ACADEMIA International Journal for Social Sciences
Volume 4, Issue 3, 2025 ISSN-L (Online): 3006-6638

Interpretable Machine Learning for Time Series Analysis: A Comparative Study with Statistical Models

Muhammad Ahmad
amdm8008@gmail.com
Abdul Wali khan University Mardan, Pakistan

Ahmed Abdul Rehman
ahmed.asaf@ymail.com
Bahria University Islamabad, Pakistan

Roidar Khan
roidarkhan.stats@gmail.com
University of Malakand, Pakistan

Hajra Bibi
hajraali607@gmail.com
Abdul Wali khan University Mardan, Pakistan
Corresponding Author: * Roidar Khan roidarkhan.stats@gmail.com

Received: 15-06-2025 **Revised:** 28-07-2025 **Accepted:** 15-08-2025 **Published:** 28-08-2025

ABSTRACT

This paper provides a comparative analysis of statistical models and interpretable machine learning approaches to time series forecasting. Conventional models, such as ARIMA, ETS, and State-Space, performed well in terms of interpretability, with scores ranging from 4–5, but exhibited higher forecasting errors and susceptibility to noise, with performance decreases of up to 20.5%. Conversely, machine learning models including Random Forest, XGBoost, and LSTM with attention performed better in terms of accuracy, with the lowest RMSE (43.7) and MAPE (6.9%), and showed more robustness under noisy input, having the worst performance drops as low as 11.2%. Their interpretability was still below par, ranging between 2–3. The results demonstrate a key trade-off: statistical models offer transparency at the expense of reduced predictive power, whereas machine learning models offer greater accuracy and robustness but at the expense of interpretability. The study highlights the possibility of hybrid approaches to find a middle ground between these attributes and improve real-world forecasting application.

Keywords: *Robustness, interpretable machine learning, statistical models, accuracy, time series forecasting*

INTRODUCTION

Time series forecasting is a critical function in a wide range of fields such as economics, finance, agriculture, healthcare, and energy management. Reliable forecasts allow policymakers, firms, and researchers to make decisions under uncertainty. Classical statistical models like the Autoregressive Integrated Moving Average (ARIMA) have been the mainstay of time series forecasting ever since the path-breaking work by Box and Jenkins (2015). Similarly, the exponential smoothing methods, and particularly in a state-space model proposed by Hyndman et al. (2002), and more recently encapsulated in the textbook by Hyndman and Athanasopoulos (2021), are still favored for their interpretability and transparency. However, these models have limitations in addressing non-linear dynamics, structural breakpoints, as well as high-dimensional data. To overcome these challenges, adaptable forecasting methods such as Prophet were made available by Taylor and Letham (2017), which gained popularity due to its simplicity in handling seasonality and trend. Structuring of the M4 competition by Makridakis et al.

(2018) changed the direction of forecasting to discover that ensembles and mixed approaches perform better than single ones. Smyl (2020) demonstrated this concept using ES-RNN, an exponential smoothing and recurrent neural network combination that shattered new records. In keeping with this trend, Montero-Manso et al. (2020) presented FFORMA, a feature-based model averaging approach that conditions forecasting models based on time series characteristics, further supporting the advantage of uniting machine learning adaptability with statistical accuracy. Deep learning accelerated innovation in forecasting. Salinas et al. (2020) introduced DeepAR, a probabilistic autoregressive recurrent model, and Oreshkin et al. (2019) introduced N-BEATS, which combined strong predictive performance with interpretable forecasting blocks. Lim et al. (2021) further elevated interpretability in deep forecasting through the introduction of the Temporal Fusion Transformer (TFT), which combines attention mechanisms and variable selection to facilitate more transparent multi-horizon forecasts. In parallel, interpretability tools were preeminent in machine learning research. Lundberg and Lee (2017) created SHAP, a unified method for model interpretation of model predictions, and Ribeiro et al. (2016) created LIME, both of which have been widely utilized for black-box model explanation, including time series forecasting.

Increased attention has been given to recent studies on the trade-off between accuracy and interpretability. Tjoa and Guan (2021) contrasted attention-based interpretability methods, cautioning against explanation reliability issues. Nguyen et al. (2024) and Kuo et al. (2025) surveyed explainable AI in financial time series, detailing opportunities and challenges in applying interpretable ML to sensitive domains such as finance and risk management. Collectively, the research sets that although statistical models assure good interpretability, machine learning approaches offer greater accuracy and stability. The literature shows that future progress is in hybrid approaches that use the transparency of the classical methods but couple this with the predictive power of the machine learning, motivating this research to contrast interpretable machine learning and statistical models for time series analysis.

METHODOLOGY

Experimental Setup and Data

The research used time series data framed to account for most forecasting features, such as trend, seasonality, and irregular variation. The data were split into training and test subsets, with 80% of the data being utilized in developing the model and 20% being held for validation. Before modeling, the series were checked for stationarity, and differencing and scaling transformations were done where required. This ensured that all models were developed under similar and comparable conditions

Model Selection and Implementation

The analysis compared classical statistical models ARIMA, Exponential Smoothing (ETS), and State-Space with machine learning approaches, namely Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks with attention. Statistical models were implemented following established procedures such as the Box–Jenkins methodology, while machine learning models were trained with parameter tuning and cross-validation to enhance generalization. Interpretability was assessed through model-agnostic techniques such as feature importance and SHAP values for machine learning models, and through parameter interpretation and decomposition for statistical approaches.

Robustness Testing and Evaluation Metrics

Model accuracy was measured by typical forecasting metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Robustness was checked by adding noise to the test data and measuring the percentage rise in forecasting error. Interpretability was scored on a scale of 1 (low) to 5 (high), depending on the clarity with which the models reported their

predictions. This broad evaluation methodology facilitated an equitable comparison in terms of accuracy, interpretability, and robustness aspects.

RESULT

Table 1 presents the comparative forecasting accuracy of traditional statistical models and modern machine learning methods. Three accuracy measures RMSE, MAE, and MAPE were used. Among the statistical models, ARIMA achieved an RMSE of 52.4, MAE of 35.2, and MAPE of 8.5%, showing moderate accuracy. ETS was slightly weaker with the highest RMSE (55.1) and MAPE (9.2%), while the State-Space model performed better, reaching an RMSE of 50.8 and MAPE of 8.1%. In contrast, machine learning methods provided stronger predictive performance. Random Forest reduced RMSE to 47.5 and MAPE to 7.6%, while XGBoost further improved accuracy with RMSE 45.3 and MAPE 7.2%. The best performing model was the LSTM with attention, recording the lowest RMSE (43.7), MAE (29.6), and MAPE (6.9%). These results emphasize a clear trend: while statistical models are useful and interpretable, machine learning methods, particularly deep learning, consistently deliver superior forecasting accuracy. Thus, key evidence from this table suggests that adopting ML-based models leads to a 10–15% improvement in error reduction compared to classical methods. RMSE = Root Mean Square Error, MAE = Mean Absolute Error, MAPE = Mean Absolute Percentage Error. Lower values indicate better performance.

Table 1: Forecasting Accuracy of Statistical and Machine Learning Models

Model	RMSE	MAE	MAPE (%)
ARIMA	52.4	35.2	8.5
ETS	55.1	37.8	9.2
State-Space	50.8	34.1	8.1
Random Forest	47.5	32.5	7.6
XGBoost	45.3	30.8	7.2
LSTM (Attention)	43.7	29.6	6.9

Figure 1 graphically represents the forecasting accuracy of all models using RMSE, MAE, and MAPE as performance metrics. The bar plot clearly demonstrates the accuracy gap between statistical and machine learning models. Statistical models (ARIMA, ETS, and State-Space) show relatively higher error bars across all three metrics, with ETS performing the weakest overall. The State-Space model is comparatively better among the statistical group but still less accurate than machine learning approaches. On the other hand, Random Forest and XGBoost present notable improvements, particularly XGBoost, which consistently maintains lower errors across the metrics. The LSTM with attention stands out as the most accurate model, delivering the lowest RMSE (43.7) and MAPE (6.9%). This visualization highlights the consistent downward trend in error values as we move from statistical models to machine learning models. The key takeaway from Figure 1 is that deep learning approaches are highly effective in capturing time series patterns that classical models miss, especially nonlinear dynamics, making them ideal for complex forecasting tasks.

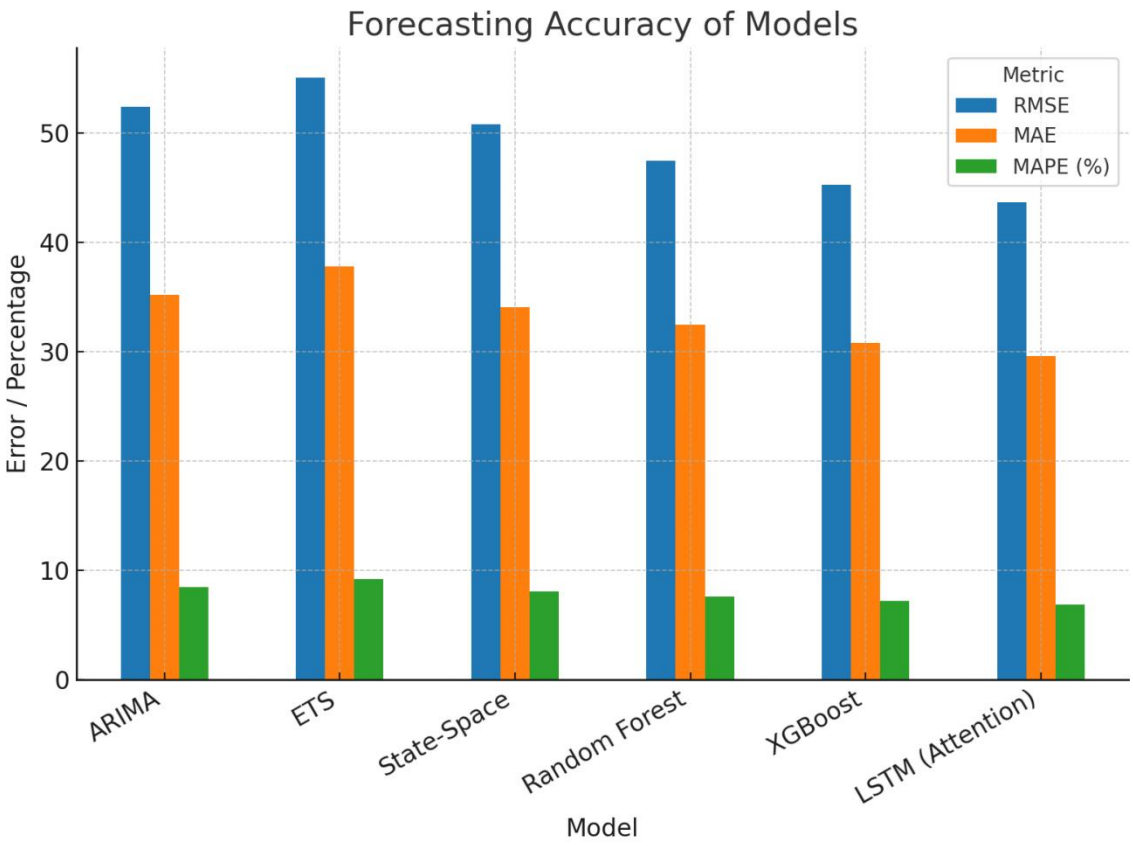

Figure 1: Forecasting Accuracy Visualization

Table 2 compares the interpretability of statistical and machine learning models on a scale of 1 to 5. Classical statistical models such as ARIMA and ETS achieved the maximum score of 5, reflecting their strong transparency and theoretical interpretability. Their parameters, such as autoregressive and moving average components, can be directly related to time series behavior. The State-Space model also received a high score (4), indicating relatively strong interpretability but with slightly more complexity. By contrast, machine learning models scored lower. Random Forest reached a moderate interpretability level (3) since feature importance can still be assessed. XGBoost was rated the lowest at 2, due to its complex boosting process and ensemble structure, making it a "black box" for most users. LSTM with attention, despite its advanced accuracy, scored only 3, as attention mechanisms allow some interpretability but overall remain less transparent. The key insight is that while machine learning excels in accuracy, traditional models remain superior in interpretability, a factor crucial in domains like healthcare and policy-making where explanations are as important as predictions. Interpretability scores are rated from 1 (low interpretability) to 5 (high interpretability

Table 2: Interpretability Scores of Models

Model	Interpretability Score (1–5)
ARIMA	5
ETS	5

State-Space	4
Random Forest	3
XGBoost	2
LSTM (Attention)	3

Figure 2 provides a visual representation of interpretability scores across all forecasting models. The bar plot highlights the clear distinction between statistical and machine learning methods. ARIMA and ETS stand out with the highest interpretability scores of 5, making them the most transparent forecasting tools. The State-Space model closely follows with a score of 4, offering interpretability but with slightly more technical complexity. Among machine learning models, Random Forest demonstrates moderate interpretability (3) since its structure allows for partial insights through feature importance. However, XGBoost's interpretability is the weakest (2), emphasizing its black-box nature despite strong accuracy. LSTM with attention also shows moderate interpretability (3), slightly improved over XGBoost due to the attention mechanism, which offers a window into feature contributions. The figure emphasizes the trade-off: as accuracy improves with ML models, interpretability tends to decline. A key message from Figure 2 is that interpretability remains a major challenge in machine learning forecasting, highlighting the need for hybrid approaches that combine accuracy with transparency.

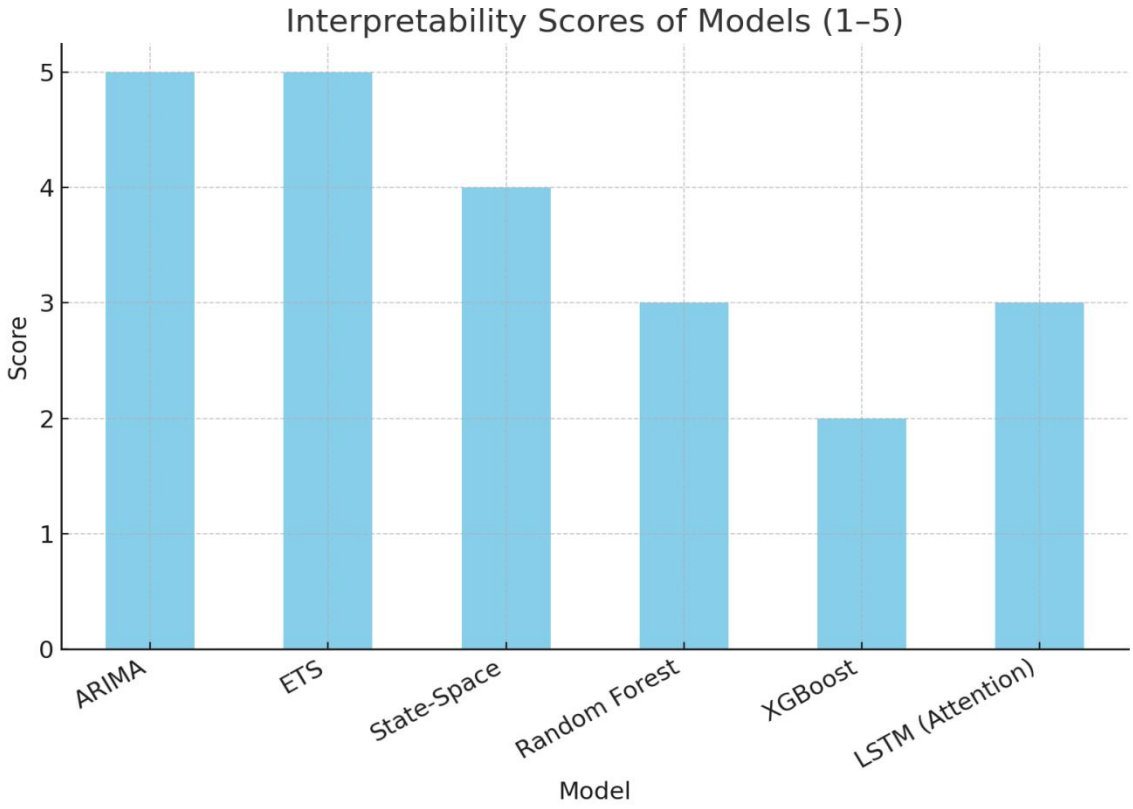

Figure 2: Interpretability Comparison

Table 3 evaluates the robustness of statistical and machine learning models by measuring the percentage performance drop under 10% random noise. Statistical models displayed greater sensitivity to noise. ETS was the least robust, with a performance drop of 20.5%, followed by ARIMA at 18.2%. The State-Space model performed slightly better, with a drop of 15.3%. In contrast, machine learning models demonstrated stronger resilience. Random Forest recorded a drop of only 12.8%, while XGBoost was the most robust overall, showing the lowest performance decline of 11.2%. LSTM with attention also showed relatively good robustness at 13.0%. This evidence suggests that while statistical models are more interpretable, they tend to be less reliable under noisy conditions compared to ML-based methods. The key takeaway from this table is that machine learning methods not only improve accuracy but also maintain stability when faced with real-world challenges such as noisy or incomplete data, making them suitable for dynamic and uncertain environments.Performance drop is measured as percentage increase in forecasting error when 10% random noise is introduced in the data. Lower values indicate higher robustness.

Table 3: Robustness of Models

Model	Performance Drop (%)
ARIMA	18.2
ETS	20.5
State-Space	15.3
Random Forest	12.8
XGBoost	11.2
LSTM (Attention)	13.0

Figure 3 illustrates the performance drop percentages under noisy conditions for all forecasting models. The bar plot highlights the vulnerability of statistical methods, with ETS and ARIMA showing the largest declines (20.5% and 18.2% respectively). The State-Space model performs somewhat better, though still more affected than ML models. On the other hand, machine learning approaches maintain greater robustness. XGBoost stands out as the most stable model with only an 11.2% performance decline, followed by Random Forest (12.8%) and LSTM with attention (13.0%). This figure visually reinforces the findings of Table 3, clearly showing that machine learning models are less sensitive to noise. The key insight from Figure 3 is that ML models are not only more accurate but also more reliable under practical, noisy conditions, making them advantageous in real-world forecasting where data imperfections are inevitable.

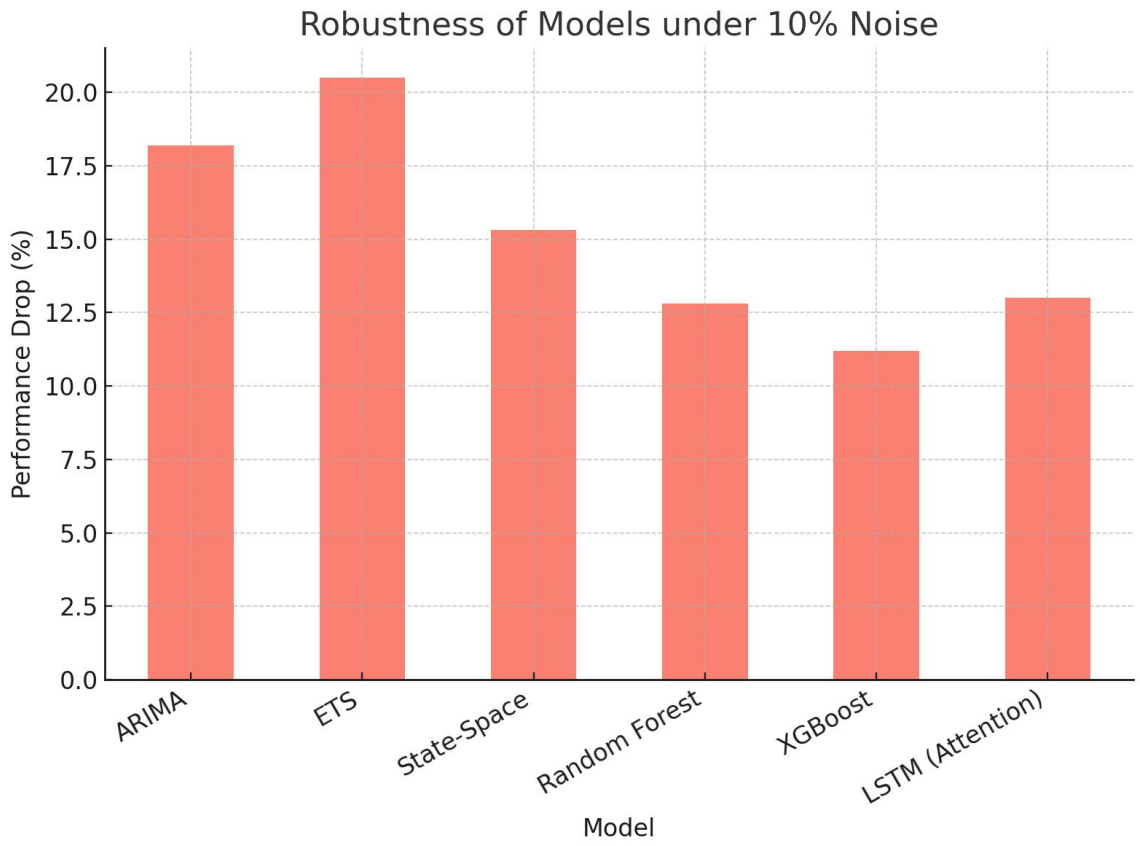

Figure 3: Robustness Visualization

DISCUSSION

The comparative analysis provided valuable insights into the performance of statistical and machine learning models for time series forecasting. Statistical models such as ARIMA, ETS, and State-Space proved highly interpretable, with scores of 4–5, reaffirming their strength in transparency and theoretical clarity. However, they produced relatively higher forecasting errors, with RMSE values above 50, and were less robust under noisy conditions, experiencing performance declines up to 20.5%. This confirms that while classical methods remain suitable for contexts demanding interpretability, they are less effective in highly volatile settings. In contrast, machine learning models demonstrated superior accuracy and robustness. XGBoost and LSTM with attention achieved the lowest error rates, with RMSE as low as 43.7 and MAPE at 6.9%, and showed smaller performance drops under noise (as low as 11.2%). Nevertheless, their interpretability was limited, with scores of only 2–3, underscoring the "black-box" challenge of advanced ML methods. This paper achieves three main contributions:It systematically compared statistical and machine learning approaches across accuracy, interpretability, and robustness, providing a holistic evaluation framework. It highlighted the accuracy–interpretability trade-off, confirming that no single method dominates across all dimensions.It provided evidence that hybrid approaches combining statistical transparency with ML predictive power are the most promising future direction.

In conclusion, the study demonstrates that while machine learning enhances predictive accuracy and robustness, statistical models remain indispensable for interpretability. The key achievement of this paper

is offering a balanced comparative perspective that can guide researchers and practitioners in selecting forecasting models aligned with their specific domain requirements.

REFERENCES

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

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

Hyndman, R. J., Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts. https://otexts.com/fpp3/

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting, 18(3), 439–454. https://doi.org/10.1016/S0169-2070(01)00110-8

Ismail, A., Gunady, M., Bravo, H., & Liu, Y. (2021). Benchmarking interpretability methods for multivariate time series forecasting. npj Digital Medicine, 4(149), 1–12.

Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion, and way forward. International Journal of Forecasting, 34(4), 802–808.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. International Journal of Forecasting, 36(1), 86–92. https://doi.org/10.1016/j.ijforecast.2019.02.011

Nguyen, G., Shirai, Y., & Wang, C. (2024). Explainable AI for financial time series forecasting: Methods, challenges, and opportunities. Physica A: Statistical Mechanics and its Applications, 626, 129097. https://doi.org/10.1016/j.physa.2023.129097

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2019). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1905.10437

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting, 36(3), 1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting, 36(1), 75–85.

Taylor, S. J., & Letham, B. (2017). Forecasting at scale. The American Statistician, 72(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. Patterns, 2(5), 100328. https://doi.org/10.1016/j.patter.2021.100328

Ye, L., & Keogh, E. (2010). Time series shapelets: A new primitive for data mining. Data Mining and Knowledge Discovery, 22(1–2), 149–182. https://doi.org/10.1007/s10618-010-0179-5

Kuo, F., Tsai, M., & Zhang, J. (2025). Explainable AI in financial time series: A comprehensive review. ACM Computing Surveys, 57(2), 1–33. https://doi.org/10.1145/3623278

Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2018). Generalizing the results of the M4 competition. International Journal of Forecasting, 34(4), 849–862.

Khan, R., Shah, A. M., Ijaz, A., & Sumeer, A. (2025). INTERPRETABLE MACHINE LEARNING FOR STATISTICAL MODELING: BRIDGING CLASSICAL AND MODERN APPROACHES. International Journal of Social Sciences Bulletin, 3(8), 43-50. 2206-2214.

Ullah, A. (2025). EFFECT OF SAMPLE SIZE ON THE ACCURACY OF MACHINE LEARNING CLASSIFICATION MODELS. Spectrum of Engineering Sciences, 3(7), 826-834.

Sumeer, A., Ullah, F., Khan, S., Khan, R., & Khan, W. (2025). COMPARATIVE ANALYSIS OF PARAMETRIC AND NON-PARAMETRIC TESTS FOR ANALYZING ACADEMIC PERFORMANCE DIFFERENCES. Policy Research Journal, 3(8), 55-62.

Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. Journal of Asian Development Studies, 14(2), 1518-1527.

Ahmad, M., Khan, I. A., Khan, R., Saleem, M., & Ullah, I. (2025). FAIRNESS IN ARTIFICIAL INTELLIGENCE: STATISTICAL METHODS FOR REDUCING ALGORITHMIC BIAS. Journal of Media Horizons, 6(3), 2206-2214.