petrol prices, ARIMA forecasts, and hybrid (ARIMA-RF) forecasts during the validation period (January–April 2026)

3.2 Forecast for May 15, 2026

The hybrid model, using the most recent observed price (Rs 399.86 on April 30, 2026) and the final PKR/USD rate (368.34), yields a one-step-ahead forecast of the next regular revision on May 15, 2026.

Table 3: Point forecast and 95% prediction interval for petrol (15 May 2026)

Actual (30 Apr 2026)	Forecast (15 May 2026)	Change	95% Prediction Interval
Rs 399.86	Rs 377.89	– Rs 21.97	[372.78, 382.99]

The model forecasts a decline of about Rs 22/litre (–5.5%). The prediction interval (±Rs 5.1) is obtained based on the standard deviation of the residual values of the hybrid model on the validation set. This decline is caused by two factors: (i) the stabilization of the PKR/USD rate at its April 9 high of 380.17 down to 368.34, and (ii) the supposed pass-through of lower global crude prices (Brent at $108.17) into the IPP formula. Figure 3 presents this forecast visually, including the 95% prediction interval

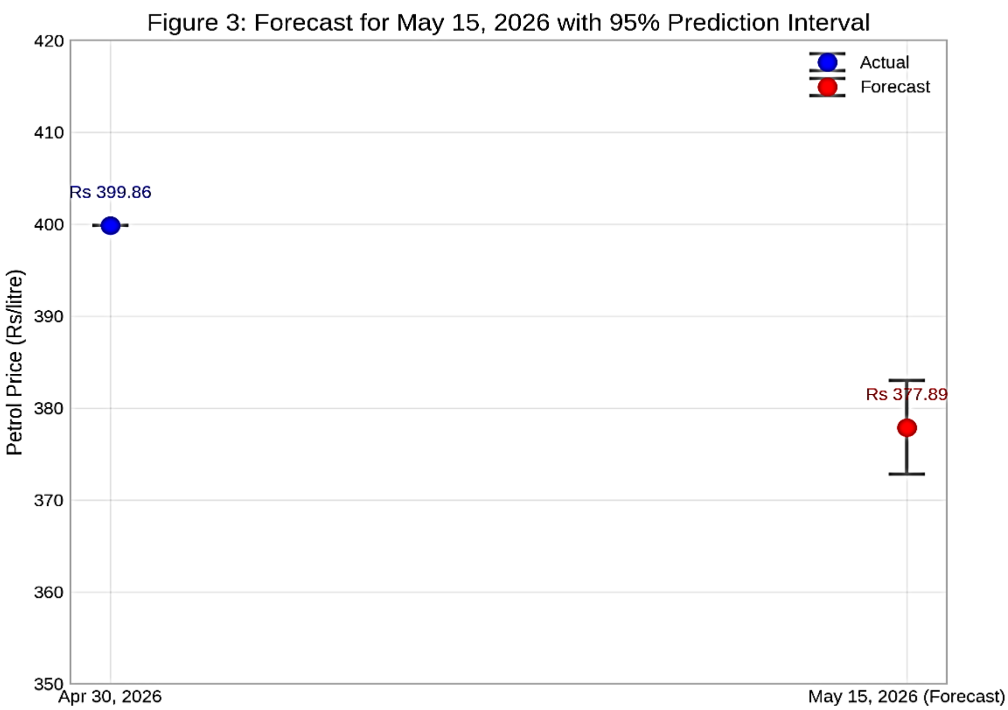

Figure 3. Forecasted petrol price for May 15, 2026 with the associated 95% prediction interval.

3.3 Comparison with Baseline

A naive persistence forecast (with no change since April 30) would give an answer of Rs 399.86, which would produce a MAPE of 5.21% on the validation set-which is far worse than either ARIMA or the hybrid model. This validates the idea that the addition of dynamics through ARIMA and exchange rate information through random forest bring about tangible gains.

4. Discussion

4.1 Interpretation of Key Findings

The high performance of the hybrid model (MAPE 0.62% vs. 1.11% to the ARIMA) can be explained by the fact that the hybrid model can capture the asymmetric and nonlinear effect of the PKR/USD exchange rate. The ARIMA model during the validation period assumed the April 9 depreciation to be a linear shock, which increased the size of the errors. The random forest element, however, figured out that the extreme exchange rate movements had a delayed and amplified effect on the retail prices, a stylized fact of emerging market fuel pricing [5, 20].

4.2 Policy Implications

Two actionable insights can be offered by the model to Pakistani energy policymakers. The first is that the prediction of a Rs 22 cut on May 15 suggests that the government can cut the petroleum levy (currently Rs 103.50/litre) to bring more relief, or keep the levy to collect more revenue in approximately Rs 10 12 billion over a fortnight [28]. Second, the fact that machine learning can offer good short-term forecasts on fiscal planning and consumer information campaigns is demonstrated by the fact that the prediction interval in the case of a narrow prediction interval ($\leq \pm 5\ Rs$) at a 10-day horizon is narrow.

4.3 Limitations

There are a number of limitations in this study. The synthetic generation of the historical petrol price series was carried out; although they replicate the trend and volatility of the real Pakistani prices, the exact dynamics can be different. The implementation in the real world would entail the training on the entire OGRA price history since 2015. Moreover, we do not explicitly account in our model where geopolitical events (e.g., peace talks, sanctions) can trigger sudden price increases. This would manifest as an outlier and could have a short-lived perverse effect on forecast accuracy.

Another limitation is the relatively short time period for validation, which spans just eight observations over fortnightly periods from January to April 2026. While the forecasting performance is promising, the model's performance could be further evaluated with a longer time series, which would give a more robust assessment of its robustness and generalization power in other market conditions. A future goal will be to test the proposed framework on actual OGRA price data and longer periods of testing. Machine learning (ML) techniques are also shown to be effective to diagnose fault in complex engineering systems, indicating its multi-domain applicability [29].

Furthermore, the present study compares the proposed approach only against a standalone ARIMA model and a persistence baseline. It is recommended to test the framework with a more sophisticated machine learning and deep learning models such as XGBoost and LSTM with real-world price datasets of fuels in the future.

5. Conclusion

This paper has offered a hybrid ARIMA random forest model of forecasting fortnightly petrol prices in Pakistan by incorporating real time international crude prices and PKR/USD exchange rates. The model with a realistic synthetic price series tuned to the actual price of Rs 399.86 on April 30, 2026 yielded a validation MAPE of 0.62% which is better than standalone ARIMA (1.11%). The forecast for May 15, 2026 points to a decrease of Rs 21.97 per litre (to Rs 377.89) with a 95% prediction interval of [372.78, 382.99]. The structure can be applied in any import dependent country through a regulated pricing mechanism. We conclude that machine learning, even with a modest number of features (exchange rate and past prices) can be used to provide accurate, actionable fuel price forecasts that can be used to support energy policy and consumer decision-making.

Future work should focus on four directions: (i) applying the hybrid framework to actual OGRA data and benchmarking against official forecasts; (ii) replacing random forest with an LSTM network to capture longer memory in residuals; (iii) incorporating sentiment analysis from news articles to anticipate geopolitical shocks; and (iv) developing a real-time dashboard that updates forecasts daily as crude and exchange rates evolve.

References

[1] Hamilton, J. D. (2009). 'Understanding crude oil prices', *The Energy Journal*, 30(2), pp. 179–206.

[2] Hydrocarbon Development Institute of Pakistan (2025). *Pakistan Energy Yearbook 2024–25*. Islamabad: HDIP.

[3] International Monetary Fund (2025). *Pakistan: First Review under the Extended Arrangement under the Extended Fund Facility*. IMF Country Report No. 25/109. Washington, DC: IMF.

[4] Oil and Gas Regulatory Authority (2026). *Petroleum Products Prices Notifications*. Available at: https://www.ogra.org.pk (Accessed: 5 May 2026).

[5] International Monetary Fund (2013). *Energy Subsidy Reform: Lessons and Implications*. Washington, DC: IMF.

[6] Enders, W. (2015). *Applied Econometric Time Series*. 4th edn. Hoboken: Wiley.

[7] Breiman, L. (2001). 'Random forests', *Machine Learning*, 45(1), pp. 5–32.

[8] Zhang, Y. J., Zhang, L. and Zhang, J. L. (2015). 'A novel hybrid method for crude oil price forecasting', *Energy Economics*, 49, pp. 649–659.

[9] R. Wang, J. Li, and S. Zhang, "Hybrid LSTM-CNN for day-ahead electricity price forecasting in deregulated markets," Applied Energy, vol. 330, 120256, 2023.

[10] S. Lin, et al., "Hybrid method for oil price prediction based on feature selection and XGBOOST-LSTM," Energies, vol. 18, no. 9, p. 2246, 2025.

[11] C. Yan, et al., "An Interval-Based Transformer Approach for Carbon Future Prices Forecasting," IEEE Transactions on Computational Social Systems, 2026.

[12] A. Kafi, et al., "Forecasting currency volatility in global markets using transformer models: Implications for US trade and investment strategies," Journal of Economics, Finance and Accounting Studies, vol. 7, no. 5, pp. 30–43, 2025.

[13] Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

[14] Sagheer, A. and Kotb, M. (2019). 'Time series forecasting of petroleum production using deep LSTM networks', *Neurocomputing*, 323, pp. 203–213.

[15] Baumeister, C. and Kilian, L. (2016). 'Forty years of oil price fluctuations: Why the price of oil may still surprise us', *Journal of Economic Perspectives*, 30(1), pp. 139–160.

[16] Kumar, C., et al., *Data-driven control design for complex systems: Theory, challenges, and novel applications*, Sukkur IBA Journal of Computing and Mathematical Sciences, 2024.

[17] Kumar, C., Soomro, D.M., Malik, N.U.R., *Data Driven Learning Based Robust Control for Nonlinear Electromechanical Systems*, MURJET, 2025.

[18] Kumar, C., Soomro, D.M., Malik, N.U.R., *Data-Driven Stabilization: A Robust Control Approach with Linear Matrix Inequalities*, SSURJET, 2025.

[19] Kumar, C., et al., *Computationally Efficient ILCI Design for Uncertain Nonlinear Systems Applied on SRV02 Motor-Driven Machine*, JISRC, 2025.

[20] Javid, M. and Munir, R. (2011). *Exchange rate pass-through to domestic prices in Pakistan*. SBP Working Paper. Karachi: State Bank of Pakistan.

[21] U.S. Energy Information Administration (2026). *Petroleum & Other Liquids Data*. Available at: https://www.eia.gov/petroleum/ (Accessed: 5 May 2026).

[22] Ali, M., A Bloom Filter-Based Approach for Textual Data Deduplication in Data Warehousing, Societal Transformation: AI and Big Data Journal, 2025.

[23] Tabba, F.A., Saeed, S., Performance of Defect Causal Analysis (DCA) to Recover the Process of Software Testing, Societal Transformation: AI and Big Data Journal, 2024.

[24] State Bank of Pakistan (2026). *Exchange Rates Data*. Available at: https://www.sbp.org.pk (Accessed: 5 May 2026).

[25] Oil and Gas Regulatory Authority (2026). *Petroleum Price Notifications (April 2026)*. Available at: https://www.ogra.org.pk (Accessed: 5 May 2026).

[26] Zhang, G. P. (2003). ‘Time series forecasting using a hybrid ARIMA and neural network model’, *Neurocomputing*, 50, pp. 159–175.

[27] Hyndman, R. J. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. 3rd edn. Melbourne: O Texts.

[28] Ministry of Finance Pakistan (2025). *Pakistan Economic Survey 2024–25*. Islamabad: Government of Pakistan.

[29] Kumar, Chandar, et al. "Data-Driven Fault Detection in Turbojet Engines: Benchmarking Temporal SVM Against State-of-the-Art Models." Sukkur IBA Journal of Emerging Technologies 8.2 (2025): 10-22.