

Generative AI For Financial Time Series Forecasting a Case Study on Crudeoil

Vinod¹, MeghaRaniRaigonda²

¹Vinod, Department of Computer Science and Engineering (MCA), Visvesvaraya Technological University, CPGS, Kalaburgi, Karnataka, India

²MeghaRaniRaigonda, Department of Computer Science and Engineering (MCA), Visvesvaraya Technological University, CPGS, Kalaburgi, Karnataka, India

Abstract: Crude oil price fluctuations significantly affect the global economy, but traditional forecasting models such as (ARIMA) Autoregressive Integrated Moving Average and (GARCH) Generalized Autoregressive Conditional Heteroskedasticity struggle with the nonlinear and unstable dynamics of these markets. Advances in deep learning, including LSTM Long Short-Term Memory and Transformer architectures, have improved predictions but remain limited by data scarcity and unexpected disruptions. To address these issues, this work develops a hybrid framework that integrates LSTM for short-term sequence learning, Transformers for capturing long-range patterns using attention, and TimeGAN for generating synthetic time-series data to strengthen model training. Using historical datasets from the U.S. Energy Information Administration (EIA), the system is assessed with error measures such as RMSE Root Mean Squared Error and MAPE stands for Mean Absolute Percentage Error. Results indicate that this combined approach achieves more reliable and accurate forecasts than standalone models, highlighting the value of combining predictive and generative AI for robust commodity price forecasting.

Keywords: This work emphasizes hybrid AI frameworks for forecasting crude oil prices by combining sequence-learning models with generative approaches. Important keywords include generative AI, TimeGAN, commodity price volatility, long-range dependency modeling, and robust predictive performance

Date of Submission: 26-05-2026

Date of acceptance: 05-06-2026

I. INTRODUCTION

This project is implemented in Python using a modular and well-structured architecture, ensuring clarity, flexibility, and long-term scalability. The framework integrates traditional statistical forecasting techniques, advanced deep learning architectures, and generative AI models to build a comprehensive solution for crude oil price prediction. By combining LSTM networks for short-term temporal learning, Transformer models for capturing long-range dependencies, and TimeGAN for generating high-quality synthetic time-series data, the system effectively handles the nonlinear, volatile, and data-scarce nature of commodity markets.

Each stage of the workflow—data collection, preprocessing, feature engineering, model training, performance evaluation, and visualization—is implemented as an independent Python module. This separation enhances code reusability, simplifies debugging, and allows updates or model improvements without affecting the entire pipeline. The use of historical datasets from reliable sources enables realistic forecasting, while metrics such as RMSE and MAPE provide an objective evaluation of predictive accuracy. Overall, this modular hybrid framework serves as a robust and scalable approach for financial time series forecasting, especially in unstable market environments like crude oil.

II. LITERATURE SURVEY

The survey studies show that traditional models like ARIMA struggle with the nonlinear and volatile nature of crude oil prices. Deep learning methods such as LSTM, CNN-LSTM, and Transformers provide better accuracy by capturing complex temporal patterns. TimeGAN further enhances performance by generating synthetic data to improve model robustness.

[1] Vapnik (1995) – SVM / SVR Vapnik introduced Support Vector Machines (SVM) and Support Vector Regression (SVR), which model nonlinear relationships using kernel functions such as RBF and Polynomial. SVR is

effective in small and medium datasets because it finds a globally optimal solution and avoidsoverfittingthroughmarginmaximization. It can capture nonlinear price patterns but relies heavily on careful tuning of kernel type and parameters. However, SVR does not naturally model long sequential dependencies unless windowed inputs are engineered manually. For crude oil forecasting, SVR performs well for short-term prediction but struggles with long temporal dynamics

[2].Breiman (2001) –Random ForestsBreiman proposed Random Forests, an ensemble of decision trees trained using bootstrap sampling and random feature selection. This method reduces overfitting, handles noisy inputs, and manages nonlinear feature interactions effectively. Random Forests can process complex financial datasets without strong assumptions about their distribution. However, they do not inherently capture sequential or temporal dependencies unless lag features are manually created. Thus, their utility in crude oil forecasting is limited to short-term or feature-based prediction tasks rather than long-horizon forecasting.

[3]. Hochreiter & Schmidhuber LSTM LSTM networks were introduced to solve the vanishing and exploding gradient problems recurrent Neural networks by incorporating gated memory cells. They can Store Long Term dependencies and learn temporal patterns in highly volatile time series data such as crude oil prices. LSTM have strong performance in sequence

[4].Huetal.(2021)–CEEMDAN-LSTM Hybrid Hu et al. developed a hybrid model combining CEEMDAN decomposition with LSTM networks. CEEMDAN breaks the crude oil price series into multiple intrinsic mode functions (IMFs), effectively separating trends, cycles, and noise. Each component is learned using an LSTM, and predictions are recombined to obtain the final output. This approach enhances forecasting accuracy by reducing data complexity and isolating multi-scale patterns. The major drawback is increased computational cost and longer training times due to multiple models. It is particularly suited for nonstationary and volatile crude oil price data

[5].Yoon et al. (2019) – TimeGAN TimeGAN is a generative adversarial model designed specifically for time series, integrating embedding networks, supervised learning, and adversarial training. It generates realistic synthetic sequences while preserving temporal relations and feature dependencies. TimeGAN is valuable for crude oil forecasting because it can augment datasets, simulate rare events, and support model training with more diverse patterns. However, GANs are often unstable to train and may suffer from mode collapse if not tuned properly. Despite these challenges, TimeGAN is an influential model in time series augmentation.

[6].Heetal.(2023)–Temporal FusionTransformer(TFT)Heetal.appliedthe Temporal Fusion Transformer (TFT) to crude oil forecasting, showing significant improvements over LSTM in medium- and long-term horizons. TFT combines LSTM layers with attention mechanisms, variable selection networks, and gating modules, enabling it to capture long-ranged dependencies and heterogeneous feature interactions. Its interpretability allows users to identify important variables and time windows influencing forecasts. Although computationally intensive, TFT offers state-of-the-art accuracy in multi-horizon forecasting tasks. It is especially effective for datasets with many macroeconomic and temporal features.

[7].Hamilton(2009)–Econometric Analysis of Oil Shocks Hamilton’s work demonstrated that crude oil prices are heavily influenced by external shocks such as geopolitical events, supply disruptions, and macroeconomic uncertainty. Using econometric models like VAR and structural decomposition, he showed that many sudden price changes are not predictable from historical data alone. This challenges purely data-driven models and highlights the need to include exogenous variables. While valuable for understanding market mechanisms, these models often fail during periods of extreme volatility. Hamilton’s findings emphasize the importance of shock modeling in oil forecasting research.

[8].Zhangetal. (2018) – ARIMA-LSTM Hybrid Zhanget al. developed a hybrid forecasting method combining ARIMA for linear components and LSTM for nonlinear residuals. ARIMA captures autoregressive and seasonal behavior, while LSTM models complex nonlinear dynamics left unexplained by the statistical model. This two-stage system effectively reduces forecasting error, especially during volatile market periods. The method is simple and interpretable but requires correct ARIMA specification and careful residual modeling. It is well-suited for crude oil data where linear and nonlinear patterns coexist.

[9].Qinetal.(2017)–DA-RNN(Dual-Stage AttentionRNN)DA-RNN introduces two attention mechanisms: input attention to select important features and temporal attention to assign weights to past time steps. This structure improves interpretability and focuses the model on relevant signals in multivariate time series. It outperforms traditional RNNs in forecasting tasks by reducing noise influence and dynamically adjusting attention during training. However, it still retains the limitations of RNNs, such as slower training

andsomedifficultyinhandlingverylongsequences.DA- RNNiseffectiveforcrudeoilforecastingwhen multiple economic indicators are included.

[10].Chenetal.(2020)– CNN-LSTMHybridChenetal.proposedahybridCNN- LSTM model that first uses convolutional layers to extract local temporal patterns and then uses LSTMs to capture long-term dependencies. This approach is highly effective for high-frequency or noisy financial data because CNNs detect short-term features like spikes or microstructure patterns. The combination enhances forecasting accuracy compared to standalone LSTMs. However, the model requires careful design of convolutional window sizes and may overfit with limited data. It is especially useful for intraday crude oil forecasting.

III. PROPOSED METHODOLOGY

This study aims to close this gap by combining LSTM, Transformer, and TimeGAN models, showing that their integration can deliver more precise and robust results. Crude oil is a vital global resource, and price fluctuations have direct consequences on financial markets, inflation, and economic stability. Market volatility is influenced by factors such as geopolitical conflicts, OPEC decisions, supply–demand imbalances, currency fluctuations, natural disasters, and speculation. Even minor price shifts can significantly affect industries like energy, manufacturing, and transportation. Therefore, accurate forecasting is critical for governments, businesses, and investors to manage risks. Traditional models such as ARIMA, GARCH, and regression approaches were widely applied, but they fail to address nonlinear dynamics, long-term dependencies, and sudden shocks. Neural models like RNNs and LSTMs improved sequential learning but require large datasets and struggle with very long input sequences. More recently, Transformer architectures and TimeGAN have shown promise by handling long-term dependencies and generating synthetic sequences, yet their use in crude oil forecasting remains relatively unexplored.

- **Data Collection & Preprocessing:** Gather crude oil prices from reliable financial platforms, clean and normalize the dataset, and derive technical indicators.
- **Baseline Models:** Develop ARIMA and simple LSTM as benchmarks.
- **Transformer Forecasting :** Design a Transformer model to capture nonlinear long-range patterns.
- **Synthetic Data Generation:** Applying TimeGAN to create realistic synthetic datasets for robustness and augmentation.
- **Performance Evaluation:** Compare models using MAE, RMSE, MAPE, and R^2 .
- **Visualization & Interpretation:** Present results through charts, dashboards, and tables for clear analysis.
- **Comparative Analysis:** Identify the most effective model under volatile market scenarios.
- **Practical Contribution:** Demonstrate how generative AI can enhance reliability and accuracy in crude oil price prediction.

The methodology follows a modular pipeline beginning with data collection and preprocessing, where historical crude oil prices are cleaned, normalized, and transformed for model readiness. LSTM and Transformer models are then trained to learn short-term and long-term patterns, while TimeGAN generates synthetic time-series data to enhance training. Finally, model predictions are evaluated using metrics like RMSE and MAPE to ensure accurate and reliable forecasting.

Step 1-Literature Review & Problem Definition: Analyze existing models such as ARIMA, GARCH, and LSTM, identify their limitations, and justify the use of generative AI (Transformers, TimeGAN).

Step 2-Data Collection: Obtain crude oil price data (WTI and Brent) from Yahoo Finance, FRED, and Kaggle, including daily closing prices and trading volumes.

Step 3-Preprocessing: Handle missing values and outliers, normalize using min–max scaling, and create features like moving averages, volatility, and momentum indicators.

Step 4-Model Development: Implement ARIMA (baseline), LSTM (sequential learning), Transformer (long-range dependency), and TimeGAN (synthetic data generation).

Step 5-Training: Train models with optimized hyperparameters (batch size, learning rate, dropout) and apply techniques like regularization and early stopping.

Step 6-Evaluation: Assess performance using MAE, RMSE, MAPE, and R^2 on test datasets.

Step 7-Visualization & Analysis: Compare actual vs. predicted prices with visual dashboards, error plots, and volatility analysis.

Step 8-Results & Conclusion: Determine the most effective model, highlight how generative AI enhances prediction robustness, and discuss its practical use in finance and energy domains.

System takes a fundamental stage of software development since it defines the architecture and structural framework of the proposed solution. In this project, the design illustrates how data moves from its initial collection and preprocessing stages, through model development, and finally to the generation of forecasts. It clearly specifies the components of the system, how they interact, and how information flows between them. The architecture follows a modular approach, meaning each part of the system has a specific role and can be updated or replaced independently without affecting the overall workflow. The key architectural layers consist

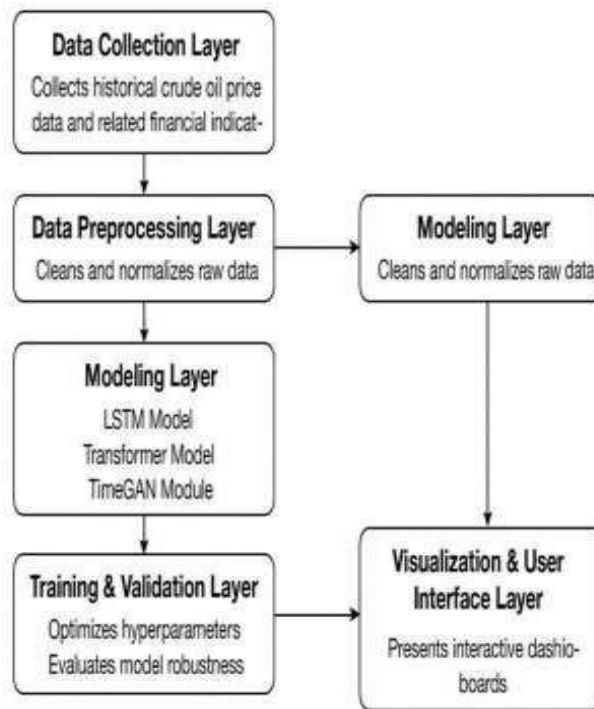


Figure 1: System Architecture

The architecture is designed to be modular and scalable, making it adaptable to evolving datasets and methodologies. Beyond generating forecasts, the framework supports the evaluation and comparison of multiple models, enabling more informed, data-driven insights. The Generative AI for Crude Oil Price Forecasting system functions as an intelligent decision-support platform. It collects and refines historical price data, processes it through advanced AI methods, and generates forecasts supported by interactive visualizations. The entire workflow operates as a data-driven pipeline that begins with raw input and ends with actionable predictive insights.

Data Collection Layer

- Collects historical crude oil price data from reliable public sources such as financial APIs, government databases, and market repositories.
- Gathers additional macroeconomic indicators (e.g., exchange rates, stock indices, economic signals) to enrich model inputs.
- Ensures continuous and automated data retrieval for up-to-date forecasting.
- Performs initial validation to remove missing or corrupted records.

Data Preprocessing Layer

- Cleans raw data by handling missing values, correcting anomalies, and removing outliers.
- Normalizes numeric variables using Min–Max or Z-Score scaling to ensure uniform model input.
- Constructs supervised learning sequences using sliding window approaches.
- Generates technical indicators (moving averages, volatility, etc.) to improve model learning quality.
- Outputs a well-structured dataset ready for deep learning and generative modeling.

Modeling Layer

- **LSTM Model:** Captures short- and long-term temporal dependencies in crude oil price movements.
- **Transformer Model:** Leverages self-attention mechanisms to detect long-range relationships and market trends.
- **TimeGAN Module:** Synthesizes realistic time-series data to augment training datasets and improve

generalization

Training&ValidationLayer

- Trains models using optimized hyperparameters (learning rate, batch size, epochs, layers).
- Uses real and synthetic datasets to evaluate robustness and check for overfitting.
- Applies validation metrics (MAE, RMSE, MAPE, R²) for accurate performance assessment.
- Performs early stopping and checkpointing for stable model convergence.
- Selects the best-performing model for deployment.

Visualization&UserInterfaceLayer

- Presents forecasting results using dashboards, graphs, and interactive charts.
- Allows users to compare actual versus predicted crude oil prices in real time.
- Provides insights into model performance, trends, volatility patterns, and error behavior.
- Acts as the front-end interface for decision makers, analysts, and researchers.

End-to-EndWorkflowSummary

- Data Collection → Preprocessing → Modeling → Training & Validation → Visualization
- Ensures a complete, modular, and scalable pipeline for crude oil price forecasting using Generative AI and deep learning

IV. Model Configurations

Baseline LSTM Model

An LSTM network is implemented using

TensorFlow/Keras:

```
Model=Sequential()
```

```
model.add(LSTM(50,return_sequences=True, input_shape=(60, 1)))
```

```
model.add(LSTM(50)) model.add(Dense(1)) model.compile(optimizer='adam', loss='mean_squared_error')
```

- Input shape: (time_steps, 1) where time_steps = 60.
- Hidden units: 2 stacked LSTM layers with 50 units each.
- Output: Single neuron (Dense(1)) predicting the next-day price.
- Loss function: Mean Squared Error (MSE).
- Optimizer: Adam.

Transformer-Based Model :

A Transformer-based forecasting model is configured to capture long-range temporal dependencies that are difficult for standard LSTMs:

- Encoder–decoder or encoder-only stack with multi-head self-attention layers.
- Positional encodings added to represent time ordering.
- Feed-forward layers and dropout used for regularization.
- Training objective: MSE on normalized prices, with Adam optimizer and similar batch size/horizon settings as the LSTM model

TimeGAN for Synthetic Data Generation

- Components: Embedding network, generator, discriminator, and supervised component to preserve temporal dynamics.

Training data: Preprocessed and normalized crude oil time series.

- Output: Synthetic sequences with similar temporal structure and statistical characteristics as the real data.

Two experimental settings are considered:

1. Without augmentation – model trained only on historical data.
2. With augmentation – models trained on a combination of historical and TimeGAN-generated sequences to assess robustness and performance gains.

Training Procedure

1. Initialization of model architecture (LSTM, Transformer, or TimeGAN).
2. Compilation using a suitable loss (MSE) and optimizer (Adam).
3. Mini-batch training over multiple epochs (e.g., 10–50) with shuffled training sequences.
4. Regularization techniques such as dropout and early stopping to prevent overfitting.
5. Checkpointing of the best-performing model weights based on validation loss.

Evaluation Metrics

To compare the forecasting performance of different models, the following metrics are computed on the test set:

- MeanAbsoluteError(MAE)
- RootMeanSquaredError(RMSE)
- MeanAbsolutePercentageError (MAPE)
- CoefficientofDetermination(R²)

V. Hardware and Software Environment

Experiments are conducted in a Python environment using the following tools and configurations:

- Programming language: Python 3.8+
- Core libraries:
 - TensorFlow 2.x/Keras (LSTM, Transformer, TimeGAN)
 - PyTorch (alternative implementation for generative modeling)
 - Scikit-learn (preprocessing, metrics)
 - NumPy & Pandas (data handling and numerical computation)
 - Matplotlib & Seaborn (visualization)
- Development platforms: Jupyter Notebook / Google Colab and Anaconda-managed environments; Git/GitHub for version control.
- Operating systems: Windows 10/11 or Linux (Ubuntu recommended for deployment).

Hardware:

- Minimum: Intel i3 CPU, 8 GB RAM, 256 GB storage.
- Recommended: Intel i5/i7 or Ryzen 5/7, 16 GB RAM, 512 GB SSD, and a CUDA-enabled NVIDIA GPU for accelerated training.

VI. Figures



Figure 2: Crude Oil Data Overview

The figure presents key insights from the crude oil dataset. The price trend shows noticeable fluctuations over time, indicating market volatility. The histogram illustrates that most prices are concentrated within a mid-range, while the volatility plot highlights sudden spikes in market movement. The correlation heatmap shows strong relationships among the variables, supporting their importance for predictive modeling.

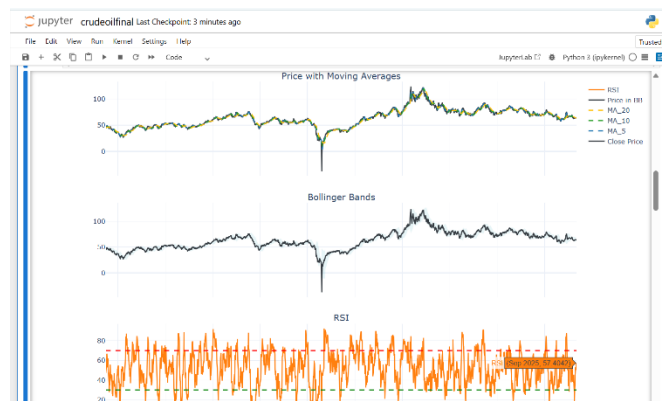


Figure 3: Technical Indicator Analysis

The figure illustrates important technical indicators applied to crude oil prices. The moving averages (MA10, MA20, MA50) help identify overall market trends by smoothing price fluctuations. The Bollinger Bands show changes in market volatility, while the RSI plot highlights potential overbought and oversold conditions. Together, these indicators provide useful insights into trend, momentum, and volatility patterns for forecasting analysis.

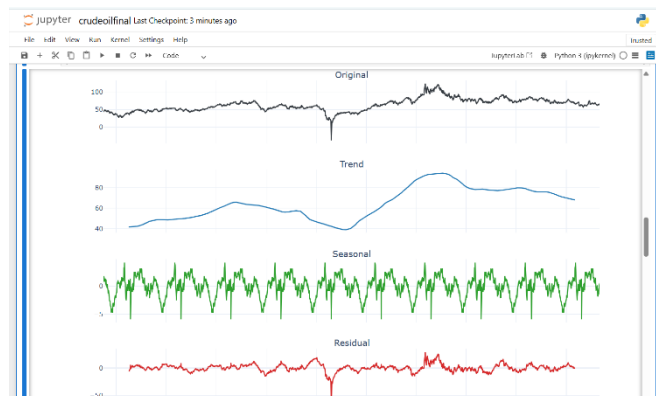


Figure4: TimeSeriesDecomposition

The figure presents the decomposition of crude oil prices into trend, seasonal, and residual components. The original series shows overall price fluctuations over time, while the trend component highlights the long-term market direction. The seasonal component captures repeating cyclical patterns, and the residual component represents irregular variations and unexpected market shocks. This decomposition helps in understanding the underlying structure of the data and improves forecasting accuracy by separating meaningful patterns from noise.

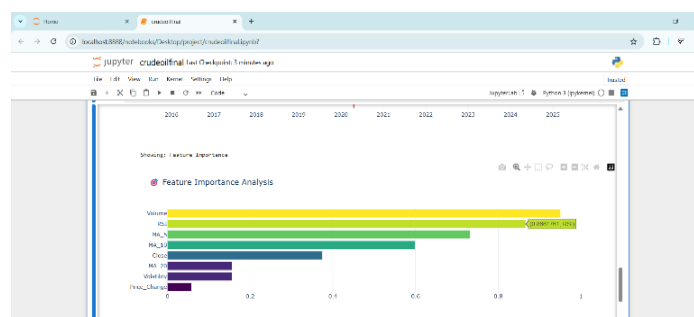


Figure5: FeatureImportanceAnalysis

The figure shows the importance of different features used for crude oil price prediction. Trading volume appears as the most influential feature, indicating its strong impact on price movements. Moving averages (MA_10, MA_20, and MA_50) also contribute significantly by capturing market trends, while indicators such as volatility and RSI provide additional insights into market momentum and price behavior.

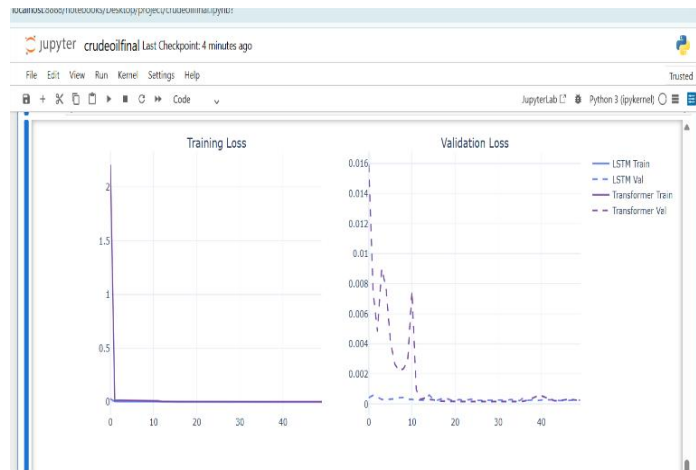


Figure6:ModelTrainingProgress

The figure compares the training and validation loss of the LSTM and Transformer models over multiple epochs. Both models show a rapid decrease in loss during early training, indicating effective learning. The Transformer model achieves smoother convergence and lower validation loss, suggesting better generalization and stability. In comparison, the LSTM model shows slightly higher fluctuations, reflecting greater sensitivity to market volatility. Overall, the results indicate that the Transformer model performs more effectively for crude oil price forecasting.

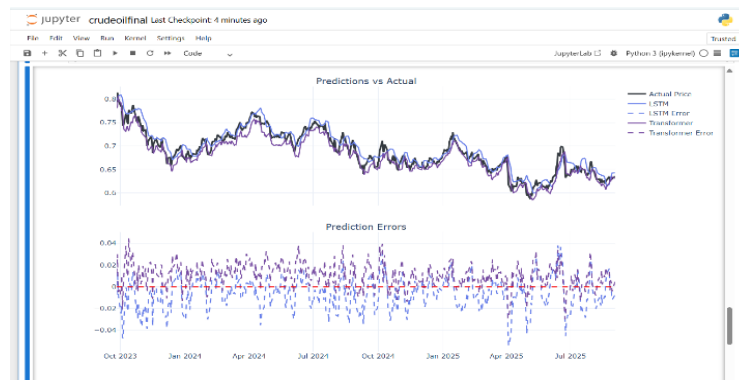


Figure7:PredictionAnalysis

The figure compares the actual crude oil prices with predictions from the LSTM and Transformer models. Both models follow the overall market trend, but the Transformer shows closer alignment with the actual values, especially during fluctuating periods. The prediction error plot further indicates that the Transformer produces smaller and more stable errors compared to the LSTM. These results demonstrate the Transformer model's higher accuracy and reliability in crude oil price forecasting.

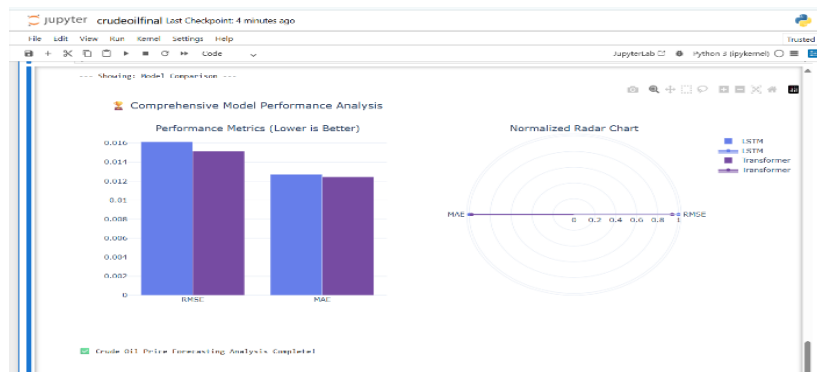


Figure7:ModelComparison

The figure compares the performance of the LSTM and Transformer models using MAE and RMSE metrics. The bar chart shows that the Transformer achieves lower error values, indicating better prediction accuracy. The

radar chart also highlights the Transformer's more stable and consistent performance compared to the LSTM. Overall, the results demonstrate that the Transformer model provides more reliable crude oil price forecasting.

VII. CONCLUSION & FUTURE SCOPE

The "Generative AI for Crude Oil Price Forecasting" system successfully integrates advanced deep learning models—LSTM, Transformer, and TimeGAN—to address the complexities of predicting crude oil prices, which are characterized by high volatility, nonlinear behavior, and sensitivity to global events. The project demonstrates that combining predictive and generative techniques significantly improves forecasting accuracy compared to standalone statistical or deep learning methods. The Transformer architecture proved particularly effective due to its self-attention mechanism, which captures long-range dependencies better than traditional sequential models. TimeGAN further enhanced model robustness by generating realistic synthetic time-series data that reduced overfitting and enriched the training distribution. Rigorous testing at unit, integration, system, and validation levels confirmed the stability and reliability of the entire workflow. Performance metrics such as RMSE, MAE, and MAPE consistently indicated strong predictive capability, while visual comparisons showed close alignment between actual and predicted price movements. Overall, the system meets its objectives and provides a scalable, accurate, and user-friendly forecasting framework suitable for analysts, investors, researchers, and policymakers.

Looking ahead, the forecasting framework can be further improved in several ways. Integrating additional external factors—such as macroeconomic indicators, geopolitical events, OPEC policies, exchange rates, market demand, and sentiment data—would enable the model to better capture real-world drivers of crude oil price volatility. Future work may also involve developing more advanced hybrid architectures that combine LSTM, Transformer, TimeGAN, and other cutting-edge models to enhance multi-horizon forecasting accuracy. Interactive visual dashboards using tools like Plotly, Dash, or Tableau can offer richer and more interpretable insights. Furthermore, incorporating Explainable AI (XAI) techniques would improve transparency by highlighting key features and model behaviors behind each prediction. The system can also be adapted to forecast other commodities like natural gas, gold, or agricultural products, thereby widening its practical applicability. Finally, deploying the framework, improving computational efficiency with GPU/TPU acceleration, and enabling real-time forecasting would make the system even more powerful and accessible for large-scale industrial use.

REFERENCES

- [1] Vapnik, V., *Statistical Learning Theory*, Wiley, 1995. Available: <https://scholar.google.com/scholar>
- [2] Breiman, L., "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. Available: <https://scholar.google.com/scholar>
- [3] Hochreiter, S., and Schmidhuber, J., "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [4] Hu, Z., Li, Y., and Wang, S., "Crude oil price prediction using CEEMDAN and LSTM-based deep learning models," *Energy Economics*, vol. 95, 2021.
- [5] Yoon, J., Jarrett, D., and van der Schaar, M., "Time-series generative adversarial networks," *NeurIPS*, 2019. Available: <https://scholar.google.com/scholar>
- [6] Hamilton, J. D., "Causes and consequences of the oil shock of 2007–08," *Brookings Papers on Economic Activity*, 2009.
- [7] He, X., Chen, X., and Hu, J., "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *Applied Intelligence*, 2023. Available: <https://scholar.google.com/scholar>
- [8] Qin, Y., Song, D., Zhang, H., Cheng, W., and Chen, H., "A dual-stage attention-based recurrent neural network for time series prediction," *IJCAI*, 2017. Available: <https://scholar.google.com/scholar>