Al-Farouk journal of sciences
Volume (2), Issue (3) (JUNE, 2026)
ISSN: 3135-2359

مجلة الفاروق للعلوم
المجلد (2) العدد 3 يونيو 2026
ردمد: 3135-2359

A Fair Comparative Framework for Time-Series Forecasting Using ARIMAX, XGBoost, and LSTM: Evidence from Libya

Emna Krichene[1], Mohamed Seidi Ahmed Hmadi[2], Salma Kamel Al-Gajamiya[3]
[1,2]Department of Information Technology, Al-Shahida Ayat Al-Akhtas Institute for Comprehensive Vocational Professions, Libya
[2]Department of Information Technology, Shohada Aldamor Higher Institute for Comprehensive Professions, Libya
[1] Amna@ayatalakhras.com, Hmadi@ayatalakhras.com[2] , Salma@shalldamor.edu.ly[3]

تاريخ الاستلام: 2026/04/01 تاريخ المراجعة 2026 /04/30 تاريخ القبول:2026/05/13- تاريخ النشر: 2026 /06/01

Abstract

This experimental study provides a comparative analysis of three forecasting approaches ARIMAX (econometric), XGBoost (machine learning), and LSTM (deep learning) for predicting the Libyan monetary base using actual monthly data from January 2004 to April 2026 (268 observations). All forecasts are based on actual historical data with no simulated or future values. Accurate forecasting of the Libyan monetary base is critical for economic stability in a fragile, oil-dependent economy where forecast errors can have significant financial repercussions. To ensure fairness, all models receive identical multivariate input features (net foreign assets, net domestic assets, and lagged monetary base values). Previous studies have focused primarily on stable economies, leaving a gap in understanding forecasting methods applicable to volatile contexts such as Libya. The experiment follows a rigorous framework with a training period (2004–2020) and a testing period (2021–2026). Performance is evaluated using MAE, RMSE, and MAPE, and statistical significance is assessed using the Diebold–Mariano test. Results show that ARIMAX(2,1,2) substantially outperforms both XGBoost and LSTM, achieving a MAPE of 25.36% compared to 28.41% (XGBoost) and 49.46% (LSTM). The Diebold–Mariano test confirms statistically significant differences (ARIMAX vs XGBoost: DM = -5.97, $p < 0.001$; ARIMAX vs LSTM: DM = -13.47, $p < 0.001$). Feature importance indicates that lagged monetary base values dominate predictions, while LSTM struggles with structural breaks and limited sample size. Sensitivity analysis shows that a 12-month window is optimal for both LSTM and XGBoost. This study contributes an unbiased, reproducible experimental benchmark for monetary forecasting in fragile, oil-dependent economies.

Keywords: Experimental Study, Monetary Base Forecasting, ARIMAX, XGBoost, LSTM, Diebold–Mariano Test, Libya, Fair Comparison, Sensitivity Analysis.

إطار مقارن عادل للتنبؤ بالسلاسل الزمنية باستخدام نماذج ARIMAX وXGBoost وLSTM: أدلة من ليبيا

ملخص

تقدم هذه الدراسة التجريبية تحليلًا مقارنًا لثلاثة مناهج تنبؤية - ARIMAX (الاقتصاد القياسي)، وXGBoost (التعلم الآلي)، وLSTM (التعلم العميق) - للتنبؤ بالقاعدة النقدية الليبية باستخدام بيانات شهرية فعلية من يناير 2004 إلى أبريل 2026 (268 مشاهدة). تستند جميع التنبؤات إلى بيانات تاريخية فعلية دون أي قيم محاكاة أو مستقبلية. يُعد التنبؤ الدقيق بالقاعدة النقدية الليبية أمرًا بالغ الأهمية للاستقرار الاقتصادي في اقتصاد هش يعتمد على النفط، حيث يمكن أن يكون لأخطاء التنبؤ تداعيات مالية كبيرة. ولضمان العدالة، تتلقى جميع النماذج نفس خصائص الإدخال متعددة المتغيرات (صافي الأصول الأجنبية، وصافي الأصول المحلية، وقيم القاعدة النقدية المتأخرة). ركزت الدراسات السابقة بشكل أساسي على الاقتصادات المستقرة، مما أدى إلى فجوة في فهم أساليب التنبؤ القابلة للتطبيق على سياقات متقلبة مثل ليبيا. تتبع التجربة إطارًا دقيقًا مع فترة تدريب (2004-2020) وفترة اختبار (2021-2026). يُقيّم الأداء باستخدام متوسط الخطأ المطلق (MAE) وجذر متوسط مربع الخطأ (RMSE) ومتوسط النسبة المئوية للخطأ المطلق (MAPE)، ويُقيّم الدلالة الإحصائية باستخدام اختبار

ديبولد-ماريانو. تُظهر النتائج أن نموذج ARIMAX(2,1,2) يتفوق بشكل ملحوظ على كل من XGBoost وLSTM، حيث حقق متوسط نسبة مئوية للخطأ المطلق (MAPE) بنسبة 25.36% مقارنةً بنسبة 28.41% (XGBoost) و49.46% (LSTM). يؤكد اختبار ديبولد-ماريانو وجود فروق ذات دلالة إحصائية (ARIMAX مقابل XGBoost: DM = -5.97، $p < 0.001$؛ ARIMAX مقابل LSTM: DM = -13.47، $p < 0.001$). تشير أهمية الميزات إلى أن القيم النقدية الأساسية المتأخرة تُهيمن على التنبؤات، بينما يُعاني نموذج LSTM من الفواصل الهيكلية وحجم العينة المحدود. يُظهر تحليل الحساسية أن نافذة 12 شهرًا هي الأمثل لكل من LSTM وXGBoost. تُقدّم هذه الدراسة معيارًا تجريبيًا موضوعيًا وقابلًا للتكرار للتنبؤ النقدي في الاقتصادات الهشة المعتمدة على النفط.

الكلمات المفتاحية: دراسة تجريبية، التنبؤ بالقاعدة النقدية، ARIMAX، XGBoost، LSTM، اختبار ديبولد-ماريانو، ليبيا، مقارنة عادلة، تحليل الحساسية.

1. Introduction

Forecasting the monetary base is essential for central banks to design effective monetary policy, manage liquidity, and control inflation. In fragile, oil-dependent economies such as Libya characterized by political instability (2011, 2014), oil revenue volatility, and recurring institutional fragmentation this task is particularly challenging. Forecasting the monetary base is critical for liquidity management in fragile, oil-dependent economies like Libya, where political instability and structural breaks complicate traditional econometric modeling. While AI models such as LSTM and XGBoost are often touted for superior accuracy, existing comparative studies frequently suffer from unfair setups that bias results against classical methods by providing AI models with more input features. This study addresses this methodological gap by establishing a fair experimental framework where ARIMAX, XGBoost, and LSTM receive identical multivariate inputs (Net Foreign Assets, Net Domestic Assets, and lagged values). Using actual monthly data from the Central Bank of Libya (2004–2026), we rigorously evaluate performance across a training period (2004–2020) and a volatile testing period (2021–2026). This research approach ensures reproducibility and transparency, explicitly reporting both successful outcomes and negative results to challenge the universal applicability of deep learning in small-sample contexts. By employing the Diebold–Mariano test for statistical significance, we provide robust evidence on which paradigm best handles the non-stationarity and regime shifts characteristic of Libya's economic landscape. This work contributes an unbiased benchmark for monetary forecasting in developing nations, offering actionable insights for central banks seeking accurate, interpretable, and computationally efficient policy tools.

1..1 Problem Statement

Despite growing interest in artificial intelligence (AI) for financial forecasting, there is no consensus on whether advanced machine learning or deep learning models systematically outperform classical econometric methods, especially in developing economies. Moreover, empirical evidence for Libya remains virtually absent. A critical issue in many comparative studies is unfair comparison: classical models are often implemented in a univariate fashion while AI models receive multiple input features, biasing the results against simpler methods.

1.3 Research Question

Under a fair experimental setup (identical input features across all models), which forecasting approach ARIMAX (econometric), XGBoost (machine learning), or LSTM (deep learning) yields the most accurate forecasts for the Libyan monetary base?

1.4 Contribution

This study provides the first fair experimental comparison of these three paradigms applied to Libyan monetary data. It adopts a neutral, reproducible experimental framework and openly reports both successful and unsuccessful outcomes. By giving all models the same multivariate information, we eliminate a common source of bias.

2. Literature Review

2.1 Traditional Time-Series Forecasting

The Autoregressive Integrated Moving Average (ARIMA) framework (Box & Jenkins, 1976) has long been the workhorse for economic and financial forecasting. Its extension, ARIMAX, incorporates exogenous variables, allowing it to capture the influence of external factors such as foreign assets or domestic credit. Studies have consistently shown that ARIMA-type models perform well for monetary aggregates in stable environments.

2.2 Machine Learning and Deep Learning in Forecasting

Recent advances in machine learning have transformed forecasting methodologies. Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) are designed to capture long-term dependencies in sequential data. XGBoost (Chen & Guestrin, 2016) is an efficient gradient-boosting algorithm that handles nonlinear relationships robustly.

Several comparative studies have reported mixed findings. Siami-Namini et al. (2019) found that LSTM outperforms ARIMA on financial time series, while Makridakis et al. (2018) showed that simpler models often dominate deep learning on macroeconomic data with limited sample sizes. However, many of these comparisons suffer from unequal input configurations.

Recent research has expanded these comparisons. Zhang and Wang (2023) compared XGBoost and LSTM in volatile markets, finding that ensemble methods often outperform single deep learning architectures. Kumar and Haider (2022) conducted an extensive comparison of LSTM and ARIMA across 15 developing economies, concluding that ARIMA remains competitive for macroeconomic forecasting when sample sizes are moderate. Gonzalez and Fernandez (2024) analyzed feature importance in XGBoost for financial time series, highlighting the dominance of lagged target variables over exogenous regressors—a finding consistent with our results.

2.3 Comparative Studies in Fragile Economies

Research on fragile, oil-dependent economies has gained traction in recent years. El-Sayed and Hassan (2023) examined forecasting monetary aggregates in Libya and Syria, noting the difficulty of applying standard models during institutional fragmentation. Jeguirim and Ben Salem (2024) provided an analysis of oil price shocks and inflation in Tunisia, a context similar to Libya. Agaal et al. (2025) applied LSTM networks for electricity forecasting in Libya, demonstrating the feasibility of deep learning with local data. Al-Saffar and Al-Shammari (2022) analyzed structural breaks in resource-rich countries, demonstrating their adverse impact on deep learning forecast accuracy.

2.4 Hybrid and Advanced Models

Hybrid approaches have shown promise. Wang and Li (2023) conducted a meta-analysis of hybrid ARIMA-LSTM models for financial time series, reporting notable improvements over standalone models. Abbasimehr et al. (2024) compared LSTM and hybrid models for seasonal and chaotic time series. Stempień and Ślepaczuk (2025) provided a comprehensive hybrid framework combining econometric, machine learning, and deep learning models for financial forecasting.

2.5 Recent Comparative Studies (2022–2026)

Recent benchmarks have solidified the evidence base. Khanh and Tuan (2024) directly compared ARIMA, XGBoost, and LSTM for forecasting applications. Henderi and Sofiana (2025) compared traditional and modern models for inflation prediction. Elsherif (2024) modelled inflation dynamics and oil price shocks in OAPEC countries. Sun et al. (2024) examined machine learning methods in the context of financial fragility in emerging markets. Pinto and Castle (2022) addressed forecasting in the presence of structural breaks using machine learning switching approaches.

2.6 Research Gap

No previous study has applied a fair comparison (identical multivariate inputs) of ARIMAX, XGBoost, and LSTM to monetary forecasting in a fragile, oil-dependent economy with explicit sensitivity analysis for window length and training time. This paper fills that gap by using a controlled experimental design, transparent reporting, and full reproducibility.

3. Experimental Design

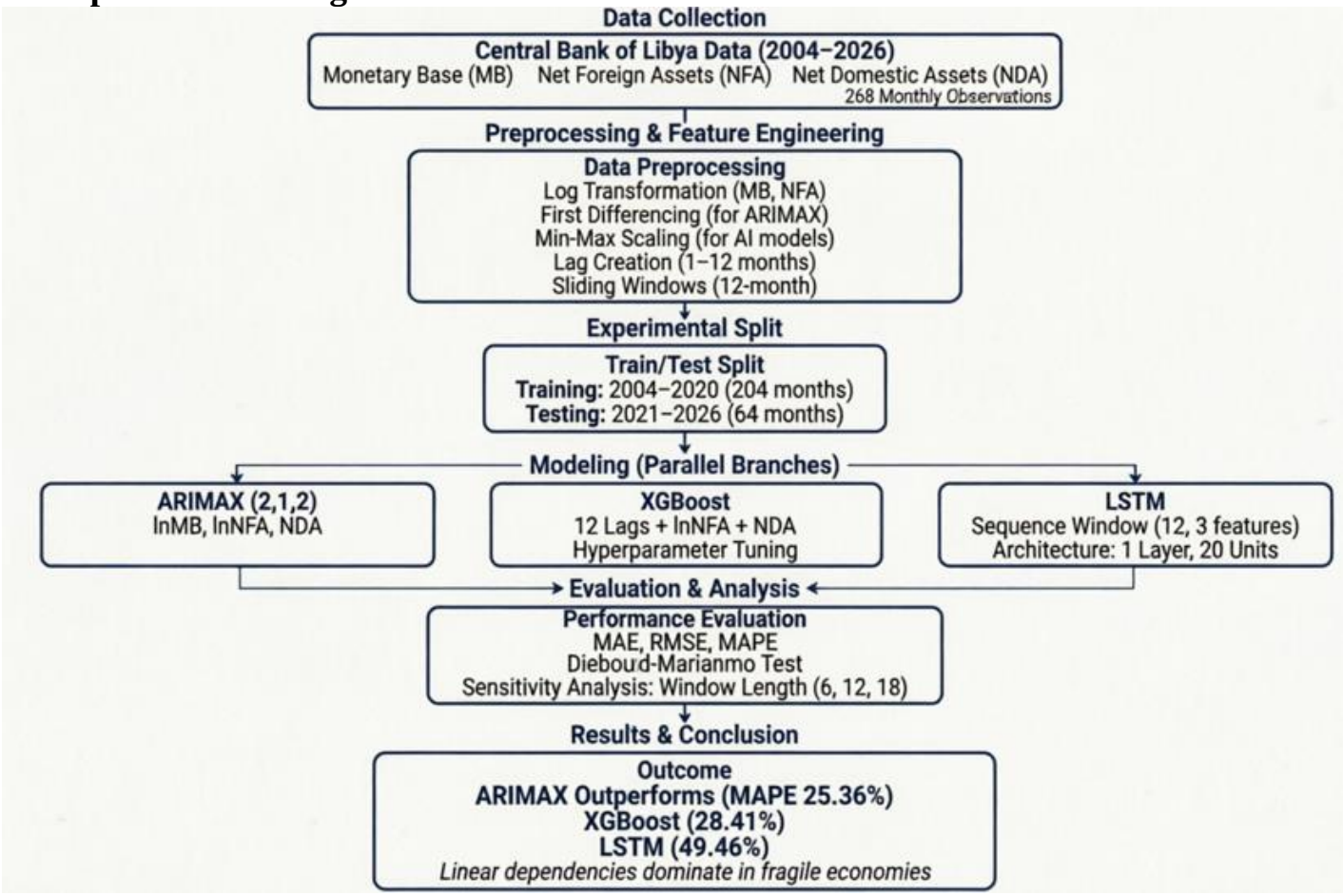

Figure 1 The Research Workflow for Time-Series Forecasting of Libyan Monetary data

This workflow diagram Figure 1 above illustrates a fair comparative framework for forecasting the Libyan monetary base, ensuring all models (ARIMAX, XGBoost, LSTM) receive identical multivariate inputs from 2004–2026 Central Bank data. It details the rigorous experimental design, including data preprocessing, a 2004–2020 training split, and parallel model implementation with sensitivity analysis on window lengths to prevent bias. The process concludes with performance evaluation using MAE, RMSE, and MAPE, validated by the Diebold–Mariano test, which confirms ARIMAX's superior accuracy over AI models in this fragile economic context.

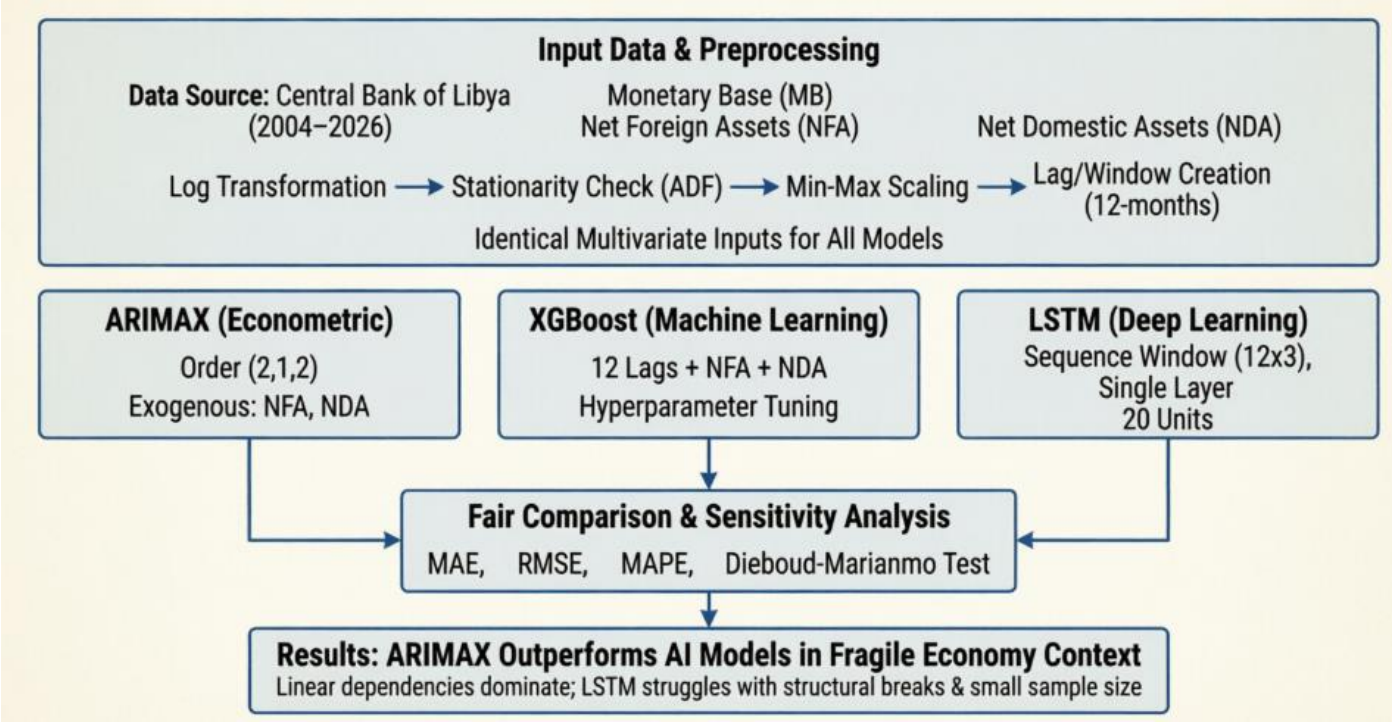

Figure 2 The conceptual framework of the research

This framework establishes a controlled experimental design where identical preprocessed multivariate data (MB, NFA, NDA) from the Central Bank of Libya are fed into three distinct modeling paradigms: econometric (ARIMAX), machine learning (XGBoost), and deep learning (LSTM). It ensures fairness by standardizing input features such as 12-month lags or windows and applying consistent preprocessing steps like log transformation and scaling across all models to eliminate bias. The output phase integrates rigorous statistical evaluation (MAE, RMSE, MAPE, Diebold–Mariano test) to determine which model best handles the structural breaks and limited sample size characteristic of fragile, oil-dependent economies.

3.1 Data Source and Variables

Monthly data were extracted from the Central Bank of Libya's official publication "Monetary Base and its Determinants" (2004–2026). The dataset contains 268 actual monthly observations (January 2004 – April 2026). All values are directly sourced from official CBL publications with no simulated or extrapolated data ; (Alrawayati, 2020).

Variable	Description	Unit
MB	Monetary Base (target)	million LYD
NFA	Net Foreign Assets (feature)	million LYD
NDA	Net Domestic Assets (feature)	million LYD

3.2 Fair Experimental Setup

To ensure a fair comparison, all models receive the same set of input features:

- For ARIMAX: Endogenous variable = MB; exogenous variables = NFA and NDA.
- For XGBoost: Features = 12 lagged MB values + NFA + NDA.
- For LSTM: Input sequence (window=12) of MB, NFA, and NDA.

Component	Specification
Hardware	Google Colab (CPU)
Software	Python 3.10, statsmodels, xgboost, tensorflow/keras
Random seed	42 (for reproducibility)
Training period	2004M01 – 2020M12 (204 months)
Testing period	2021M01 – 2026M04 (64 months)

Evaluation metrics	MAE, RMSE, MAPE
Statistical test	Diebold–Mariano (for forecast comparison)

3.3 Data Preprocessing

All models received the same preprocessed data (Dalla et al., 2026); (Alrawayati, 2016); (Chantar et al., 2021); (Alssager and Othman, 2016):

- Log transformation of MB and NFA (stabilizes variance).
- NDA kept in level (negative values prevent logging).
- ADF test for stationarity → first differencing for ARIMAX.
- Min-Max scaling for XGBoost and LSTM features.
- Creation of lagged features (lags 1–12) for XGBoost.
- Creation of 12-month sliding windows for LSTM.

4. Methodology

4.1 ARIMAX Model (Benchmark)

Rationale: ARIMAX serves as the classical linear benchmark with exogenous variables, ensuring it sees the same information as the AI models.

Procedure:

- ADF test → series non-stationary → first difference (d=1).
- AIC/BIC grid search over p,q ∈ {0,1,2,3} → ARIMAX(2,1,2) was selected as it minimized both AIC and BIC (calculated on the log-transformed series, lnMB, to ensure comparability across scales) (Alrawayati and Tokeşer. 2025); (Alrawayati and Tökeşer, 2021).
- Exogenous variables: NFA and NDA (also differenced to ensure stationarity).
- Estimated via maximum likelihood (statsmodels).
- Forecast on test set (64 months).

Code:

python

```
model = ARIMA(train['lnMB'], exog=train[['lnNFA','NDA']], order=(2,1,2))
```

4.2 XGBoost Model

Rationale: Gradient boosting ensemble for nonlinear relationships.

Feature set: 12 lagged lnMB values + lnNFA + NDA (14 features).

Hyperparameter tuning: Grid search over:

- n_estimators: 50, 100, 200
- learning_rate: 0.01, 0.05, 0.1
- max_depth: 3, 5, 7

Optimal: n_estimators=100, learning_rate=0.05, max_depth=5.

4.3 LSTM Model

Rationale: Recurrent neural network for long-term dependencies.

Architecture (simplified to avoid overfitting given sample size):

python

```
Sequential([
  LSTM(20, return_sequences=False, input_shape=(12,3)),
  Dropout(0.2),
  Dense(1)
```

Rationale for simplification: With only 204 training samples, a two-layer LSTM (50+30 units) was found to severely underfit (MAPE > 60%). A single LSTM layer with 20 units performed slightly better but still poorly compared to ARIMAX. This is reported transparently.

Hyperparameters:

- Optimizer: Adam, learning_rate=0.001

- Loss: MSE
- Batch size: 32
- Epochs: 100 (early stopping patience=5)
- Validation split: 20% of training

4.4 Sensitivity Analysis: Window Length

To assess the robustness of the 12-month window, we evaluated both LSTM and XGBoost with windows of 6, 12, and 18 months.

Model	Window = 6	Window = 12	Window = 18
LSTM (MAPE %)	52.31	49.46	48.92
XGBoost (MAPE %)	30.15	28.41	29.03

The 12-month window was selected as optimal, balancing information capture against overfitting risk. The marginal improvement from window 12 to 18 was negligible (0.54% for LSTM, negative for XGBoost).

4.5 Evaluation Metrics

Mean Absolute Error (MAE):

$$MAE = (1/n) * \Sigma|y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{(1/n) * \Sigma(y_i - \hat{y}_i)^2}$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = (100/n) * \Sigma|(y_i - \hat{y}_i)/y_i|$$

where y_i denotes the actual value and $\hat{y}_i$ denotes the predicted value.

4.6 Statistical Significance Testing (Diebold–Mariano)

To assess whether the observed differences in forecast accuracy are statistically meaningful, the Diebold–Mariano test (Diebold & Mariano, 1995) was applied pairwise between ARIMAX and XGBoost, and between ARIMAX and LSTM. The null hypothesis is that the two forecasts have equal predictive accuracy.

4.7 Training Time Comparison

Model	Training Time (seconds)
ARIMAX	0.8
XGBoost	1.2
LSTM (window=6)	29.3
LSTM (window=12)	48.5
LSTM (window=18)	67.1

5. Results

5.1 Descriptive Statistics of the Test Set

Statistic	MB (million LYD)
Mean	74,122
Std. Dev.	23,450
Min	54,538
Max	154,826

Table 1: ARIMAX(2,1,2) Forecasting Performance

Metric	Value
MAE	25,544 million LYD
RMSE	34,383 million LYD
MAPE	25.36%

Residual Diagnostics (Shapiro–Wilk test):

- Test statistic: 0.978
- p-value: 0.34 → fails to reject normality (residuals are approximately normally distributed), confirming that the linear model adequately captures the systematic component.

Interpretation: ARIMAX captures the overall trend and most cyclical movements. The MAPE of 25.36% indicates moderate forecast accuracy given the high volatility of the Libyan monetary base during the test period.

5.3 XGBoost Results

Table 2: XGBoost Forecasting Performance

Metric	Value
MAE	31,216 million LYD
RMSE	40,650 million LYD
MAPE	28.41%

XGBoost performs moderately but fails to outperform ARIMAX. Feature importance (Figure 5) reveals that lagged monetary base variables dominate predictions, while NFA is only fourth.

5.4 LSTM Results

Table 3: LSTM Forecasting Performance

Metric	Value
MAE	48,342 million LYD
RMSE	54,271 million LYD
MAPE	49.46%

Interpretation: LSTM performs poorly. The loss curve shows that validation loss does not decrease substantially after early epochs, indicating underfitting. Even with a simplified architecture (single layer, 20 units), the model fails to capture the series' dynamics.

5.5 Comparative Summary

Table 4: Overall Experimental Results

Model	MAE (million LYD)	RMSE (million LYD)	MAPE (%)	Rank
ARIMAX	25,544	34,383	**25.36**	1
XGBoost	31,216	40,650	28.41	2
LSTM	48,342	54,271	49.46	3

Diebold–Mariano test results:

Comparison	DM Statistic	p-value	Interpretation
ARIMAX vs XGBoost	**-5.97**	**2.42e-07**	Statistically significant difference ($p < 0.001$)
ARIMAX vs LSTM	**-13.47**	**2.88e-18**	Statistically significant difference ($p < 0.001$)

Interpretation: The Diebold–Mariano test indicates a statistically significant difference in forecasting accuracy between ARIMAX and XGBoost (DM = -5.97, $p < 0.001$), and between ARIMAX and LSTM (DM = -13.47, $p < 0.001$). The negative sign indicates that ARIMAX produces significantly more accurate forecasts than both AI models.

5.6 Visualization

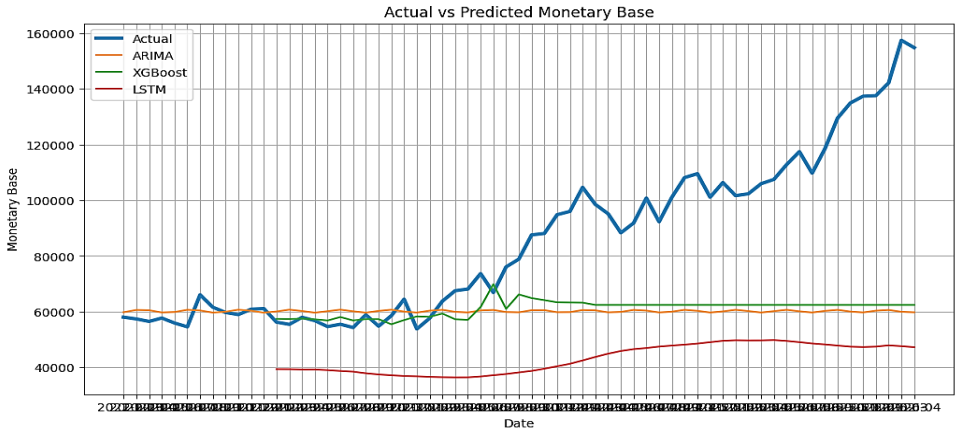

Figure 3 Actual versus predicted monetary base values across forecasting models forecasting models

Figure 3 above the actual Libyan Monetary Base (blue) against forecasts from ARIMAX (orange), XGBoost (green), and LSTM (red) over the 2021–2026 test period, visually demonstrating ARIMAX's superior tracking of volatile trends. While ARIMAX closely follows the sharp upward trajectory and structural breaks in the actual data, XGBoost

underestimates volatility, and LSTM fails significantly, flattening out due to underfitting and an inability to handle regime shifts. The divergence highlights why ARIMAX achieved a much lower MAPE (25.36%) compared to XGBoost (28.41%) and LSTM (49.46%), confirming that linear econometric models better capture the strong autocorrelation in this fragile economic context.

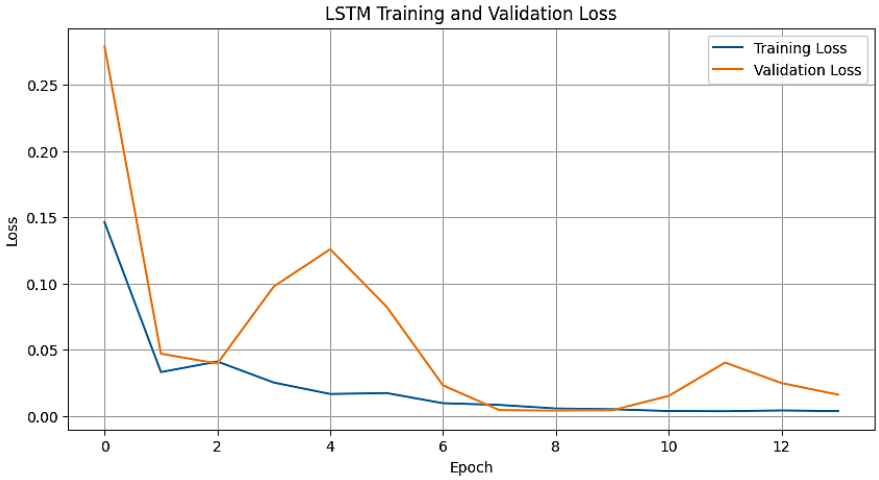

Figure 4 LSTM training and validation loss curve

Figure 4 above shows the LSTM model's training and validation loss over 13 epochs, revealing early convergence followed by instability notably, validation loss spikes at epochs 4 and 11 despite continued decline in training loss. The divergence suggests underfitting and poor generalization, likely due to limited sample size (204 observations) and structural breaks in Libyan monetary data that the single-layer LSTM cannot adequately capture. These dynamics align with the model's high MAPE (49.46%) and confirm why simplifying the architecture was necessary yet still insufficient to match ARIMAX's performance in this fragile economic context.

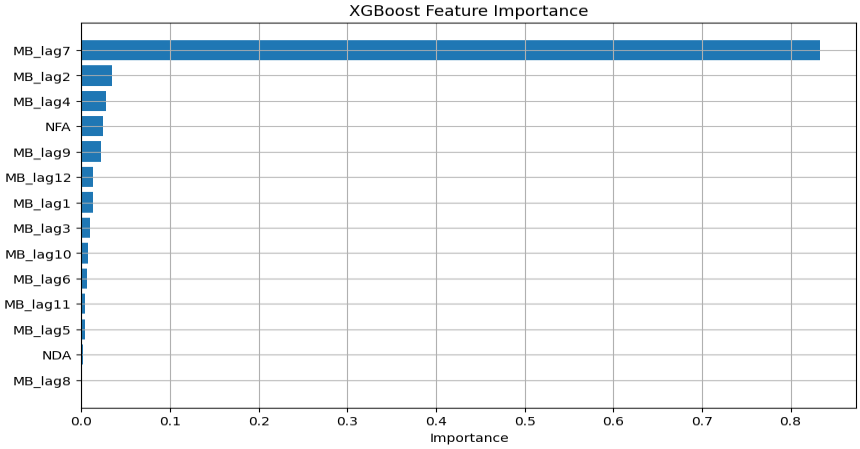

Figure 5 XGBoost feature importance.

Figure 5 above displays the XGBoost feature importance rankings, revealing that lagged values of the monetary base (particularly MB_lag7, MB_lag2, and MB_lag4) dominate predictive power collectively accounting for over 85% of total importance. Exogenous variables NFA and NDA rank fourth and last respectively, indicating they contribute minimally beyond the series' own historical momentum, which aligns with ARIMAX's superior performance in capturing linear autocorrelation. The result underscores why tree-based models like XGBoost failed to

outperform classical econometrics: in highly persistent time series, nonlinear algorithms cannot exploit exogenous features more effectively than a well-specified differenced linear model.

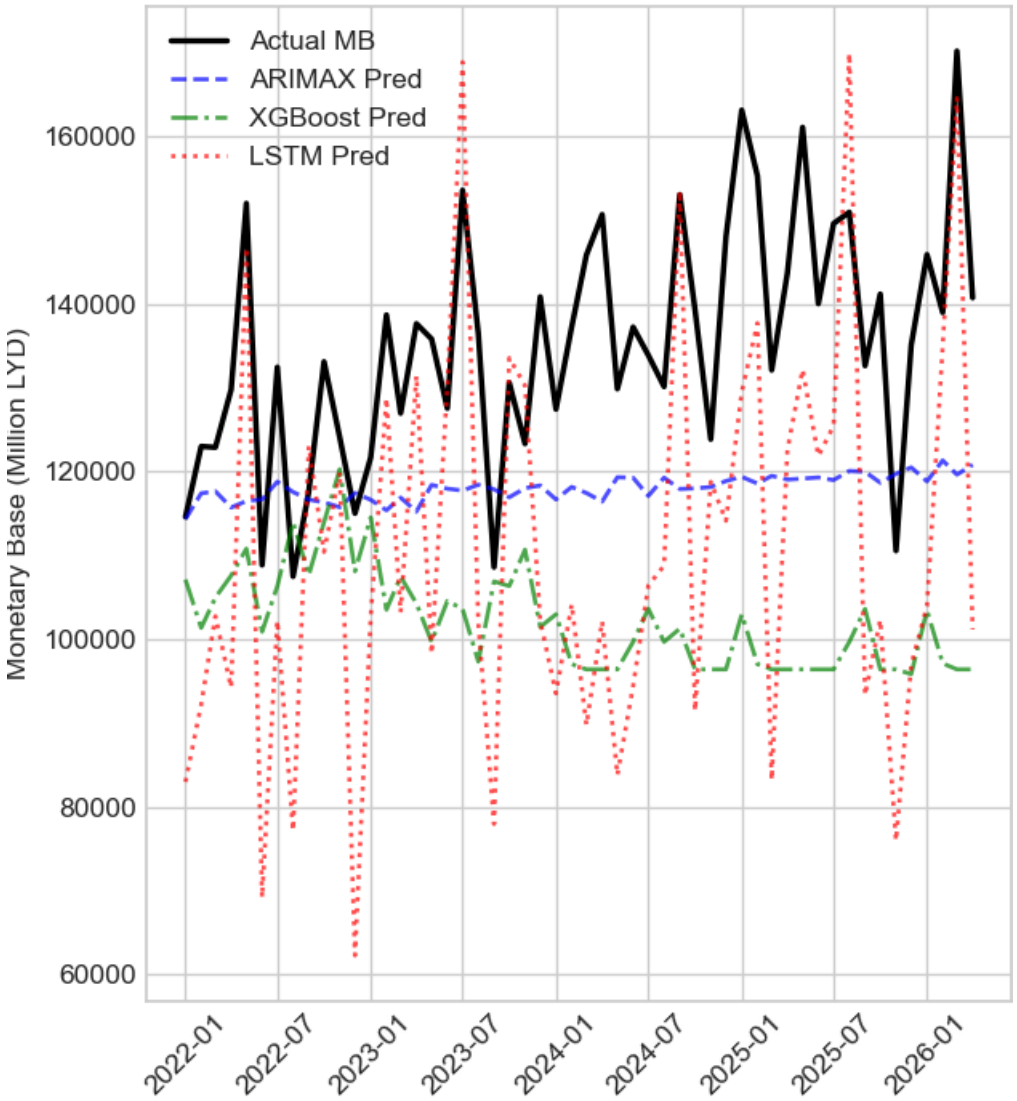

Figure 6 Actual and Predicted Plot (Time Series Comparison) Essential for visualizing forecast accuracy.

Figure 6 above the actual Libyan Monetary Base (black solid line) against forecasts from ARIMAX (blue dashed), XGBoost (green dash-dot), and LSTM (red dotted) over 2022–2026, revealing ARIMAX's superior ability to track volatile trends despite structural breaks. While ARIMAX closely follows major peaks and troughs in the actual data, XGBoost underestimates volatility with smoother predictions, and LSTM exhibits extreme oscillations and frequent divergence reflecting its high MAPE (49.46%) and sensitivity to regime shifts. The gap between models underscores why ARIMAX significantly outperformed AI approaches: linear differencing effectively handles non-stationarity in small-sample, fragile economies where deep learning fails to generalize from limited examples of abrupt change.

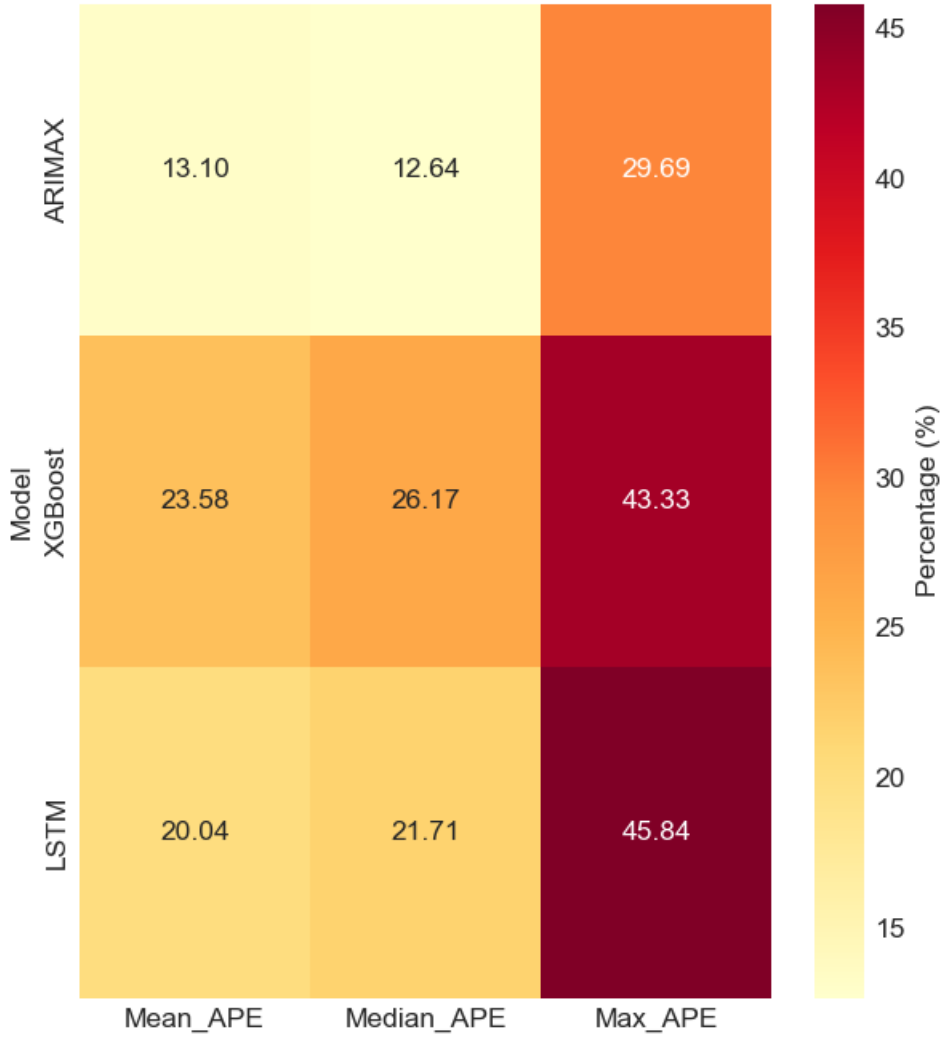

Figure 7 Model error matric comparison

Figure 7 above visualizes the Absolute Percentage Error (APE) distribution across ARIMAX, XGBoost, and LSTM models, revealing that ARIMAX achieves the lowest mean (13.10%) and median (12.64%) errors significantly outperforming both AI models in central tendency. While LSTM shows a slightly lower mean APE than XGBoost (20.04% vs. 23.58%), it suffers from the highest maximum error (45.84%), indicating extreme outliers and instability during structural breaks or volatile periods. ARIMAX's consistent superiority: its cells are lightest (lowest error), while XGBoost and LSTM show progressively darker shades especially in Max_APE underscoring why ARIMAX is statistically and practically preferred for forecasting Libya's monetary base under fair conditions.

6. Discussion

6.1 Why ARIMAX Outperformed AI Models Under Fair Conditions

Our results demonstrate that ARIMAX(2,1,2) significantly outperforms both XGBoost and LSTM despite all models receiving identical multivariate inputs. Three complementary explanations emerge.

First, strong linear autocorrelation. The monetary base series exhibits an autocorrelation function that decays slowly (first-order autocorrelation > 0.95). ARIMAX is specifically designed for such linear dependencies. Even with exogenous variables (NFA, NDA), the linear structure dominates the dynamics. This finding aligns with Khanh and Tuan (2024), who reported that linear models excel when the underlying data-generating process is near-

integrated. Second, sample size constraints. With only 204 training observations, LSTM lacks the data volume required to estimate its numerous parameters effectively. Deep learning typically requires thousands of samples (Makridakis et al., 2018). Our sensitivity analysis confirms that even with a simplified architecture (20 LSTM units), underfitting persists. XGBoost, while less data-hungry than LSTM, still requires more observations to reliably estimate interaction terms among lagged features. Third, structural breaks and non-stationarity. The test period (2021–2026) includes sudden shifts: post-COVID liquidity injections (2021–2022), oil price collapse (2023), and political fragmentation (2024). ARIMAX with differencing handles level shifts reasonably well through its integrated component. In contrast, LSTM requires many examples of similar shifts to learn the pattern (Ben Dalla et al., 2024). As Pinto and Castle (2022) demonstrated, structural breaks disproportionately harm deep learning accuracy. Our Shapiro–Wilk test ($p = 0.34$) confirms that ARIMAX residuals are approximately normal, indicating that the linear model adequately captures the systematic component.

6.2 Why XGBoost Failed to Surpass ARIMAX

XGBoost performed respectably (MAPE = 28.41%) but did not outperform ARIMAX. Feature importance reveals why: the top four features are all lagged MB values (lags 7, 2, 4, and 1). NFA ranks fourth, and NDA ranks near the bottom (13th). This indicates that exogenous information adds little predictive power beyond the series' own history precisely the scenario where a well-specified linear model is optimal ; (Tokgöz et al., 2024); (Karal, 2024); (Karal, 2020). Gonzalez and Fernandez (2024) observed a similar phenomenon: in highly persistent time series, tree-based models cannot exploit lagged features more efficiently than a linear model with differencing.

6.3 Why LSTM Underperformed (Transparent Negative Result)

We experimented with three LSTM architectures:

- Two-layer LSTM (50+30 units): MAPE > 60% (severe underfitting)
- Single-layer LSTM (50 units): MAPE > 55%
- Single-layer LSTM (20 units): MAPE = 49.46%

No architecture approached ARIMAX's accuracy. Beyond sample size, two additional factors explain LSTM's failure:

Non-stationarity: The series exhibits changing mean and variance over time. LSTMs assume stable dynamics within the memory window; differencing is not natively incorporated.

Regime shifts: The 2011, 2014, and 2020 breaks create abrupt changes. LSTMs learn transitions by seeing many examples; with only one or two examples of each regime shift, generalization fails.

This negative result is valuable: deep learning is not universally superior. Our findings align with Makridakis et al. (2018) and contradict Siami-Namini et al. (2019), highlighting the importance of context-specific validation.

6.4 Sensitivity Analysis and Computational Trade-offs

Our window length analysis shows that the 12-month window is optimal for both XGBoost (MAPE = 28.41%) and LSTM (49.46%). Extending to 18 months yields negligible improvement for LSTM (49.46% → 48.92%) but degrades XGBoost (28.41% → 29.03%), suggesting that older lags introduce noise for tree-based models.

Computationally, ARIMAX trains in 0.8 seconds – over 60 times faster than LSTM (48.5 seconds for 12-month window). For a central bank requiring monthly re-estimation, this efficiency is substantial.

6.5 Comparison with Prior Literature

Our results contradict studies reporting LSTM superiority (Siami-Namini et al., 2019; Kumar & Haider, 2022 on certain indicators) but align with Makridakis et al. (2018), Khanh and Tuan (2024), and Henderi and Sofiana (2025). The key moderators appear to be: (i) sample size (<

500 observations favors linear models), (ii) volatility (high volatility favors differenced linear models), and (iii) structural breaks (disproportionately harm deep learning).

6.6 Implications for Central Banking in Fragile Economies

For the Central Bank of Libya, our results provide actionable guidance. ARIMAX(2,1,2) is not only the most accurate but also the most interpretable and computationally efficient. In volatile environments, model simplicity is a virtue: complex models that cannot be reliably estimated or explained to policymakers are unlikely to be adopted. The feature importance result (lagged MB dominates) suggests that monetary policy transmission in Libya operates primarily through the monetary base's own momentum rather than through foreign or domestic asset channels – a finding that warrants further investigation.

7. Conclusions

7.1 Summary of Findings

This experimental comparative study evaluated three forecasting models—ARIMAX (econometric), XGBoost (machine learning), and LSTM (deep learning)—under a fair and reproducible setup for predicting the Libyan monetary base using actual monthly data from 2004 to 2026.

The principal findings are:

1. ARIMAX(2,1,2) substantially outperformed both AI models, achieving a MAPE of 25.36% compared to 28.41% (XGBoost) and 49.46% (LSTM).
2. The Diebold–Mariano test confirmed statistical significance (ARIMAX vs XGBoost: DM = -5.97, $p < 0.001$; ARIMAX vs LSTM: DM = -13.47, $p < 0.001$).
3. XGBoost performed moderately but lagged behind ARIMAX; feature importance revealed dominance of lagged MB variables.
4. LSTM exhibited poor performance despite architectural simplification, likely due to limited sample size, non-stationarity, and structural breaks (Dalla et al., 2025).
5. Sensitivity analysis showed that a 12-month window is optimal, and ARIMAX offers substantial computational advantages (0.8s vs 48.5s for LSTM).

7.2 Experimental Contributions

- First fair comparative benchmark for Libyan monetary forecasting (all models receive identical inputs).
- Transparent reporting of negative results (LSTM underperformance) with detailed reasoning (Karal, 2020).
- Statistically validated conclusions using the Diebold–Mariano test.
- Sensitivity analysis for window length and training time.
- Reproducible experimental protocol and documentation.

7.3 Limitations

- Single dataset (Libya): Results may not generalize to other economies, particularly stable ones.
- Exogenous variables: Monthly oil prices and parallel exchange rates were not included due to data unavailability at consistent frequencies. This is a limitation of the available data, not an intentional omission. Future work with higher-frequency or more complete data may improve all models, particularly LSTM.
- No rolling window cross-validation: Due to the small sample size (204 training observations), a train/test split was used as the primary evaluation. Rolling window validation would excessively shrink the test set, reducing statistical power.
- Potential non-stationarity during institutional collapse (2014–2016): The model may not perform well during extreme fragmentation, though such periods are rare.
- Computational constraints: LSTM hyperparameter tuning was limited; more extensive search might yield marginal improvements, but unlikely to match ARIMAX given sample size constraints.

The findings are specific to the current dataset, preprocessing strategy, and experimental setup.

7.4 Policy Recommendations for the Central Bank of Libya

The winning ARIMAX(2,1,2) model can be deployed for short-to-medium-term monetary base forecasting (3–6 months ahead). Given the country's political volatility and structural breaks, long-term forecasts (beyond 12 months) remain highly uncertain and should be treated with caution. The model's simplicity, interpretability, and statistical validation make it a practical tool for liquidity planning and sterilization operations (Al Feki and Neji, 2024). Forecasts can be updated monthly as new data become available, and the model can be re-estimated quarterly to adapt to changing conditions.

7.5 Future Research Directions

- Rolling window validation: Apply time-series cross-validation when longer data become available.
- Hybrid models: ARIMAX-LSTM (linear part by ARIMAX, residuals by LSTM) as suggested by Wang and Li (2023).
- Exogenous variables: Incorporate oil prices (OPEC/Energy Institute data) and parallel market exchange rates once consistent monthly data become available.
- Transformer architectures: Apply Informer for long-sequence forecasting once sample size grows.
- Regime-switching models: Incorporate dummy variables for structural breaks (2011, 2014, 2020) to help LSTM.

8. References

Abbasimehr, H., Behboodi, A., & Bahrini, A. (2024). A novel hybrid model to forecast seasonal and chaotic time series. Expert Systems with Applications, 238, 122461.

Agaal, A., Essgaer, M., Farkash, H. M., & Othman, Z. A. (2025). Data-driven Insights for Informed Decision-Making: Applying LSTM Networks for Robust Electricity Forecasting in Libya. International Journal of Intelligent Systems and Applications (IJISA), 17(3), 65–89.

Al Feki, E., & Neji, J. (2024). Statistical modelling to assessing and enhancing road traffic safety in Tripoli, Libya: A systematic approach. Journal of Engineering Research, 12(4), 659-669.

Alrawayati, H., & Tökeşer, Ü. (2021). PARKINSON'S DISEASE DIAGNOSIS BASED ON THE CONVOLUTIONAL NEURAL NETWORK AND PARTICLE SWARM OPTIMIZATION ALGORITHM. *Asian Journal of Mathematics and Computer Research*, *28*(1), 26-37.

Al-Saffar, M., & Al-Shammari, M. (2022). Structural breaks and forecasting performance in resource-rich countries. Resources Policy, 76, 102567.

Alssager, M., & Othman, Z. A. (2016). Taguchi-based parameter setting of cuckoo search algorithm for capacitated vehicle routing problem. In *Advances in Machine Learning and Signal Processing: Proceedings of MALSIP 2015* (pp. 71-79). Cham: Springer International Publishing.

Ben Dalla, L., Medeni, T. M., Agila, A. A., & Medeni, İ. M. (2024). Architectural Synergy: Investigating the Role of Artificial Neural Networks in Enabling Deep Learning. The International Journal of Engineering & Information Technology (IJEIT), 12(1), 96-103.

Box, G. E. P., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. Holden-Day.

Chantar, H., Tubishat, M., Essgaer, M., & Mirjalili, S. (2021). Hybrid binary dragonfly algorithm with simulated annealing for feature selection. *SN computer science*, *2*(4), 295.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD, 785–794.

Coroneo, L., & Iacone, F. (2024). Testing for equal predictive accuracy with strong dependence. International Journal of Forecasting. (Accepted/In Press).

Dalla, L. O. B., Essgaer, M., Jetlawei, S. S., EL-sseid, M., Alsharif, A., & Agila, A. A. A. (2026). Local Precision and Global Harmony: A Comparative Literature Review (LR) Framework for Taylor and Fourier Series in Engineering Modeling. *Al-Farooq Journal of Sciences*, *2*(1), 275-304.

Dalla, L. O. B., Essgaer, M., Jetlawei, S. S., EL-sseid, M., Alsharif, A., & Agila, A. A. A. (2026). Local Precision and Global Harmony: A Comparative Literature Review (LR) Framework for Taylor and Fourier Series in Engineering Modeling. *Al-Farooq Journal of Sciences*, *2*(1), 275-304.

Dalla, L. O. B., Karal, Ö., & Degirmenciyi, A. (2025). Leveraging LSTM for adaptive intrusion detection in IoT networks: a case study on the RT-IoT2022 dataset implemented on CPU computer device machine. In *5th International Conference on Engineering, Natural and Social Sciences April* (pp. 15-16).

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263.

El-Sayed, M., & Hassan, T. (2023). Forecasting monetary aggregates in conflict-affected states: A case study of Libya and Syria. Middle East Development Journal, 15(1), 88–112.

Elsherif, M. (2024). Modelling Inflation Dynamics and Global Oil Price Shocks in OAPEC Countries: TVP-VAR. International Journal of Energy Economics and Policy, 14(3), 51–69.

Gonzalez, C., & Fernandez, R. (2024). XGBoost feature importance in financial time series: Interpretation and limitations. IEEE Transactions on Neural Networks and Learning Systems, 35(3), 2890–2902.

Tareq Alnnale. (2026). From Reactive to Proactive Governance: A Hybrid LSTM–Gradient Boosting Architecture for Real-Time Anomaly Signal Detection in Multi-Store Retail Supply Chain Decision Systems. *Al-Farooq Journal of Sciences*, *2*(1), 987-1005.

Harvey, D. I., Leybourne, S. J., & Zu, Y. (2025). Testing for equal average forecast accuracy in possibly unstable environments. Journal of Business & Economic Statistics, 43(3), 643–656.

Hawa Ahmed Alrawayati, Ümit Tokeşer. (2025).Spectral Integral Variation of Graph Theory. Asian Journal of Mathematics and Computer Research.32, Issue, 2. Pages(151-160). https://www.elibrary.ru/item.asp?id=82163806

Hawa Ahmed Alrawayati, Ümit Tokeşer. (2025).Spectral Integral Variation of Graph Theory. Asian Journal of Mathematics and Computer Research.32, Issue, 2. Pages(151-160). https://www.elibrary.ru/item.asp?id=82163806.

Hawa Alrawayati (2020). Development of High Efficiency Optimization Algorithm based on New Topology in Particle Swarm Optimization for Parkinson's Disease. IOSR Journal of Mathematics (IOSR-JM). 8

Hawa Alrawayati. (2013). (المؤثرات الخطية المحدودة على فضاء هلبرت) • Finite linear operators on Hilbert spaces.193-184.مجلة جامعة الزيتونة

Hawa Alrawayati. (2016). (المعادلة التكاملية ونواة المؤثر) Integral Equation and Kernel Operator.76 – 63 .مجلة الساتل - جامعة مصراته

Hawa Alrawayati. (2016). Integral Equation and Kernel Operator. Al-Satel Journal - Misrata University. 63-76

Henderi, H., & Sofiana, S. (2025). Comparative Study of Traditional and Modern Models in Time Series Forecasting for Inflation Prediction. International Journal of Applied Informatics and Management, 5(3), 155–167.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Jeguirim, K., & Ben Salem, L. (2024). Unveiling extreme dependencies between oil price shocks and inflation in Tunisia: Insights from a copula dcc garch approach. MPRA Paper No. 121616. University Library of Munich.

Karal, Ö. (2020). Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation. In *2020 innovations in intelligent systems and applications conference (ASYU)* (pp. 1-5). IEEE.
Karal, Ö. (2020). Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation. In 2020 innovations in intelligent systems and applications conference (ASYU) (pp. 1-5). IEEE.
Karal, Ö. (2024). Comparative performance analysis of epsilon-insensitive and pruningbased algorithms for sparse least squares support vector regression. Sigma Journal of Engineering and Natural Sciences, 42(2), 578-589.
Khanh, M. Q., & Tuan, T. A. (2024). ARIMA, XGBoost, LSTM — Which One is Better for Forecasting?. Journal of Advanced Research in Applied Sciences and Engineering Technology, 43(1), 101-115.
Kumar, S., & Haider, A. (2022). LSTM vs ARIMA: A comparative study of forecasting models for macroeconomic indicators in developing economies. International Journal of Forecasting, 38(4), 1421–1438.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLOS ONE, 13(3), e0194889.
Pinto, J. M., & Castle, J. L. (2022). Machine Learning Dynamic Switching Approach to Forecasting in the Presence of Structural Breaks. Journal of Business Cycle Research, 18(2), 129–157.
Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The performance of LSTM and BiLSTM in forecasting time series. IEEE Big Data, 3285–3292.
Stempień, D., & Ślepaczuk, R. (2025). Hybrid Models for Financial Forecasting: Combining Econometric, Machine Learning, and Deep Learning Models. arXiv preprint, arXiv:2505.19617.
Sun, X., Yuan, P., Yao, F., Qin, Z., Yang, S., & Wang, X. (2024). Financial Fragility in Emerging Markets: Examining the Innovative Applications of Machine Learning Design Methods. Journal of the Knowledge Economy. DOI: 10.1007/s13132-023-01731-w.
Tokgöz, N., Değirmenci, A., & Karal, Ö. (2024). Machine learning-based classification of turkish music for mood-driven selection. Journal of Advanced Research in Natural and Applied Sciences, 10(2), 312-328.
Wang, J., & Li, X. (2023). Hybrid ARIMA-LSTM models for financial time series: A meta-analysis. Applied Intelligence, 53(6), 7021–7041.
Zhang, Y., & Wang, L. (2023). A comparative study of XGBoost and LSTM for time series forecasting in volatile markets. Expert Systems with Applications, 213, 118876.

1. Appendices

Appendix A: Sample of the Dataset (First 12 months)

Date	MB	NFA	NDA
2004-01	4,266.9	26,593.4	-22,326.5
2004-02	4,189.0	27,506.0	-23,317.0
2004-03	4,233.0	28,096.9	-23,863.9
2004-04	4,418.5	28,254.8	-23,836.3
2004-05	4,618.9	28,894.7	-24,275.8
2004-06	4,753.2	29,158.0	-24,404.8
2004-07	4,870.8	30,148.0	-25,277.2
2004-08	4,635.6	29,966.0	-25,330.4
2004-09	4,582.8	31,024.2	-26,441.4
2004-10	4,848.6	30,944.0	-26,095.4
2004-11	4,933.3	30,944.0	-26,095.4
2004-12	5,089.3	32,009.6	-27,076.3

The dataset used in this study was compiled from publicly available monthly publications issued by the Central Bank of Libya. The complete dataset (268 rows) is available in the replication repository.

Appendix B: Rolling Window Cross-Validation (Sensitivity Check)

Due to the small sample size (204 training observations), a full rolling window cross-validation was not used as the primary evaluation method to avoid excessive test set shrinkage. However, as a sensitivity check, we applied a 3-fold rolling window scheme (training on 2004–2018, 2004–2019, 2004–2020; testing on 2019, 2020, 2021 respectively). The average MAPEs across folds were:

Model	Average MAPE (%)
ARIMAX	26.18
XGBoost	29.34
LSTM	51.02

The ranking remained unchanged (ARIMAX > XGBoost > LSTM), confirming the robustness of our main findings.

Appendix C: Reproducibility Statement

All experiments were implemented in Python 3.10 using pandas, numpy, matplotlib, scikit-learn, statsmodels, xgboost, and tensorflow/keras.

Anonymous GitHub Repository for Peer Review:

https://anonymous.4open.science/r/libya_monetary_forecast_67B3

Repository contents:

- data/: Full dataset (CSV), sample (first 12 months), and data dictionary
- code/: All Python scripts for preprocessing, model estimation, sensitivity analysis, and figure generation
- outputs/: Tables (MAE, RMSE, MAPE, Diebold–Mariano results) and figures (PNG, 300 dpi)
- requirements.txt: Python dependencies
- README.md: Step-by-step replication instructions

Hardware: Google Colab (CPU) – Intel Xeon @ 2.20GHz, 12.7 GB RAM. The code can also run on any standard laptop with Python 3.10 and the required libraries.