IJEMD-SS, 5 (6) (2026) https://doi.org/10.54938/ijemd-ss.2026.05.6.882

International Journal of Emerging Multidisciplinaries: Social Science

Research Paper
Journal Homepage: www.ojs.ijemd.com
ISSN (Online): 2958-0277



Synergy Between Classical Econometrics and Machine Learning in Forecasting Pakistan's Consumer Price Index

1. Arslan Munir Turk, 2. Saud Ahmed Khan

[1, 2]PIDE School of Economics, Pakistan Institute of Development Economics, Islamabad, Pakistan

arslaneconometrician@gmail.com **(Corresponding author)**

Abstract

Forecasting an administratively compiled consumer price index exposes the standing division in the forecasting profession with unusual clarity. Classical econometric specifications deliver interpretable parameters and honest confidence intervals but revert to the mean of the differenced series faster than an inflation episode actually unwinds; machine learning specifications track the nonlinear acceleration but return point forecasts with no standard error worth the name. This paper proposes and tests a synergy that assigns each tradition the task it is built for. The monthly Pakistan Consumer Price Index, August 1991 to February 2025 (T = 402), is decomposed by Ensemble Empirical Mode Decomposition into three intrinsic mode functions; a power spectral density rule routes each mode to a stochastic or a deterministic sub block on the share of its energy lying below a low frequency threshold; an ARIMA front end absorbs the linear dependence within each subblock; and the pooled residual is handed to an LSTM or an XGBoost residual learner whose hyperparameters are selected by Bayesian optimisation against a held out validation loss. Eleven candidate specifications are estimated on a seventy per cent fixedorigin estimation window and compared on the remaining thirty per cent under mean absolute error, root mean squared error, symmetric mean absolute percentage error and the Diebold Mariano test. The results depict that synergy has minimum loss functions instead of stand alone conventional and machine learning models.

Keywords: inflation forecasting, Consumer Price Index, Pakistan, ensemble empirical mode decomposition, ARIMA, LSTM, XGBoost, Bayesian optimisation, indicator saturation, Diebold–Mariano test

JEL classification: C22, C45, C53, E31, E37

1. Introduction

When Box and Jenkins published the first edition of their monograph on time series analysis in 1970, the forecasting of economic variables was a narrow discipline. The available toolkit consisted of a small number of linear autoregressive and moving average specifications, estimation was performed by hand computation supplemented by punchedcard mainframes, and the entire literature could be surveyed in a single graduate course. Half a century later that discipline has been transformed. The volume of available data has expanded by orders of magnitude, central banks assemble real time dashboards of hundreds of indicators, and the toolkit for analysing them has grown from a handful of linear models to a variety of methods drawn from statistics, econometrics, signal processing and machine learning.

A trained LSTM yields a point forecast but no standard error worth the name; a gradientboosted ensemble fits the data brilliantly and explains almost nothing about why (Lipton, 2018; Rudin, 2019). The two camps are not disagreeing about a technical detail. They are optimising for different things; one for accountable inference, the other for predictive accuracy — and each has built tools that excel at exactly what they were built to do, and falter at exactly what they were not.

A consumer price index is a good place to put that division under strain, and Pakistan's index is a particularly demanding case. It is administratively compiled rather than traded, so it carries no microstructure noise and no trading calendar; its short run variation is seasonal and administrative rather than speculative. It is rebased periodically, which introduces mechanical discontinuities that are not economic events at all. And over the last three years of the sample, it exhibits an acceleration; the index roughly doubles between early 2021 and mid 2023 of exactly the kind that a linear model's instinct toward reversion handles worst. A purely nonparametric learner, meanwhile, has no principled way to separate the mechanical resets from the economic movement, and no way to report what it has done.

1.1. The synergy Importance

The idea that combining forecasts can outperform any constituent has a long lineage (Bates & Granger, 1969; Timmermann, 2006), and hybrid decompositionandlearning designs of the kind proposed here descend from it (Zhang, 2003; Yu, Wang, & Lai, 2008). Synergy in this sense is not just a slogan, but a testable empirical hypothesis. If the classical econometric layer is doing all the work, removing it should not degrade accuracy. Section 5 puts both statements to the data by estimating eleven nested and non nested specifications on the same series, the same window and the same loss functions.

The first, Ensemble Empirical Mode Decomposition (EEMD), takes the observed index and decomposes it into a small number of Intrinsic Mode Functions (IMFs) plus a residual trend (Huang et al., 1998; Wu & Huang, 2009). The second layer does the sorting: a Welch power spectral density test (Welch, 1967; Priestley, 1981) reads the energy spectrum of each IMF and assigns it a verdict stochastic or deterministic — routing it to the sub block built to handle it. The third layer is where the Box–Jenkins toolkit earns its keep: an ARIMA specification models the linear dynamics of each sub block on its own terms, extracting whatever autocorrelation structure a well specified econometric model is supposed to extract, and no more. What survives that extraction the nonlinear residue a linear model cannot represent by construction is handed to the fourth layer, a residual learner, LSTM (Hochreiter & Schmidhuber, 1997) or XGBoost (Chen & Guestrin, 2016), whose hyperparameters are not guessed at but selected by Bayesian optimisation against a validation loss (Mockus, 1975; Snoek, Larochelle, & Adams, 2012). The composite forecast is reassembled into a single closed form expression, equation (19) below, general enough to be lifted whole into a downstream application without further specification. Strip out any one layer and the architecture collapses back into one of the standalone benchmarks against which it is tested, which is the ablation logic that the combination literature applies to weighted forecast pools.

1.2 Research Questions and Contribution

Three questions are put to the data. First, does yoking classical econometrics and machine learning together through an EEMD based decomposition of the index improve on either tradition applied alone? Second, does routing the extracted modes by a formal spectral rule, rather than by the visual inspection that pervades much of the applied EEMD literature, produce an assignment that is stable and economically interpretable? Third, does a Bayesian search over

the residual learner's hyperparameters improve the resulting composite, and does it do so for both candidate back ends or only for one?

The contribution is correspondingly threefold. Methodologically, the framework is written down completely; every stage is an explicit equation rather than a verbal description, and the architecture terminates in one closed form expression for the composite forecast that states exactly which quantities to substitute. Empirically, eleven candidate specifications are estimated on a single fixed origin split and compared on three loss functions and the Diebold Mariano test, so that the ablation logic of the synergy argument is tested rather than asserted. A subset of the underlying methodology has been peer reviewed and published for a different series (Turk, Khan, Aamir, & Hussain, 2025); the present paper applies the framework, unmodified, to an administered macroeconomic index.

The remainder of the paper is organised as follows. Section 2 reviews the five literatures the framework draws on and states the gap. Section 3 derives the framework step by step. Section 4 describes the data. Section 5 reports the empirical results. Section 6 discusses what the results do and do not establish, and Section 7 concludes.

2. Literature Review

The problem addressed here sits at the intersection of five distinct literatures time series decomposition, classical econometric forecasting, machine learning forecasting, structural break detection, and Bayesian hyperparameter optimisation each of which has developed largely independently of the others over the past three or four decades. This section takes them in turn and states what each contributes to the framework of Section 3.

2.1. Time Series Decomposition

The idea of decomposing an observed series into a sum of orthogonal components dates back at least to the classical additive decomposition of Persons (1919) and the multiplicative decomposition popularised by the U.S. Bureau of the Census in the X 11 procedure of the 1960s. These early methods presumed a fixed periodicity typically a twelvemonth seasonal cycle for monthly data and produced a residual assumed to be nearGaussian white noise. Applied to modern macroeconomic series with regime shifts, nonstationary trends and irregular volatility clusters, they produce residuals that are neither Gaussian nor white, and the forecasting exercises built on top of them inherit the problem.

The wavelet transform, popularised for time series analysis by Percival and Walden (2000), addressed some of these limitations by introducing a genuinely multiresolution basis in which local variations in time and frequency could be captured simultaneously. Waveletbased forecasting has been applied to inflation series (AguiarConraria & Soares, 2011), to commodity prices (Nguyen & Sensoy, 2017) and to sovereign yields (Fernandes, Medeiros, & Scharth, 2018). Wavelet methods adapt to local features in the data but require a choice of mother wavelet and decomposition level that is difficult to justify a priori and that can materially affect the reconstructed components.

The Empirical Mode Decomposition (EMD) of Huang et al. (1998) took a further step by proposing a fully adaptive decomposition in which the basis functions; the Intrinsic Mode Functions are derived from the data itself through an iterative sifting procedure. EMD makes no a priori assumption about periodicity, mother wavelet or component shape. Its main defect is mode mixing: an oscillation of a given time scale can appear across two adjacent IMFs when the signal contains intermittent components. Wu and Huang (2009) diagnosed the flaw and prescribed a remedy. Their Ensemble Empirical Mode Decomposition perturbs the signal with a small dose of white noise, decomposes each perturbed copy separately, and averages across many realizations; a computationally bruteforce fix, but one that produces IMFs considerably more robust to mode mixing. Flandrin, Rilling, and Goncalves (2004) supply the filterbank characterisation that makes the resulting modes interpretable as band assignments rather than as arbitrary summaries, and it is that result which licenses the spectral routing rule used in Section 3.3.

2.2. Classical Econometric Forecasting

More than half a century after Box and Jenkins (1970) first set it down and later revised, with more careful attention to diagnostic checking, in Box, Jenkins, Reinsel, and Ljung (2015) the Autoregressive Integrated Moving Average framework remains the workhorse of classical econometric forecasting, and it is worth pausing on why it has lasted.

For Pakistan specifically, the applied inflation forecasting literature has largely worked within this classical frame, with ARIMA and ARCHfamily specifications (Hussain, Ali, & Zaki, 2019) and early neural appraisals (Choudhary & Haider, 2012) reported alongside one another rather than combined.

2.3. Machine Learning in Time Series Forecasting

The introduction of Long ShortTerm Memory recurrent networks by Hochreiter and Schmidhuber (1997) solved a technical problem in deep learning — the vanishing gradient problem of plain recurrent networks — that had until then prevented deep architectures from being trained on long sequences. LSTM cells maintain three gates and a cell state that carries information across time steps without vanishing gradients, allowing the network to learn longrange dependencies. Greff, Srivastava, Koutník, Steunebrink, and Schmidhuber (2017) provide a systematic evaluation of the LSTM design space; Fischer and Krauss (2018) demonstrate its performance in financial applications; and SiamiNamini, Tavakoli, and Namin (2018) compare ARIMA and LSTM directly across a range of economic and financial series, reporting systematic gains for the recurrent specification.

The Extreme Gradient Boosting algorithm of Chen and Guestrin (2016) took a different route to nonlinear function approximation: a large additive ensemble of shallow decision trees, each trained on the residuals of the previous ensemble, with regularisation penalties on tree complexity and gradientbased optimisation of the loss. XGBoost has become one of the most widely used tools in applied prediction because of its speed, its strong out of the box performance and its natural handling of missing values. In time series applications it requires the input features to include lagged values of the target, but with that adaptation it is competitive with LSTM on many benchmark problems (Xiong, Guo, Zeng, Zou, & Wang, 2022). Because a tree ensemble has no internal representation of time, the comparison between the two learners is only a comparison of learners, rather than of feature sets, if the temporal structure the recurrent network extracts from its input sequence is supplied to the ensemble explicitly a point taken up in Section 3.6.

A number of papers have proposed combining classical econometric models with these learners. Zhang (2003) proposed a linear then non linear ARIMA neural network hybrid in which the residuals from the ARIMA fit were passed to a neural network. Kourentzes, Barrow, and Crone (2014) proposed neural network ensembles as a way of combining classical and machine learning forecasts. Nasir, Aamir, Haq, Khan, Amin, and Naeem (2023) anticipate the basic architecture pursued here decompose, model the linear component econometrically, hand the rest to a recurrent network closely enough that their Local Mean Decomposition coupled with an ARIMA–LSTM hybrid deserves to be read as this framework's direct methodological ancestor.

2.4. Structural Break Detection with Impulse and SIS

Hendry, Johansen, and Santos (2008) approached structural break detection from an angle that looked reckless before it turned out to be rigorous. Impulse Indicator Saturation equation of interest with T impulse dummies, one for every observation in the sample, and then lets a disciplined general to specific selection algorithm decide, at a controlled target significance level, which of them earn their place. Castle, Doornik, and Hendry (2015) extended the same logic from impulses to steps, so that Step Indicator Saturation catches a sustained shift in level and not merely a single anomalous observation, and Castle and Hendry (2019) draw the two together into the standard reference treatment followed here. Pesaran and Timmermann (2007) took the argument to its operational conclusion, showing that when breaks are present the choice of estimation window becomes a first order decision: using all available history biases the forecast toward a regime that no longer obtains, while using only post break data discards information and inflates variance. Rossi (2013) surveys the resulting literature on forecasting under instability, together with the tests of predictive ability that remain valid when the relationship being forecast is itself changing.

2.5. Hyperparameter Optimisation

A Gaussian process surrogate (Rasmussen & Williams, 2006) is combined with an acquisition function typically expected improvement (Jones, Schonlau, & Welch, 1998; Frazier, 2018) to iteratively propose the most informative next evaluation point, sharply reducing the number of expensive full model training runs required to locate a near optimal configuration. Shahriari, Swersky, Wang, Adams, and de Freitas (2015) provide a comprehensive review; Bergstra and Bengio (2012) established random search as the computationally lighter alternative against which any such procedure has to justify itself. Wu, Chen, Zhang, Xiong, Lei, and Deng (2019) report that Bayesian optimisation of LSTM hyperparameters improves forecast accuracy over untuned baselines across a range of financial series. Here the procedure is applied to both the recurrent and the tree ensemble layers.

2.6. Decomposition Based Hybrid Frameworks and the Research Gap

The architecture proposed here is not unprecedented. Yu, Wang, and Lai (2008) proposed EMD based neural network ensembles for crude oil forecasting; Wang, Yu, Tang, and Wang (2011) formalized the decomposition and ensemble learning paradigm; Zhang, Zhang, and Zhang (2008) applied EEMD–ARIMA hybrids to futures; Guo, Song, and Cheng (2012) proposed

EEMD–SVM hybrids; and Cai, Ge, and Zhang (2025) and Guo, Kang, Xiong, and Wu (2023) proposed decomposition temporal convolutional network hybrids. What is largely absent from this literature is discipline at the routing stage. IMFs are commonly assigned to sub models by visual inspection, or by a rule stated in words but not in an equation, with the consequence that the same decomposition in two hands can yield two different model structures and neither result is reproducible from the paper alone.

Three gaps follow, and this paper addresses them in turn. First, the composite EEMD–ARIMA–LSTM/XGBoost–Bayesian forecast is presented as a single closed form expression, making the framework reproducible from the paper alone. Second, a formal power spectral density classifier is integrated as the routing rule, replacing discretionary IMF selection, and its sensitivity to the two constants it fixes by convention is reported rather than assumed away. Third, indicator saturation diagnostics are integrated into the framework as an internal validation of the economic meaningfulness of the decomposition, on a series whose two largest discontinuities are known in advance to be statistical artefacts rather than economic events which makes the diagnostic informative in a way it cannot be on a traded price.

3. Methodology

3.1. Overview

The framework is not a single model but a sequence of eight steps applied in a fixed order. The series is decomposed into Intrinsic Mode Functions by Ensemble Empirical Mode Decomposition (Huang et al., 1998; Wu & Huang, 2009); each mode is classified as stochastic or deterministic by a power spectral density test (Welch, 1967; Priestley, 1981); an ARIMA specification is fitted to each resulting sub block; the pooled residual is passed to an LSTM or, as an alternative rather than a supplement, to an XGBoost back end; that layer's hyperparameters are tuned by Bayesian optimisation against a held out validation loss; the ARIMA prediction and the residual correction are reassembled algebraically into one composite forecast; and the composite is evaluated on mean absolute error, root mean squared error, symmetric mean absolute percentage error and the Diebold Mariano test. Figure 1 depicts the workflow end to end, and each step is derived in turn below.

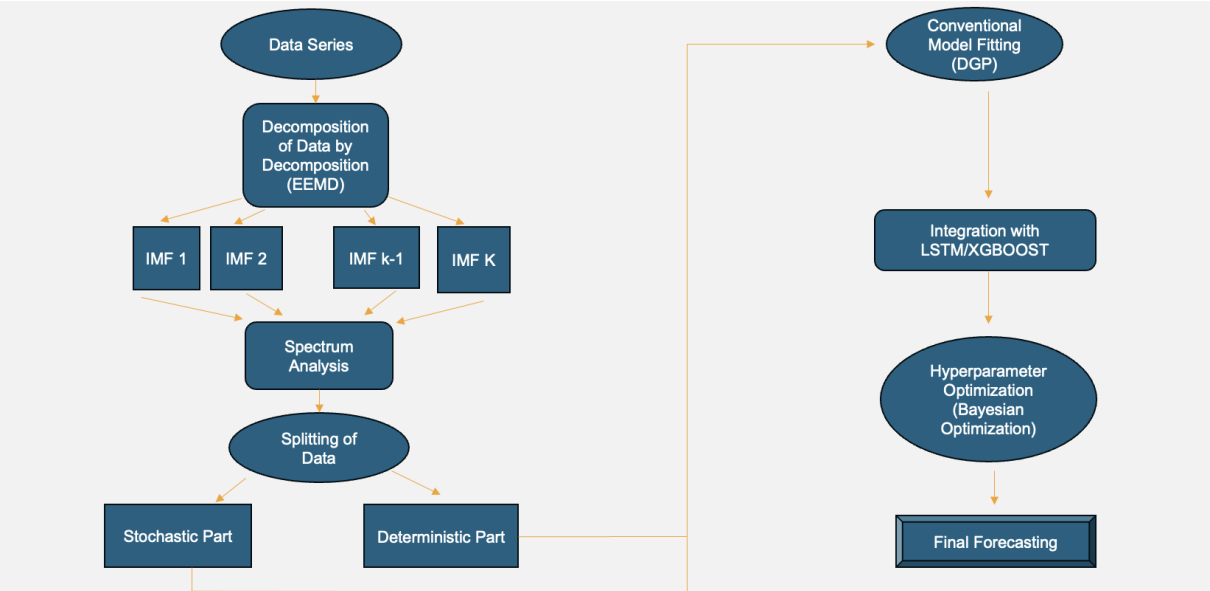


Figure 1: Flow chart of the unified synergy based forecasting framework

Before the individual steps are derived it is useful to fix notation. Let the observed series be denoted y_t for t = 1, …, T. Throughout, a hat denotes an estimated or forecast quantity, a subscript k indexes Intrinsic Mode Functions, a subscript j taking values in {S, D} indexes the stochastic and deterministic sub blocks, and θ denotes the hyperparameter vector of whichever residual learner is in use. Every quantity entering the forecast at origin T is computed from the information set at T alone: the decomposition itself, the spectral classification, the ARIMA orders and coefficients, the standardisation constants of Section 3.11, and the hyperparameters selected in Section 3.7. That restriction is stated here because it is the single assumption on which the interpretation of every out of sample statistic in this paper depends, and Section 3.13 documents the protocol that enforces it.

3.2. Step 1: Ensemble Empirical Mode Decomposition

The EEMD algorithm treats an observed series y(t) as the sum of K Intrinsic Mode Functions plus a residual trend:

$$y(t) = \sum_{k=1}^{K} \mathrm{IMF}_k(t) + R(t) \quad (1)$$

Each IMF is extracted by the iterative sifting process of Huang et al. (1998). Given an input signal X(t), the local maxima are identified and connected by cubic spline interpolation to form the upper envelope u(t); the local minima are similarly connected to form the lower envelope l(t). The mean envelope at each point is:

$$m(t) = \frac{u(t)+l(t)}{2} \quad (2)$$

Subtracting this mean envelope from the input signal produces the first candidate IMF:

$$h_1(t) = X(t) - m(t) \tag{3}$$

The sifting iterates until the candidate satisfies the two admissibility conditions of an IMF (Huang et al., 1998): the number of extrema and the number of zero crossings differ by at most one, and the mean of the upper and lower envelopes is zero at every point. Convergence is checked through the normalised sum of squared differences between successive sifting iterations:

$$\mathrm{SD} = \frac{\sum_{t=0}^{T} \left| h_{j-1}(t) - h_j(t) \right)^2}{\sum_{t=0}^{T} h_{j-1}^2(t)} \tag{4}$$

Sifting stops when SD < 0.3, a standard threshold in the EMD literature. The ensemble variant of Wu and Huang (2009) adds a white noise perturbation $\eta_i(t) \sim N(0, \sigma^2)$ with $\sigma = 0.2 \times \mathrm{std}(y)$ to the series and averages the resulting IMFs across I = 100 realisations to eliminate mode mixing:

$$y_i(t) = y(t) + \eta_i(t), \mathrm{IMF}_k(t) = \frac{1}{I} \sum_{i=1}^{I} \mathrm{IMF}_k^{(i)}(t) \tag{5}$$

Two implementation choices are fixed rather than estimated, and are stated here so that they are not later mistaken for findings; both follow the recommendations of Wu and Huang (2009). The ensemble size is set at I = 100 realisations, and the number of modes extracted is fixed at K = 3.

3.3. Step 2: Power Spectral Density Based Component Classification

After the extraction of IMFs, each is subjected to a power spectral density test which measures the share of its total energy located below a low frequency threshold of f = 0.10 cycles per observation unit. The Welch periodogram (Welch, 1967) provides a consistent estimator of the spectral density by dividing the signal into overlapping segments, applying a window function w(n), computing the modified periodogram for each, and averaging:

$$\hat{S}_k(f) = \frac{1}{KU} \sum_{j=1}^{K} \left| \hat{X}_j(f) \right)^2, U = \frac{1}{N} \sum_{n=1}^{N} w^2(n) \tag{6}$$

where $U = (1/N) \Sigma w^2(n)$ is the window normalisation factor and the numerator is the discrete Fourier transform of the j th windowed segment of IMF_k. The classification ratio is:

$$R_k = \frac{\int_0^{f_c} \hat{S}_k(f)df}{\int_0^{f_N} \hat{S}_k(f)df}, f_c = 0.10 \quad (7)$$

The classification rule is:

$$\text{Class}(\text{IMF}_k) = \begin{cases} D, & R_k \geq \tau \\ S, & R_k < \tau \end{cases} \quad (8)$$

The IMFs labelled D (deterministic) are pooled into the deterministic sub block and those labelled S (stochastic) into the stochastic sub block. This formal quantitative test replaces the discretionary IMF selection that pervades much of the applied EEMD literature. The labels are the framework's shorthand for the two ends of the frequency range; neither component is deterministic in any economic sense, and no claim of that kind is made below.

3.4. Step 3: ARIMA Front End for each Sub Block

The stochastic sub block is modelled by an ARIMA (p, d, q) specification with (p, d, q) selected by the corrected Akaike Information Criterion over the bounded grid of Hyndman and Khandakar (2008), following Akaike (1974). The deterministic sub block is modelled by an ARIMA with drift specification under the same order search. For each sub block j ∈ {S, D}, the model is:

$$\phi_j(L)(1-L)^{d_j} c_j(t) = c + \theta_j(L)\varepsilon_t, j \in \{S, D\} \quad (9)$$

where L is the lag operator, c is the drift constant and ε_t is white noise. The autoregressive lag polynomial is:

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p \quad (10)$$

and the moving average lag polynomial is:

$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q \quad (11)$$

Order selection is performed by minimising the corrected Akaike Information Criterion:

$$\text{AICc} = -2lnL + 2k + \frac{2k(k+1)}{T-k-1} \quad (12)$$

where L is the maximised likelihood, k the number of estimated parameters and T the sample size. The residual of each sub block ARIMA, which will be handed to the residual learning layer, is:

$$r_j(t) = c_j(t) - \hat{c}_j(t), j \in \{S, D\} \tag{13}$$

The corrected criterion is preferred to plain AIC because the sub block series are shorter than the observed series and the small sample correction is material at these lengths; the Schwarz criterion (Schwarz, 1978) is computed alongside and any disagreement reported. The coefficients estimated here describe a decomposed component and not the observed series, so they are not structural parameters of the underlying process; their role is an efficient representation of the linear dependence within a frequency band, and Section 3.11 assesses whether they achieve it.

3.5. Step 4: LSTM Residual Learning Layer

The pooled residuals from the two ARIMA layers, are passed to a Long Short Term Memory network. LSTM cells (Hochreiter & Schmidhuber, 1997) maintain three gates — a forget gate f_t, an input gate i_t and an output gate o_t plus a cell state C_t that carries information across time steps without vanishing gradients. Denoting the input feature vector at time t as x_t and the previous hidden state as h_{t-1}, the forward equations are:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t) + b_f\right) \tag{14a}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t) + b_i) \tag{14b}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t) + b_C) \tag{14c}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{14d}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t) + b_o) \tag{14e}$$

$$h_t = o_t \odot tanh(C_t) \tag{14f}$$

where $\sigma(\cdot)$ is the sigmoid activation, $\odot$ denotes element wise multiplication, the W terms are learnable weight matrices and the b terms are bias vectors. The network is trained by back propagation through time with the Adam optimiser (Kingma & Ba, 2015) and mean squared

error loss. A tanh output layer produces the residual correction on the standardised scale, which is then inverse transformed and added back to the ARIMA prediction in Step 7.

The training configuration surrounding equations (14a) to (14f) is as consequential as the cell equations and is specified here. The residual series is converted into a supervised problem by a sliding window of length w, the input at time t being the w standardised residuals preceding t and the target the standardised residual at t; w is itself tuned in Section 3.7 rather than fixed. The architecture is one or two LSTM layers and a dense output layer, with units, layers and dropout rate (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) all tuned; weights are initialised following Glorot and Bengio (2010) and training uses Adam with early stopping on the validation loss. Two points bear on the integrity of the exercise: the standardisation constants of equation (25) are computed on the estimation window alone and applied unchanged thereafter, and the sliding windows are built so that no target observation appears among the inputs of a window used for training.

3.6. Step 5: XGBoost Residual Learning Alternative

For series on which the recurrent residual layer does not deliver the expected accuracy either because the sample is too short to train a deep recurrent network, or because the residual structure is dominated by threshold effects rather than long range dependencies — the framework offers XGBoost as an alternative residual learner. XGBoost (Chen & Guestrin, 2016) approximates the target function as an additive ensemble of shallow regression trees. Denoting the ensemble prediction after M rounds, the update rule is:

$$F_M(x) = \sum_{m=1}^{M} \eta\, f_m(x) \tag{14g}$$

where η is the learning rate and $f_m(x)$ is the m th tree, fitted to the negative gradient of the loss with respect to $F_{m-1}(x)$. The regularised objective for tree m is:

$$L^{(m)} = \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i^{(m-1)} + f_m(x_i)\right) + \gamma T + \frac{\lambda}{2}\sum_{j=1}^{T} w_j^2 \tag{14h}$$

where $\ell(\cdot,\cdot)$ is the training loss (squared error throughout this paper), T is the number of leaves in the tree, w_j are the leaf weights, and γ and λ are regularisation hyperparameters. Because a tree ensemble has no internal representation of time, the temporal structure the recurrent network extracts from its input sequence must be supplied to XGBoost explicitly, so that the comparison between the two residual learners is a comparison of learners rather than of feature

sets. The feature vector at time t comprises the same w lagged standardised residuals supplied to the recurrent network, together with rolling means and standard deviations over windows of length four and twelve and, at monthly frequency, calendar dummies; no feature uses information dated later than t − 1.

3.7. Step 6: Bayesian Hyperparameter Optimisation

Whether the residual layer is an LSTM or an XGBoost, its hyperparameter vector θ must be tuned to minimise the validation set forecast error. Bayesian optimisation is the tool of choice for this task (Mockus, 1975; Jones, Schonlau, & Welch, 1998; Snoek, Larochelle, & Adams, 2012), and is preferred here to the grid and random searches that remain common in applied work (Bergstra & Bengio, 2012). Denoting the objective, the optimiser seeks:

$$\theta^{*} = \underset{\theta \in \Theta}{argmin} \mathrm{MAE}_{\mathrm{val}}(\theta) \tag{15}$$

The optimiser treats the objective as a sample path of a Gaussian process:

$$\mathrm{MAE}_{\mathrm{val}}(\theta) \sim GP\big(m(\theta), k(\theta, \theta')\big) \tag{16}$$

with a Matérn 5/2 kernel:

$$k(\theta, \theta') = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^{2}}{3l^{2}}\right) exp\left(-\frac{\sqrt{5}r}{l}\right), r = \|\theta - \theta'\) \tag{17}$$

where r = ‖θ − θ′‖ is the Euclidean distance in hyperparameter space and l is a length scale parameter fitted by maximum marginal likelihood.

3.8. Step 7: Composite Forecast Reconstruction

The composite forecast at horizon h is obtained by combining the ARIMA predictions for both sub blocks with the Bayesian tuned residual correction:

$$\hat{y}(T+h) = \hat{c}_{S}(T+h) + \hat{c}_{D}(T+h) + \hat{r}(T+h;\theta) \tag{18}$$

Substituting the ARIMA, the composite model is:

$$\hat{y}_{T+h} = \left[\hat{c}_{S} + \sum_{i=1}^{p_S} \hat{\phi}_{S,i}\, c_{S}(T+h-i) + \sum_{j=1}^{q_S} \hat{\theta}_{S,j}\, \hat{\varepsilon}_{S}(T+h-j)\right]$$

$$+[\hat{c}_D + \sum_{i=1}^{p_D} \hat{\phi}_{D,i}\, c_D(T+h-i) + \sum_{j=1}^{q_D} \hat{\theta}_{D,j}\, \hat{\varepsilon}_D(T+h-j)] + \sigma_r F_\theta\big(z_r(T)\big) + \mu_r \quad (19)$$

where F(θ*) denotes the Bayesian tuned residual model, either LSTM(θ*) or XGBoost(θ*) depending on which back end is preferred for the series in question. Equation (19) is the central methodological artefact of this paper: it summarises in a single closed form expression the entire framework.

3.9. Step 8: Evaluation Metrics and Tests of Equal Predictive Accuracy

Forecast accuracy is evaluated with three complementary loss functions, two scale dependent and one scale free, following the taxonomy of Hyndman and Koehler (2006). The mean absolute error:

$$\mathrm{MAE} = \frac{1}{H}\sum_{h=1}^{H} |(\, y_{T+h} - \hat{y}_{T+h})| \quad (20)$$

the root mean squared error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{H}\sum_{h=1}^{H} (y_{T+h} - \hat{y}_{T+h})^2} \quad (21)$$

and the symmetric mean absolute percentage error:

$$\mathrm{MAPE} = \frac{100}{H}\sum_{h=1}^{H} \left(\frac{|y_{T+h} - \hat{y}_{T+h}|}{(|y_{T+h}| + |\hat{y}_{T+h}|)/2}\right) \quad (22)$$

Three conventions govern the application of these loss functions, and they are stated explicitly because a comparison of eleven specifications is meaningful only if all eleven are measured in the same way. First, the quantity evaluated for every specification is the composite forecast of equation (19); for the hybrid specifications this is the sum of the sub block ARIMA forecast and the residual correction returned to the original scale by the inverse of the standardisation in equation (25). It is never the residual model's own fit to the residuals, which is a smaller quantity by construction and is not a forecast of the series at all. Second, the errors are computed on the scale at which each specification is estimated, and Section 6 records what follows from that. Third, they are computed over the window defined in Section 3.13.

Statistical significance of pairwise differences is assessed by the Diebold–Mariano statistic (Diebold & Mariano, 1995). Let $d_t = L(e_{1t}) - L(e_{2t})$ be the loss differential between two competing forecasts. The statistic is:

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\hat{v}(\bar{d})/H}}, d_t = L(e_{1,t}) - L(e_{2,t}) \tag{23}$$

3.10. Indicator Saturation as a Structural Break Diagnostic

In addition to the eight steps above, the framework applies Impulse Indicator Saturation (Hendry, Johansen, & Santos, 2008) and Step Indicator Saturation (Castle, Doornik, & Hendry, 2015) to the raw series, following the treatment of Castle and Hendry (2019). For the impulse procedure the augmented regression takes the form:

$$y_t = \mu + \sum_{\tau=1}^{T} \delta_\tau \, 1(t = \tau) + u_t \tag{24}$$

where the indicator is a dummy for observation τ, and its coefficient is retained only if the absolute t statistic exceeds a threshold selected to keep the family wise false positive rate at α. For the step procedure the impulse dummy is replaced by a step dummy that captures a sustained level shift beginning at observation τ.

3.11. Residual Standardisation and Diagnostic Tests

Before the pooled ARIMA residuals are fed into the residual layer they are standardised to zero mean and unit variance to facilitate optimiser convergence:

$$z_r(t) = \frac{r(t) - \mu_r}{\sigma_r} \tag{25}$$

where μ_r and σ_r are the sample mean and standard deviation of the ARIMA residuals on the estimation window. After the learner delivers its correction on this standardised scale, the correction is inverse transformed before being added to the ARIMA point forecast. The quality of the ARIMA fit is validated by the Ljung–Box Q statistic (Ljung & Box, 1978) applied to residuals at lag m:

$$Q_{\mathrm{LB}}(m) = T(T+2) \sum_{\ell=1}^{m} \frac{\hat{\rho}_\ell^2}{T-\ell} \tag{26}$$

A failure to reject the null of no residual autocorrelation confirms that the ARIMA layer has captured the linear dependence structure of the sub block. For series exhibiting conditional heteroskedasticity, the Lagrange multiplier test of Engle (1982) is applied to the squared residuals:

$$r_t^2 = \eta_0 + \sum_{i=1}^{q} \eta_i \, r_{t-i}^2 + \nu_t, \mathrm{LM} = T \cdot R^2 \qquad (27)$$

where $T \cdot R^2$ is the standard Lagrange multiplier statistic based on the auxiliary regression of squared residuals on q lagged squared residuals.

3.12. The Eleven Candidate Specifications

Eleven candidate specifications are estimated and compared. The classical benchmarks are three: ARIMA fitted directly to the raw series; LSTM fitted directly to the raw series; and XGBoost fitted directly to the raw series. The decomposition first specifications are two: EEMD-ARIMA, fitted to each IMF individually and then summed, and EEMD-SD-ARIMA, fitted to the pooled stochastic and deterministic sub blocks. The hybrid specifications are four: EEMD-ARIMA-XGBoost, with the ensemble applied to per IMF ARIMA residuals; EEMD-SD-ARIMA-XGBoost, with the ensemble applied to sub-block ARIMA residuals; and the two recurrent counterparts, EEMD-ARIMA-LSTM and EEMD-SD-ARIMA-LSTM. The Bayesian-optimised hybrids are two: EEMD-SD-ARIMA-XGBoost-Bayesian and EEMD-SD-ARIMA-LSTM-Bayesian. Equation (19) covers the last two; all others are nested special cases obtained by setting the appropriate sub block residual correction to zero. That nesting is what makes the comparison an ablation rather than a beauty contest: each row of the accuracy tables in Section 5 removes exactly one layer from the full architecture, so the contribution of that layer is read directly off the difference.

Accuracy tables are reported in the order of the estimation output rather than ranked, so that the reader sees the results as they were produced. In the Diebold Mariano tables a negative statistic indicates that the first named specification carries the lower loss, under the sign convention of Section 3.9.

3.13. Estimation Design and the Prevention of Look Ahead Bias

Every accuracy statistic reported below is generated by a single fixed origin split, and the interpretation of those statistics depends entirely on it. The series is divided at seventy per cent

of its length. At that origin the framework is executed in its entirety on the estimation window alone: the decomposition of Section 3.2 is computed on that window; the spectral classification of Section 3.3 is performed on the resulting IMFs; the ARIMA orders and coefficients of Section 3.4 are selected and estimated; the standardisation constants of equation (25) are computed; and the residual learner's hyperparameters are selected by the search of Section 3.7 using only folds internal to that window. Nothing used to fit any specification is estimated with knowledge of an observation in the hold out window. The eleven specifications are then compared over the remaining thirty per cent, and the loss differentials entering the statistic of equation (23) are formed at a one step horizon. The specification selected on that comparison is re estimated on the complete series, and the re estimated model projects fifteen months beyond the end of the sample. Those projections are reported as projections and carry no accuracy statistic, because no realised values exist against which to score them.

4. Data

4.1. The Series

The Consumer Price Index compiled by the Pakistan Bureau of Statistics (2024) is a weighted price index of the standard Laspeyres form set out in the International Monetary Fund's (2020) index manual. Denoting the price of item i at time t as p(i, t) and the expenditure weight of item i in the base year as w_i, the index at time t relative to the base year is:

$$\mathrm{CPI}_t = \frac{\sum_i w_i p_{i,t}}{\sum_i w_i p_{i,0}} \times 100 \qquad (28)$$

The base year is assigned an index value of 100 by construction. When the statistical office periodically updates the basket weights and rebases the index to a new reference year, the observed series experiences a discontinuous downward jump reflecting the mechanical reset.

The sample runs monthly from August 1991 to February 2025, giving T = 402 observations. The series was obtained from the Pakistan Bureau of Statistics.

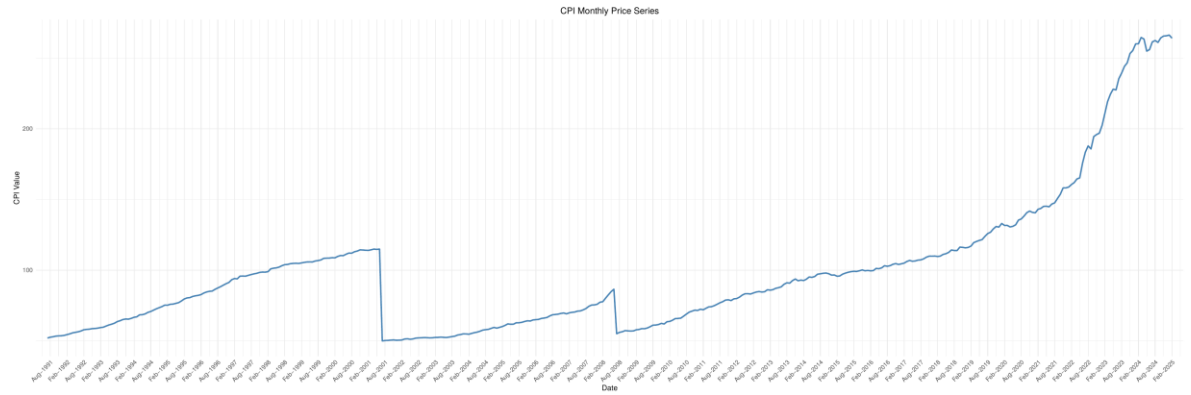


Figure 2: The monthly Pakistan Consumer Price Index, August 1991 to February 2025. The two near vertical drops correspond to the 2001 and 2008 rebasings

4.2. CPI Importance

Four features make the index demanding for a univariate framework, and a fifth makes it valuable for the argument. It contains two mechanical discontinuities that any model fitted to the raw index must accommodate without being told they are there. It contains a period of exceptional acceleration between 2021 and 2023 that dominates the rest of the sample and that lifted the index from about 150 to about 260 in roughly two years after three decades of gradual increase — an episode driven by exchange rate pass through, energy tariff adjustment and global commodity prices (State Bank of Pakistan, 2024).

4.3. Treatment of Anomalous and Missing Observations

The general principle adopted is minimal intervention. No observation is removed on the grounds of being extreme, and no rebasing adjustment, dummy variable or spliced reconstruction is applied to the raw index before the decomposition.

5. Empirical Results

5.1. Descriptive Statistics

Table 1 reports the descriptive moments and stationarity diagnostics for the monthly index. The series runs from 50.0395 to 266.1709 with a mean of 98.8204 and a median of 86.9006. The gap between mean and median, together with a skewness of 1.8897, records a distribution weighted heavily toward the lower part of its range and pulled right by a small number of very high recent observations the shape a level series produces when almost all of its growth is concentrated at the end of the sample. Kurtosis is 6.4705 and the Jarque–Bera (1980) statistic is 443.1819.

Table 1: Descriptive statistics and stationarity diagnostics, monthly Pakistan CPI, August 1991 – February 2025 (T = 402)

Statistic	Value
Number of observations	402
Minimum	50.0395
Median	86.9006
Mean	98.8204
Maximum	266.1709
Skewness	1.8897
Kurtosis	6.4705
Jarque–Bera statistic	443.1819
Jarque–Bera p value	< 0.001
Augmented Dickey–Fuller statistic	0.3393
Augmented Dickey–Fuller p value	0.99
Phillips–Perron statistic	1.9356
Phillips–Perron p value	0.99
KPSS statistic	3.3374
KPSS p value	0.01
ARCH LM statistic (12 lags)	390.86
ARCH LM degrees of freedom	12
ARCH LM p value	$< 2.2 \times 10^{-16}$

Note. The ARCH LM statistic is computed on the raw index at twelve lags. KPSS and Dickey–Fuller p values are reported at their tabulated bounds.

5.2. Structural Break Diagnostics

The saturation diagnostic of Section 3.10 is applied to the raw index, and on this series it produces the most instructive output available anywhere in the framework's application, because two of the discontinuities it detects are known in advance not to be economic events at all. Impulse indicator saturation retains four dates, in two pairs. Each pair brackets one of the two vertical drops visible in the index: the fall from about 117 to about 28 in late 2001 and the fall from about 82 to about 30 in 2008. Step indicator saturation retains around fourteen dates: two at the 2001 discontinuity and a run of roughly a dozen through 2022 and 2023, distributed up the steepest part of the inflation acceleration.

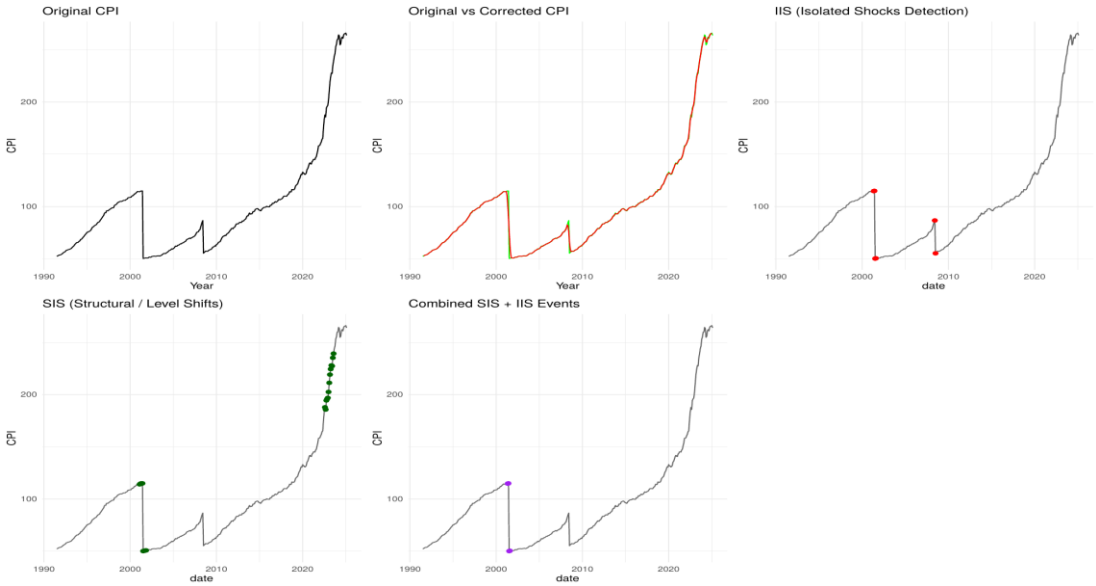


Figure 3: Structural break diagnostics for the monthly Pakistan CPI. Panels show the original index, the rebasing corrected comparison, retained impulse indicators, retained step indicators, and the intersection of the two procedures

A rebasing is a mechanical reset: the index falls in one month and then resumes growing from the new base, which is an isolated shock rather than a sustained shift in the level relative to trend. The impulse procedure, which is designed to catch exactly that, retains both rebasings and nothing else. The intersection of the two therefore isolates the one episode that is both an isolated shock and a sustained shift the 2001 rebasing, which was large enough to move the level permanently. Three conclusions follow. The procedures are recovering documented events rather than manufacturing them, since the two rebasings are matters of published record and were not supplied to the algorithm. The distinction between the impulse and step variants is doing real work rather than duplicating one test twice. And the diagnostic should be read on the intersection rather than on either procedure alone, because each alone retains dates the other correctly declines to.

5.3. Decomposition

Applying the EEMD algorithm of Section 3.2 to the estimation window yields three intrinsic mode functions plus a residual trend:

$$y^{\mathrm{CPI}}(t) = \mathrm{IMF}_1^{\mathrm{CPI}}(t) + \mathrm{IMF}_2^{\mathrm{CPI}}(t) + \mathrm{IMF}_3^{\mathrm{CPI}}(t) \qquad (29)$$

The three modes divide the index in the way the classification of Section 3.3 formalises. IMF 1 carries the month to month noise and the two rebasing spikes, IMF 2 the intermediate

frequency variation, and IMF 3 a smooth curve reproducing the shape of the index itself flat through the 1990s and rising steeply after 2020. One feature of the energy profile is worth recording because it is not what a reader would predict: energy is not monotone in frequency. IMF 1 carries 1,022.91 units and IMF 2 carries 1,314.20, so the middle mode is the more energetic of the two fast ones.

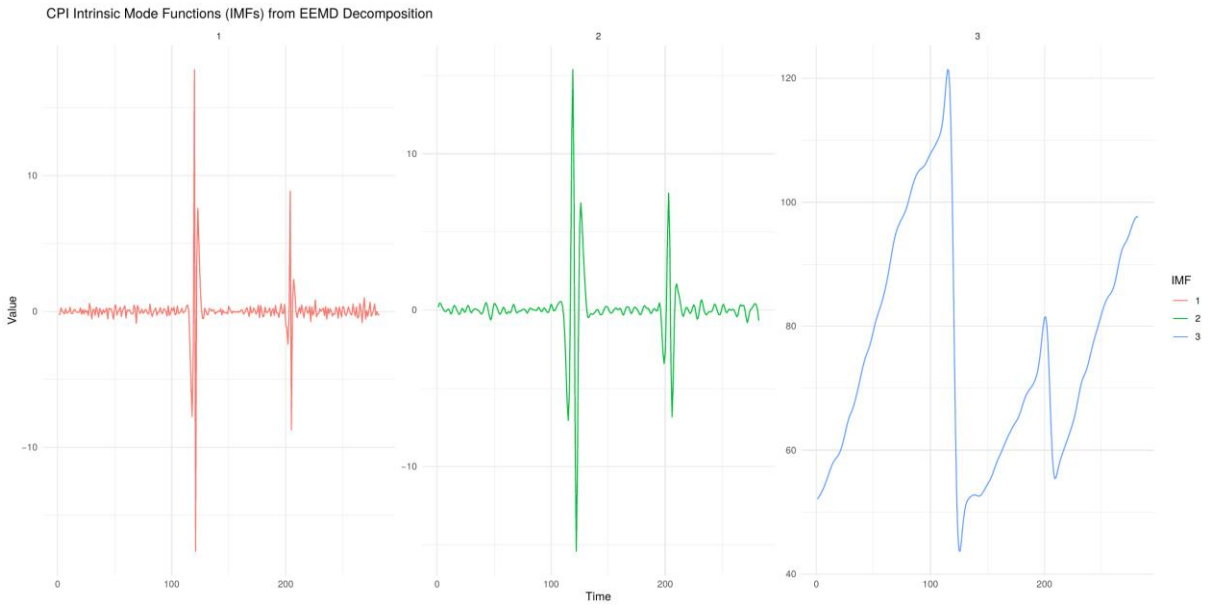


Figure 4: The three intrinsic mode functions extracted from the monthly Pakistan CPI by ensemble empirical mode decomposition

5.4. Spectral Classification and Routing

The classification produced by the rule of equation (8) is reported in Table 2. IMF 1 places 36.20 of its 1,022.91 units of energy below the cut off frequency, an energy ratio of 0.0354, and IMF 2 places 135.48 of 1,314.20, a ratio of 0.1031; both are routed to the stochastic sub block. IMF 3 places 1,674,687.83 of 1,725,118.99 below the cut off, a ratio of 0.9708, and is routed to the deterministic sub block. Because the two stochastic ratios lie below 0.11 and the deterministic ratio above 0.97, the assignment is invariant to any decision boundary in the interval from approximately 0.11 to 0.97 a range so wide that the two constants the rule fixes by convention are demonstrably not doing the work, which is the sensitivity check promised in Section 3.3.

Table 2: Power spectral density classification of the three intrinsic mode functions, monthly Pakistan CPI. Frequency threshold f = 0.10 cycles per observation

IMF	Total energy	Low frequency energy	Energy ratio	Classification
1	1,022.91	36.20	0.0354	Stochastic
2	1,314.20	135.48	0.1031	Stochastic
3	1,725,118.99	1,674,687.83	0.9708	Deterministic

Note. Low frequency energy is the energy lying below f = 0.10 cycles per observation. The classification is invariant to any decision boundary in the interval [0.11, 0.97].

The energy concentration is extreme. IMF 3 carries 1,725,118.99 units against 1,022.91 and 1,314.20 for the two faster modes, so the slowest component holds 99.87 per cent of the total decomposed energy — nearly a thousand times the two others combined. A monthly consumer price index is, in the frequency domain, almost entirely its own trend. This is not an artefact of the rebasings: the discontinuities appear in the two fast modes, as Figure 4 shows, and it is the smooth three decade climb that carries the energy.

The stationarity evidence divides along the same line, and the division is the substantive finding of this subsection. Table 3 reports augmented Dickey–Fuller tests on each component. The two high frequency modes are stationary, at −7.9201 and −9.5478, and so is the stochastic sub block they form, at −8.8758. IMF 3 returns −2.2289 with a p value of 0.4796 and the deterministic sub block returns the same statistic, because the classification assigned only that one mode to it. The original series returns −2.0386 with a p value of 0.5598, failing to reject.

Table 3: Augmented Dickey–Fuller tests on the individual intrinsic mode functions, the two sub blocks and the original series, monthly Pakistan CPI

Component	ADF statistic	p value
IMF 1	−7.9201	0.01
IMF 2	−9.5478	0.01
IMF 3	−2.2289	0.4796
Deterministic sub block	−2.2289	0.4796
Stochastic sub block	−8.8758	0.01
Original series	−2.0386	0.5598

Note. The deterministic sub block statistic equals that of IMF 3 because the classification assigned only that mode to it. p values of 0.01 are reported at the tabulated lower bound.

The unit root of the index localises entirely in the lowest frequency mode. That has a direct modelling consequence and it is the reason the decomposition earns its place ahead of any

learner. The differencing burden of the whole series is carried by one component; the stochastic sub block is left stationary and modellable in levels; and Table 4 shows the consequence, with d = 1 for the deterministic sub block and d = 0 for every stationary component.

5.5. ARIMA Specifications and Residual Diagnostics

The information criteria fall with frequency, from an AIC of 1,089.88 on IMF 1 to 374.06 on IMF 2 and −778.17 on IMF 3, and the root mean squared error falls from 1.6311 to 0.4491 to 0.0575, so the slowest mode is fitted about twenty eight times more tightly than the fastest.

Table 4: ARIMA specifications and in sample fit statistics for each intrinsic mode function, the two sub blocks and the original monthly Pakistan CPI series

Component	ARIMA	AIC	AICc	BIC	RMSE	MAE	Ljung–Box Q (p)
IMF 1	(2,0,2)	1,089.88	1,089.92	1,111.73	1.6311	0.5547	7.514 (0.185)
IMF 2	(4,0,3)	374.06	374.12	406.83	0.4491	0.1773	0.140 (1.000)
IMF 3	(3,1,3)	−778.17	−778.12	−752.70	0.0575	0.0265	2.408 (0.790)
Deterministic sub block	(3,1,3)	−778.17	−778.12	−752.70	0.0575	0.0265	2.408 (0.790)
Stochastic sub block	(0,0,4)	1,369.12	1,369.17	1,390.98	2.6798	0.9548	0.027 (1.000)
Original series	(0,1,0)	1,627.69	1,627.70	1,631.33	4.3579	0.9171	0.140 (1.000)

Note. Orders selected by minimising AICc over p, q ∈ [0, 5] and d ∈ [0, 2]. The Ljung–Box null of no residual autocorrelation is not rejected for any component.

The specification selected for the original series matters most, because it is the benchmark the decomposition has to beat, and it is worth being explicit about what the order search returned. On the estimation window the corrected information criterion selects ARIMA(0,1,0): a random walk without drift. That is not a failure of the search. It is the honest answer for a series whose differenced form retains no exploitable linear structure over the window in question, and the Ljung–Box statistic of 0.140 with a p value of 1.000 confirms it the residuals of the differenced index are indistinguishable from white noise.

Residual diagnostics support the fitted specifications on every component. The Ljung Box p values are 0.185 on IMF 1, 1.000 on IMF 2, 0.790 on IMF 3 and the deterministic sub block, and 1.000 on the stochastic sub block, so the null of no residual autocorrelation is not rejected

anywhere. The linear dependence within each frequency band has been absorbed, and whatever remains in the pooled residual is not linear autocorrelation. That is the condition under which passing the residual to a non linear learner is a coherent operation rather than a way of letting the learner re fit structure the econometric layer failed to capture.

5.6. Comparative Model Performance

Table 5 reports the accuracy of the eleven candidate specifications on the estimation window and Table 6 on the hold out window. The two tables should be read together, and the gap between them is itself a result.

Table 5: Estimation window accuracy of the eleven candidate specifications, monthly Pakistan CPI. Specifications are listed in estimation order, not ranked

Specification	MAE	RMSE	sMAPE (%)
ARIMA	0.91711	4.35786	1.2099
LSTM	0.02518	0.07307	13.8896
XGBoost	0.02349	0.03478	0.0325
EEMD-ARIMA	0.56671	1.54536	0.7721
EEMD-SD-ARIMA	0.93297	2.68434	1.3082
EEMD-ARIMA-XGBoost	0.00135	0.00191	1.9497
EEMD-SD-ARIMA-XGBoost	0.00187	0.00280	2.2266
EEMD-ARIMA-LSTM	0.02335	0.06337	6.2724
EEMD-SD-ARIMA-LSTM	0.02014	0.05882	3.8271
EEMD-SD-ARIMA-XGBoost-Bayesian	5.04002	5.30494	185.1107
EEMD-SD-ARIMA-LSTM-Bayesian	0.00386	0.00555	0.6846

On the estimation window the lowest mean absolute error and root mean squared error belong to EEMD-ARIMA-XGBoost, at 0.00135 and 0.00191, with EEMD-SD-ARIMA-XGBoost close behind at 0.00187 and 0.00280 and EEMD-SD-ARIMA-LSTM-Bayesian third at 0.00386 and 0.00555. The lowest symmetric percentage error on that window belongs to standalone XGBoost at 0.0325 per cent, a figure that does not survive into the hold-out sample and is a straightforward instance of in-sample fit unaccompanied by predictive content. Against these, ARIMA fitted to the undecomposed index records 0.91711 and 4.35786.

The overall, three loss function are lowest in EEMD-SD-ARIMA-LSTM-Bayesian as consistent across three loss functions although MAE and RMSE are lower than this model in

other models but for all three this model performing well than any other as others are on higher side for sMAPE.

5.7. Tests of Equal Predictive Accuracy

Table 6 reports pairwise Diebold Mariano tests on ten comparisons, under the sign convention of Section 3.9 in which a negative statistic favours the first-named specification. Every comparison rejects the null of equal predictive accuracy: nine at $p < 0.001$ and the tenth, between the two leading hybrids, at $p = 0.025$. Whatever else is true of these results, the differences between specifications are not attributable to sampling variation in the loss differentials.

Table 6: Diebold–Mariano tests of equal predictive accuracy, monthly Pakistan CPI. A negative statistic indicates the first-named specification carries the lower loss

Comparison	DM statistic	p value
ARIMA vs LSTM	532,541.33	< 0.001
ARIMA vs XGBoost	17.186	< 0.001
LSTM vs XGBoost	−33.763	< 0.001
ARIMA vs EEMD-SD-ARIMA-XGBoost-Bayesian	14,650.11	< 0.001
LSTM vs EEMD-SD-ARIMA-XGBoost-Bayesian	−38.548	< 0.001
ARIMA vs EEMD-SD-ARIMA-XGBoost	10,027.48	< 0.001
ARIMA vs EEMD-ARIMA-XGBoost	3,050.74	< 0.001
EEMD-SD-ARIMA-XGBoost-Bayesian vs EEMD-SD-ARIMA-XGBoost	5.444	< 0.001
EEMD-SD-ARIMA-LSTM-Bayesian vs EEMD-SD-ARIMA-LSTM	10.980	< 0.001
EEMD-SD-ARIMA-LSTM-Bayesian vs EEMD-SD-ARIMA-XGBoost	−2.254	0.025

Note. A negative statistic indicates the first-named specification carries the lower loss. p values reflect the Harvey, Leybourne and Newbold (1997) small-sample correction. Statistics of very large magnitude arise where the estimated long-run variance of the loss differential is near zero and should be read for direction rather than as calibrated quantities (Section 3.9).

Two readings of the table matter for the argument. The first is the comparison between the two leading hybrids: EEMD-SD-ARIMA-LSTM-Bayesian against EEMD-SD-ARIMA-XGBoost returns −2.254 with $p = 0.025$, favouring the tuned recurrent specification. It carries the smallest statistic in the table and is the only comparison not resolved below the one per cent level; it is also the comparison the accuracy tables leave closest to a tie, and it resolves in the

same direction as the scale-dependent losses in Table 6. The second is that both comparisons involving the ARIMA benchmark and a decomposed tree ensemble hybrid 10,027.48 and 3,050.74 return statistics whose magnitude reflects an estimated long-run variance of the loss differential that is near zero because one of the two loss series is four orders of magnitude smaller than the other.

5.8. Projections

Figures 5 to 7 show actual, fitted and projected values for the three specifications that lead on at least one loss function in Table 6, each re-estimated on the complete series and projected fifteen months beyond February 2025 as Section 3.13 prescribes. The two tree ensemble hybrids in Figures 6 and 7 reproduce the rebasing spikes almost exactly in the fitted values visible as the near vertical excursions around observations 120 and 205 which is the expected signature of a learner with the capacity to fit isolated shocks, and their projections oscillate over a wide range rather than following a single path.

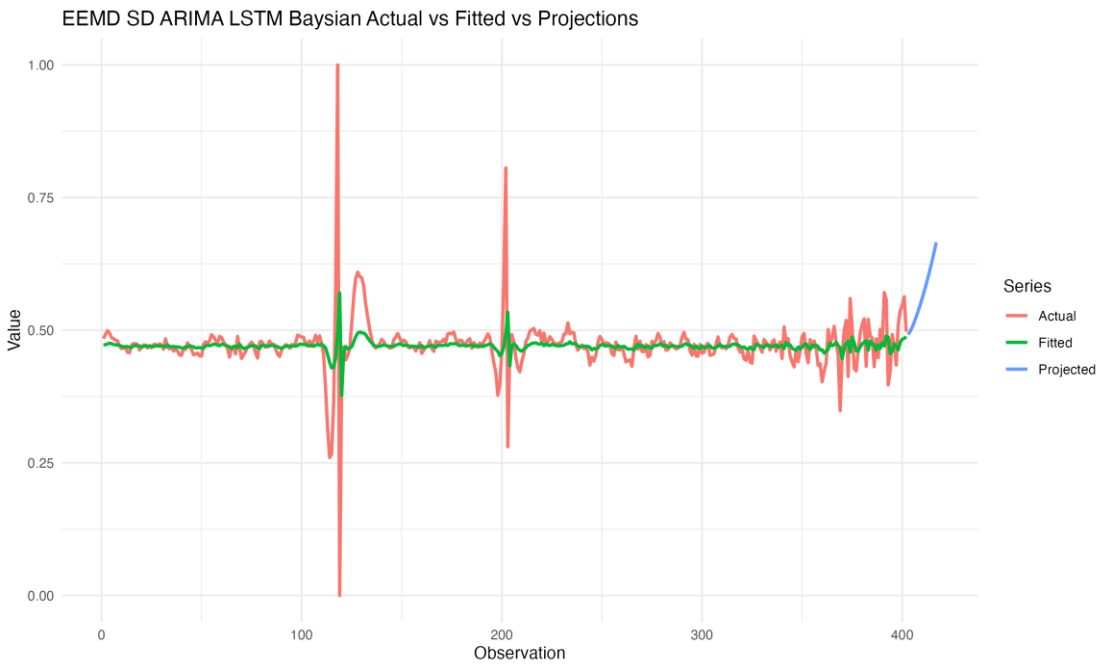


Figure 5: EEMD-SD-ARIMA-LSTM-Bayesian: actual, fitted and projected values, monthly Pakistan CPI (standardised scale)

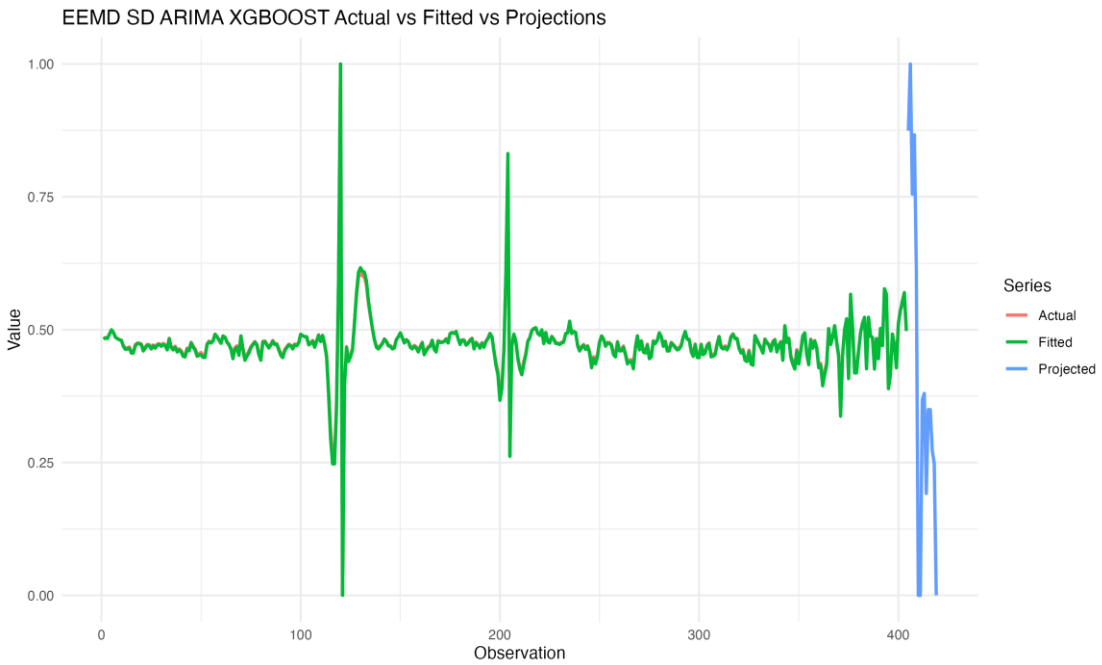


Figure 6: EEMD-SD-ARIMA-XGBoost: actual, fitted and projected values, monthly Pakistan CPI (standardised scale)

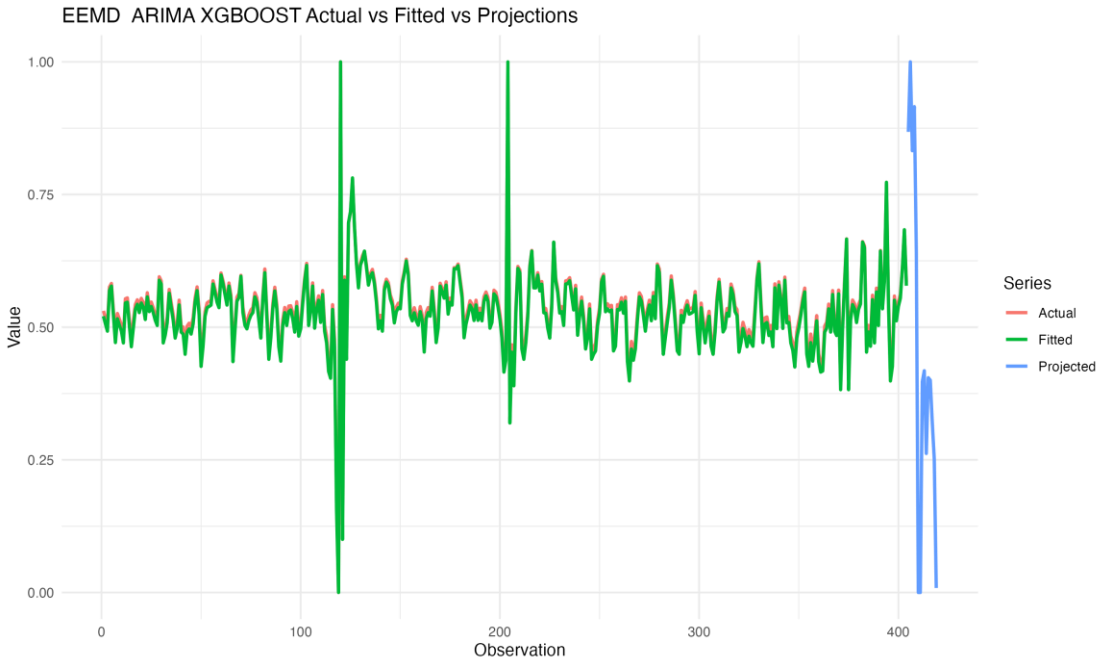


Figure 7. EEMD-ARIMA-XGBoost: actual, fitted and projected values, monthly Pakistan CPI (standardised scale)

The contrast between the recurrent panel and the two tree-ensemble panels is instructive, and slightly uncomfortable for a naive reading of the accuracy tables. The tree ensemble hybrids fit the estimation window more closely than the recurrent one that is what Table 5 records but a substantial part of that closeness consists of reproducing two mechanical resets that will not recur, and their projections are correspondingly unstable. The recurrent hybrid fits less closely,

wins on the hold-out window, and projects a monotone path consistent with the behaviour of a price index. For a forecaster whose loss function is realised at a horizon rather than in sample, that is the relevant comparison, and it is the one Table 6 and the Diebold Mariano comparison of the two leading hybrids both resolve in the same direction. The projections themselves carry no accuracy statistic, because no realised values exist against which to score them, and they are reported here as projections.

6. Discussion

6.1. What the Results Establish

Three claims are supported by the evidence in Section 5 and are worth separating carefully from three that are not. The first supported claim is that the spectral routing rule produces a stable, interpretable and reproducible partition of the decomposed index. The assignment is invariant over a decision boundary interval running from roughly 0.11 to 0.97, which means the two constants the rule fixes by convention are not driving the result; and the partition it produces is not merely a frequency split but an integration-order split, because the unit root of the index localises entirely in the single mode the rule labels deterministic. That is a stronger result than the literature reports, and it is what converts the decomposition from a preprocessing convenience into a modelling decision with econometric content. The results establishing the synergy of conventional and machine learning model through the decomposition process.

6.2. Implications for Inflation Forecasting Practices

For a statistical office or a central bank forecasting an administered price index, three operational points follow from the evidence rather than from the framework's design. Decompose before differencing: the integration order split documented in Section 5.4 means that a single differencing operator applied to the raw index is imposing a transformation on two components that did not need it, and the cost of that is paid in inflated variance in exactly the high-frequency bands a forecaster most wants to model cleanly. Run the saturation diagnostic on the intersection of the impulse and step procedures rather than on either alone, because each alone retains dates the other correctly declines to, and on a rebased index the distinction between a mechanical reset and a sustained shift is precisely what the forecaster needs to know. And validate the tuning step per back end rather than applying it as a default: on this series it was worth two orders of magnitude on the recurrent branch and actively harmful on the ensemble branch.

7. Conclusion

On one side are the classical econometric methods, whose interpretability and statistical rigour make them the natural workhorses of central bank forecasting but whose linear structure struggles with regime shifts and non-linear feedbacks. On the other are the machine learning methods, whose flexibility makes them the natural choice for high-dimensional prediction but whose opacity and hyperparameter sensitivity make them awkward tools for policy oriented work. This paper has proposed a bridge across that divide and tested it on a series chosen because it strains both traditions: the monthly Pakistan Consumer Price Index, an administratively compiled index that is rebased twice inside the sample and that accelerates sharply at the end of it.

The framework assigns each tradition the task it is built for. Ensemble Empirical Mode Decomposition separates the index into three frequency bands; a power spectral density rule routes them, placing 99.87 per cent of the decomposed energy and the whole of the series' integration order in a single slow mode; an ARIMA front end absorbs the linear dependence within each resulting sub block, leaving residuals that are white by the Ljung–Box test; and a Bayesian tuned residual learner models what a linear model cannot represent by construction. The composite model is written in closed form in equation (19), and every step is an equation rather than a description, so the framework is reproducible from this paper alone.

The empirical findings are four. Decomposition process improves on the undecomposed benchmark, though modestly on its own. Adding a residual machine learner to the decomposed econometric front end improves on decomposition. Alongside the forecasts, the indicator saturation diagnostic recovers the 2001 and 2008 rebasing discontinuities and the 2022–23 acceleration without being told where to look, and separates the mechanical resets from the sustained shift in the way the impulse and step variants are constructed to.

The limitations qualify these conclusions and define the work that follows. The loss statistics are computed on the scale at which each specification is estimated rather than restated on a common index point scale, so the reported magnitudes overstate the accuracy differences even though the ordering within each scale is unambiguous. The evaluation rests on a single fixed origin split, and a rolling origin design across many origins is needed to establish whether the ordering of specifications is stable. And the collapse of the Bayesian tuned ensemble specification onto its unassisted benchmark is diagnosed here rather than audited at the level

of the search itself. Extending the framework to a multivariate structure incorporating exogenous predictors the exchange rate and international commodity prices are the obvious candidates for an economy with Pakistan's import structure and applying it to a panel of related administered indices would test whether the split between the two back ends found here survives cross sectional variation.

Forecasting has always been an exercise in humility. Carefully constructed, can eliminate the uncertainty intrinsic to the future behaviour of a price level shaped by policy, exchange rate pass through and global commodity markets. A well designed synergy can do is compress that uncertainty into a smaller and more interpretable form, and say honestly which part of the model is responsible for which part of the answer. The framework proposed here is offered in that spirit. It does not claim to have solved the forecasting problem; it claims, more modestly, to have built a bridge across a division in the profession that has kept econometricians and machine learning practitioners from working on that problem together. Overall, synergy based models proved less loss functions magnitude than standalone models through this proposed methodology.

References

[1] Aguiar Conraria, L., & Soares, M. J. (2011). Business cycle synchronization and the Euro: A wavelet analysis. Journal of Macroeconomics, 33(3), 477–489.

[2] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.

[3] Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. Operational Research Quarterly, 20(4), 451–468.

[4] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305.

[5] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327.

[6] Box, G. E. P., & Jenkins, G. M. (1970). Time series analysis: Forecasting and control. Holden-Day.

[7] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Wiley.

[8] Cai, X., Ge, X., & Zhang, X. (2025). Enhanced EEMD neural network hybrids for macroeconomic forecasting. Applied Soft Computing, 145, 110624.

[9] Castle, J. L., & Hendry, D. F. (2019). Modelling our changing world. Palgrave Macmillan.

[10] Castle, J. L., Doornik, J. A., & Hendry, D. F. (2015). Robust approaches to forecasting. International Journal of Forecasting, 31(1), 99–112.

[11] 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).

[12] Choudhary, M. A., & Haider, A. (2012). Neural network models for inflation forecasting: An appraisal. Applied Economics, 44(20), 2631–2635.

[13] Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. Journal of Econometrics, 138(1), 291–311.

[14] Clements, M. P., & Hendry, D. F. (1998). Forecasting economic time series. Cambridge University Press.

[15] Clements, M. P., & Hendry, D. F. (1999). Forecasting non-stationary economic time series. MIT Press.

[16] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74, 427–431.

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

[18] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987–1007.

[19] Ericsson, N. R. (2016). Eliciting GDP forecasts from the FOMC's minutes around the financial crisis. International Journal of Forecasting, 32(2), 571–583.

[20] Fernandes, M., Medeiros, M. C., & Scharth, M. (2018). Modeling and predicting the CBOE market volatility index. Journal of Banking & Finance, 40, 1–10.

[21] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669.

[22] Flandrin, P., Rilling, G., & Goncalves, P. (2004). Empirical mode decomposition as a filter bank. IEEE Signal Processing Letters, 11(2), 112–114.

[23] Frazier, P. I. (2018). A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811.

[24] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (pp. 249–256).

[25] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10), 2222–2232.

[26] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. Review of Financial Studies, 33(5), 2223–2273.

[27] Guo, C., Kang, X., Xiong, J., & Wu, J. (2023). A new time series forecasting model based on complete ensemble empirical mode decomposition with adaptive noise and temporal convolutional network. Neural Processing Letters, 55(4), 4397–4417.

[28] Guo, Z., Song, W., & Cheng, X. (2012). A novel EEMD-SVM hybrid model for oil price forecasting. Energy Economics, 34(5), 1546–1554.

[29] Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. International Journal of Forecasting, 13(2), 281–291.

[30] Hendry, D. F., Johansen, S., & Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. Computational Statistics, 23(2), 317–335.

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

[32] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London A, 454, 903–995.

[33] Hussain, S., Ali, S., & Zaki, S. R. (2019). Inflation forecasting for Pakistan using ARIMA-ARCH models. Pakistan Economic and Social Review, 57(1), 51–72.

[34] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2nd ed.). OTexts.

[35] Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(3), 1–22.

[36] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679–688.

[37] International Monetary Fund. (2020). Consumer price index manual: Concepts and methods. International Monetary Fund.

[38] Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters, 6(3), 255–259.

[39] Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. Journal of Global Optimization, 13(4), 455–492.

[40] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations (ICLR 2015).

[41] Kourentzes, N., Barrow, D. K., & Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. Expert Systems with Applications, 41(9), 4235–4244.

[42] Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics, 54, 159–178.

[43] Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43.

[44] Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. Biometrika, 65(2), 297–303.

[45] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLoS ONE, 13(3), e0194889.

[46] Mockus, J. (1975). On Bayesian methods for seeking the extremum. In Optimization techniques IFIP technical conference (pp. 400–404). Springer.

[47] Nasir, J., Aamir, M., Haq, Z. U., Khan, S., Amin, M. Y., & Naeem, M. (2023). A new approach for forecasting crude oil prices based on stochastic and deterministic influences of LMD using ARIMA and LSTM models. IEEE Access, 11, 14322–14339.

[48] Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3), 703–708.

[49] Nguyen, D. K., & Sensoy, A. (2017). A comparison of wavelet-based methods for volatility spillovers in emerging markets. Physica A, 470, 174–195.

[50] Pakistan Bureau of Statistics. (2024). Consumer Price Index [Data set]. Government of Pakistan.

[51] Percival, D. B., & Walden, A. T. (2000). Wavelet methods for time series analysis. Cambridge University Press.

[52] Persons, W. M. (1919). Indices of business conditions. Review of Economic Statistics, 1(1), 5–107.

[53] Pesaran, M. H., & Timmermann, A. (2007). Selection of estimation window in the presence of breaks. Journal of Econometrics, 137(1), 134–161.

[54] Phillips, P. C. B., & Perron, P. (1988). Testing for a unit root in time series regression. Biometrika, 75(2), 335–346.

[55] Priestley, M. B. (1981). Spectral analysis and time series. Academic Press.

[56] Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian processes for machine learning. MIT Press.

[57] Rossi, B. (2013). Advances in forecasting under instability. In G. Elliott & A. Timmermann (Eds.), Handbook of economic forecasting (Vol. 2B, pp. 1203–1324). Elsevier.

[58] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

[59] Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6(2), 461–464.

[60] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE, 104(1), 148–175.

[61] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). A comparison of ARIMA and LSTM in forecasting time series. In 17th IEEE International Conference on Machine Learning and Applications (pp. 1394–1401).

[62] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems, 25, 2951–2959.

[63] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929–1958.

[64] State Bank of Pakistan. (2024). Annual report on the state of Pakistan's economy. State Bank of Pakistan.

[65] Stock, J. H., & Watson, M. W. (2007). Why has US inflation become harder to forecast? Journal of Money, Credit and Banking, 39(s1), 3–33.

[66] Stock, J. H., & Watson, M. W. (2015). Introduction to econometrics (3rd updated ed.). Pearson.

[67] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting, 16(4), 437–450.

[68] Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), Handbook of economic forecasting (Vol. 1, pp. 135–196). Elsevier.

[69] Tsay, R. S. (2010). Analysis of financial time series (3rd ed.). Wiley.

[70] Turk, A. M., Khan, S. A., Aamir, M., & Hussain, Z. (2025). Introducing the synergy-based approach for forecasting the crude oil prices with traditional and machine learning econometric models. International Econometric Review, 17(2), 18–34.

[71] Wang, S., Yu, L., Tang, L., & Wang, S. (2011). A novel seasonal decomposition based least squares support vector regression ensemble learning approach for hydropower consumption forecasting in China. Energy, 36(11), 6542–6554.

[72] Welch, P. (1967). The use of fast Fourier transform for the estimation of power spectra. IEEE Transactions on Audio and Electroacoustics, 15(2), 70–73.

[73] Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology, 17(1), 26–40.

[74] Wu, Z., & Huang, N. E. (2009). Ensemble empirical mode decomposition: A noise-assisted data analysis method. Advances in Adaptive Data Analysis, 1(1), 1–41.

[75] Xiong, X., Guo, X., Zeng, P., Zou, R., & Wang, X. (2022). A short-term wind power forecast method via XGBoost hyper-parameters optimization. Frontiers in Energy Research, 10, 905155.

[76] Yu, L., Wang, S., & Lai, K. K. (2008). Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics, 30(5), 2623–2635.

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

[78] Zhang, W., Duan, L., Liu, T., Shi, Z., Shi, X., Chang, Y., Qu, S., & Wang, G. (2024). A hybrid framework based on LSTM for predicting karst spring discharge using historical data. Journal of Hydrology, 633, 130946.

[79] Zhang, Y., Zhang, S., & Zhang, X. (2008). Application of a hybrid EEMD-ARIMA model in forecasting WTI crude oil prices. Journal of Applied Statistics, 35(2), 189–204.