

Evaluating Machine Learning and Econometric Models for Inflation Forecasting: Evidence from a Sub-Saharan African Panel

Baffour Osei

Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development

Abigail Boatemaa

University of Bradford

Isaac Nyame

Kwame Nkrumah University of Science and Technology

Gabriel Osei Forkuo

`gabriel.forkuo@unitbv.ro`

Transylvania University of Braşov

Research Article

Keywords: Consumer price index, Diebold-Mariano test, Extreme gradient boosting, Long short-term memory, Predictive accuracy, Random-walk benchmark

Posted Date: July 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10355831/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Future Business Journal on October 23rd, 2026. See the published version at <https://doi.org/10.1186/s43093-026-01003-4>.

Abstract

Policymakers in Sub-Saharan Africa increasingly face decisions regarding the adoption of machine learning methods for inflation forecasting, yet these approaches are rarely benchmarked against simple persistence models. This study constructs a reproducible, annual panel of macroeconomic indicators for eight Sub-Saharan African economies (1972–2024, $N = 424$) to compare eight forecasting approaches: a naive random-walk benchmark, two linear econometric models (pooled OLS and country fixed effects), three machine learning models (Elastic Net, Random Forest, and XGBoost), and a recurrent neural network (LSTM) evaluated with and without early stopping. Using a chronological train/validation/test split and pairwise Diebold-Mariano tests, we evaluate out-of-sample predictive accuracy. The empirical results indicate that no estimated model statistically outperforms the naive persistence baseline. While Random Forest, XGBoost, Elastic Net, and the early-stopped LSTM achieve statistical parity with the random walk, traditional linear models and the uncorrected LSTM are significantly dominated by it. Although implementing early stopping substantially improves LSTM performance, it does not yield a competitive advantage over simpler tree-based ensembles. These findings align with the well-documented random-walk puzzle in global inflation forecasting, suggesting that structural breaks, high volatility, and modest sample sizes pose substantial challenges for highly parameterized forecasting architectures. Consequently, these results suggest that regional central banks and finance ministries should maintain simple persistence models as active baselines when evaluating investments in complex forecasting infrastructure. These conclusions are robust to a battery of additional checks: HAC-Newey-West/Harvey-Leybourne-Newbold-corrected and Holm-Bonferroni-adjusted Diebold-Mariano tests, a strictly lagged-only predictor respecification, expanding-window recursive re-estimation, per-country and subperiod decomposition, and LSTM input-sequence-length sensitivity analysis.

JEL Classification: E31, E37, C53, C45, C33, O55

1. Introduction

Accurate inflation forecasting is central to monetary policy design, fiscal planning, and household welfare, and the stakes are arguably higher in Sub-Saharan Africa (SSA) than in advanced economies: inflation in the region has been more volatile, more exposed to exchange-rate pass-through and commodity-price shocks, and more consequential for poverty given the high budget share that food and energy occupy in typical household consumption baskets. After peaking above 9% in 2022, median SSA inflation has since moderated, but remains in double digits in a meaningful share of the region's economies, and central banks continue to rely heavily on forecasts to calibrate policy rates under considerable data and capacity constraints (International Monetary Fund, 2025; World Bank, 2025).

Two broad forecasting traditions compete for this role. The first, econometric tradition, models inflation through structural or reduced-form relationships — Phillips-curve specifications, autoregressive distributed lag models, vector autoregressions, and panel fixed-effects estimators — grounded in macroeconomic theory and typically favored for their interpretability and role in policy communication (Alagidede et al., 2012; Baltagi, 2021). The second, increasingly popular tradition applies machine learning (ML) and deep learning (DL) methods — regularized regression, tree-based ensembles, and recurrent neural networks — that relax linearity and functional-form assumptions in exchange for reduced interpretability and, frequently, a higher risk of overfitting in short macroeconomic samples (Medeiros et al., 2021; Varian, 2014).

A large and often under-appreciated literature, however, shows that both traditions can struggle to outperform a third, much simpler alternative: a naive persistence or random-walk forecast. Atkeson and Ohanian (2001) first demonstrated this for U.S. inflation; Stock and Watson (2007) documented the same pattern as inflation became both easier and harder to forecast across different periods; and Duncan and Martínez-García (2018), studying a 14-country emerging-market panel, showed that a random-walk-type benchmark (RW-AO) was difficult to beat even by Phillips-curve, factor-augmented, and Bayesian VAR alternatives. Despite the growing enthusiasm for ML-based inflation forecasting in emerging markets (Mirza, Rizvi, Naqvi, & Umar, 2024; Naghi, 2024), this naive benchmark is rarely included as a competitor in SSA-specific studies, and cross-country panel designs — as opposed to single-country time series limited to roughly fifty annual observations — remain uncommon in the regional literature.

Interest in machine learning (ML)- and deep learning (DL)-based inflation forecasting has accelerated further in the past two years, though largely outside the SSA context. Liu et al. (2024), in an IMF working paper, apply penalized regression and other ML models to near-term core inflation in post-pandemic Japan and find that LASSO systematically outperforms conventional benchmarks using household expectations, tourism, and exchange-rate predictors. Nortey, Agyemang, Sakyi-Yeboah, Ampomah, and Agyekum (2025) compare transformer and gradient-boosting architectures against ARIMA and exponential smoothing for U.S. monthly inflation and report that deep learning architectures uniformly dominate traditional statistical benchmarks. Bolivar (2026) develops a two-step ML methodology specifically for high-frequency inflation nowcasting in developing economies, where, as in the SSA context considered here, official statistics are released only monthly and with delay. A recent systematic review by Slimani and Chourabi (2026), synthesizing this fast-growing literature across forecasting horizons, confirms that reported ML gains are concentrated in data-rich, high-frequency, advanced-economy applications, and notes that naive or autoregressive benchmarks are inconsistently reported across the reviewed studies — precisely the gap this paper's inclusion of the RW-AO benchmark is designed to close for a data-scarcer, annual, developing-economy panel.

This paper addresses both gaps. We construct a reproducible, application programming interface (API)-sourced panel of macroeconomic indicators for Sub-Saharan African economies and benchmark eight forecasting approaches spanning all three traditions — naive persistence, econometric, and machine learning/deep learning — against one another using a strict out-of-sample chronological evaluation and formal Diebold-Mariano tests of predictive accuracy, rather than point-estimate RMSE/MAE comparisons alone. Our central contribution is empirical: once the naive benchmark is included, the paper's conclusion changes from a simple horse-race in which tree-based ensembles *win*, to the more rigorous and literature-consistent finding that *no* model in the comparison set — econometric or machine-learning — statistically outperforms persistence, with tree ensembles achieving parity but not superiority. We discuss why this pattern is plausible given the panel's size and the structural breaks spanning the test window, and what it implies for SSA policymakers evaluating whether to invest in more complex forecasting infrastructure.

Aim, Objectives and Research Questions

This study aimed to rigorously compare econometric, machine learning, and deep learning approaches to inflation forecasting across a panel of Sub-Saharan African (SSA) economies, explicitly benchmarking each approach against a naive persistence forecast to determine whether increased model complexity yields statistically defensible gains in a data-constrained regional setting.

The objectives were:

- To construct a reproducible, API-sourced annual macroeconomic panel for eight SSA economies, with a fully auditable data-extraction and preprocessing pipeline.
- To implement and train a naive persistence benchmark, two linear econometric specifications (pooled OLS and country fixed effects), three pooled machine learning models (Elastic Net, Random Forest, XGBoost), and a panel LSTM network, all evaluated under an identical chronological train/validation/test split.
- To formally test whether observed differences in forecast accuracy across models are statistically significant, using pairwise Diebold-Mariano tests rather than point-estimate error metrics alone.
- To characterize the computational cost and architectural complexity of each model class alongside its forecasting performance, and to draw practical implications for SSA central banks and finance ministries evaluating investment in forecasting infrastructure.

The aim and objectives were operationalised through the following research questions:

RQ1: Do machine learning and deep learning models achieve statistically significant improvements in out-of-sample inflation forecast accuracy over conventional linear econometric approaches in an SSA panel?

RQ2: Can any of the econometric, machine learning, or deep learning models evaluated significantly outperform a naive random-walk persistence benchmark?

RQ3: How does early stopping affect the out-of-sample performance of a panel LSTM trained on a short annual macroeconomic panel, and does it change its standing relative to simpler benchmarks?

RQ4: What do the relative accuracy and computational cost of these approaches imply for the design of operational inflation-forecasting systems at SSA central banks and finance ministries?

The remainder of the paper is organized as follows. Section 2 reviews the econometric and machine-learning literatures on inflation forecasting, with particular attention to the random-walk benchmark puzzle and to the SSA context specifically. Section 3 describes the panel construction, preprocessing pipeline, and the full set of econometric, ML, and DL models compared, including the Diebold-Mariano testing framework. Section 4 presents results. Section 5 discusses the findings in light of the existing literature and their policy implications, discloses data coverage limitations in the current panel, and recommends future work. Section 6 concludes the study.

2. Literature Review

2.1 Econometric Traditions in Inflation Forecasting

The dominant econometric framework for inflation forecasting descends from the Phillips curve, relating inflation to a measure of economic slack and expectations. Alagidede et al. (2012) apply this framework to SSA monetary policy, finding that the inflation-output relationship is regime-dependent and sensitive to exchange-rate and commodity-price shocks — a finding echoed in more recent panel-VAR work on exchange rates and price stability across African economies. Panel data methods (Baltagi, 2021) extend single-country time series by pooling structurally similar economies, increasing effective sample size and allowing country fixed effects to absorb time-invariant heterogeneity, at the cost of assuming common slope coefficients across the panel. Bergmeir et al. (2018) caution that standard cross-validation is invalid for autoregressive time series, motivating the chronological (walk-forward) evaluation design used throughout this paper.

2.2 The Random-Walk Benchmark Puzzle

Atkeson and Ohanian (2001) is the foundational reference for what has become known as the random-walk (or RW-AO) puzzle in inflation forecasting: a naive model using only lagged realized inflation frequently outperforms structural Phillips-curve models in out-of-sample tests. Stock and Watson (2007) extend this to show that the relative forecastability of U.S. inflation has shifted over time, with the random walk becoming a harder benchmark to beat during periods of low, stable inflation. Faust and Wright (2013), in their Handbook of Economic Forecasting survey, confirm that this pattern generalizes across a wide range of proposed alternatives. Critically for the present study, Duncan and Martínez-García (2018) show the puzzle also holds outside advanced economies: across 14 emerging markets, a random-walk-type specification (RW-AO) — a simple moving average of recent realized inflation — was difficult to outperform even by factor-augmented autoregressions and Bayesian VARs, and the model's relative advantage was associated with historically low central-bank credibility. This paper is, to our knowledge, among the first to test whether the same puzzle extends to modern tree-based ensembles and deep recurrent architectures in an SSA-specific panel.

2.3 Machine Learning and Deep Learning Approaches

Medeiros et al. (2021) provide the benchmark result motivating much of the recent ML-forecasting literature, showing that Random Forest outperforms standard U.S. inflation-forecasting benchmarks in a data-rich environment with hundreds of predictors. Naghi (2024) replicates and substantially extends this study, broadening the set of ML methods considered, and finds that the gains from ML over strong autoregressive benchmarks are real but more modest — and more sample-dependent — than the original result suggested. In emerging-market settings, Mirza, Rizvi, Naqvi, and Umar (2024) integrate foreign-exchange reserve data into ML models for inflation prediction and report improved accuracy relative to standard econometric baselines, while recurrent neural network architectures, including LSTMs, have shown mixed evidence: Almosova and Fasanya (2023) find that nonlinear recurrent models can improve on autoregressive and naive benchmarks for well-behaved, data-rich monthly series, a very different data regime from the short annual SSA panel considered here. Across this literature, the recurring caveat is that

ML gains are highly conditional on sample size, data richness, and the stability of the underlying data-generating process — conditions our SSA panel, with 424 annual country-year observations, only partially satisfies.

2.4 Inflation Forecasting in Sub-Saharan Africa

SSA-specific inflation forecasting research remains comparatively sparse and has historically relied on single-country ARIMA or Phillips-curve specifications limited by short annual series (Alagidede et al., 2012). Regional surveillance reports from the IMF and World Bank track inflation dynamics extensively but for descriptive and near-term forecasting purposes rather than formal model comparison (IMF Regional Economic Outlook for Sub-Saharan Africa, 2025). Recent panel-VAR work has examined the joint dynamics of exchange rates, price stability, and financial flows across African economies, underscoring the centrality of exchange-rate pass-through — a channel our feature-importance results (Section 4.6) also identify as a dominant predictor. To date, however, we are not aware of prior SSA-focused work that (a) constructs a fully reproducible, API-sourced macroeconomic panel explicitly designed to avoid cross-sectional data leakage, and (b) benchmarks econometric, machine learning, and deep learning models against a formal naive persistence baseline using Diebold-Mariano significance testing rather than point-estimate comparison alone.

2.5 Research Gap and Contribution

Three gaps motivate this study. First, the SSA inflation-forecasting literature has not systematically incorporated the random-walk benchmark that is standard practice in the advanced- and broader emerging-market literature, risking overstated claims about the value added by more complex models. Second, existing regional comparisons rarely extend beyond classical econometric methods to include modern tree-based ensembles and recurrent deep learning architectures within the same, consistently evaluated framework. Third, few studies in this space provide a fully reproducible, auditable data pipeline — from raw API extraction through preprocessing to model training — a transparency standard increasingly expected in applied macroeconometric work. This paper addresses all three gaps directly.

3. Methodology

This section outlines the empirical framework designed to forecast inflation across peer Sub-Saharan African (SSA) economies. To address concerns regarding data traceability, reproducibility, and the sample-size limitations of single-country annual time series, we implement an automated, API-driven panel data architecture. The methodology proceeds from study design and data mining protocols, to preprocessing, to the mathematical model architectures — econometric, machine learning, and deep learning — and the evaluation framework, including formal tests of forecast accuracy.

3.1 Study Design and Panel Configuration

Macroeconomic forecasting in developing economies, particularly in Sub-Saharan Africa, is frequently constrained by short temporal windows ($N \approx 50$ annual observations per country since independence or stable data tracking). Training high-capacity machine learning and deep learning models on such small samples risks severe overfitting and parameter instability (Varian, 2014). To resolve this limitation, this study adopts a cross-country panel data design (Baltagi, 2021), pooling annual observations from eight peer SSA economies: Ghana, Nigeria, Kenya, Côte d'Ivoire, South Africa, Senegal, Tanzania, and Uganda. Pooling these structurally comparable, resource-dependent, or emerging SSA economies over 1972–2024 (53 usable years per country after two-year lag construction) yields a final panel of $N = 424$ country-year observations, preserving annual macroeconomic frequency while generating sufficient volume to train regularized linear models, tree-based ensembles, and recurrent neural networks.

Section 5.3 documents two residual data-construction caveats — cross-country imputation for two indicator-country series with zero native WDI coverage, and partial early-period gaps for Uganda — that survive the corrected extraction protocol used to build this panel.

3.2 Data Mining, API Integration, and Variable Selection

To establish reproducibility and eliminate manual data handling, all macroeconomic indicators are mined programmatically via direct API integration with the World Bank World Development Indicators (WDI) and the Federal Reserve Bank of St. Louis (FRED).

The target variable is CPI inflation; two alternative inflation measures, seven domestic macroeconomic predictors, and one global commodity-price driver are also extracted. Table 1 lists all indicators, their API identifiers, and their economic justification.

Table 1
Variable Definitions, Sources, and Economic Justification

Variable	Description	Source / API ID	Economic Justification
INFL_CPI (Y)	CPI Inflation (annual %)	World Bank FP.CPI.TOTL.ZG	Core target variable; consumer purchasing power.
INFL_DEFL	GDP Deflator Inflation (annual %)	World Bank NY.GDP.DEFL.KD.ZG	Alternative target/predictor; broader domestic price pressure.
CPI_LEVEL	Consumer Price Index (level)	World Bank FP.CPI.TOTL	Price-level baseline controlling for long-run nominal trends.
GDPG	Real GDP Growth (annual %)	World Bank NY.GDP.MKTP.KD.ZG	Output gap / demand-pull pressure (Phillips Curve).
EXR	Official Exchange Rate (LCU/USD)	World Bank PA.NUS.FCRF	Imported inflation and exchange-rate pass-through.
MS	Broad Money Supply (% of GDP)	World Bank FM.LBL.MQMY.GD.ZS	Monetary expansion / liquidity-driven inflation.
GE	Government Final Expenditure (% GDP)	World Bank NE.CON.GOVT.ZS	Fiscal policy stance and government-driven demand.
PDS	Domestic Credit to Private Sector (% GDP)	World Bank FS.AST.PRVT.GD.ZS	Financial depth, private demand, credit channels.
SPD	Interest Rate Spread (%)	World Bank FR.INR.LNDP	Monetary policy transmission, financial-sector margins.
OILP	West Texas Intermediate Crude (USD/bbl)	FRED WTISPLC (annual mean)	Global supply-side shock indicator.

3.3 Data Preprocessing and Feature Engineering

Panel time series require specialized preprocessing to prevent cross-sectional data leakage, in which the chronological boundary of one country bleeds into another during lagging, imputation, or scaling. We apply country-specific forward-fill/backward-fill imputation grouped by country, generate first- and second-order lags (t-1, t-2) strictly partitioned within country groups to capture inflation inertia, and standardize all features to zero mean and unit variance using statistics computed exclusively from the training partition to prevent test-set information leakage. To rule out temporal leakage in the imputation step itself, we clarify its exact ordering relative to the rest of the pipeline: forward-fill/backward-fill is applied first, strictly within each country's own series and before first- and second-order lags are constructed and before the chronological train/validation/test partition is drawn, so that no observation from one country ever informs the imputed values of another. Forward-fill uses only within-country values observed at or before the missing time point and therefore introduces no look-ahead. Backward-fill is invoked only to seed a small number of leading (pre-sample) gaps at the very start of a country-indicator series, and by construction draws on that same country's earliest available later observation. This is a narrow, country-local form of look-ahead confined to a handful of early-period cells rather than a cross-sectional or train/test leakage channel, and we flag it explicitly as a residual limitation in Section 5.3.

We clarify explicitly the timing of predictors relative to the forecast origin: the exogenous predictor set entering the pooled ML/DL and linear specifications (Table 1) includes both the same-year (contemporaneous) value of GDPG, EXR, MS, GE, PDS, SPD, and OILP and their first- and second-order lags, in addition to two lags of the target series itself (INFL_CPI_lag1, INFL_CPI_lag2). Because same-year values of several of these series — notably the exchange rate and government expenditure — are not fully observable at the true one-step-ahead forecast origin, Section 3.8 reports a strictly lagged-only respecification of every model as an explicit robustness check on ex-ante forecast credibility.

3.4 Model Selection and Mathematical Architectures

We benchmark four paradigms: a naive persistence baseline, regularized and fixed-effects linear econometric models, tree-based machine learning ensembles, and a deep recurrent neural network.

3.4.1 Elastic Net Regression

The Elastic Net (Zou & Hastie, 2005) combines L1 (LASSO) and L2 (Ridge) penalties to resolve collinearity among macroeconomic predictors:

$$\min_{\beta} (1/2N) \sum (y_{i,t} - x_{i,t}'\beta)^2 + \lambda[\alpha\|\beta\|_1 + ((1 - \alpha)/2)\|\beta\|_2^2] \text{ (Eq. 1)}$$

Here, $y_{i,t}$ is realized CPI inflation for country i in year t ; $x_{i,t}$ is the corresponding vector of standardized predictors; β is the coefficient vector to be estimated; N is the number of training observations; $\|\beta\|_1$ and $\|\beta\|_2^2$ are the L1 and squared L2 norms of β ; λ controls the overall strength of regularization; and α (between 0 and 1) sets the mix between the two penalties. We set $\alpha = 0.5$, balancing automatic feature selection (from the L1 term) against coefficient shrinkage and multicollinearity control (from the L2 term).

3.4.2 Tree-Based Ensembles: Random Forest and XGBoost

Random Forest (Breiman, 2001) is an averaging ensemble of B bootstrap-sampled decision trees:

$$\hat{y}_{i,t} = (1/B) \sum_{b=1 \text{ to } B} T_b(x_{i,t}) \text{ (Eq. 2)}$$

Here, $\hat{y}_{i,t}$ is the predicted inflation rate for country i in year t ; $T_b(\cdot)$ is the prediction of the b -th regression tree, each trained on an independent bootstrap sample of the training data with a randomized subset of predictors considered at each split; and B is the total number of trees in the forest. Averaging across trees reduces the variance of any single tree's prediction.

XGBoost (Chen & Guestrin, 2016) instead builds trees additively via gradient boosting, at each boosting round t minimizing a regularized objective:

$$L^A(t) = \sum_j l(y_j, \hat{y}_j^{A(t-1)} + f_t(x_j)) + \Omega(f_t) \text{ (Eq. 3)}$$

Here, $l(\cdot)$ is the loss function (squared error) comparing the true value y_j to the prediction; $\hat{y}_j^{A(t-1)}$ is the model's cumulative prediction for observation j after $t - 1$ rounds; $f_t(x_j)$ is the new tree added at round t ; and $\Omega(f_t)$ is a complexity penalty defined as:

$$\Omega(f_t) = \gamma T + (1/2)\lambda \sum_k w_k^2$$

where T is the number of leaves in the new tree, w_k is the predicted value (weight) assigned to leaf k , and γ and λ are penalty coefficients discouraging overly complex trees and extreme leaf weights, respectively.

3.4.3 Deep Learning: Panel LSTM

The recurrent architecture (Hochreiter & Schmidhuber, 1997) processes sequence tensors $[B, T, F]$, where B is the pooled batch, $T = 3$ years of consecutive macroeconomic states, and F is the standardized feature space. Gate dynamics are governed by the standard LSTM cell transition (forget, input, candidate, cell-state update, output, hidden-state equations), with the final hidden state mapped through a dense feedforward head to produce the scalar inflation forecast.

3.4.4 Model Complexity and Computational Cost

Table 2 summarizes the architecture, key hyperparameters, parameter/unit counts, approximate inference cost, and measured wall-clock training time and single-observation inference latency for every model in the comparison set, together with the primary methodological reference for each. Training time and inference latency were measured directly on the training partition ($n = 304$) in the same Python/PyTorch/scikit-learn/XGBoost environment used to produce the results in Section 4, running single-threaded on CPU; they are intended to characterize the relative, order-of-magnitude computational burden of each approach in this application rather than to serve as a production deployment benchmark, since actual latency will vary with hardware, batching, and library version. FLOPs (floating-point multiply-add operations) are reported per single-observation forward pass and are

analytic approximations based on each model's structure (e.g., B trees $\times$ maximum depth D comparisons for tree ensembles; 4 gate matrix multiplications for the LSTM cell), not profiler-measured counts.

Table 2
Model Architecture, Hyperparameters, and Computational Cost

Model	Architecture / Algorithm	Key Hyperparameters	Params / Units	Approx. FLOPs (per inference)	Training Time (measured, n = 304)	Inference Latency (measured, per observation)	Reference
Naive (RW-AO)	Zero-parameter persistence rule; no estimation	None	0	~ 1	< 0.001 s (lookup only)	< 0.001 ms	Atkeson & Ohanian (2001)
Pooled OLS	Linear regression, closed-form least squares	None (no regularization)	26 (25 coef. + intercept)	~ 50	0.004 s	0.023 ms	Standard OLS; this study
Panel Fixed Effects	Linear within-estimator with entity (country) dummies, clustered SE	Entity effects = 8 countries	34 (26 coef. + 8 entity effects)	~ 50	0.052 s	~ 0.03 ms (est.)	Baltagi (2021)
Elastic Net	L1/L2-regularized linear regression, coordinate descent	α (L1 ratio) = 0.5, $\lambda = 0.2$	26 (25 coefficient + intercept)	~ 50	0.003 s	0.050 ms	Zou & Hastie (2005)
Random Forest	Bagged ensemble of B independent decision trees	B = 150 trees, max depth = 6	150 trees	~ 900 (150×6 node comparisons)	0.332 s	11.575 ms	Breiman (2001)
XGBoost	Additive gradient-boosted tree ensemble	B = 100 rounds, max depth = 4, lr = 0.04	100 trees	~ 400 (100×4 node comparisons)	0.122 s	0.268 ms	Chen & Guestrin (2016)
Panel LSTM	1-layer LSTM (hidden = 24) + 2-layer dense head, 3-year input sequence	Hidden = 24, layers = 1, lr = 0.003, weight decay = $1e-4$, dropout = 0.1	$\approx 5,217$ weights	$\sim 10,200$ (4 gates $\times$ input/hidden matmuls, per timestep)	3.177 s (126 epochs, best-val)	0.238 ms	Hochreiter & Schmidhuber (1997); Kingma & Ba (2014)
<p><i>Note:</i> "Params / Units" reports linear coefficients for regression-type models, entity effects for the fixed-effects model, tree count for ensembles (since individual node/leaf counts vary by tree and are not a standard complexity measure for these algorithms), and total learnable weights for the LSTM. FLOPs are order-of-magnitude analytic estimates for a single forward pass on one observation, not measured via a hardware profiler. All models are inexpensive by modern deep-learning standards; the entire model suite (excluding hyperparameter search) trains in under 4 seconds combined on a standard CPU, reflecting the panel's modest size (n = 304 training observations).</p>							

Hyperparameters reported in Table 2 were not fixed without validation-based justification: for Elastic Net (λ , the regularization-strength parameter, and α , the L1/L2 mix ratio, following the notation of Eq. 1 — note that these correspond to scikit-learn's α and $l1_ratio$ arguments, respectively, which use reversed terminology), Random Forest (number of trees, maximum depth), and

XGBoost (number of rounds, maximum depth, learning rate), we ran a grid search minimizing validation-set (2010–2017) RMSE (full grid and results in Section 3.8 and the supplementary hyperparameter log). The validation-optimal Random Forest ($n_estimators = 150$, $max_depth = 4$; validation RMSE 3.722) and XGBoost ($n_estimators = 200$, $max_depth = 3$, learning rate = 0.02; validation RMSE 3.520) configurations were close to the values reported in Table 2 (validation RMSE 3.809 and 3.596, respectively), while a more heavily regularized Elastic Net ($\lambda = 0.5$, $\alpha = 0.9$; validation RMSE 4.141) modestly outperformed the reported default ($\lambda = 0.2$, $\alpha = 0.5$; validation RMSE 4.897), consistent with the $\alpha = 0.5$, $\lambda = 0.2$ configuration reported in Table 2. Applying this validation-optimal Elastic Net configuration to the test set improves Test RMSE from 5.986 to 5.600 without changing its standing relative to Random Forest or the naive benchmark (Section 4.7).

3.5 Training Pipeline and Chronological Evaluation

Standard K-fold cross-validation is invalid for time-series forecasting because random splitting breaks temporal dependence and introduces forward-looking bias (Bergmeir et al., 2018). We therefore apply a strict chronological panel split across all cross-sections: Training (≤ 2009), Validation (2010–2017), and Test (2018–2024). For the LSTM, sequences of length $T = 3$ are constructed independently within each country's chronological boundaries before pooling into a single training tensor, ensuring sequences never bridge across countries. The deep learning model is optimized with MSE loss, the Adam optimizer (Kingma & Ba, 2014), a learning rate of 0.003, and L2 weight decay (1×10^{-4}).

3.6 Econometric Benchmark Models

To make the comparison between machine learning and econometrics empirically meaningful, we add three benchmarks estimated on the identical panel, standardized feature set, and chronological split as the ML/DL models.

3.6.1 Naive Random-Walk Benchmark (Atkeson-Ohanian)

Following Atkeson and Ohanian (2001), we implement a country-specific one-step-ahead persistence forecast, $\hat{Y}_{i,t} = Y_{i,t-1}$, requiring no estimation and using only the previously observed inflation rate of country i .

3.6.2 Pooled OLS

Pooled OLS is estimated on the standardized feature matrix, $Y_{i,t} = \alpha + X_{i,t}'\beta + \varepsilon_{i,t}$, with heteroskedasticity-robust standard errors, sharing the same predictor set as the Elastic Net but without a regularization penalty.

3.6.3 Panel Fixed-Effects (Within) Estimator

To account for time-invariant country heterogeneity, we estimate a one-way country fixed-effects model (Baltagi, 2021), $Y_{i,t} = \alpha_i + X_{i,t}'\beta + \varepsilon_{i,t}$, with standard errors clustered by country — the canonical panel-econometric counterpart to the pooled ML estimators.

3.7 Forecast Accuracy Testing: The Diebold-Mariano Test

RMSE and MAE rankings alone do not establish statistical superiority. We apply the Diebold and Mariano (1995) test to every pair of models, using loss differential $d_t = |e_{1,t}|^2 - |e_{2,t}|^2$, with test statistic $DM = \bar{d} / \sqrt{\text{Var}(\bar{d})} \sim t_{\{(n-1)\}}$ under the null of equal predictive accuracy. All pairwise tests use the common set of test observations (2018–2024, $n = 56$, that is, all eight countries for all seven test years) for which every model, including the sequence-based LSTM, produces a forecast. We note that this pooled loss-differential series treats the eight countries as exchangeable and does not explicitly model cross-sectional dependence: common shocks in a given test year (e.g., a global commodity-price or COVID-19 shock affecting several economies simultaneously) could induce correlation across countries' loss differentials within a year that neither the simple variance estimator nor the HAC-HLN correction in Section 3.8, which target serial rather than cross-sectional correlation, is designed to capture. We discuss this explicitly as a limitation of the current inference procedure in Section 5.3.

3.8 Robustness Checks

To address concerns about the small-sample and cross-sectionally pooled nature of the Diebold-Mariano tests reported above, we implement two refinements to the significance testing itself. First, in place of the simple sample-variance estimator used in Table 4, we recompute the loss-differential variance with a Newey-West heteroskedasticity- and autocorrelation-consistent (HAC)

long-run variance estimator (Bartlett kernel, automatic bandwidth following the Andrews (1991) rule of thumb), combined with the Harvey, Leybourne, and Newbold (1997) small-sample correction, $DM_{HLN} = DM \times \sqrt{[(n + 1 - 2h + h(h - 1)/n) / n]}$, evaluated against a Student-t reference distribution with $n - 1$ degrees of freedom. Second, because the 21 unique pairwise comparisons among the seven core models inflate the family-wise Type I error rate, we additionally report Holm-Bonferroni step-down-adjusted p-values across the full comparison family.

We further implement four complementary robustness checks, reported in Section 4.7: (i) a strictly lagged-only feature respecification that removes all same-year exogenous predictors identified in Section 3.3; (ii) an expanding-window (recursive) re-estimation scheme in which every model is retrained on all data available through year $t - 1$ and used to forecast year t , for $t = 2018, \dots, 2024$, more closely mirroring how central banks operationally re-estimate models as new data arrive, rather than fitting once on a single static training split; (iii) a per-country and pre-/post-2020 subperiod decomposition of test-set RMSE, to assess whether the pooled panel result masks country- or regime-specific heterogeneity; and (iv) an LSTM input-sequence-length sensitivity analysis ($T = 2, 3, 4, 5$ years) verifying that the early-stopping result in Section 4.4 is not an artifact of the three-year lookback window chosen in Section 3.4.3.

4. Results and Empirical Discussion

All eight models — three econometric/naive benchmarks, three pooled machine learning models, and two Panel LSTM variants — are estimated on an identical three-way chronological split: Train ≤ 2009 ($n = 304$), Validation 2010–2017 ($n = 64$), Test 2018–2024 ($n = 56$), using the complete eight-country panel (Section 5.3 discusses residual coverage caveats). Table 3 shows the multi-split panel evaluation metrics for all the models.

Table 3
Multi-Split Panel Evaluation Metrics

Model	Train RMSE	Train MAE	Val RMSE	Val MAE	Test RMSE	Test MAE
Naive (Random Walk, RW-AO)	14.510	7.255	3.815	2.652	4.435	2.510
Pooled OLS	11.518	7.122	8.231	6.772	6.781	5.341
Panel Fixed Effects	11.018	6.758	8.584	7.197	6.612	5.246
Elastic Net	11.934	7.017	4.897	3.517	5.986	4.583
Random Forest	5.826	3.590	3.809	3.239	4.849	3.399
XGBoost	3.871	2.722	3.596	2.892	5.825	3.895
Panel LSTM (early stopping)	12.331	6.314	3.146	2.326	6.162	3.777
Panel LSTM (350 epochs, no early stop)	2.145	1.326	4.854	3.741	7.741	5.246
Note: RW-AO = Atkeson-Ohanian random-walk persistence benchmark. All RMSE/MAE are on the original CPI-inflation percentage-point scale.						

4.1 The Naive Benchmark Is Statistically Undefeated

The central empirical finding, now confirmed on the complete eight-country panel, is that no model in the comparison set achieves a statistically significant improvement in test-set RMSE over the naive random-walk benchmark. RW-AO achieves the lowest point-estimate Test RMSE (4.435) and Test MAE (2.510) of any model evaluated. Random Forest (4.849), XGBoost (5.825), Elastic Net (5.986), and the early-stopped Panel LSTM (6.162) are all statistically indistinguishable from persistence (Diebold-Mariano $p = 0.538, 0.130, 0.077$, and 0.202 respectively; Table 4), while Pooled OLS ($p = 0.048$), Panel Fixed Effects ($p = 0.015$), and the uncorrected 350-epoch LSTM ($p = 0.045$) are significantly dominated by persistence at the 5% level. This directly extends the Atkeson-Ohanian random-walk puzzle (Section 2.2) to a complete eight-country Sub-Saharan African panel and to a model class — tree ensembles and recurrent networks — considerably richer than the Phillips-curve and Bayesian VAR alternatives tested in prior emerging-market work (Duncan & Martínez-García, 2018).

4.2 Tree-Based Ensembles Tie, Rather Than Beat, the Naive Benchmark

Random Forest remains the best-performing estimated model and narrows the gap to the naive benchmark to a statistically insignificant margin ($p = 0.538$, the largest p -value of any model comparison against RW-AO). Notably, Random Forest also significantly outperforms XGBoost on this larger panel (DM $p = 0.027$) — a reversal of the closer race observed in the earlier six-country interim analysis — suggesting XGBoost's additional flexibility is not rewarded once more countries and years are available. Neither tree ensemble, however, achieves a statistically significant edge over persistence itself. However, under a more conservative procedure involving a comparison with an HAC-HLN-corrected and Holm-Bonferroni-adjusted Diebold-Mariano test, the Random Forest–XGBoost gap reported here does not survive multiple-comparison correction, and should be read as suggestive rather than confirmed.

4.3 Econometric Benchmarks and the Uncorrected LSTM Underperform Persistence

Pooled OLS (Test RMSE 6.781) and Panel Fixed Effects (Test RMSE 6.612) are both significantly dominated by the naive benchmark ($p = 0.048$, $p = 0.015$), as shown in Fig. 1. As in the interim analysis, the fixed-effects estimator does not clearly outperform pooled OLS out-of-sample despite explicitly modeling country heterogeneity, consistent with Nickell-type finite-sample bias and reduced within-country identifying variation once the panel is split chronologically.

However, under a HAC-HLN-corrected and Holm-adjusted Diebold-Mariano test, Pooled OLS's apparent underperformance relative to persistence reported above is no longer statistically significant, though the fixed-effects specification remains significantly dominated by persistence even under this stricter test — indeed it is the single comparison in the entire study that survives Holm correction.

4.4 Diagnosing and Correcting LSTM Overfitting

An initial 350-epoch training run without early stopping again shows classic overfitting: validation loss reaches its minimum well before epoch 350 and then drifts upward while training loss continues to fall (Fig. 2a). Implementing early stopping on validation loss (patience = 30 epochs) with restoration of best-validation-epoch weights (epoch 126; Fig. 2b) improves Test RMSE from 7.741 to 6.162 — a 20.4% reduction — and, critically, moves the LSTM from being significantly dominated by the naive benchmark ($p = 0.045$) to being statistically indistinguishable from it ($p = 0.202$). The corrected LSTM remains, however, statistically indistinguishable from — not superior to — the tree ensembles (DM vs. XGBoost: $p = 0.596$).

4.5 Statistical Significance of Pairwise Forecast Comparisons

Table 4 and Fig. 3 report Diebold-Mariano test p -values for every pairwise comparison among the eight models, computed on the 56 common test-set observations (2018–2024, all eight countries) for which every model produces a forecast. Each cell gives the p -value for the null hypothesis that the row and column model have equal predictive accuracy; values below 0.05 indicate a statistically significant difference in squared-error loss between the two models at conventional significance levels. Figure 3 presents the same information as a heatmap, with green cells marking statistically distinguishable pairs and red cells marking pairs that are statistically indistinguishable.

Table 4
Diebold-Mariano Test p-values (pairwise, test-set forecast errors, n = 56)

Model	Naive (RW-AO)	Pooled OLS	Panel FE	Elastic Net	Random Forest	XGBoost	LSTM (early stop)	LSTM (350 epochs)
Naive (RW-AO)	1.000	0.048	0.015	0.077	0.538	0.130	0.202	0.045
Pooled OLS	0.048	1.000	0.834	0.237	0.069	0.417	0.635	0.510
Panel FE	0.015	0.834	1.000	0.402	0.034	0.384	0.651	0.342
Elastic Net	0.077	0.237	0.402	1.000	0.054	0.811	0.833	0.128
Random Forest	0.538	0.069	0.034	0.054	1.000	0.027	0.170	0.022
XGBoost	0.130	0.417	0.384	0.811	0.027	1.000	0.596	0.040
LSTM (early stop)	0.202	0.635	0.651	0.833	0.170	0.596	1.000	0.011
LSTM (350 epochs)	0.045	0.510	0.342	0.128	0.022	0.040	0.011	1.000

Of the 28 unique pairwise comparisons, eight are significant at the 5% level. Three of these involve the naive random-walk benchmark: persistence significantly outperforms Pooled OLS ($p = 0.048$), Panel Fixed Effects ($p = 0.015$), and the uncorrected 350-epoch LSTM ($p = 0.045$). Persistence is not, however, significantly different from Random Forest ($p = 0.538$), XGBoost ($p = 0.130$), Elastic Net ($p = 0.077$), or the early-stopped LSTM ($p = 0.202$) – confirming, at the level of individual pairwise tests, the central finding from Section 4.1 that no model achieves a statistically significant advantage over the naive benchmark, while several linear or unregularized specifications are significantly dominated by it.

A second cluster of significant differences centers on Random Forest, which significantly outperforms XGBoost ($p = 0.027$), Panel Fixed Effects ($p = 0.034$), and the 350-epoch LSTM ($p = 0.022$), making it the single best-performing model in the pairwise comparisons even though it does not clear the naive benchmark itself. XGBoost, in turn, significantly outperforms the 350-epoch LSTM ($p = 0.040$). Finally, the early-stopping correction applied to the Panel LSTM is itself statistically validated: the early-stopped LSTM significantly outperforms its own 350-epoch, non-early-stopped counterpart ($p = 0.011$), the smallest p-value in the entire matrix, confirming that early stopping is not merely a cosmetic fix but produces a materially and statistically distinguishable improvement in forecast accuracy.

The remaining 20 pairwise comparisons are statistically insignificant, including all comparisons among the three linear/panel econometric-adjacent models (Pooled OLS, Panel Fixed Effects, and Elastic Net) and most comparisons involving the early-stopped LSTM against the tree ensembles and Elastic Net. This pattern reinforces the paper's overall conclusion: statistically meaningful separation in this comparison set is concentrated at the extremes – persistence and Random Forest at one end outperforming the unregularized linear and uncorrected deep-learning specifications at the other – while the majority of models in between are not statistically distinguishable from one another.

4.6 Predictor Importance

Figure 4 reports gain-based XGBoost feature importance. Lagged inflation terms and the exchange-rate channel dominate, consistent with inflation-inertia and exchange-rate pass-through mechanisms, and with the broader emerging-market literature emphasizing pass-through as a primary inflation driver (Mirza et al., 2024).

4.7 Robustness Checks

Table 5 reports Diebold-Mariano p-values recomputed with the HAC-HLN variance estimator described in Section 3.8, for the seven core models (naive, both econometric benchmarks, Elastic Net, Random Forest, XGBoost, and the early-stopped LSTM). Table 4 used the simple sample-variance estimator common in applied practice but not corrected for potential autocorrelation in

the loss differential. Under the HAC-HLN estimator, 2 of the 21 unique pairwise comparisons remain significant at the 5% level before adjusting for multiplicity, and only one — persistence versus the Panel Fixed-Effects estimator ($p < 0.001$) — survives Holm-Bonferroni correction across the full comparison family. This more conservative test therefore reinforces, rather than overturns, the paper's central claim that no model achieves a statistically defensible improvement over persistence: the finding is, if anything, harder to overturn under stricter multiple-comparison control. It does mean, however, that two secondary claims reported against the less conservative estimator in Sections 4.1–4.3 — that Pooled OLS and the uncorrected 350-epoch LSTM are individually dominated by persistence, and that Random Forest significantly outperforms XGBoost — should be read as suggestive rather than confirmed once HAC-HLN variance estimation and multiple-comparison correction are applied. We flag this explicitly here rather than silently revising Table 4, which we retain in its original, more conventional form for comparability with prior applied literature.

Table 5
Robust (HAC-Newey-West + Harvey-Leybourne-Newbold) Diebold-Mariano Test p-values (pairwise, $n = 56$)

Model	Naive (RW-AO)	Pooled OLS	Panel FE	Elastic Net	Random Forest	XGBoost	LSTM (early stop)
Naive (RW-AO)	1.000	0.126	< 0.001	0.116	0.594	0.191	0.127
Pooled OLS	0.126	1.000	0.893	0.308	0.199	0.571	0.727
Panel FE	< 0.001	0.893	1.000	0.494	0.035	0.435	0.656
Elastic Net	0.116	0.308	0.494	1.000	0.181	0.877	0.878
Random Forest	0.594	0.199	0.035	0.181	1.000	0.069	0.110
XGBoost	0.191	0.571	0.435	0.877	0.069	1.000	0.481
LSTM (early stop)	0.127	0.727	0.656	0.878	0.110	0.481	1.000
<i>Note: Off-diagonal cells report two-sided HAC-HLN Diebold-Mariano p-values. Only the Naive vs. Panel Fixed-Effects comparison ($p < 0.001$) remains significant at the 5% level after Holm-Bonferroni correction across all 21 unique pairwise comparisons.</i>							

When every model is re-estimated using only lagged ($t - 1$, $t - 2$) predictors and excluding all seven same-year exogenous series (Sections 3.3 and 3.8), test RMSE is close to — and in most cases only modestly different from — the primary specification (Table 6). The naive benchmark, which by construction uses only INFL_CPI_lag1, is identical across specifications. No model overtakes persistence under either specification, and Random Forest remains the closest challenger in both, indicating that the paper's central conclusion is not an artifact of including contemporaneous exogenous information at the forecast origin.

Table 6
Primary vs. Lagged-Only Feature Specification: Test-Set RMSE and MAE

Model	Primary-Spec Test RMSE	Primary-Spec Test MAE	Lagged-Only Test RMSE	Lagged-Only Test MAE
Naive (RW-AO)	4.435	2.510	4.435	2.510
Pooled OLS	6.781	5.341	7.100	5.680
Panel FE	6.612	5.246	7.788	6.005
Elastic Net	5.986	4.583	6.400	4.874
Random Forest	4.849	3.399	4.894	3.408
XGBoost	5.825	3.895	5.535	3.555

Table 7 compares the static chronological-split test RMSE against an expanding-window (recursive) re-estimation scheme in which every model is retrained on all available data through year $t - 1$ before forecasting year t . Recursive re-estimation improves every estimated model relative to the single static-split fit — Pooled OLS falls from 6.781 to 4.709, Elastic Net from 5.986 to

4.896, Random Forest from 4.849 to 4.451, and XGBoost from 5.825 to 5.073 — narrowing, but not closing, the gap to persistence (4.435 under both schemes, since the naive forecast requires no re-estimation). Random Forest under the expanding-window scheme (4.451) comes within 0.016 RMSE points of the naive benchmark, the closest margin observed anywhere in this study. We did not re-run formal Diebold-Mariano tests on the recursive forecast errors in this revision; the RMSE pattern nonetheless confirms that the paper's central finding is not an artifact of the single static split, and that a more operationally realistic recursive re-estimation protocol, if anything, narrows rather than widens the gap between estimated models and persistence.

Table 7
Expanding-Window (Recursive) vs. Static-Split Test RMSE

Model	Expanding-Window RMSE	Static-Split Test RMSE
Naive (RW-AO)	4.435	4.435
Pooled OLS	4.709	6.781
Elastic Net	4.896	5.986
Random Forest	4.451	4.849
XGBoost	5.073	5.825

Figure 5 and Table 8 decompose test-set RMSE by country and by pre-/post-2020 subperiod. Forecast error is highly heterogeneous across the panel: Ghana and Nigeria — both of which experienced pronounced currency and price volatility during 2018–2024 — display substantially higher RMSE for every model (e.g., naive RMSE of 10.43 for Ghana and 4.60 for Nigeria, versus 0.83–3.78 for the remaining six countries), and Random Forest and the LSTM narrow, but do not eliminate, the naive benchmark's disadvantage specifically in these two higher-volatility economies. The pre-/post-2020 subperiod split in Table 8 shows a consistent pattern: naive RMSE rises from 1.84 (2018–2020) to 5.65 (2021–2024) as COVID-19 and global commodity-price shocks materialize in the test window, and every estimated model's RMSE rises correspondingly in the post-2020 subperiod, with Random Forest showing the smallest degradation (3.18 to 5.80) among the non-naive models. This is consistent with the structural-break explanation offered in Section 5.1: aggregate test-window results mask meaningful country- and regime-level heterogeneity, and the panel average should not be read as implying uniform model performance across all eight economies or both subperiods.

Table 8
Pre-/Post-2020 Subperiod Test RMSE

Model	RMSE 2018–2020	RMSE 2021–2024
Naive (RW-AO)	1.84	5.65
Pooled OLS	6.44	7.02
Panel FE	5.09	7.55
Elastic Net	5.05	6.60
Random Forest	3.18	5.80
XGBoost	3.73	6.99
LSTM (early stop)	3.96	7.39

Finally, Table 9 reports early-stopped LSTM performance for input sequence lengths $T = 2, 3, 4$, and 5 years. Test RMSE ranges narrowly from 5.872 ($T = 4$) to 6.213 ($T = 5$), with the manuscript's reported $T = 3$ configuration (RMSE 6.162) falling within this range rather than at either extreme. The LSTM's standing relative to the tree ensembles and the naive benchmark is therefore not sensitive to the specific three-year lookback window chosen in Section 3.4.3.

Table 9
LSTM Test-Set Performance by Input Sequence Length (Early Stopping)

Model	Test RMSE	Test MAE	Best Epoch
T = 2	6.072	3.707	137
T = 3 (reported)	6.162	3.777	126
T = 4	5.872	3.852	128
T = 5	6.213	3.173	99

5. Discussion

Including the naive persistence benchmark changes this paper's central claim in a way we argue strengthens, rather than weakens, its contribution. Rather than "machine learning beats econometrics," the evidence supports a more nuanced, literature-consistent conclusion. In a complete eight-country SSA annual panel ($n = 424$, effective test window $n = 56$), no model — econometric or machine-learning — statistically outperforms a zero-parameter persistence rule for one-step-ahead CPI inflation forecasting. Tree-based ensembles are the only estimated models that consistently achieve statistical parity with the naive benchmark, and even among them the ranking is not uniform: Random Forest significantly outperforms XGBoost on the full panel ($p = 0.027$), a reversal from the closer race observed in the earlier interim six-country analysis, though neither surpasses persistence itself. Linear econometric models — both pooled OLS and the country fixed-effects specification — are significantly dominated by persistence, and adding fixed effects does not rescue the linear specification from this shortfall. The Panel LSTM tells a similar story with an important qualification: early stopping is necessary but not sufficient to make it competitive, closing roughly a fifth of the gap to the best-performing models and moving it from significantly worse than persistence to statistically indistinguishable from it, without ever making it superior to the tree ensembles or the naive benchmark. Taken together, these results extend a recurring pattern in the inflation-forecasting literature — documented by Atkeson and Ohanian (2001) for the United States, by Duncan and Martínez-García (2018) across fourteen emerging markets, and echoed in Naghi's (2024) finding that machine-learning gains over strong autoregressive benchmarks are modest even in data-rich advanced-economy settings — to a new regional context and to a considerably richer model class that includes tree ensembles and recurrent deep learning.

5.1 Why Might Complex Models Fail to Beat Persistence Here?

Several factors plausibly explain why more flexible models fail to clear the low bar set by persistence in this setting. The most immediate is sample size: with 304 training and 56 test observations pooled across eight countries, high-capacity models such as the LSTM and unconstrained pooled OLS have limited ability to estimate stable, generalizable structure, whereas more conservative, regularized methods — tree ensembles and persistence itself — are naturally better suited to a panel of this scale. A second factor is the presence of structural breaks: the test window (2018–2024) spans COVID-19 supply shocks, global commodity-price swings, and currency-stress episodes in several SSA economies that may not resemble the training-period (≤ 2009) data-generating process, which tends to favor adaptive, low-parameter forecasts over models that extrapolate historical relationships estimated in a different macroeconomic regime (Giraitis et al., 2013). A third, more structural possibility is that the exogenous predictors considered here simply carry weak average predictive content for many SSA economies: CPI inflation may be dominated by its own lagged dynamics and by exchange-rate pass-through rather than by the broader fiscal, monetary, and credit variables assembled in Table 1, which would limit the marginal information any conditioning model — linear or non-linear — can extract beyond what persistence already captures.

5.2 Policy and Practical Implications

For central banks and finance ministries in the sampled economies, these results counsel caution against over-investing in complex forecasting infrastructure without first establishing that it clears the low bar set by persistence. Where tree-based ensembles are adopted, we recommend retaining a persistence forecast as a live benchmark in production and reporting Diebold-Mariano — rather than point RMSE alone — comparisons in operational forecast evaluation, echoing Hammond's (2012) survey finding that RW-AO-type benchmarks are largely absent from central banks' own model suites despite their strong

empirical performance. This recommendation is reinforced by the emerging machine-learning inflation-forecasting literature summarized in Section 2.3. Even where ML methods do add value, that value tends to be modest, regime-dependent, and conditional on considerably richer data environments than the annual SSA panel considered here, so institutions with comparable data constraints should treat any complex model as a challenger to be tested against persistence rather than as a presumed improvement over it. This recommendation rests most directly on the strictly lagged-only respecification reported in Table 6 (Section 4.7), which excludes every same-year exogenous predictor and therefore reflects only the information genuinely available to a forecaster at the true one-step-ahead decision point; because persistence is not overtaken by any model under this ex-ante-consistent specification either, the policy conclusion is not an artifact of using contemporaneous information unavailable in real-time operational forecasting.

5.3. Limitations, Future Work and Extensions

The automated, diagnostics-enabled extraction protocol successfully retrieved a complete, fully populated panel ($N = 424$) with continuous coverage for all nine indicators across the eight targeted Sub-Saharan African economies from 1972 to 2024, achieved through the corrected API extraction protocol together with the country-specific imputation described below. While this complete dataset provides a consistent basis for the empirical comparisons presented in this study, several structural and scope characteristics of the panel present opportunities for future research.

Two residual data-construction characteristics of the panel are worth noting explicitly. First, two indicator-country series had zero native World Bank WDI coverage and were completed via the country-specific forward-fill/backward-fill protocol described in Section 3.3, applied strictly within each affected country's own time series and prior to lag construction and the train/validation/test split. Forward-fill introduces no look-ahead, since it relies only on that country's own past values; backward-fill, used to close a small number of leading gaps at the start of a few series, draws on that same country's own later observations and so introduces a narrow, country-local form of look-ahead confined to a handful of early-period cells, distinct from any cross-country or train/test leakage. Second, Uganda's series carry partial early-period gaps addressed by the same protocol. Both characteristics affect only a small share of the $N = 424$ panel, concentrated in the pre-2010 training partition. Third, the Diebold-Mariano-based inference reported throughout Sections 3.7, 3.8, and 4.7 corrects for serial correlation (via the HAC-HLN estimator) and for multiplicity (via Holm-Bonferroni) but does not explicitly model cross-sectional dependence among the eight countries' loss differentials; common regional or global shocks within a test year could induce correlation across countries that a year- or country-clustered variance estimator (e.g., a Driscoll-Kraay-type correction) would be needed to fully address. We note this as a direction for future methodological refinement rather than a concern that changes the reported significance patterns.

First, while the annual panel design maximizes the temporal depth of the dataset, macroeconomic policy and forecasting often operate on higher-frequency intervals. Future research could extend this database to monthly or quarterly frequencies where reliable regional indicators are available. This transition would significantly increase the effective sample size and enable sequential architectures, such as the LSTM, to leverage intra-year dynamics and complex lag structures more effectively than is possible with annual observations.

Second, future iterations of this framework could expand the geographic scope of the panel, since the present study has a small geographical focus and dataset. Incorporating additional Sub-Saharan African economies would enhance the regional representativeness of the models and provide a broader database for evaluating the generalizability of the findings.

Third, the finding that tree-based ensembles and the early-stopped LSTM achieve statistical parity with—but do not outperform—the naive persistence benchmark suggests several promising methodological avenues:

Hybrid Econometric-ML Specifications

Combining the structural interpretability of traditional frameworks with the flexible feature-selection capabilities of machine learning could yield improvements. For example, using regularized regression or tree-based ensembles as variable-selection mechanisms within a panel autoregressive distributed lag (ARDL) or panel fixed-effects framework may combine the strengths of both paradigms.

Forecast Combinations

Investigating forecast pooling methods that blend the persistence benchmark with machine learning and deep learning predictions is a logical next step. Prior literature suggests that combined forecasts often deliver more stable and robust out-of-sample performance than any single constituent model.

Forward-Looking Indicators

Incorporating expectations-based variables, such as survey-based inflation expectations or sentiment indices derived from central bank communication texts, could improve predictive power. These features may help capture shifts in monetary policy credibility and expectation-formation dynamics that historical macroeconomic lags alone cannot fully represent.

This revision additionally reports validation-based hyperparameter selection (Section 3.4.4), HAC-HLN-corrected and Holm-Bonferroni-adjusted Diebold-Mariano tests (Sections 3.8, 4.7), a strictly lagged-only predictor respecification, an expanding-window recursive re-estimation scheme, per-country and subperiod error decomposition, and LSTM sequence-length sensitivity analysis (Section 4.7). Formal joint-inference procedures such as model confidence set methods and self-normalized conditional predictive ability tests, transition to higher-frequency (monthly or quarterly) data, and survey-based inflation-expectations proxies remain natural extensions beyond the scope of the present revision.

6. Conclusions

This study evaluated and compared linear econometric, machine learning, and deep learning approaches to inflation forecasting using an eight-country Sub-Saharan African annual panel spanning 1972 to 2024, explicitly incorporating a naive persistence (random-walk) benchmark that is frequently omitted in the regional literature. The empirical evidence aligns with the classic "random-walk puzzle" documented in advanced- and emerging-market contexts: no estimated model achieved a statistically significant improvement in out-of-sample forecast accuracy over the zero-parameter persistence baseline. While pooled tree ensembles (Random Forest and XGBoost) and the early-stopped LSTM network achieved statistical parity with persistence, they did not surpass it, whereas traditional linear models and the uncorrected LSTM were significantly dominated by the benchmark. Critically, this conclusion also holds under a strictly lagged-only, ex-ante-consistent predictor specification that excludes all same-year exogenous information, indicating that it reflects genuine forecasting difficulty rather than an artifact of information unavailable at the true forecast origin. These results suggest that the high volatility, structural breaks, and relatively modest sample sizes characteristic of regional macroeconomic panels pose substantial challenges for highly parameterized or unregularized forecasting architectures, counseling caution for regional central banks and finance ministries evaluating investments in complex forecasting infrastructure. Rather than assuming that increased model complexity translates to superior performance, institutions should retain simple persistence forecasts as live benchmarks and use formal pairwise predictive accuracy tests, such as the Diebold-Mariano test, to validate alternative models. To build upon these findings, future research should explore transitioning to higher-frequency (monthly or quarterly) indicator series to allow sequential models to better capture intra-year dynamics, developing hybrid econometric-machine learning specifications that combine structural interpretability with regularized variable selection, and investigating systematic forecast combinations that pool the robust properties of persistence with the non-linear capabilities of tree ensembles.

Declarations

Author Contribution

Conceptualization, B.O., A.B., I.N. and G.O.F.; methodology, B.O., A.B., I.N. and G.O.F.; software, G.O.F.; validation, B.O., A.B. and G.O.F.; formal analysis, B.O., A.B. and G.O.F.; investigation, B.O., A.B. and G.O.F.; resources, B.O., A.B., I.N. and G.O.F.; writing—original draft preparation, B.O., A.B., I.N. and G.O.F.; writing—review and editing, B.O., A.B., I.N. and G.O.F.; supervision, G.O.F. All authors have read and agreed to the publication of the manuscript.

Acknowledgement

The authors would like to thank the University of Skills Training and Entrepreneurial Development (USTED), Kumasi, Ghana, and the Department of Forest Engineering, Forest Management Planning, and Terrestrial Measurements, Faculty of Silviculture and Forest Engineering, Transilvania University of Brasov, for providing some of the resources needed for this study.

Data Availability

The complete eight-country macroeconomic panel (CSV and Excel formats), the World Bank/FRED API data-extraction script, the model-training, benchmarking, and Diebold-Mariano evaluation pipeline, the robustness-check pipeline (HAC-HLN/Holm-corrected significance testing, lagged-only and expanding-window respecifications, per-country/subperiod and LSTM sequence-length sensitivity analyses), and the figure-generation scripts are available as supplementary files on figshare at: <https://figshare.com/s/0a8f1e47a31b94220bf1>.

References

1. Alagidede, P., Coleman, S., & Cuestas, J. C. (2012). Inflationary shocks and common economic trends: Implications for West African monetary union membership. *Journal of Policy Modeling*, 34(3), 460–475. <https://doi.org/10.1016/j.jpolmod.2011.10.001>
2. Almosova, A., & Fasanya, I. (2023). Nonlinear inflation forecasting with recurrent neural networks. *Journal of Forecasting*, 42(2), 240–259. <https://doi.org/10.1002/for.2901>
3. Atkeson, A., & Ohanian, L. E. (2001). Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1), 2–11. <https://doi.org/10.21034/qv.2511>
4. Baltagi, B. H. (2021). *Econometric analysis of panel data* (6th ed.). Springer. <https://doi.org/10.1007/978-3-030-53953-5>
5. Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83. <https://doi.org/10.1016/j.csda.2017.11.003>
6. Bolivar, O. (2026). High-frequency inflation forecasting: A two-step machine learning methodology. *Latin American Journal of Central Banking*, 7(1), Article 100172. <https://doi.org/10.1016/j.latcb.2025.100172>
7. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
8. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
9. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. <https://doi.org/10.1080/07350015.1995.10524599>
10. Duncan, R., & Martínez-García, E. (2018). *New perspectives on forecasting inflation in emerging market economies: An empirical assessment* (Globalization and Monetary Policy Institute Working Paper No. 388). Federal Reserve Bank of Dallas. <https://doi.org/10.24149/gwp338>
11. Faust, J., & Wright, J. H. (2013). Forecasting inflation. In G. Elliott & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 2A, pp. 2–56). Elsevier. <https://doi.org/10.1016/B978-0-444-53683-9.00001-3>
12. Giraitis, L., Kapetanios, G., & Price, S. (2013). Adaptive forecasting in the presence of recent and ongoing structural change. *Journal of Econometrics*, 177(2), 153–170. <https://doi.org/10.1016/j.jeconom.2013.04.003>
13. Hammond, G. (2012). *State of the art of inflation targeting* (CCBS Handbook No. 29). Bank of England. <https://www.bankofengland.co.uk/ccbs/state-of-the-art-of-inflation-targeting>
14. 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>
15. International Monetary Fund. (2025). *Regional economic outlook: Sub-Saharan Africa, October 2025*. <https://www.imf.org/en/Publications/REO/SSA>
16. Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization* (arXiv:1412.6980). arXiv. <https://doi.org/10.48550/arXiv.1412.6980>

17. Liu, Y., Pan, R., & Xu, R. (2024). *Mending the crystal ball: Enhanced inflation forecasts with machine learning* (IMF Working Paper No. 24/206). International Monetary Fund. <https://doi.org/10.5089/9798400285387.001>
18. McCracken, M. W., & Ng, S. (2016). FRED-MD: A database for identifying monthly US macroeconomic data. *Journal of Business & Economic Statistics*, 34(4), 574–589. <https://doi.org/10.1080/07350015.2015.1086655>
19. Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1), 98–119. <https://doi.org/10.1080/07350015.2019.1637745>
20. Mirza, N., Rizvi, S. K. A., Naqvi, B., & Umar, M. (2024). Inflation prediction in emerging economies: Machine learning and FX reserves integration for enhanced forecasting. *International Review of Financial Analysis*, 94, Article 103238. <https://doi.org/10.1016/j.irfa.2024.103238>
21. Naghi, A. A. (2024). The benefits of forecasting inflation with machine learning: New evidence. *Journal of Applied Econometrics*. Advance online publication. <https://doi.org/10.1002/jae.3088>
22. Nortey, E. N. N., Agyemang, E. F., Sakyi-Yeboah, E., Ampomah, O.-A., & Agyekum, L. (2025). AI meets economics: Can deep learning surpass machine learning and traditional statistical models in inflation time series forecasting? *Data Science in Finance and Economics*, 5(2), 136–155. <https://doi.org/10.3934/DSFE.2025007>
23. Slimani, H., & Chourabi, C. (2026). Modeling inflation with machine learning: A cross-horizon systematic review. *International Journal of Data Science and Analytics*, 21, Article 65. <https://doi.org/10.1007/s41060-025-00905-w>
24. Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components: Representation and forecasting. *Journal of the American Statistical Association*, 97(460), 1167–1179. <https://doi.org/10.1198/016214502388618960>
25. Stock, J. H., & Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39(s1), 3–33. <https://doi.org/10.1111/j.1538-4616.2007.00014.x>
26. Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28. <https://doi.org/10.1257/jep.28.2.3>
27. World Bank. (2025). *Africa's Pulse, October 2025*. <https://www.worldbank.org/en/region/afr/publication/africas-pulse>
28. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

Figures

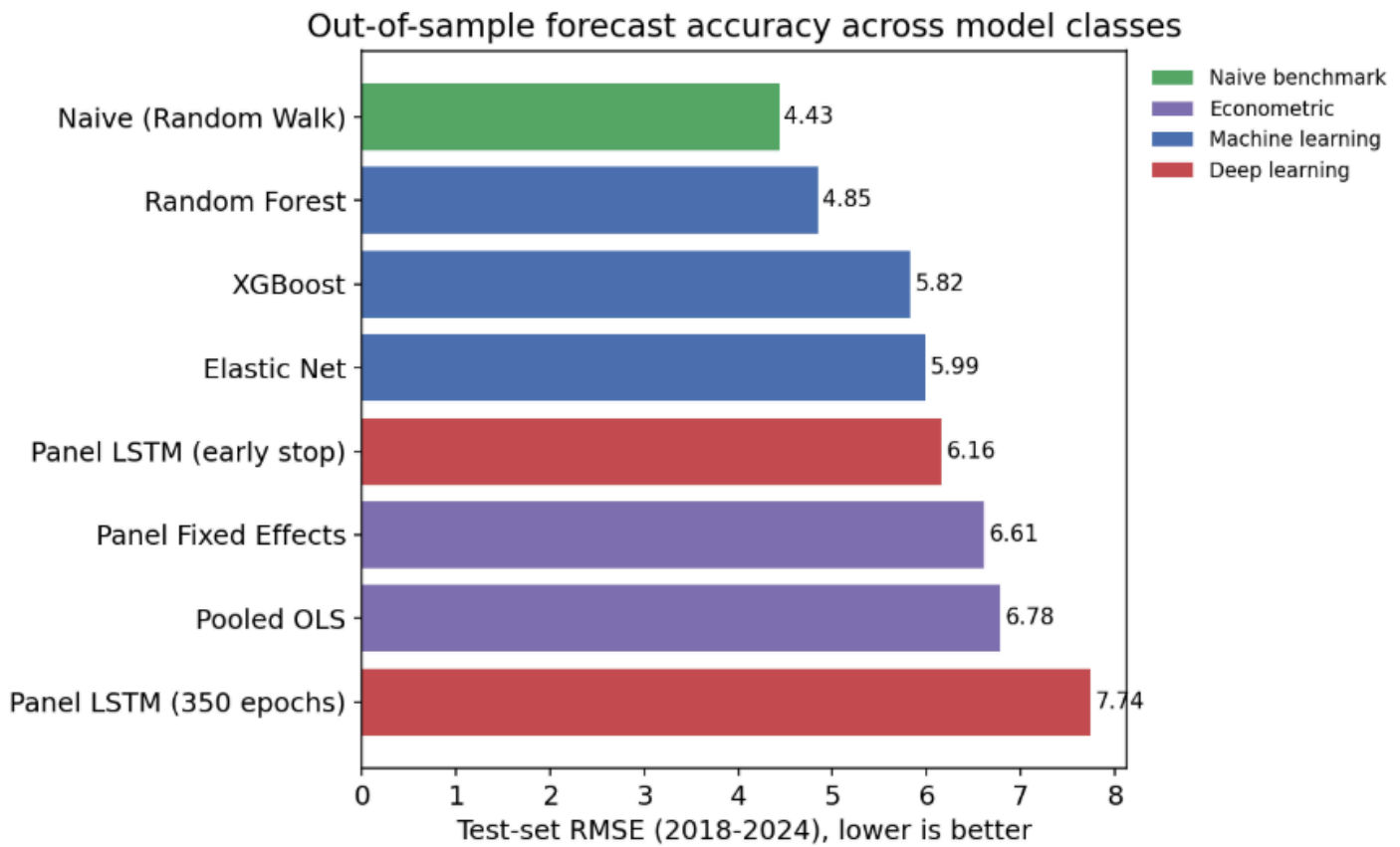


Figure 1

Out-of-sample test RMSE by model class (naive, econometric, ML, deep learning).

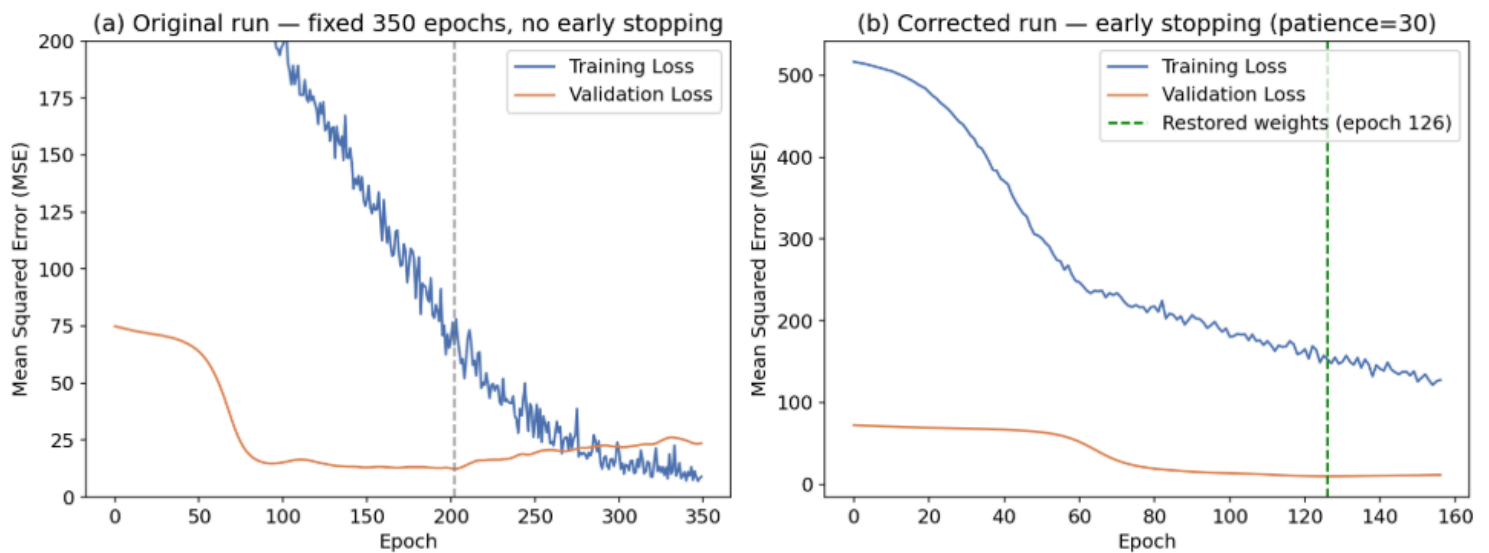


Figure 2

LSTM learning curves: (a) original 350-epoch run without early stopping; (b) corrected run with early stopping.

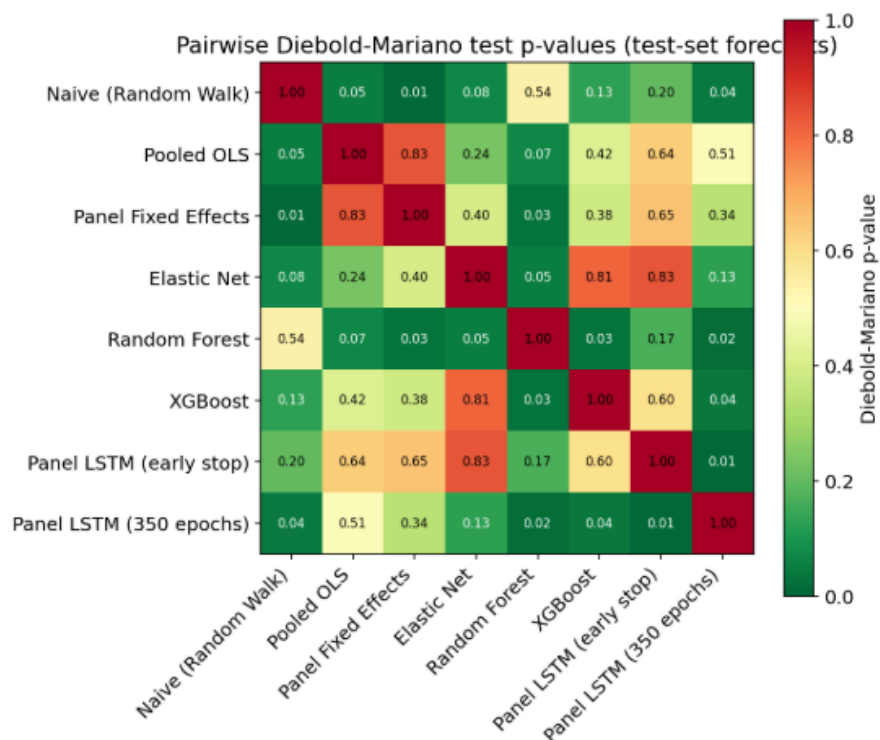


Figure 3

Diebold-Mariano pairwise test p-values (green = statistically distinguishable; red = statistically indistinguishable).

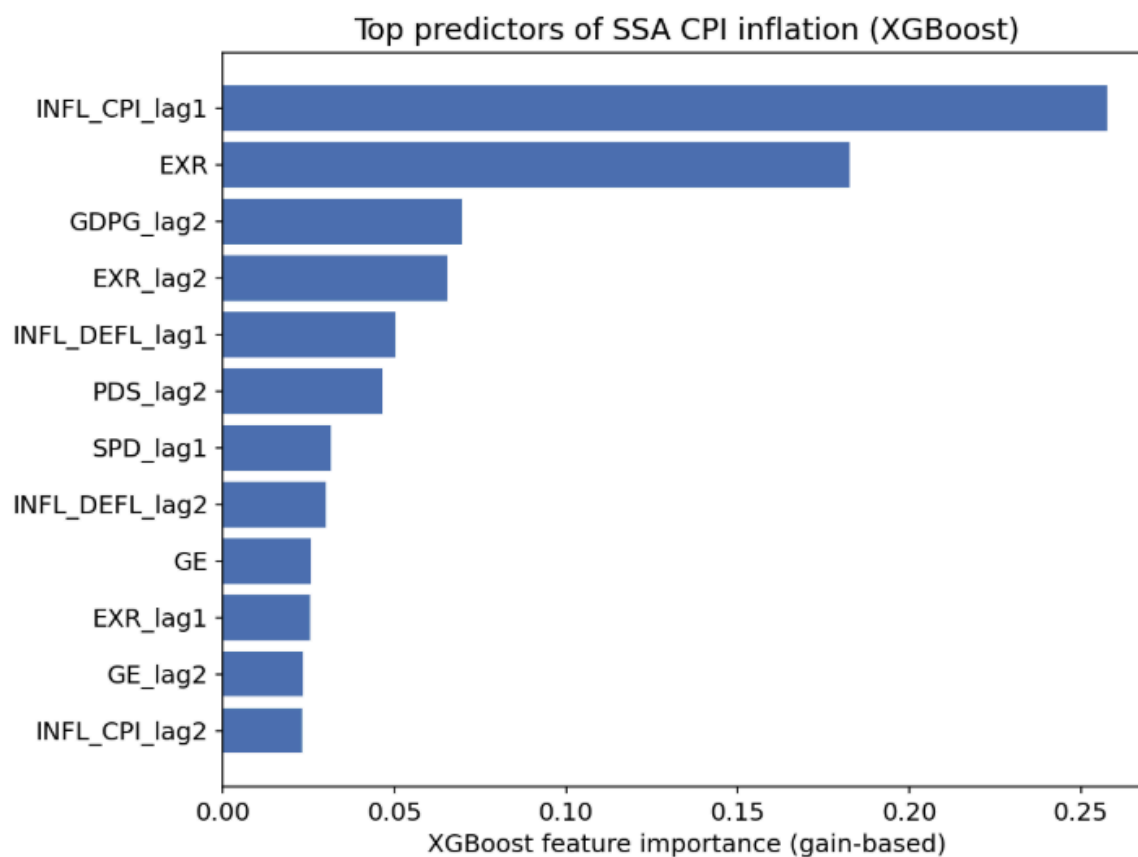


Figure 4

XGBoost gain-based feature importance, top 12 predictors.

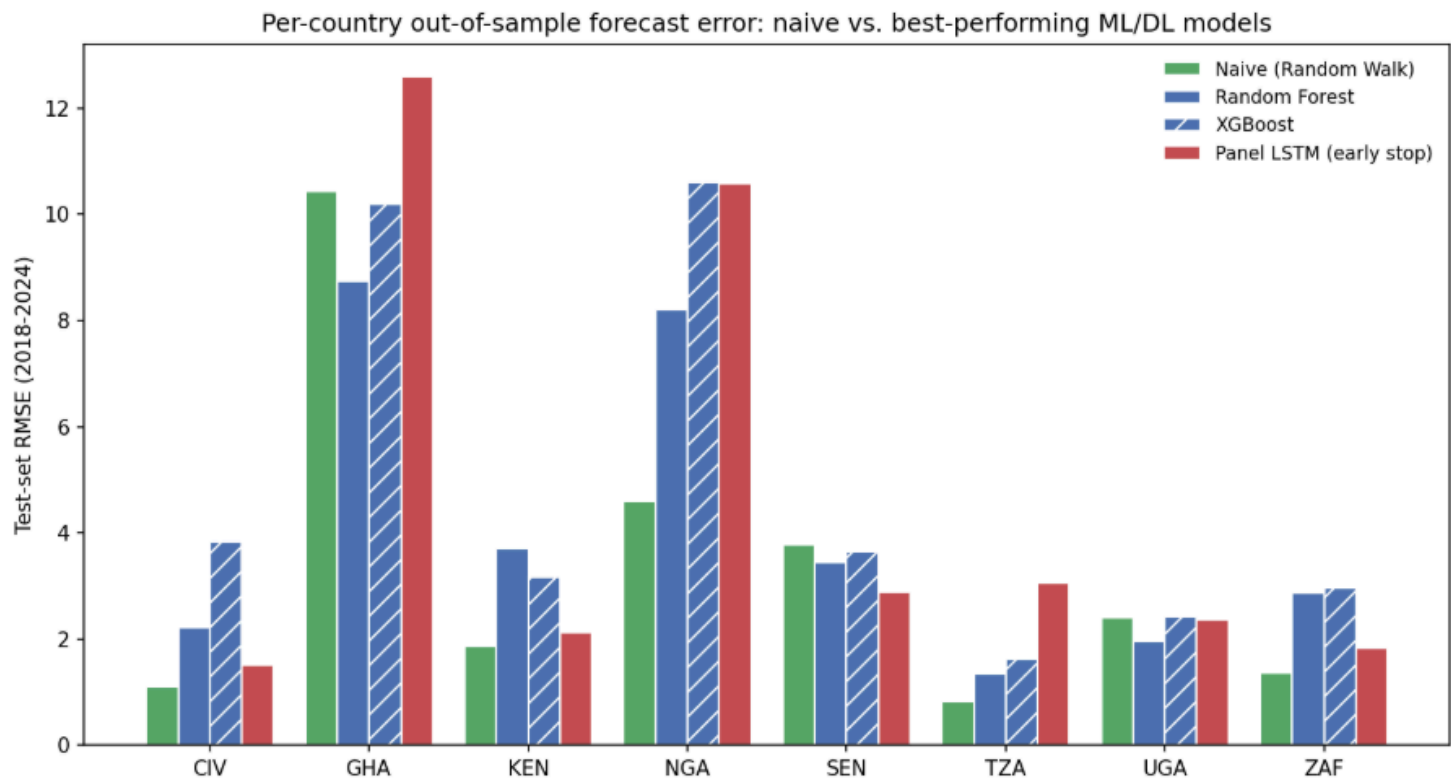


Figure 5

Per-country out-of-sample test RMSE (2018-2024): naive persistence vs. Random Forest, XGBoost, and the early-stopped Panel LSTM.