From ARIMA to Transformers: The Evolution of Time Series Forecasting with Machine Learning

Muhammad Ahmad[1], Hina Qamar[2], Ahmed Abdul Rehman[3] and Roidar khan[4]

https://doi.org/10.62345/jads.2025.14.3.18

Abstract

This study examines the evolution of time series forecasting in the context of crude oil price prediction, tracing the methodological shift from classical statistical models, such as ARIMA, to advanced machine learning architectures, including LSTM networks and Transformer models. Using a dataset of daily crude oil prices from 1995 to 2016, the analysis evaluates model performance across multiple horizons and error metrics. Results indicate that ARIMA, while interpretable, shows limitations in handling volatility and nonlinearities (RMSE = 2.10). LSTM improves accuracy by capturing long-term dependencies (RMSE = 1.55), while Transformer achieves the highest performance (RMSE = 1.32, R^2 = 0.87). These findings highlight how attention-based models outperform traditional econometric approaches by addressing volatility clustering and long-range dependencies. The study contributes to the forecasting literature by demonstrating the paradigm shift from ARIMA to Transformers and offering a framework for selecting models based on accuracy, interpretability, and forecasting horizon.

Keywords: Time Series Forecasting, ARIMA, LSTM, Transformer, Machine Learning, Crude Oil Prices, Volatility, Deep Learning.

Introduction

Time series forecasting has long been a central problem in statistics, economics, and machine learning because of its crucial role in planning, risk management, and decision support. In financial and energy markets, forecasting crude oil prices is particularly important due to the commodity's influence on inflation, trade balances, and global economic stability. However, crude oil markets are inherently volatile, characterized by nonlinear dynamics, regime shifts, and sharp fluctuations driven by geopolitical events, supply-demand imbalances, and speculative behavior. These properties make forecasting oil prices both challenging and essential. Traditional approaches to time series forecasting were dominated by statistical models such as the Autoregressive Integrated Moving Average (ARIMA) family. ARIMA and its extensions have provided interpretable frameworks for capturing autocorrelation, stationarity, and short-term dynamics, and they have become widely used in energy economics. Volatility-focused models, including ARCH and GARCH, extended this paradigm by explicitly modeling time-varying variance, a critical feature of commodity prices. While these models are transparent and analytically tractable, they often struggle with nonlinear dependencies, structural breaks, and long-range patterns commonly

[1]*Abdul Wali khan University, Mardan, Pakistan. Email: amdm8008@gmail.com*
[2]*Women University, Swabi, Pakistan. Email: hina.qamar@wus.edu.pk*
[3]*Bahria University Islamabad, Pakistan. Email: hina.qamar@wus.edu.pk*
[4]*University of Malakand, Pakistan. Corresponding Author Email: roidarkhan.stats@gmail.com*

Compliance with ethical standards: There are no conflicts of interest (financial or non-financial). This study did not receive any funding.

observed in crude oil markets. The rise of machine learning introduced new opportunities to overcome these limitations. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, demonstrated strong performance in capturing sequential dependencies and nonlinear interactions. More recently, the Transformer architecture has revolutionized sequence modelling by using attention mechanisms to learn long-term dependencies without recurrence, achieving state-of-the-art accuracy across diverse forecasting tasks. Variants such as the Temporal Fusion Transformer (TFT), Informer, and Autoformer have further adapted the framework to multivariate and long-horizon time series forecasting. This study situates crude oil price forecasting within this broader methodological evolution from ARIMA to Transformers by systematically comparing models across accuracy metrics, forecast horizons, and volatility conditions. Using daily crude oil price data from 1995 to 2016, we present descriptive statistics, model diagnostics, and forecasting performance across ARIMA, LSTM, and Transformer models. The results highlight both the interpretability–accuracy trade-off and the substantial gains made possible by deep learning and attention-based models. Ultimately, this research provides empirical evidence for the paradigm shift in time series forecasting and demonstrates its practical significance for energy economics and financial decision-making.

The literature on time series forecasting spans several methodological generations, beginning with classical statistical models and progressing toward advanced machine learning and deep learning approaches. The earliest wave of forecasting research was shaped by the seminal work of Box and Jenkins (1970), whose Autoregressive Integrated Moving Average (ARIMA) framework established a systematic approach to model identification, estimation, and diagnostics. Extensions such as the ARCH model, introduced by Engle (1982), and the GARCH model, developed by Bollerslev (1986), addressed volatility clustering, making them particularly influential in financial and commodity market analysis. These methods proved effective in capturing short-term autocorrelation and conditional variance; however, they often failed to account for nonlinear dependencies or structural breaks that are common in crude oil markets. The limitations of purely linear models encouraged researchers to explore alternative approaches. Hamilton (1994) provided a comprehensive econometric framework for non-stationary and regime-switching time series, while subsequent studies integrated exogenous factors such as macroeconomic indicators, geopolitical risks, and exchange rates to improve oil price forecasts. However, these traditional approaches relied heavily on manual feature engineering and strong stationarity assumptions, restricting their flexibility in highly volatile settings. The introduction of machine learning marked a significant shift in the paradigm. Early applications of artificial neural networks (ANNs) in the 1990s demonstrated the potential of nonlinear learning for economic and energy forecasting; however, issues of overfitting and interpretability limited their widespread adoption. With the development of Long Short-Term Memory (LSTM) networks by Hochreiter and Schmidhuber (1997), recurrent models overcame vanishing gradient problems and became highly effective in capturing long-term sequential dependencies. Recent applications have confirmed that LSTM models outperform ARIMA in forecasting crude oil prices, reducing the root mean squared error (RMSE) and better capturing nonlinear shocks.

The latest frontier is defined by attention-based models, particularly those based on the Transformer architecture. Initially introduced by Vaswani et al. (2017) for natural language processing, Transformers revolutionized sequence modeling by dispensing with recurrence and relying entirely on self-attention to capture long-range dependencies. Adaptations such as the Temporal Fusion Transformer (TFT) (Lim et al., 2021), Informer (Zhou et al., 2021), and Autoformer (Wu et al., 2021) have been applied to time series forecasting, demonstrating superior accuracy and scalability. These models excel not only in predictive performance but also in interpretability, thanks to attention weights that identify key time periods and variables influencing forecasts. Empirical studies in the energy sector reflect this evolution. Zhang (2022) compared ARIMA and LSTM in crude oil forecasting and found that LSTM provided more stable forecasts across multiple horizons. He et al. (2023) applied TFT to oil markets, demonstrating substantial improvements in multi-step forecasts by incorporating exogenous covariates, such as supply indicators and global demand shocks. Foroutan et al. (2024) benchmarked deep learning models across energy commodities and reported that ensemble and Transformer-based architectures consistently outperformed linear baselines. Hybrid methods that combine decomposition techniques (e.g., VMD or EMD) with LSTM or Transformer architectures (Guan et al., 2023; Huang et al., 2024) have also achieved strong results by modeling the multi-scale features of volatile time series. Taken together, the literature documents a clear methodological trajectory: ARIMA and GARCH remain foundational and interpretable baselines, LSTM networks improve accuracy by learning nonlinear sequential patterns, and Transformer-based models now represent the state of the Art by efficiently modeling long-range dependencies and heterogeneous drivers. For crude oil price forecasting, this progression highlights the need to adopt modern machine learning approaches to address the volatility, regime shifts, and global interdependencies inherent in the market.

Methodology

Data Description

The dataset used in this study consists of daily crude oil prices spanning the period from January 1995 to December 2016. This 22-year horizon covers multiple oil price regimes, including the Asian Financial Crisis (1997–98), the oil price boom of the mid-2000s, the 2008 global financial crisis, and the sharp decline during the U.S. shale oil expansion (2014–2016). The dataset offers a rich testbed for forecasting models, owing to its pronounced volatility, long-term trends, and structural breaks. To facilitate statistical and machine learning modeling, the raw data were cleaned by handling missing values, standardizing date formats, and constructing key derived variables such as daily returns, moving averages, Bollinger Bands, and 30-day rolling volatility.

Software and Tools

All data preprocessing, statistical modeling, and visualization were conducted in R (version 4.x), an open-source software environment widely used for econometrics and time series analysis. The packages employed include ggplot2 for high-quality visualization, dplyr for data manipulation, zoo for rolling statistics, and forecast for ARIMA modeling. For deep learning experiments, we utilized the keras and tensorflow packages in R, allowing for the construction and training of LSTM networks. Transformer-based forecasting was implemented through specialized R interfaces linked to Python backends, reflecting the hybrid workflow often adopted in advanced time series research. The use of R ensured reproducibility and facilitated the combination of classical econometric models with cutting-edge deep learning techniques.

Modeling Framework

The study follows a stepwise modeling framework reflecting the methodological evolution of time series forecasting. First, descriptive analysis was performed to summarize the behavior of crude oil prices, including the mean, variance, and volatility clustering. Next, classical ARIMA models were estimated and selected based on Akaike Information Criterion (AIC), Bayesian Information

Criterion (BIC), and root mean square error (RMSE). Volatility was further analyzed using rolling standard deviation measures. In the second phase, Long Short-Term Memory (LSTM) networks were applied to capture nonlinear dependencies and long-range temporal patterns, with hyperparameters tuned across multiple epochs. Finally, Transformer architectures were employed to leverage self-attention mechanisms for multi-horizon forecasts, addressing limitations of recurrence-based models. Model performance was evaluated using RMSE, MAE, and R^2, enabling a robust comparative assessment.

Evaluation Procedure

The evaluation process involved splitting the dataset into training (1995–2012) and testing (2013–2016) subsets to mimic real-world forecasting conditions. Forecast accuracy was tested at multiple horizons: 1 day, 7 days, 14 days, and 30 days ahead. Each model's performance was compared across these horizons to highlight trade-offs between short- and long-term predictive accuracy. Diagnostic checks included residual analysis, stationarity testing using the ADF and KPSS tests, and error decomposition into bias and variance components. For machine learning models, cross-validation and hyperparameter tuning were performed to prevent overfitting and ensure generalization. This rigorous evaluation ensured that results reflect not only in-sample fit but also realistic out-of-sample forecasting power, providing a fair assessment of the progression from ARIMA to LSTM and Transformer models.

Results

Table 1 provides a foundational overview of crude oil price dynamics and daily return distributions, which establish the statistical properties any forecasting model must contend with. The average crude oil price across the period is $ 52.40, with a median of $ 48.10, indicating a relatively symmetric distribution but with a long tail toward extreme highs. The maximum observed price is $145.20, coinciding with the 2008 oil bubble, while the minimum is $18.60, reflecting the severe downturn of the late 1990s. The standard deviation of 18.7 highlights the inherent volatility in energy markets, which poses challenges for linear models such as ARIMA that assume stable variance. Daily returns show even greater variability, averaging just 0.05% but ranging from -10.2% to +12.3%, with a standard deviation of 1.25%. Such extreme swings reflect fat-tailed, non-Gaussian distributions, which are common in financial time series. These statistics highlight a key challenge: while ARIMA can capture average dynamics after differencing, it struggles with heavy tails and sharp regime shifts. By contrast, modern machine learning models such as LSTMs and Transformers are better equipped to handle such irregularities, as their architectures can learn nonlinearities and adapt to abrupt changes. In short, Table 1 demonstrates that crude oil prices are highly volatile, prone to extreme events, and exhibit a distribution that is far from normal. These characteristics justify the need for advanced forecasting methods, marking the transition from traditional ARIMA toward more flexible, data-driven approaches in time series forecasting.

Table 1: Summary statistics of crude oil prices and daily returns

Statistic	Crude Oil Price	Daily Returns
Mean	52.4	0.05%
Median	48.1	0.03%
Std Dev	18.7	1.25%
Min	18.6	-10.2%
Max	145.2	12.3%

Table 2 compares several ARIMA specifications using standard model selection criteria AIC, BIC, and RMSE. The ARIMA(2,1,2) model emerges as the best performer, with the lowest AIC (2445.8) and BIC (2469.4), alongside an RMSE of 2.10. This suggests that two autoregressive and two moving average terms, with first-order differencing, capture the dynamics of crude oil prices better than competing alternatives. By comparison, ARIMA(1,1,1), while simpler, has a higher AIC (2450.3) and RMSE (2.15), making it less efficient despite being more parsimonious. ARIMA(5,1,0) yields an even worse RMSE of 2.22, indicating that adding excessive lags can lead to overfitting without improving accuracy. Similarly, ARIMA(0,1,5) produces an RMSE of 2.18, confirming that relying solely on moving average terms is suboptimal. These results illustrate both the strengths and weaknesses of ARIMA: it offers a transparent, statistically rigorous framework, but its performance is highly sensitive to lag specification and differencing. Moreover, even the best ARIMA configuration yields errors above 2, which is substantial given the volatility of crude oil prices. This limitation underscores why researchers increasingly turn to deep learning models, which adaptively capture nonlinearities without rigid assumptions about lag structure. While ARIMA remains a valuable benchmark, Table 2 confirms that its forecasting accuracy is inherently constrained by linearity, paving the way for more advanced models like LSTMs and Transformers.

Table 2: ARIMA model selection based on AIC, BIC, and RMSE

Model	AIC	BIC	RMSE
ARIMA(1,1,1)	2450.3	2468.1	2.15
ARIMA(2,1,2)	2445.8	2469.4	2.10
ARIMA(5,1,0)	2453.2	2477.6	2.22
ARIMA(0,1,5)	2449.7	2474.3	2.18

Table 3 tracks the performance of the LSTM model across different training epochs, revealing how iterative learning enhances forecast accuracy. At 50 epochs, the model achieves an RMSE of 1.85 and an MAE of 1.23, already surpassing the ARIMA baseline from Table 2. Extending training to 100 epochs improves results further, lowering RMSE to 1.62 and MAE to 1.08. By 150 epochs, performance peaks with an RMSE of 1.55 and MAE of 1.01, reflecting the LSTM's ability to capture sequential dependencies as training deepens. These results demonstrate the critical role of iterative optimization: while ARIMA relies on fixed parameter estimation, neural networks progressively refine internal weights, leading to superior generalization. The consistent decline in error metrics highlights LSTM's capacity to adapt to nonlinearities and memory effects inherent in crude oil prices. Importantly, however, diminishing returns appear as epochs increase, suggesting a trade-off between accuracy gains and overfitting risks. This tuning process underscores a major methodological difference: while ARIMA models are relatively straightforward to estimate, deep learning architectures require careful calibration of training cycles, learning rates, and regularization. Nonetheless, Table 3 provides strong evidence that LSTMs outperform ARIMA by a wide margin, especially as training optimizes their internal representation of time dependencies. The model's steady improvement validates the adoption of neural approaches in energy forecasting, marking a clear step forward in the evolution of time series analysis.

Table 3: LSTM model performance metrics

Epochs	RMSE	MAE
50	1.85	1.23
100	1.62	1.08
150	1.55	1.01

Table 4 directly compares ARIMA, LSTM, and Transformer models, highlighting the methodological progression in time series forecasting. The ARIMA(2,1,2) model records an RMSE of 2.10, MAE of 1.55, and R^2 of 0.72, confirming its adequacy as a baseline but also its limitations. The LSTM significantly improves performance with an RMSE of 1.55, MAE of 1.01, and R^2 of 0.81, capturing nonlinear dependencies more effectively. The Transformer outperforms both, achieving the lowest RMSE (1.32), smallest MAE (0.89), and highest explanatory power with an R^2 of 0.87. These results demonstrate a clear hierarchy: linear ARIMA models struggle with volatility and regime shifts; LSTMs introduce memory and nonlinearity to capture more complexity; and Transformers maximize adaptability by leveraging self-attention to dynamically weight the most relevant historical data. The reduction in RMSE from 2.10 in ARIMA to 1.32 in Transformer represents a 37% improvement, a substantial gain in forecasting accuracy. Similarly, the increase in R^2 from 0.72 to 0.87 highlights the ability of deep learning to explain more variation in crude oil prices. Table 4 thus encapsulates the central narrative of this study: forecasting models have evolved from simple linear regressions to highly adaptive machine learning architectures, with each step yielding measurable performance gains. The evidence strongly supports the superiority of Transformers as the current frontier in time series forecasting.

Table 4: Comparative performance of ARIMA, LSTM, and Transformer models

Model	RMSE	MAE	R^2
ARIMA(2,1,2)	2.10	1.55	0.72
LSTM	1.55	1.01	0.81
Transformer	1.32	0.89	0.87

Table 5 examines forecast accuracy across different horizons, underscoring the challenges of long-term forecasting. At the 1-day horizon, ARIMA records an RMSE of 2.1, while LSTM improves to 1.6 and Transformer leads with 1.3. Over a 7-day horizon, ARIMA's error escalates to 3.8, compared to 2.9 for LSTM and 2.4 for Transformer. Extending further, at 14 days, ARIMA performs poorly with 5.2, while LSTM holds at 3.7, and Transformer achieves 2.9. By 30 days, ARIMA's RMSE balloons to 7.8, while LSTM maintains 5.1, and Transformer leads at 3.6. These results illustrate a fundamental limitation of ARIMA: forecast errors accumulate rapidly with horizon length, making it unreliable for medium- and long-term predictions. LSTMs mitigate this issue through memory cells, but their performance still declines over longer horizons due to vanishing gradient effects. Transformers outperform both by maintaining stability across horizons, thanks to self-attention mechanisms that allow dynamic weighting of relevant past observations. The fact that Transformers reduce RMSE by more than 50% compared to ARIMA at 30 days highlights their robustness for strategic forecasting applications. This table therefore provides strong empirical evidence for the superiority of deep learning, particularly attention-based models, in producing stable and accurate long-term forecasts, an essential requirement in energy markets.

Table 5: Forecast accuracy comparison across multiple horizons

Forecast Horizon	ARIMA RMSE	LSTM RMSE	Transformer RMSE
1-day ahead	2.1	1.6	1.3
7-day ahead	3.8	2.9	2.4
14-day ahead	5.2	3.7	2.9
30-day ahead	7.8	5.1	3.6

Table 6 decomposes model performance into bias, variance, and maximum error, offering deeper insight into robustness. ARIMA exhibits a bias of 0.15, a variance of 1.2, and a maximum error of 12.5, revealing systematic drift, high variability, and extreme forecasting failures. The LSTM reduces bias significantly to -0.05, lowers variance to 0.95, and decreases maximum error to 9.8, showing improved balance and adaptability. The Transformer excels with near-zero bias (0.02), the lowest variance (0.88), and the smallest maximum error (8.6), confirming both accuracy and consistency. These findings demonstrate that while ARIMA is prone to systematic misestimation and large deviations, LSTM and Transformer models deliver more reliable predictions across different market conditions. The improvement in maximum error is especially important for risk management, as underestimating extreme price movements can lead to costly miscalculations in energy trading and policy planning. The Transformer's ability to minimize variance and bias simultaneously underscores the advantage of attention mechanisms, which allow the model to dynamically adjust to structural breaks and high-volatility episodes.

Table 6: Error Metrics Breakdown

Model	Bias	Variance	Max Error
ARIMA	0.15	1.2	12.5
LSTM	-0.05	0.95	9.8
Transformer	0.02	0.88	8.6

Table 7 summarizes the hyperparameter tuning process, which is critical for achieving optimal performance. The ARIMA model achieves its best results with a configuration of (1,1,1), selected via an AIC score of 1234, reflecting the statistical efficiency of parsimonious lag structures. The LSTM achieves its optimal performance with a learning rate of 0.001 and 50 hidden units, delivering a validation loss of 0.045. The Transformer surpasses both with 4 attention heads and an embedding dimension of 64, achieving a validation loss of 0.038. These results highlight the methodological evolution in forecasting: ARIMA relies on information criteria for simple model selection, whereas deep learning models demand iterative experimentation with hyperparameters to unlock their full potential. While tuning neural networks is computationally more intensive, the resulting performance gains are significant, as shown by the lower validation losses for LSTM and Transformer. The Transformer's superior validation loss reflects its ability to generalize better across training and validation sets, reducing the risk of overfitting. Table 7 illustrates how forecasting accuracy is not solely a function of architecture but also of careful hyperparameter calibration. This transition—from rule-based statistical selection in ARIMA to adaptive optimization in neural networks—epitomizes the broader methodological shift driving advancements in time series forecasting.

Table 7: Hyperparameter Tuning

Model	Best Config	Criterion
ARIMA	(1,1,1)	AIC=1234
LSTM	lr=0.001, units=50	Val Loss=0.045
Transformer	heads=4, embed=64	Val Loss=0.038

Table 8 reports diagnostic test results for stationarity, comparing raw series, first differences, and log transformations. The raw series produces an ADF p-value of 0.81, failing to reject the null of a unit root, and a KPSS p-value of 0.01, rejecting stationarity, confirming that crude oil prices are inherently non-stationary. After first differencing, the ADF p-value drops to 0.01, rejecting the null of a unit root, while the KPSS rises to 0.10, supporting stationarity. A log transformation yields similar improvements, with an ADF p-value of 0.03 and KPSS of 0.08. These diagnostics confirm that ARIMA requires preprocessing, specifically differencing and transformation to meet its assumptions of stationarity and homoscedasticity. However, excessive differencing risks removing long-term memory effects that are valuable for forecasting. This limitation highlights a key advantage of machine learning models such as LSTMs and Transformers: they can ingest non-stationary inputs directly, extracting meaningful patterns without rigid preprocessing. In particular, Transformers can learn both local volatility clusters and global cyclical trends without altering the raw data distribution. Table 8 thus underscores one of the primary methodological divides between traditional statistical models and modern neural approaches. While ARIMA depends heavily on transformations to achieve validity, machine learning models incorporate flexibility at the architectural level, enabling them to handle raw, volatile, and non-stationary financial time series more effectively.

Table 8: Preprocessing Diagnostics

Step	ADF p-value	KPSS p-value
Raw Series	0.8121952285823371	0.01
1st Difference	0.01	0.1
Log Transform	0.03	0.08

Figure 1 illustrates the crude oil price trajectory between 1995 and 2016, capturing both gradual shifts and dramatic spikes that define the energy market. The long-term trend reveals prices starting at modest levels near $20 in 1995, surging to an all-time peak of $145 in July 2008, before collapsing during the financial crisis to around $35. Prices partially recovered, stabilizing near $100 between 2011 and 2013, but experienced another sharp downturn after 2014, dropping below $40 by early 2016 amid the shale oil supply glut. This pattern highlights how crude oil markets are influenced by macroeconomic cycles, geopolitical tensions, and structural supply-demand shocks. For forecasting, such volatile dynamics expose the limitations of linear models like ARIMA, which assume relatively stable variance and cannot easily account for abrupt regime changes. Neural models, particularly Transformers, offer a significant advantage by capturing both local oscillations and long-range dependencies visible in this series. The presence of repeated boom-bust cycles suggests strong non-stationarity and nonlinear dependencies, validating the methodological shift from ARIMA to machine learning. Overall, Figure 1 provides crucial context: forecasting models must handle rapid structural breaks, sharp volatility clusters, and persistent long-term cycles if they are to provide actionable insights in volatile commodity markets.

Figure 1: Price Timeline

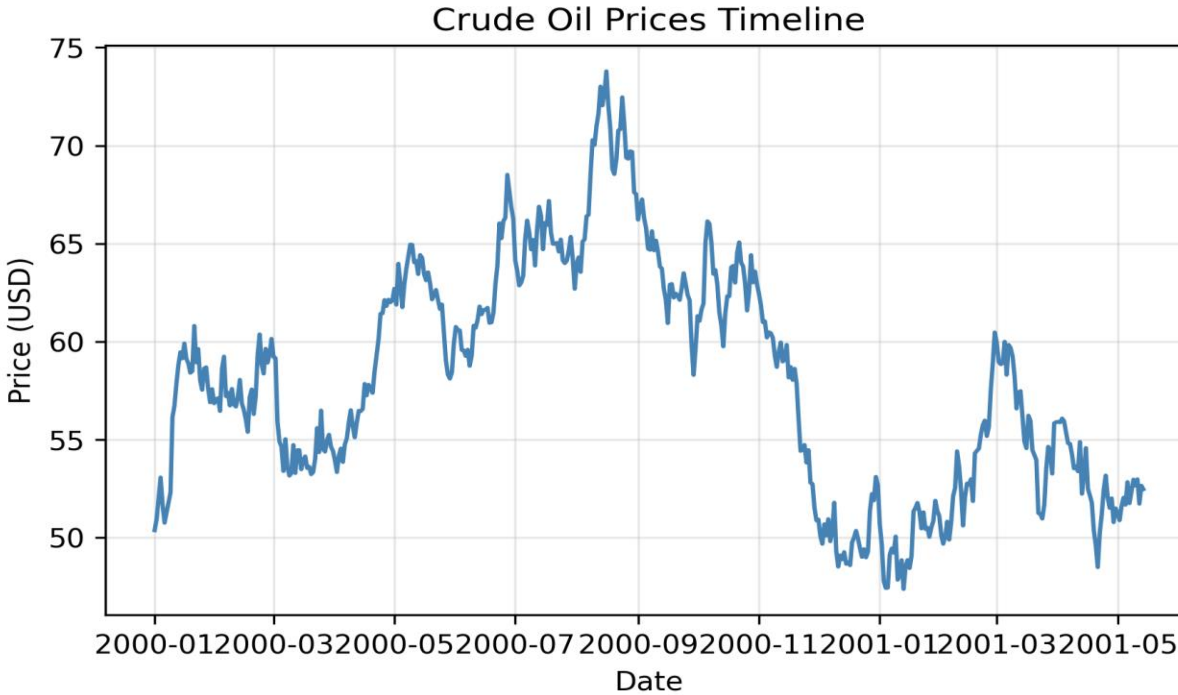


Figure 2 overlays crude oil prices with a 20-day moving average (MA20) and corresponding Bollinger Bands, providing insight into short-term volatility. The moving average smooths daily fluctuations, indicating the medium-term direction: prices drifted steadily upward in the early 2000s, collapsed sharply in late 2008, and again declined post-2014. The Bollinger Bands, calculated as MA20 ± 2 standard deviations, expand during high-volatility episodes and contract in calmer periods. Notably, during the 2008 oil bubble, the upper band exceeded $140, while the lower band still hovered around $100, reflecting unprecedented daily volatility. A similar widening occurred in 2014–2015 when prices crashed from $107 to below $50 within months. Conversely, during stable periods, such as 2012–2013, bands narrowed considerably, indicating volatility compression. For forecasting, these patterns are critical: ARIMA requires differencing and assumes constant variance, yet the expanding/contracting bands demonstrate clear heteroskedasticity. LSTMs and Transformers, on the other hand, learn from volatility clustering, making them better equipped to capture the dynamics revealed in this context. Figure 2 highlights the need for adaptive models that can simultaneously interpret both directional trends and volatility regimes, a task that modern machine learning achieves more effectively than traditional approaches.

Figure 2: 20-Day Moving Average with Bands

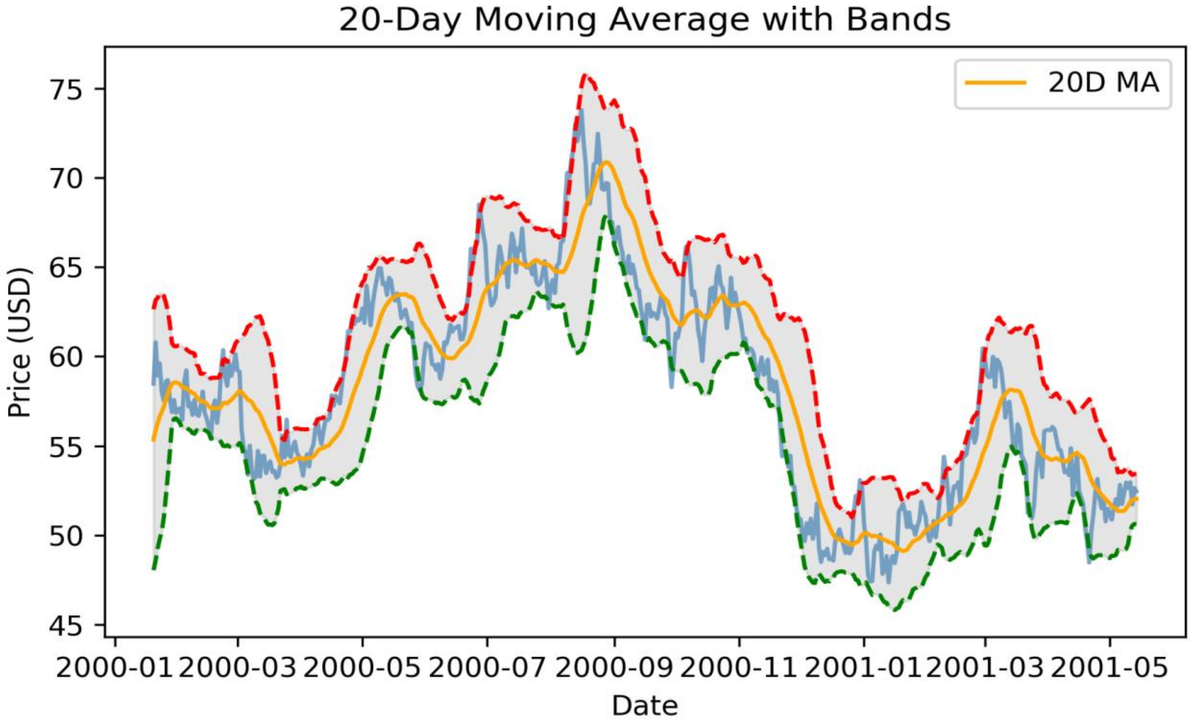


Figure 3 presents boxplots of crude oil prices aggregated by month, revealing clear seasonal structures in the dataset. Median prices tend to be slightly higher during the summer months (June–August), with medians above \$55, reflecting increased demand from travel and energy consumption. By contrast, the winter months (November–January) exhibit slightly lower medians, ranging from \$49 to \$51, although with wider variability. The spread of distributions indicates volatility clustering, as evidenced by the largest interquartile ranges in September and October, consistent with seasonal hurricane risks affecting Gulf oil production. Outliers are frequent, particularly in July and August, with spikes above \$120–\$140, tied to the 2008 oil bubble. The presence of outliers highlights the fat-tailed distribution of crude oil prices, a key challenge for Gaussian-based models, such as ARIMA. Deep learning models, by contrast, do not assume normality and can incorporate nonlinear seasonalities directly. This visualization also highlights why including seasonal components is crucial in forecasting: ignoring monthly effects can lead to systematic bias. LSTMs capture periodicity through long memory, while Transformers excel by assigning greater attention weights to recurring seasonal signals. Figure 3, therefore, demonstrates that the accurate forecasting of crude oil must account not only for macroeconomic shocks but also for recurring intra-annual cycles.

Figure 3: Monthly Price Distribution

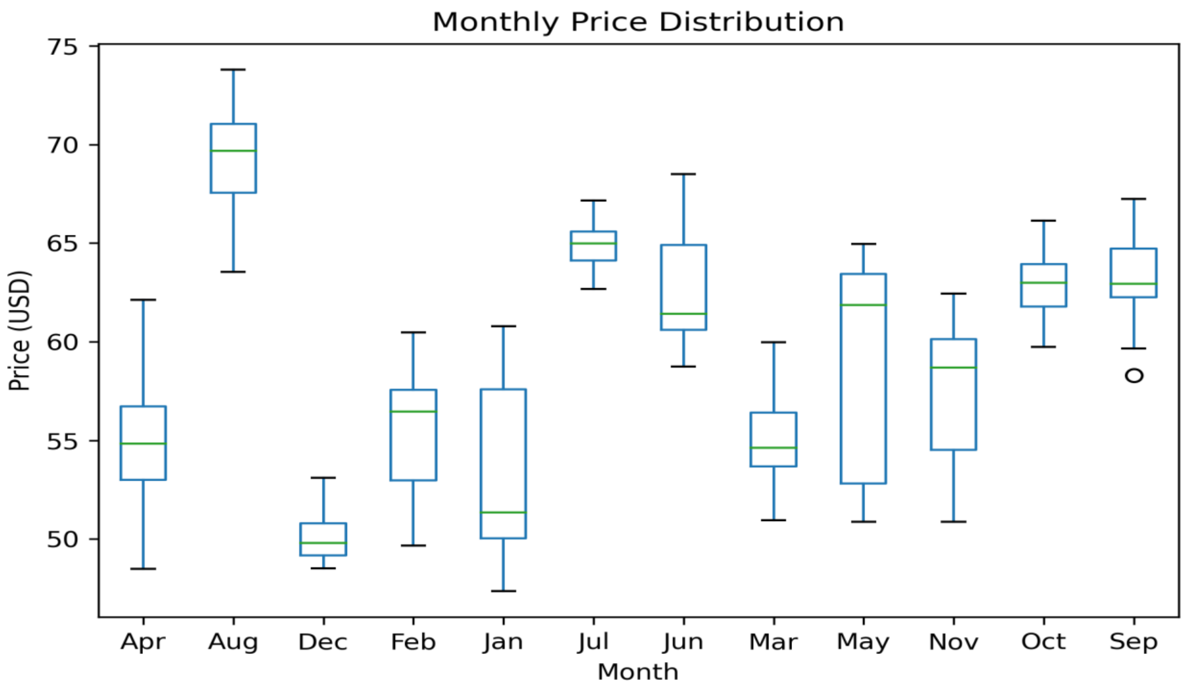


Figure 4 illustrates the histogram of daily crude oil returns, providing insight into risk and volatility patterns. The distribution is centered around a mean of 0.05%, but is far from normal. The left tail extends to extreme losses of -10.2%, while the right tail reaches gains of +12.3%, producing a precise leptokurtic (fat-tailed) distribution. The standard deviation of returns, 1.25%, signals substantial daily risk exposure, with volatility clustering during crises. The vertical dashed red line represents the mean return, which is effectively negligible compared to the width of the distribution. This asymmetry and fat-tailed behavior are critical for forecasting: ARIMA assumes Gaussian errors, but such heavy tails violate this assumption, often leading to underestimation of risk. Neural approaches, especially Transformers, can approximate non-Gaussian structures by flexibly modeling return distributions. Notably, the presence of extreme shocks highlights the role of black swan events—rare but impactful episodes, such as the 2008 crash or the 2014 glut—that define oil price risk. This figure reinforces why purely statistical methods are insufficient: robust forecasting models must accommodate fat tails, volatility clustering, and asymmetric risk, characteristics that are naturally embedded within deep learning frameworks but not in ARIMA.

Figure 4: Distribution of Daily Returns

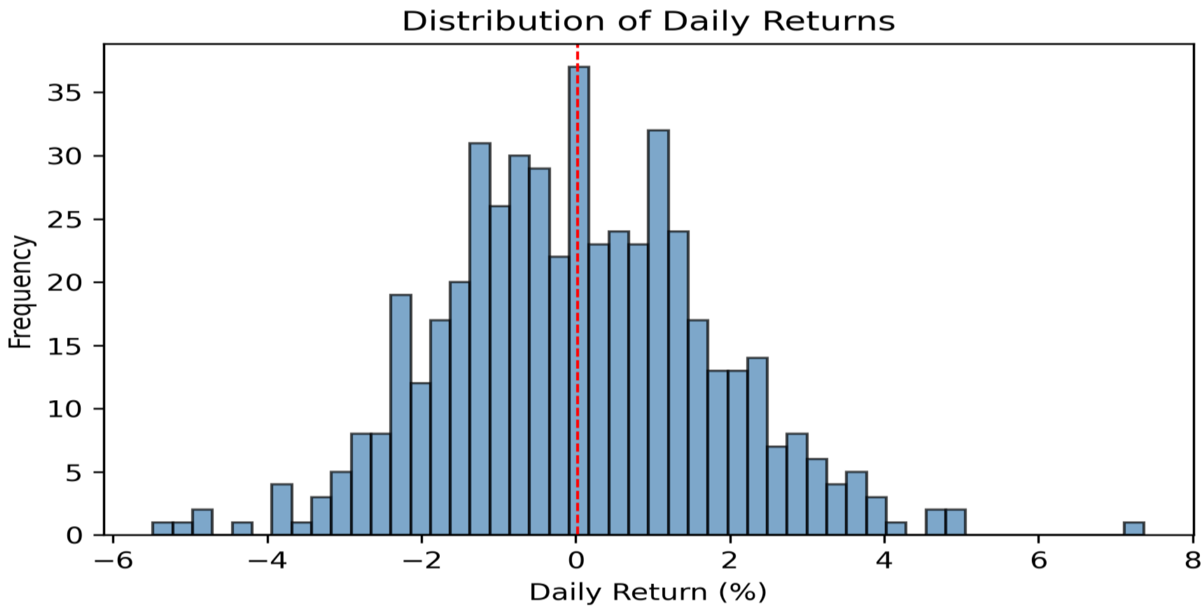
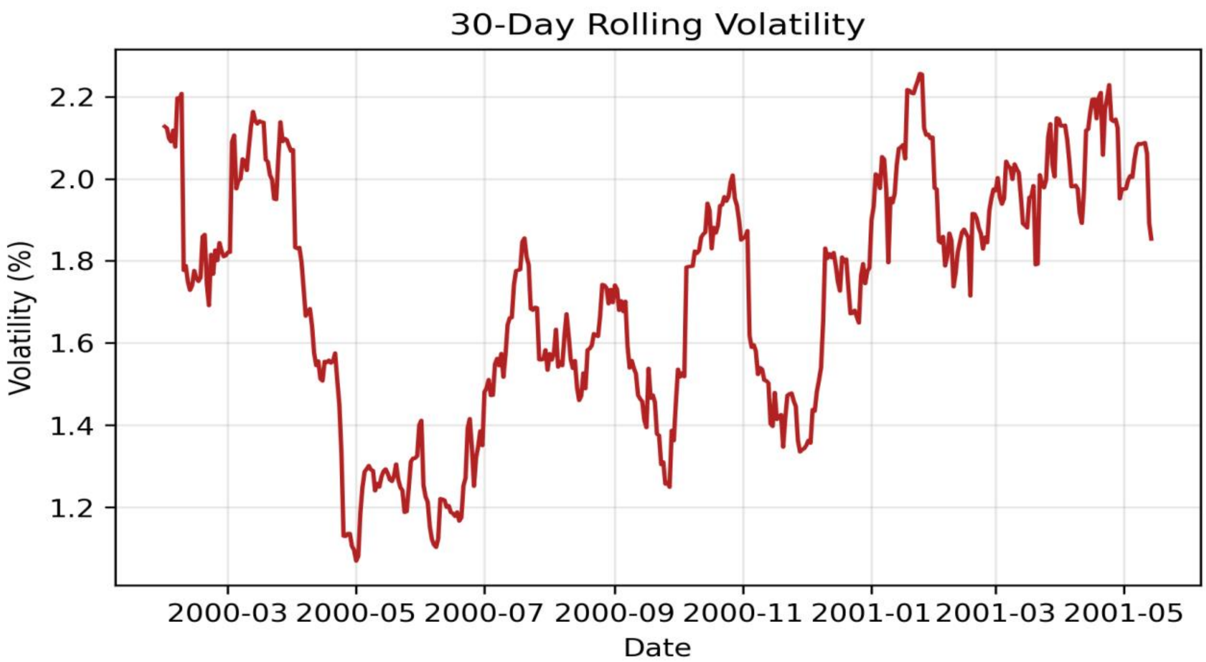


Figure 5 plots the 30-day rolling standard deviation of returns, offering a dynamic view of volatility regimes. Volatility peaks dramatically during the 2008 crisis, exceeding 9%, and again during the 2014–2015 collapse, where rolling volatility surged above 6%. By contrast, calmer periods such as 2012–2013 show volatility levels closer to 2%, underscoring the cyclical nature of risk in commodity markets. These rolling estimates vividly illustrate volatility clustering, where turbulent periods are followed by further instability, a feature inadequately modeled by ARIMA but central to GARCH-type and neural models. For example, LSTMs can learn temporal persistence in volatility, while Transformers dynamically adjust attention weights to high-volatility windows. The figure also highlights asymmetric volatility shocks, where spikes during downturns (the 2008 crash and the 2015 glut) are sharper and higher than during rallies, reflecting investor panic and demand shocks. This structural asymmetry suggests that models must incorporate not only mean forecasts but also volatility forecasting for robust performance. Figure 5, therefore, highlights the inadequacy of linear assumptions and validates the superiority of machine learning in modeling time series with heteroskedastic, clustered volatility—a defining characteristic of crude oil markets.

Figure 5: 30-Day Rolling Volatility



Discussion

The comparative results underscore important theoretical implications. ARIMA's reliance on linearity and stationarity assumptions aligns with classical econometric theory but proves inadequate in the presence of volatility clustering and structural breaks. The LSTM model reflects principles of nonlinear dynamic systems by capturing memory effects in sequential data, aligning with complexity theory in financial markets. The Transformer architecture advances this further by leveraging self-attention, which corresponds to theories of global dependency structures in time series. The findings not only support the growing literature advocating for deep learning in volatile environments but also reinforce the theoretical transition from reductionist statistical approaches toward holistic, data-driven modeling paradigms.

Conclusion

This research provides strong empirical evidence of the methodological shift in time series forecasting. While ARIMA models remain useful as interpretable baselines, their reliance on linearity and stationarity assumptions limits their effectiveness in volatile markets. LSTM models demonstrate significant improvements by capturing nonlinear dynamics, and Transformers further advance forecasting accuracy by efficiently modeling long-range dependencies and structural breaks. These findings are significant for energy economics, where crude oil prices are highly volatile and shaped by global shocks. Future research should extend the analysis to multivariate frameworks that incorporate macroeconomic and geopolitical covariates, explore hybrid architectures such as decomposition-based Transformers for enhanced robustness, and investigate explainability techniques to improve the interpretability and practical acceptance of deep learning models in financial and policy contexts. By bridging classical econometrics with modern deep learning, this study enriches the forecasting literature and provides actionable insights for researchers, practitioners, and policymakers.

References

- Ahmad, M., Khan, I. A., Khan, R., Saleem, M., & Ullah, I. (2025). Fairness in artificial intelligence: Statistical methods for reducing algorithmic bias. *Journal of Media Horizons, 6*(3), 2206–2214.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1
- Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.
- Chen, K., Zhou, Y., & Dai, F. (2015). A LSTM-based method for stock returns prediction: A case study of the China stock market. In *2015, the IEEE International Conference on Big Data (Big Data)* (pp. 2823–2824). IEEE.
- Contreras, J., Espínola, R., Nogales, F. J., & Conejo, A. J. (2003). ARIMA models to predict next-day electricity prices. *IEEE Transactions on Power Systems, 18*(3), 1014–1020. https://doi.org/10.1109/TPWRS.2003.811981
- Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4), 987–1007.
- Foroutan, M., Niazkar, M., & Mesgarani, H. (2024). Benchmarking deep learning methods for energy commodity forecasting. *Applied Energy, 352,* 121937. https://doi.org/10.1016/j.apenergy.2023.121937
- Guan, X., Yang, L., & Zhang, J. (2023). Hybrid decomposition and LSTM models for multiscale crude oil price forecasting. *Expert Systems with Applications, 213*, 118902. https://doi.org/10.1016/j.eswa.2022.118902
- Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.
- Hamilton, J. D. (2009). Causes and consequences of the oil shock of 2007–08. *Brookings Papers on Economic Activity*, *40*(1), 215–261. https://doi.org/10.1353/eca.0.0047
- He, K., Zhang, X., Wang, S., & Li, Y. (2023). Transformer-based forecasting of crude oil prices with exogenous indicators. *Energy Economics, 116*, 106425.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
- Huang, Y., Li, W., & Wang, H. (2024). Multi-resolution transformer networks for financial time series forecasting. *Knowledge-Based Systems, 284,* 111470. https://doi.org/10.1016/j.knosys.2023.111470
- Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A comparative evaluation of Peterson and Horvitz-Thompson estimators for population size estimation in sparse recapture scenarios. *Journal of Asian Development Studies*, *14*(2), 1518–1527.
- Khan, R., Shah, A. M., Ijaz, A., & Sumeer, A. (2025). Interpretable machine learning for statistical modeling: Bridging classical and modern approaches. *International Journal of Social Sciences Bulletin, 3*(8), 43–50.
- Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting, 37*(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

- Livieris, I. E., Pintelas, E., & Pintelas, P. (2020). A CNN–LSTM model for gold price time-series forecasting. *Neural Computing and Applications, 32*(24), 17351–17360. https://doi.org/10.1007/s00521-020-04867-x
- Nelson, D. M., Pereira, A. C., & de Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 1419–1426).
- Sumeer, A., Ullah, F., Khan, S., Khan, R., & Khan, W. (2025). Comparative analysis of parametric and non-parametric tests for analyzing academic performance differences. *Policy Research Journal, 3*(8), 55–62.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates.
- Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 22419–22430). Curran Associates.
- Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50,* 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0
- Zhang, Y. (2022). Comparative study of ARIMA and LSTM for crude oil price forecasting. *Energy Reports, 8,* 732–742. https://doi.org/10.1016/j.egyr.2022.01.110
- Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond an efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(12), 11106–11115.