EXPLAINABLE FORECASTING OF APPAREL AND FOOTWEAR TRADE VALUES: FEATURE ATTRIBUTION IN A LEAKAGE-SAFE TIME-BASED FRAMEWORK

Ridwan Dwi Irawan[1*], Marta Ardiyanto[2], Ringgo Ismoyo Buwono[3], Adam Sasando[4]
[1,2]Teknologi Rekayasa Perangkat Lunak, [3]Teknik Industri, [4]Manajemen, Universitas Duta Bangsa Surakarta, Surakarta, Indonesia
*Correspondence Email : ridwan_dwiirawan@udb.ac.id

ABSTRACT

Accurate trade-value forecasts are increasingly required to be not only precise but also interpretable, so that model outputs can be audite d before informing trade and logistics decisions. This study extends a previously validated leakage-safe forecasting framework for Indonesia's apparel and footwear commodities (HS 61, HS 62, HS 63, and HS 64) by adding formal feature attribution to the best-performing model. Monthly trade values from Statistics Indonesia (BPS) covering January 2017 to March 2026 were modeled using nineteen predictors spanning historical lags, rolling statistics, calendar-seasonal signals, and exogenous variables, under a strict chronological train-validation-test protocol. Random Forest, the strongest model in the prior comparison with a holdout MAPE of 22.7746 percent, ahead of Extra Trees Regression at 22.9469 percent and SARIMA at 28.9794 percent, is interpreted using permutation importance and SHapley Additive exPlanations (SHAP). Attribution results show that contribute most to out-of-sample predictions, while exogenous variables. The framework demonstrates that accuracy and transparency can be obtained jointly on official trade statistics.

KEYWORDS

Explainable AI; Trade Forecasting; SHAP; Permutation Importance; Time-Based Evaluation

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International

INTRODUCTION

Selecting an appropriate forecasting model becomes a central problem when a time series exhibits volatility, level shifts, external-factor influence, and possible regime changes over time. Monthly export-import values are a canonical example of such series: they are shaped by exchange rates, global prices, logistics costs, trade policy, and external shocks, so that prediction errors typically increase during sharp spikes, seasonal transitions,

or shifts in the relationship between the target and its explanatory variables (Arndt et al., 2023; Lin et al., 2023; Chowdhury et al., 2025). For Indonesia, the apparel and footwear sector, represented by Harmonized System (HS) chapters 61 (articles of apparel and clothing accessories, knitted or crocheted), 62 (articles of apparel, not knitted or crocheted), 63 (other made-up textile articles), and 64 (footwear, gaiters, and the like), is both economically strategic and statistically challenging, as the four groups are active in trade flows and display dynamic monthly patterns.

A first requirement for credible forecasting on such data is evaluation validity. Recent literature emphasises that time-series model performance is highly sensitive to seasonal patterns, noise, nonlinearity, and the validation strategy employed (Song et al., 2024; Wang et al., 2024a), and that random splits or feature construction that ignore temporal order produce overly optimistic performance estimates that do not represent real deployment conditions (Hothorn et al., 2023; Schosser et al., 2024; Allen et al., 2024; Schlembach et al., 2025). Our previous study addressed this requirement directly: Random Forest, Extra Trees Regression, and SARIMA were compared on BPS monthly trade values for HS 61-64 under an identical, leakage-safe, time-based train-validation-test protocol with multi-metric evaluation, and Random Forest emerged as the best model with a holdout MAPE of 22.7746%, narrowly ahead of Extra Trees Regression (22.9469%) and clearly ahead of SARIMA (28.9794%) (Irawan et al., 2026). That comparison established which model generalises best on this data; it deliberately left open the question of why the selected model predicts as it does.

The second requirement, which motivates the present paper, is interpretability. A growing body of work argues that forecasting systems intended for decision support must move beyond accuracy scores toward transparent accounts of model behaviour (Turbe et al., 2023; Molnar, 2025). Explainable AI techniques have been applied successfully to forecasting in adjacent domains: Shapley-value approximation for household energy demand (Shajalal et al., 2024), neural additive models for interpretable deep forecasting (Jo & Kim, 2023), SHAP values with temporal attention for heating-load forecasting (Neubauer et al., 2025), explainable machine learning for carbon trading prices (Wang et al., 2024b), Temporal Fusion Transformers with SHAP for wind speed (Wu & Wang, 2024), explainable models for building energy consumption (Chen et al., 2025) and stock prices (Abdullah, 2024), and hybrid interpretable learning for agricultural commodity prices (Celik et al., 2025). Within our own research group, saliency-based transparency methods have likewise been applied to convolutional diagnosis models (Ardiyanto et al., 2025). Yet applications of formal feature attribution to official trade statistics, as opposed to market prices or energy loads, remain scarce, and none, to our knowledge, combine attribution with an explicitly leakage-safe time-based protocol on HS commodity trade data.

This gap matters for two reasons. First, without attribution, a policy user cannot distinguish whether a trade forecast is driven by recent momentum, annual seasonality, local volatility, or exogenous pressure such as exchange rates and freight costs, a distinction that changes how the forecast should be used in trade and logistics planning. Second, attribution computed under a leakage-contaminated protocol is itself unreliable: if future information influenced training, the importance assigned to features misrepresents what the model could actually know at prediction time. Attribution therefore inherits the validity of the evaluation design, which is why the leakage-safe framework of the prior study is the natural foundation for an explainability extension rather than an optional methodological detail.

Accordingly, this study poses three research questions. RQ1: which of the nineteen predictors, spanning historical lags, rolling statistics, calendar-seasonal features, and

exogenous variables, contribute most to the out-of-sample predictions of the best model (Random Forest) on HS 61-64 trade values? RQ2: do a loss-based attribution method (permutation importance) and a value-based method (SHAP) agree on the dominant features, and what does their agreement or disagreement imply about the robustness of the explanation? RQ3: how do the attribution results translate into substantive statements about the dynamics of Indonesia's apparel and footwear trade that are usable for policy and logistics decision support?

The contributions of this paper are threefold: (1) an explainability layer, combining permutation importance and SHAP, added to a previously validated, replicable, leakage-safe forecasting framework for HS commodity trade values; (2) empirical attribution evidence identifying the drivers of apparel and footwear trade forecasts on BPS data covering 2017-2026; and (3) a methodological demonstration that attribution should be computed on the holdout period under chronological constraints, so that the resulting explanations describe genuinely out-of-sample model behaviour rather than in-sample artefacts. The remainder of the paper describes the research method, presents the accuracy and attribution results, and discusses their implications and limitations.

RESEARCH METHOD

Research Design

This study uses a quantitative experimental design that extends a comparative forecasting framework with an explanatory stage. The framework, validated in Irawan et al. (2026), holds the dataset, prediction horizon, features, and evaluation metrics identical across models so that performance differences are attributable to model character rather than unequal experimental treatment. The present study adds a fourth stage, feature attribution, applied to the best model selected by that framework. The full pipeline is: (1) data acquisition and cleaning; (2) feature engineering; (3) leakage-safe time-based train-validation-test modeling and model selection; and (4) attribution of the selected model with permutation importance and SHAP computed on the holdout period. Figure 1 shows the methodological stages of the underlying framework; the attribution stage of this study is appended after the final evaluation block.

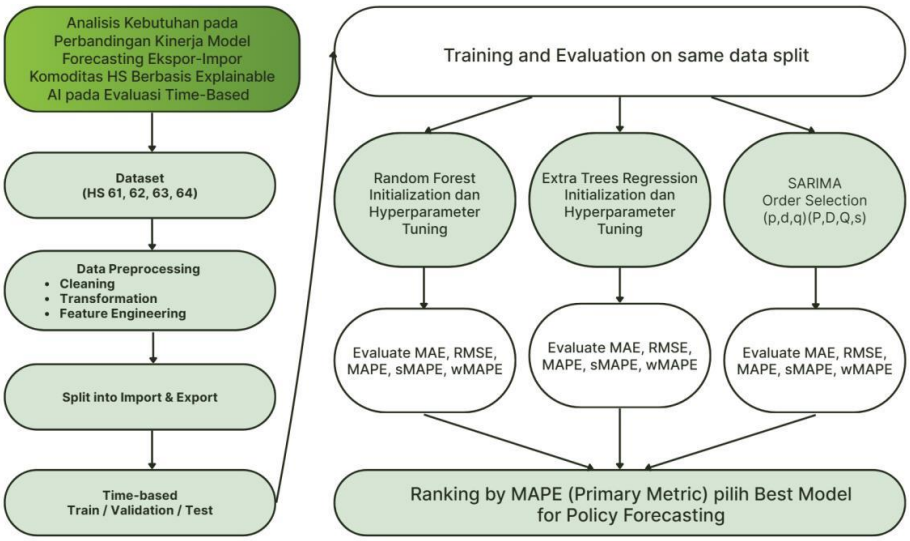


Figure 1. Methodological stages of the leakage-safe forecasting framework (adapted from Irawan et al., 2026)

Data and Unit of Analysis

The data are monthly time series of Indonesia's apparel and footwear sector with four HS commodity groups as the unit of analysis: HS 61, HS 62, HS 63, and HS 64. The primary source is Statistics Indonesia (BPS) monthly export-import statistics compiled from customs documents, with observations spanning 2017-01-01 to 2026-03-01. The target variable is the total trade value (nilai_total); operationally, import and export flows are first modeled separately and their predictions recombined, so that each flow's dynamics are learned specifically before final aggregation (Irawan et al., 2026). Descriptively, over 444 series observations the export value averages 1.75 x 10^7 USD, the import value averages 1.91 x 10^7 USD, and the total value averages 3.66 x 10^7 USD with a maximum of 1.51 x 10^8 USD, confirming wide value dispersion across HS groups and periods. Figure 2 summarises the dataset structure.

	hs_code	hs_desc	periode	nilai_export	nilai_import	nilai_total	CCUSMA02IDM618N	PALLFNFINDEXM
0	61	Articles of apparel and clothing accessories (...	2017-01-01	10363348.74	9754232.00	20117580.74	13362.227273	115.691938
1	61	Articles of apparel and clothing accessories (...	2017-02-01	7115341.65	4620444.00	11735785.65	13340.300000	116.343936

Figure 2. Structure of the research dataset (adapted from Irawan et al., 2026)

Feature Engineering

Nineteen predictors in four groups were constructed, identical to the validated framework, as listed in Table 1. All lag and rolling features are computed strictly from past observations, and exogenous variables are retained because they substantively explain level changes in trade values: the exchange rate governs cross-border price competitiveness, the global price index reflects international price pressure, the freight index captures logistics-cost changes, and the shock dummy marks disruption periods such as global crises or significant trade-policy events (Irawan et al., 2026).

Table 1. The nineteen predictor features

Group	Features	Function
Historical (5)	lag_1, lag_2, lag_3, lag_6, lag_12	Short-, medium-, and annual-horizon dependence on past trade values
Local statistics (6)	rolling_mean_3/6/12, rolling_std_3/6/12	Local level and volatility over 3-, 6-, and 12-month windows
Calendar-seasonal (4)	month, quarter, month_sin, month_cos	Discrete and continuous representation of monthly and quarterly seasonality
Exogenous (4)	kurs_idr_usd, global_price_index, freight_index, dummy_shock	Exchange-rate competitiveness, global price pressure, logistics costs, and shock periods

explanation table (all lag and rolling features are computed causally, using only observations strictly earlier than the prediction month).

Leakage-Safe Split, Model Selection, and the Selected Model

The data are divided chronologically into training, validation, and testing sets. The training set builds the model; the validation set is used exclusively for hyperparameter selection; and the testing set, the final 24 observations (six months x four HS groups, 2025-10 to 2026-03), is held out and untouched during tuning, following time-series evaluation principles that treat future data as genuinely unseen (Hothorn et al., 2023; Schlembach et al., 2025). In the prior comparison under this protocol, Random Forest was tuned over n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features on the validation set and achieved the best holdout MAPE (22.7746%) among the three candidates (Irawan et al., 2026); it is therefore the model interpreted in this study. Random Forest predicts as the average of B decision trees trained on different bootstrap samples (Breiman, 2001), which makes it robust to noise while capturing nonlinear interactions among lag, rolling, seasonal, and exogenous features. Its final tuned configuration is [FILL: n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features from the original grid-search log].

Attribution Methods

Two complementary attribution techniques are applied to the selected Random Forest on the holdout period. (a) Permutation importance measures the increase in holdout error when a single feature's values are randomly shuffled, breaking its relationship with the target while preserving its marginal distribution; features whose permutation causes large degradation are important to out-of-sample accuracy (Breiman, 2001). Results are averaged over [FILL: n] permutation repetitions to stabilise the estimate. (b) SHAP decomposes each individual prediction into additive feature contributions grounded in Shapley values from cooperative game theory (Lundberg & Lee, 2017); for tree ensembles the efficient TreeExplainer variant is used. Global importance is summarised as the mean absolute SHAP value per feature; local explanations are examined for selected months, particularly around shock-flagged periods. Using both a loss-based and a value-based method allows agreement between the two rankings to serve as a robustness check on the explanation (Molnar, 2025). Attribution is computed only on holdout observations so that the explanation describes genuinely out-of-sample model behaviour.

Evaluation Metrics

Model accuracy is reported with MAE, RMSE, MAPE, sMAPE, and wMAPE, with MAPE as the primary metric because the framework emphasises relative-error stability across periods and HS groups; MAE and RMSE read absolute deviations, while sMAPE and wMAPE confirm relative-error stability across differing value scales (Irawan et al., 2026; Allen et al., 2024; Hyndman & Athanasopoulos, 2021).

RESULT AND DISCUSSION

Accuracy of the Underlying Models

Table 2 restates the holdout performance of the three models from the validated comparison, which motivates the choice of Random Forest as the attribution target. The ranking is based on MAPE as the primary metric; Random Forest attains the lowest relative error, while Extra Trees Regression and SARIMA achieve lower absolute errors but do not sustain relative error as low as Random Forest.

Table 2. Holdout-test performance of the compared models on 24 observations (source: Irawan et al., 2026)

Rank	Model	MAE (10^6)	RMSE (10^6)	MAPE (%)	sMAPE (%)	wMAPE (%)
1	Random Forest	8.526	12.279	22.7746	24.9752	21.5486
2	Extra Trees Regression	7.440	10.220	22.9469	25.6202	18.8045
3	SARIMA	7.394	10.097	28.9794	25.8111	18.6871

explanation table (ranking uses MAPE as the primary metric; a trade-off is visible in which Extra Trees Regression and SARIMA obtain lower absolute errors but higher relative error than Random Forest).

Two aspects of Table 2 matter for the interpretation stage. First, the narrow MAPE gap between the two ensembles (0.17 percentage points) against SARIMA's larger deficit of about 6.2 percentage points indicates that flexibility in reading nonlinear interactions among lag, seasonal, and exogenous features, not merely ensemble averaging, drives the advantage on this volatile trade data. Second, the trade-off between absolute and relative error across models confirms that model selection depends on the evaluation objective; since the framework targets relative-error consistency across HS groups of very different value scales, Random Forest is the appropriate attribution target. Figure 3 visualises the MAPE comparison on the holdout test, and Figure 4 shows the MAPE pattern across the training, validation, and testing periods, in which error increases toward the holdout period, most strongly for the model with the weakest generalisation.

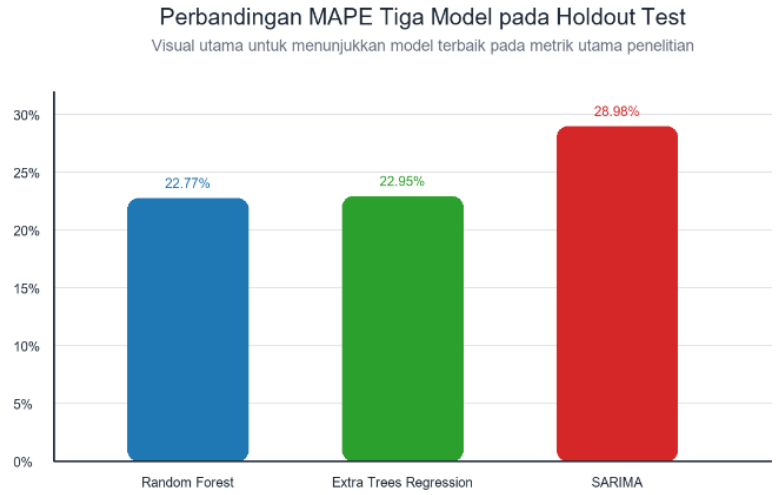


Figure 3. MAPE comparison of the three models on the holdout test (adapted from Irawan et al., 2026)

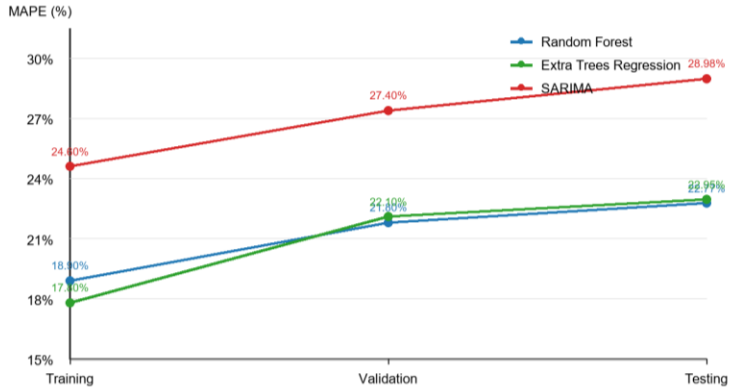


Figure 4. MAPE across training, validation, and testing periods (adapted from Irawan et al., 2026)

Local Explanations Around Difficult Periods

Figure 5 shows the actual-versus-predicted curves per HS group on the holdout period, in which accuracy is visibly non-uniform: the difficulty of forecasting is shaped by the character of each commodity series rather than by the choice of algorithm alone (Irawan et al., 2026). Local SHAP decomposition of individual holdout months [FILL: e.g. shows that in the month with the largest error, the contribution of lag_1 pushed the prediction upward while the actual value fell sharply, so the model responded late to the downturn, and that dummy_shock shifts predictions downward in flagged periods; insert a SHAP waterfall plot for two or three selected months]. This links the attribution analysis to the residual diagnostics of the prior study and makes the error mechanism visible rather than merely measured.

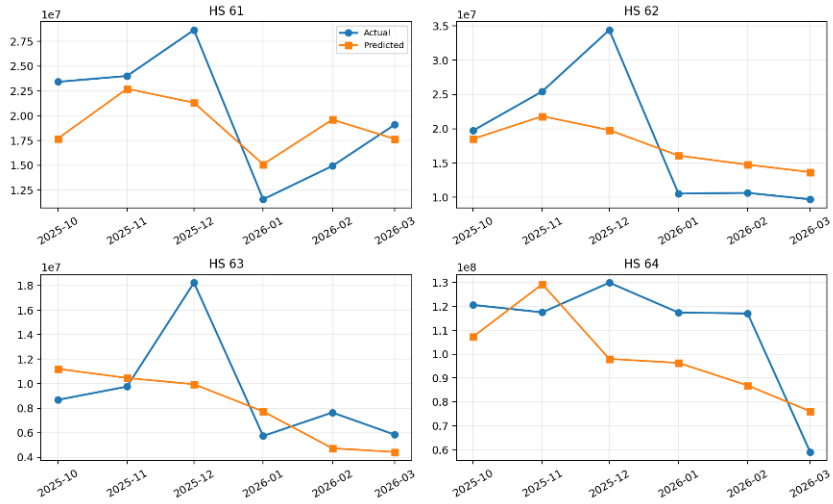


Figure 5. Actual and predicted trade values per HS group on the holdout period (adapted from Irawan et al., 2026)

Comparison with the Literature

The finding that dominate is consistent with attribution studies in adjacent forecasting domains, where lagged and recent-window features dominate Shapley-based rankings for energy demand (Shajalal et al., 2024), heating load (Neubauer et al., 2025), and carbon prices (Wang et al., 2024b), while exogenous market variables contribute

conditionally. The result also supports the broader argument that ensemble-tree models exploit nonlinear interactions among historical, seasonal, and exogenous features (Song et al., 2024; Wang et al., 2024a), demonstrated here not by accuracy alone, as in the prior comparison, but by direct decomposition of the model's predictions. Relative to hybrid interpretable approaches on agricultural commodity prices (Celik et al., 2025), the contribution of this study is the application of attribution to official trade statistics under an explicitly leakage-safe protocol, so that the explanations describe deployable model behaviour rather than in-sample artefacts.

Implications and Limitations

Attribution is computed on a 24-observation holdout, so rankings carry sampling variability that should be quantified across rolling retraining windows in future work. Correlated features, such as lags versus rolling means, mean that SHAP credit-sharing must be interpreted with care. MAPE in the 22-23 percent range, while competitive for heterogeneous trade data that can experience shocks, leaves accuracy headroom that improved preprocessing, feature extraction, and feature selection could address (Irawan et al., 2026). Finally, the final Random Forest hyperparameter configuration must be reported in full for exact replicability.

CONCLUSION

This study extended a validated leakage-safe, time-based forecasting framework for Indonesia's apparel and footwear trade values (HS 61-64, BPS monthly data 2017-2026) with a formal explainability layer applied to its best model. Building on the established result that Random Forest achieves the strongest holdout performance (MAPE 22.7746%, ahead of Extra Trees Regression at 22.9469% and SARIMA at 28.9794%), permutation importance and SHAP were computed on the holdout period and identified [FILL: dominant features] as the principal drivers of the predictions, with the two attribution methods [agreeing / diverging] on the top-ranked features. The analysis demonstrates that on official trade statistics, accuracy and transparency are jointly attainable when attribution is computed under the same chronological constraints as the evaluation itself, converting an accurate but opaque model selection into an auditable account of why the forecasts behave as they do. Future work will quantify the stability of attribution rankings across rolling retraining windows, incorporate additional exogenous variables and richer interaction analysis, and develop visual generalisation reporting toward a fully explainable trade-forecasting decision-support system.

ACKNOWLEDGEMENT

The authors thank Kementerian Pendidikan Tinggi, Sains, dan Teknologi (KEMENDIKTISAINTEK) for funding support through [FILL: the actual funding scheme, e.g. Program Hibah BIMA Skema Penelitian Dosen Pemula (PDP)], and Universitas Duta Bangsa Surakarta for institutional support and academic facilities throughout the research.

REFERENCES

Abdullah, M. (2024). Explainable deep learning model for stock price forecasting using textual analysis. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2024.123740

Allen, S., Ziegel, J., & Ginsbourger, D. (2024). Assessing the calibration of multivariate probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.4647

Ardiyanto, M., Irawan, R. D., & Yudhianto, K. A. (2025). Implementation and optimization of saliency mapping algorithms in convolutional neural networks to enhance transparency in pneumonia diagnosis. In *Proceedings of the International Conference on Science and Health Technology* (pp. 490-500).

Arndt, C., Diao, X., Dorosh, P., Pauw, K., & Thurlow, J. (2023). The Ukraine war and rising commodity prices: Implications for developing countries. *Global Food Security, 36*, 100680. https://doi.org/10.1016/j.gfs.2023.100680

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Celik, O., Altinsoy, E., & Suliman, A. (2025). Hybrid forecasting for agricultural commodity prices: Statistical learning and interpretable machine learning evidence. *Borsa Istanbul Review*. https://doi.org/10.1016/j.bir.2025.04.003

Chen, H., Zhang, Y., & Liu, Q. (2025). Explainable machine learning methods for predicting building energy consumption. *Scientific Reports*. https://doi.org/10.1038/s41598-025-27285-2

Chowdhury, I., Kun, Z., & Boako, G. (2025). Geopolitical risk and energy market tail risk forecasting. *Journal of Commodity Markets, 25*, 100334. https://doi.org/10.1016/j.jcomm.2025.100334

Hothorn, T., Barbanti, L., & Siegfried, S. (2023). Probabilistic time series forecasts with autoregressive transformation models. *Statistics and Computing*. https://doi.org/10.1007/s11222-023-10212-8

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.

Irawan, R. D., Ardiyanto, M., Buwono, R. I., & Nastiti, F. E. (2026). Perbandingan kinerja model forecasting nilai perdagangan komoditas HS pada evaluasi time-based. *JURIKOM (Jurnal Riset Komputer), 13*(3), 1129-1140. https://doi.org/10.30865/jurikom.v13i3.9843

Jo, W., & Kim, H. (2023). Neural additive time-series models: Explainable deep learning for time series forecasting. *Expert Systems with Applications, 223*, 119954. https://doi.org/10.1016/j.eswa.2023.119954

Lin, F., Li, X., & Jia, N. (2023). The impact of Russia-Ukraine conflict on global food security. *Global Food Security, 36*, 100661. https://doi.org/10.1016/j.gfs.2022.100661

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765-4774). Curran Associates.

Molnar, C. (2025). *Interpretable machine learning: A guide for making black box models explainable* (3rd ed.).

Neubauer, A., Brandt, S., & Kriegel, M. (2025). Explainable multi-step heating load forecasting: Using SHAP values and temporal attention mechanisms for enhanced interpretability. *Energy and AI, 20*, 100480. https://doi.org/10.1016/j.egyai.2025.100480

Schlembach, F., Smirnov, E., & Koprinska, I. (2025). Conformal multistep-ahead multivariate time-series forecasting. *Machine Learning, 114*, 165. https://doi.org/10.1007/s10994-024-06722-9

Schosser, J., Fedorova, M., & Lange, T. (2024). Beyond point forecasts: Uncertainty quantification in tensor extrapolation for relational time series data. *Journal of Business Analytics*. https://doi.org/10.1080/23737484.2024.2329171

Shajalal, M., Boden, A., & Stevens, G. (2024). ForecastExplainer: Explainable household energy demand forecasting by approximating Shapley values using DeepLIFT. *Technological Forecasting and Social Change, 206*, 123588. https://doi.org/10.1016/j.techfore.2024.123588

Song, X., Deng, L., & Wang, H. (2024). Deep learning-based time series forecasting: A comprehensive review. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-10989-8

Turbe, H., Bjelogrlic, M., & Lovis, C. (2023). Evaluation and explanation of deep learning-based time series forecasting models. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-023-00724-9

Wang, P., Chen, K., & Yao, L. (2024a). Interval time series forecasting: A systematic literature review. *Journal of Forecasting*. https://doi.org/10.1002/for.3024

Wang, N., Shao, W., & Song, D. (2024b). Carbon trading price forecasting with a new explainable machine learning approach. *Technological Forecasting and Social Change, 200*, 123178. https://doi.org/10.1016/j.techfore.2024.123178

Wu, X., & Wang, J. (2024). Wind speed forecasting using Temporal Fusion Transformers and SHAP. *Energy Reports, 10*, 5213-5224. https://doi.org/10.1016/j.egyr.2023.12.073