

A Proposed Hybrid CNN–LSTM Framework for Forecasting Botswana’s Rough-Diamond Export Unit Values with Explainable AI

Kabelo Kelepile^{1*}, Nelson Lungu²

¹Student, Africa Research University, Zambia

²Lecturer, Electrical and Electronics, University of Lusaka, Zambia

DOI: <https://doi.org/10.47772/IJRISS.2026.100900304>

Received: 17 September 2026; Accepted: 22 September 2026; Published: 09 October 26 2026

ABSTRACT

Advanced forecasting models can represent complex nonlinear and temporal relationships, but their black-box nature may limit transparency and practical application in policy and strategic decision-making. This concern is particularly relevant to Botswana, where rough-diamond export performance has significant implications for fiscal planning, foreign-exchange earnings and macroeconomic stability. This paper proposes a unified and reproducible framework for forecasting Botswana’s rough-diamond export unit values, measured in United States dollars per carat, by integrating a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) model with SHapley Additive exPlanations (SHAP). The proposed framework is designed to combine macroeconomic, diamond-market, country-level and global-supply indicators covering 2004–2025. Variables reported at different temporal frequencies will be harmonised to a monthly frequency while preserving temporal ordering, aggregation consistency and information availability at each forecast origin. The forecasting performance of the proposed CNN–LSTM model will be evaluated against naïve and appropriately specified econometric benchmarks, including Autoregressive Integrated Moving Average (ARIMA) and Generalised Autoregressive Conditional Heteroskedasticity (GARCH)-family models. Evaluation will employ walk-forward validation, Root Mean Squared Error, Mean Absolute Error, Mean Absolute Percentage Error and pairwise Diebold–Mariano tests, supported by robustness and sensitivity analyses. SHapley Additive exPlanations is pre-specified as a post-hoc interpretation stage within the analytical pipeline and will be used to quantify the predictive contributions of input variables without interpreting these associations as causal effects. As a proposed-framework paper, the study does not report completed empirical forecasting or SHAP results. Its contribution lies in establishing a methodologically coherent foundation for subsequent empirical evaluation of forecasting accuracy and interpretability in Botswana’s diamond-dependent economy.

Keywords: Rough-diamond export unit value; Botswana; CNN–LSTM; SHAP; time-series forecasting; temporal harmonisation

BACKGROUND

Forecasting commodity-related export outcomes is important for fiscal planning, investment decisions, resource management and macroeconomic policy, particularly in economies whose public revenues and external balances are sensitive to fluctuations in international commodity markets. Such outcomes may exhibit volatility, nonlinear behaviour, structural changes and sensitivity to domestic and global economic conditions, making reliable forecasting methodologically challenging. Traditional econometric approaches, including ARIMA, Vector Autoregression (VAR) and GARCH models, remain important because of their established statistical foundations, relative transparency and interpretability. However, their ability to represent complex nonlinear relationships and changing market dynamics may be limited when the underlying data-generating process departs from their modelling assumptions. The forecasting literature therefore emphasises evaluating alternative methods under consistent out-of-sample conditions rather than assuming the superiority of a

particular modelling class (Makridakis et al., 2020; Cerqueira et al., 2020; Petropoulos et al., 2022). Cerqueira et al. (2020), in particular, demonstrate the importance of evaluation procedures that preserve temporal ordering when assessing forecasting performance.

Advances in machine learning and deep learning have expanded the range of methods available for financial and commodity time-series forecasting. Unlike conventional linear specifications, deep neural networks can learn complex representations and nonlinear relationships from multivariate data. Long Short-Term Memory (LSTM) networks are particularly suited to sequential data because of their capacity to represent longer-term temporal dependencies, while Convolutional Neural Networks (CNNs) can extract local and short-term patterns from time-series sequences. The growing application of these architectures has encouraged the development of hybrid CNN–LSTM models that combine local feature extraction with sequential learning (Hochreiter & Schmidhuber, 1997; Sezer et al., 2020). Reviews of financial forecasting research document the increasing application of CNN, LSTM and related deep-learning architectures, while the hybrid-model literature identifies the complementary capabilities of CNN and LSTM approaches for forecasting complex time-series behaviour.

The representational flexibility of deep-learning models, however, introduces an important methodological challenge: interpretability. Complex models are frequently characterised as black-box systems because the relationships through which multiple inputs generate individual forecasts are not immediately transparent. This presents a particular concern when forecasts are intended to inform economic or policy decisions, because predictive accuracy alone may not provide decision-makers with sufficient understanding of the variables associated with a forecast. Explainable Artificial Intelligence has emerged in response to this limitation by providing methods for interpreting complex predictive systems (Barredo Arrieta et al., 2020). Among these approaches, SHAP, developed by Lundberg and Lee (2017), provides a theoretically grounded method for attributing model predictions to individual input features. Applications of explainability to financial time-series forecasting further demonstrate the potential for Explainable Artificial Intelligence to complement predictive modelling by improving the transparency of model outputs (Çelik et al., 2023).

The relationship between predictive performance and interpretability is especially relevant to commodity-dependent economies. In such settings, changes in commodity-market conditions may be associated with export earnings, fiscal revenues, foreign-exchange availability and broader macroeconomic planning. Botswana provides an important application because of the longstanding significance of rough-diamond exports to its economy. Forecasting Botswana's rough-diamond export outcomes therefore requires a framework capable of representing predictive relationships between international market conditions, domestic economic indicators and export performance while retaining sufficient transparency for policy interpretation.

Importantly, Botswana's rough diamonds should not be treated as a homogeneous commodity with a single, directly observable market price. This study therefore focuses specifically on Botswana's rough-diamond export unit value, expressed in United States dollars per carat. The export unit value is derived by dividing the value of rough-diamond exports by the corresponding volume of exported carats. It represents an average value per exported carat rather than the transaction price of a standardised or homogeneous diamond product. Consequently, movements in the observed export unit value may reflect not only changing market conditions but also changes in the size, quality and overall assortment composition of the rough diamonds exported during a particular period. The proposed framework will therefore interpret and forecast export unit value as a composition-sensitive export measure and not as a homogeneous rough-diamond market price. This distinction will be maintained consistently throughout the variable definitions, data-processing procedures, model specifications, figures, tables, interpretation and conclusion.

The methodological challenge extends beyond model selection. Potential predictors are obtained from different national and international sources and may be reported at different temporal frequencies. A coherent forecasting framework must therefore specify procedures for temporal harmonisation, preserve the chronological ordering of observations during model training and evaluation, compare the proposed model with appropriate benchmarks, and provide interpretable explanations of model outputs. Appropriate time-series evaluation is particularly important because conventional random cross-validation can violate temporal

dependence and produce misleading estimates of forecasting performance. Out-of-sample and temporally ordered evaluation strategies provide a more appropriate basis for assessing model generalisation in non-stationary time-series environments (Cerqueira et al., 2020).

Despite substantial methodological advances, deep learning, econometric benchmarking, temporal data harmonisation and explainability are frequently addressed as separate components of the forecasting process. Deep-learning studies often prioritise predictive performance, while explainability is commonly introduced after model estimation to interpret an already-developed predictive system (Barredo Arrieta et al., 2020; Lundberg & Lee, 2017). Furthermore, evidence that hybrid architectures perform well in other financial and commodity-forecasting applications does not establish that they will provide superior forecasts of Botswana's rough-diamond export unit values. This creates a need for a coherent and reproducible methodological framework through which conventional benchmarks and hybrid deep-learning models can be evaluated under clearly specified data, information and forecasting conditions, with explainability pre-specified as a post-hoc interpretation stage of the analytical process.

Accordingly, this study proposes a unified framework for forecasting Botswana's rough-diamond export unit values by integrating hybrid deep learning, econometric benchmarking, temporal data harmonisation and explainable artificial intelligence within a single analytical architecture. The framework is designed to combine a hybrid CNN–LSTM model with macroeconomic, diamond-market, country-level and global-supply indicators. It will provide a structured basis for evaluating the forecasting performance of the hybrid model against appropriately specified ARIMA and GARCH-family benchmarks under temporally consistent out-of-sample conditions. After the CNN–LSTM model has been developed, fitted and evaluated, SHAP will be applied as a post-hoc interpretation method to quantify and interpret the predictive contributions of individual input variables. These contributions will be interpreted as predictive associations and not as evidence of causal effects.

The study addresses five linked analytical objectives:

- To identify relevant macroeconomic predictors of Botswana's rough-diamond export unit values.
- To identify relevant diamond-market predictors of Botswana's rough-diamond export unit values.
- To evaluate the forecasting performance of ARIMA and GARCH-family benchmark models for Botswana's rough-diamond export unit values.
- To develop and evaluate a hybrid CNN–LSTM model for forecasting Botswana's rough-diamond export unit values.
- To explain the fitted CNN–LSTM forecasts using SHAP and quantify the predictive contributions of the input variables.

RELATED LITERATURE

Commodity-related forecasting has traditionally relied on econometric and statistical time-series methods, including ARIMA, VAR and GARCH models. These approaches remain valuable because of their established statistical foundations, relatively transparent model structures and interpretability. ARIMA models are particularly suited to representing linear temporal dependence and autocorrelation in the conditional mean of a time series, while GARCH-family models provide an established framework for modelling time-varying conditional variance and volatility clustering. However, commodity-market data may exhibit nonlinear relationships, structural changes, volatility clustering and interactions among multiple economic and market variables that are difficult to represent within a single conventional specification. Forecasting research therefore increasingly emphasises rigorous out-of-sample evaluation and comparison of alternative modelling approaches rather than assuming that one model class will perform consistently across different datasets and forecasting environments (Makridakis et al., 2020; Cerqueira et al., 2020; Petropoulos et al., 2022).

The increasing availability of economic and financial data, together with advances in computational methods, has expanded the application of machine learning and deep learning to time-series forecasting. Long Short-Term Memory networks are particularly relevant because their recurrent architecture enables them to represent sequential dependencies and retain information across extended temporal horizons. Convolutional Neural Networks, although originally developed primarily for image-related applications, have also been adapted to time-series forecasting because convolutional filters can extract informative local patterns from sequential data. Reviews of financial time-series forecasting demonstrate the growing application of these architectures and highlight their ability to represent nonlinear and complex temporal relationships that may be difficult to specify explicitly using conventional models (Hochreiter & Schmidhuber, 1997; Sezer et al., 2020).

Hybrid architectures have subsequently been developed to exploit the complementary capabilities of different deep-learning models. In a CNN–LSTM framework, the CNN component can extract local and short-term features from an input sequence, while the LSTM component represents longer-term sequential dependencies. This combination provides a theoretically useful architecture for forecasting environments in which both short-term fluctuations and persistent temporal patterns may be relevant. Studies of hybrid deep-learning approaches report promising forecasting performance across financial and related time-series applications, although performance remains dependent on data characteristics, model specification, hyperparameter selection and evaluation design (Sezer et al., 2020). Consequently, the value of a hybrid CNN–LSTM architecture for forecasting Botswana's rough-diamond export unit values must be established through subsequent empirical comparison with appropriate benchmark models rather than presumed from its architectural complexity.

Despite their predictive flexibility, deep-learning models introduce an important interpretability challenge. Their internal representations and nonlinear interactions can make it difficult to determine how individual predictors contribute to particular forecasts. This limitation is especially relevant when model outputs are intended to support economic and policy decisions, for which transparency and interpretability may be required alongside predictive accuracy. Explainable Artificial Intelligence has therefore emerged as an important methodological field concerned with making complex predictive models more interpretable and accountable (Barredo Arrieta et al., 2020).

Among the methods developed for this purpose, SHAP, introduced by Lundberg and Lee (2017), provides a theoretically grounded approach for attributing model predictions to input features using concepts derived from cooperative game theory. SHAP can provide global information about the relative importance of predictors across observations and local explanations showing how particular variables contribute to individual model predictions. Its application to financial time-series modelling demonstrates how explainability can complement predictive modelling by improving the interpretation of otherwise complex machine-learning outputs (Çelik et al., 2023). SHAP nevertheless remains a post-hoc interpretation method because it is applied after a model has been fitted and predictions have been generated. Its inclusion can be pre-specified as an integral component of the research design, but SHAP values must be interpreted as predictive contributions rather than causal effects.

A further methodological issue concerns the temporal structure of forecasting data. Commodity forecasting frequently requires combining variables obtained from different sources and reported at different frequencies. Without appropriate harmonisation, mixed-frequency observations may undermine temporal consistency across model inputs. Temporal disaggregation provides a mechanism for estimating higher-frequency representations of lower-frequency variables while maintaining consistency with the originally observed aggregates. The Denton method provides an established approach to temporal disaggregation and benchmarking by minimising changes in the movement of an estimated high-frequency series subject to defined aggregation constraints (Denton, 1971). However, the method does not independently recover genuine within-period movements unless an appropriate high-frequency indicator is available. The choice of indicator, Denton variant and aggregation constraint must therefore be specified and justified for each disaggregated series.

Evaluation design is equally important. Standard random cross-validation procedures may be inappropriate for time-series forecasting because they can disrupt chronological ordering and allow future information to influence model development. Cerqueira et al. (2020) demonstrate that performance-estimation procedures for time-series forecasting must account for temporal dependence and non-stationarity. Walk-forward or rolling-origin evaluation provides an alternative in which models are trained using information available up to a particular forecast origin and evaluated on subsequent observations, more closely approximating real forecasting conditions. Forecast accuracy can then be assessed using complementary measures such as Root Mean Squared Error, Mean Absolute Error and Mean Absolute Percentage Error. Pairwise Diebold–Mariano tests can provide a statistical basis for examining whether observed differences in forecast-loss performance between competing models are significant, provided that the loss function, forecast horizon and variance correction are clearly specified (Diebold & Mariano, 1995; Hyndman & Athanasopoulos, 2021).

Botswana's Diamond Dependence and the Importance of Rough-Diamond Export Forecasting

Botswana provides a particularly important setting for commodity-related forecasting because diamonds have historically occupied a dominant position in the country's export earnings, fiscal revenues and broader macroeconomic performance. The International Monetary Fund reported that diamond exports accounted for almost 90% of Botswana's goods exports over the decade preceding its 2024 assessment. The decline in diamond exports during 2023 consequently contributed to weaker economic activity, fiscal pressures and deterioration in the external position (International Monetary Fund, 2024). This concentration means that fluctuations in rough-diamond export outcomes can have implications extending beyond the mining industry to government revenue, foreign-exchange availability, public expenditure and national economic planning.

The significance of diamonds to Botswana does not, however, imply that rough-diamond export performance can be represented adequately by a general commodity-price series. Rough diamonds are heterogeneous natural products whose value varies according to characteristics such as size, shape, colour, clarity, modelled polished yield and broader assortment composition. Export parcels reported during different periods may therefore contain materially different combinations of goods. Changes in aggregate export outcomes may reflect both market conditions and changes in the quality and composition of the diamonds exported. A forecasting framework for Botswana must recognise this heterogeneity and avoid treating rough diamonds as a standardised commodity with a single directly observable international price.

Botswana's exposure to changing global demand conditions also supports the inclusion of both domestic and international predictive indicators. Rough-diamond export outcomes may contain predictive information associated with exchange-rate movements, global economic activity, luxury consumption conditions, international liquidity, competitor production and changes in natural-diamond supply. Country-level indicators, including foreign-exchange reserves and domestic macroeconomic conditions, may provide additional information about Botswana's external position. These variables must nevertheless be treated as potential predictors rather than causal determinants, and their usefulness must be established through temporally ordered out-of-sample evaluation.

Kimberley Process Rough-Diamond Data

The Kimberley Process Certification Scheme provides one of the principal internationally comparable sources of rough-diamond trade and production statistics. Its public statistical system reports participating countries' rough-diamond imports, exports and production in both value and carat volume, together with Kimberley Process certificate counts. The database includes annual summary tables, semi-annual production statistics and quarterly import and export tables. This structure makes Kimberley Process data potentially valuable for constructing a longitudinal measure of Botswana's rough-diamond export unit value.

The characteristics of Kimberley Process data also introduce methodological considerations. The publicly available statistics are aggregated at the participant and reporting-period levels rather than presented as observations for homogeneous individual stones or assortments. The Kimberley Process further advises users to consult the accompanying explanatory notes and indicates that public statistics may be released with a delay

and subsequently revised (Kimberley Process, n.d.). Therefore, a forecasting design must account for publication timing, revisions, missing observations and the information that would genuinely have been available at each historical forecast origin.

The distinction between production, imports, exports and re-exports is also important when interpreting rough-diamond trade statistics. International rough-diamond trade operates through interconnected producing, sorting and trading centres, and Kimberley Process participation influences access to formal rough-diamond markets and recorded trade flows (Borsky & Leiter, 2022). Consequently, Botswana's reported export value and exported carats must be examined for definitional consistency across time. Data-quality checks should consider whether changes in reporting coverage, trade classification or the composition of exported parcels could create apparent movements unrelated to a uniform change in rough-diamond market prices.

Rough-Diamond Export Unit Value and Composition Effects

The dependent variable proposed in this study is Botswana's rough-diamond export unit value, expressed in United States dollars per carat. For a given period (t), it is defined as:

$$U_t = \frac{V_t}{C_t}$$

where U_t represents Botswana's rough-diamond export unit value for period t , measured in United States dollars per carat, V_t represents the total value of rough-diamond exports during period t , measured in United States dollars, and C_t represents the total volume of rough-diamond exports during quarter t , measured in carats.

An export unit value is not equivalent to the market price of a homogeneous product. Unit values constructed from aggregate trade value and quantity can change because of movements in underlying transaction prices, changes in product quality, or changes in the composition of the exported basket. Silver (2009) demonstrates that trade unit-value indices may differ materially from true price indices when the quality or product mix of imports and exports changes. This concern is particularly relevant to rough diamonds because the exported assortment can vary considerably in value per carat according to the mix of sizes, qualities and other valuation characteristics.

The proposed forecasting target must therefore be interpreted as a composition-sensitive measure of average rough-diamond export value per carat. An increase in the measure does not necessarily mean that every category of rough diamond experienced an equivalent price increase. It may instead reflect the export of a relatively higher-value assortment, changing market conditions, or a combination of both. Similarly, a reduction in export unit value may reflect weaker market conditions, a lower-value export mix, or both mechanisms. The framework will consequently avoid referring to the dependent variable simply as "diamond price" and will maintain the term "Botswana's rough-diamond export unit value" throughout the data definitions, models, figures, tables, interpretation and conclusion.

Temporal disaggregation of this target also requires special treatment. Quarterly export unit values should not be divided or summed as if they were additive quantities. Export value and exported carat volume must instead be disaggregated separately, after which monthly export unit values can be calculated as the ratio of the estimated monthly export value to the estimated monthly carat volume. The resulting monthly estimates must satisfy the quarterly ratio identity:

$$U_q = \frac{\sum_{m=1}^3 V_{q,m}}{\sum_{m=1}^3 C_{q,m}} = \frac{V_q}{C_q}$$

where U_q is the quarterly rough-diamond export unit value in USD per carat; $V_{q,m}$ are rough-diamond export values for months 1, 2 and 3; and $C_{q,m}$ are rough-diamond export volumes in carats for months 1, 2 and 3.

This approach preserves the distinction between the additive components, namely export value and exported carats, and their non-additive ratio, the export unit value.

Existing Diamond-Price and Export-Unit-Value Research

Existing applications of machine learning to diamond valuation demonstrate that nonlinear models can represent complex relationships between product characteristics and value. For example, Kigo et al. (2023) evaluated several supervised machine-learning algorithms for predicting diamond prices using stone-level characteristics such as carat weight, cut, colour, clarity and physical dimensions. Such research is relevant because it illustrates the heterogeneity of diamonds and the nonlinear relationship between observable characteristics and individual-stone prices.

However, individual-diamond pricing studies address a different analytical problem from the one considered in this paper. They generally use cross-sectional observations of polished diamonds and seek to predict the price of an individual stone from its gemmological characteristics. They do not forecast an aggregated national time series of rough-diamond export value per carat, do not necessarily model macroeconomic and international-market indicators, and do not confront the same mixed-frequency, publication-lag and assortment-composition challenges. Their findings therefore cannot be transferred directly to Botswana's rough-diamond export unit-value forecasting problem.

Research on international rough-diamond trade provides additional evidence about the structure of recorded diamond flows. Borsky and Leiter (2022) examined the relationship between the Kimberley Process Certification Scheme and international rough-diamond trade values, demonstrating the relevance of institutional arrangements to recorded rough-diamond market access and trade. Nevertheless, this literature focuses primarily on trade relationships and institutional effects rather than forecasting country-level rough-diamond export unit values through time.

The available literature therefore contains relevant but fragmented evidence. General commodity-forecasting studies examine econometric and deep-learning models; diamond-pricing studies commonly predict individual polished-diamond prices from product characteristics; trade studies examine rough-diamond flows and institutional arrangements; and unit-value research demonstrates the measurement consequences of product-composition changes. Limited attention has been given to bringing these evidence streams together in a forecasting framework specifically designed for Botswana's aggregate rough-diamond export unit value.

Research Gap and Positioning of the Proposed Framework

Taken together, the literature identifies several relevant methodological and domain-specific developments: econometric benchmarking, hybrid deep learning, temporal harmonisation, rigorous time-series evaluation, explainable artificial intelligence, international rough-diamond trade measurement and the interpretation of composition-sensitive export unit values. These components, however, are generally investigated separately.

Conventional econometric models provide comparatively transparent benchmarks but may have limitations in representing complex nonlinear predictive relationships. Hybrid deep-learning models offer greater representational flexibility but introduce interpretability and reproducibility challenges. Explainable Artificial Intelligence provides mechanisms for interpreting complex predictions but is normally applied after model fitting.

Kimberley Process statistics provide rough-diamond export value and carat-volume data, but the resulting unit-value measure is sensitive to assortment composition, data revisions and publication timing. Existing diamond-pricing studies, meanwhile, largely address individual polished-diamond valuation rather than aggregate national rough-diamond export-unit-value forecasting.

The resulting research gap is both methodological and contextual. There remains limited research specifying how Botswana's rough-diamond export value and exported carats can be transformed into a temporally consistent, composition-aware export-unit-value forecasting target; how that target can be combined with

macroeconomic, diamond-market, country-level and global-supply predictors without information leakage; how a hybrid CNN–LSTM model can be evaluated against appropriate econometric and naïve benchmarks under walk-forward conditions; and how the fitted forecasts can subsequently be interpreted using SHAP without treating predictive contributions as causal effects.

The present study addresses this gap by proposing an integrated forecasting framework for Botswana’s rough-diamond export unit values. The framework will combine a hybrid CNN–LSTM architecture with appropriate benchmark models, temporal data harmonisation, walk-forward validation, forecast-accuracy evaluation and post-hoc SHAP interpretation within a consistent analytical pipeline. Rather than assuming that the hybrid architecture will outperform conventional alternatives, its forecasting performance will be evaluated empirically during the subsequent implementation of the framework.

After the CNN–LSTM model has been developed, fitted and evaluated, SHAP will be applied to quantify and interpret the predictive contributions of macroeconomic, diamond-market, country-level and global-supply indicators. In this way, the proposed framework brings together prediction, benchmarking, evaluation, temporal harmonisation and interpretability within a unified methodology designed to support forecasting rigour and policy-relevant interpretation in Botswana’s diamond-dependent economy.

METHODOLOGY

Overview of the Methodological Framework

Figure 3.1 presents the proposed framework for forecasting Botswana’s monthly rough-diamond export unit values, measured in United States dollars per carat. The framework comprises six sequential stages: data acquisition; temporal harmonisation; preprocessing and sequence construction; model development; out-of-sample evaluation; and post-hoc SHAP interpretation. This ordering follows established principles of temporally ordered forecast evaluation and interpretable predictive modelling (Cerqueira et al., 2020; Lundberg and Lee, 2017). It is specified prospectively and does not presume that the hybrid model will outperform the benchmarks.

The process begins with quarterly rough-diamond export value and corresponding exported carat volume from the Kimberley Process (Kimberley Process, n.d.). These additive components will be temporally disaggregated separately to monthly frequency using the specified Denton procedure, subject to aggregation constraints and indicator-series assumptions (Denton, 1971; International Monetary Fund, 2017). The procedure will not be applied directly to export unit value because that measure is a non-additive ratio.

$$V_{q,1} + V_{q,2} + V_{q,3} = V_q$$

$$C_{q,1} + C_{q,2} + C_{q,3} = C_q$$

For month m within quarter q , the monthly export unit value will be calculated as:

$$U_{q,m} = V_{q,m} / C_{q,m}$$

The quarterly identity will then be validated as:

$$U_q = (V_{q,1} + V_{q,2} + V_{q,3}) / (C_{q,1} + C_{q,2} + C_{q,3}) = V_q / C_q$$

The monthly target will be aligned with macroeconomic, diamond-market, country-level and global-supply predictors according to observation date, publication frequency, release delay and availability at each forecast origin. Missing-data treatment, transformations, lag construction and scaling parameters will be estimated from the training data only. Chronological ordering will be preserved throughout.

Proposed Forecasting Framework

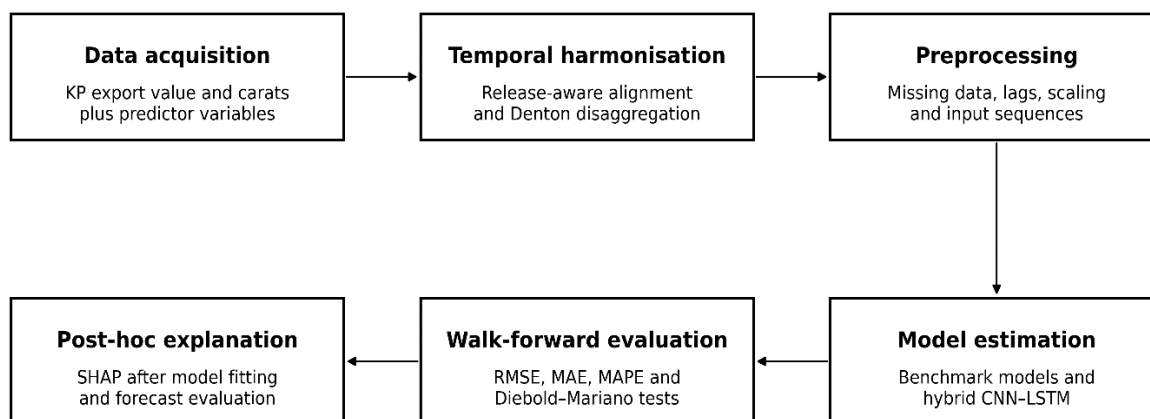


Figure 3.1: Proposed framework for forecasting Botswana’s rough-diamond export unit values

The hybrid CNN–LSTM will use multivariate input sequences. The CNN will extract local patterns, the LSTM will represent longer-term temporal dependencies, and a dense layer will generate the specified level forecast (Hochreiter and Schmidhuber, 1997; Sezer et al., 2020). The benchmark hierarchy will include naïve and seasonal-naïve forecasts, ARIMA, ARIMAX and an appropriately specified ARMA–GARCH model. Standard GARCH output will be interpreted as conditional variance and will not be compared directly with level forecasts (Engle, 1982; Bollerslev, 1986).

All directly comparable level forecasts will use the same target definition, forecast horizon, test period and forecast origins. RMSE, MAE and MAPE will summarise forecast errors, and pairwise Diebold–Mariano tests will assess forecast-loss differentials (Diebold and Mariano, 1995). SHAP will be applied only after the CNN–LSTM has been fitted and evaluated, and its values will be interpreted as predictive contributions rather than causal effects (Lundberg and Lee, 2017; Barredo Arrieta et al., 2020).

Table 3.1: Model Comparison Framework

Model	Information set	Output	Role	Strength	Limitation
ARIMA	Historical unit values	Unit-value level forecast	Primary traditional benchmark	Established, transparent and interpretable; effectively captures linear time-series patterns	Limited ability to represent nonlinear relationships and does not incorporate external predictors
ARMA–GARCH	Historical unit and residual innovations	Unit-value level forecast from the ARMA mean equation, plus conditional variance forecast from GARCH	Benchmark for modelling both expected unit values and time-varying volatility	Captures serial dependence, volatility clustering and changing forecast uncertainty	Relies on parametric assumptions and may not capture complex nonlinear relationships

Multivariate CNN–LSTM	Historical unit values and lagged macroeconomic, commodity-market and global-demand predictors	Unit-value level forecast	Principal proposed forecasting model	Captures short-term local patterns through CNN layers and longer-term temporal relationships through LSTM layers	More complex, data- and tuning-intensive, and requires post-hoc interpretation such as SHAP
------------------------------	--	---------------------------	--------------------------------------	--	---

Standard GARCH primarily forecasts conditional variance rather than the unit-value level. Therefore, this study applies an ARMA–GARCH specification in which the ARMA mean equation generates the unit-value forecast and the GARCH variance equation models time-varying volatility. This makes its mean forecast comparable with the forecasts produced by ARIMA and the proposed CNN–LSTM model. Because the CNN–LSTM uses additional predictors while ARIMA and ARMA–GARCH rely mainly on historical unit values, differences in forecasting performance will be interpreted with appropriate caution.

Design Principles of the Hybrid Forecasting Framework

The proposed framework is governed by five design principles that support coherent implementation, transparent comparison and policy-relevant interpretation. Explainability is planned from the outset but remains a post-hoc stage applied to the fitted CNN–LSTM model.

Principle	Operational definition	Methodological role	Contribution
Integrated modelling architecture	Preprocessing, modelling, benchmarking and evaluation are specified in one pipeline.	Creates an auditable sequence of analytical decisions.	Reproducibility and procedural consistency.
Pre-specified post-hoc explainability	SHAP is planned in advance and applied after model fitting.	Links fitted forecasts to systematic interpretation without changing SHAP’s post-hoc status.	Transparent predictive attribution.
Temporal integrity	Ordering, release lags and forecast-origin availability are preserved.	Reduces look-ahead bias and leakage.	More credible out-of-sample assessment.
Multi-scale representation	CNN extracts local features; LSTM represents longer dependencies.	Allows complementary temporal representations.	Flexible modelling of nonlinear sequences.
Policy-oriented interpretation	Outputs retain the export-unit-value definition and non-causal language.	Connects forecast evidence to Botswana’s fiscal context.	Improved interpretability and policy relevance.

These principles do not constitute evidence of achieved accuracy or explainability. Rather, they define the conditions under which the proposed framework will later be implemented, evaluated and interpreted.

METHODOLOGICAL POSITIONING

The proposed framework is positioned as an integrated predictive methodology rather than a causal model. Its novelty does not arise from claiming that CNN, LSTM, SHAP, ARIMA, GARCH or temporal disaggregation are individually new. The contribution lies in specifying how these established components can be combined coherently for Botswana's composition-sensitive rough-diamond export unit value.

First, explainability is incorporated into the research design from the outset. SHAP nevertheless remains post-hoc because it is calculated from predictions generated by the trained CNN–LSTM (Lundberg and Lee, 2017). Its implementation, reference dataset, aggregation across features and lags, and interpretation rules will be specified before evaluation.

Second, the framework distinguishes procedural consistency from identical information sets. All models will share the same target, forecast origins, horizon and evaluation period, but their inputs differ by design. ARIMA and naïve models are univariate; ARIMAX and the multivariate CNN–LSTM use lagged predictors; standard GARCH models conditional variance; and ARMA–GARCH produces both a conditional-mean level forecast and a conditional-variance forecast. Results will therefore be interpreted by comparison class rather than as a single undifferentiated ranking.

Third, temporal integrity and reproducibility are enforced throughout. Raw-data versions, transformations, indicator choices, disaggregation constraints, release lags, forecast origins, hyperparameters and random seeds will be recorded. The test period will not be used for hyperparameter selection. These controls follow the requirement that time-series performance estimation preserve temporal order and exclude future information (Bergmeir et al., 2018; Cerqueira et al., 2020).

Taken together, these features provide a structured basis for later evaluating predictive performance and interpreting fitted forecasts. They do not constitute empirical evidence of model superiority at the framework stage.

System Architecture of the Hybrid Forecasting Framework

Figure 3.2 defines the operational architecture in terms of the data objects passed between components. The first input object contains quarterly rough-diamond export value and exported carats. A separate predictor object contains macroeconomic, diamond-market, country-level and global-supply series, together with release dates and revision information.

The temporal-harmonisation component produces estimated monthly export value and monthly exported carats under separate aggregation constraints. Their ratio produces the monthly rough-diamond export-unit-value target. Predictor series are transformed, lagged and aligned according to their real-time availability. The output is a monthly analytical table containing the target, admissible lagged predictors and metadata required to reproduce each transformation.

Within each training window, preprocessing parameters are estimated using training observations only. The resulting input tensor X has dimensions [samples, sequence length, features], while the target vector y contains the corresponding future export-unit values for the specified horizon. The CNN block transforms X into local feature maps; the LSTM block converts those maps into a temporal hidden representation; and the dense output layer generates the level forecast.

The benchmark branch receives the information set appropriate to each model. The evaluation branch receives aligned actual values and forecasts indexed by model, forecast origin and horizon. It calculates error measures and pairwise forecast-loss tests. Only after evaluation does the SHAP branch receive the fitted CNN–LSTM,

its background/reference sample and evaluation sequences. It returns local feature-by-lag contributions and aggregated global and temporal importance summaries.

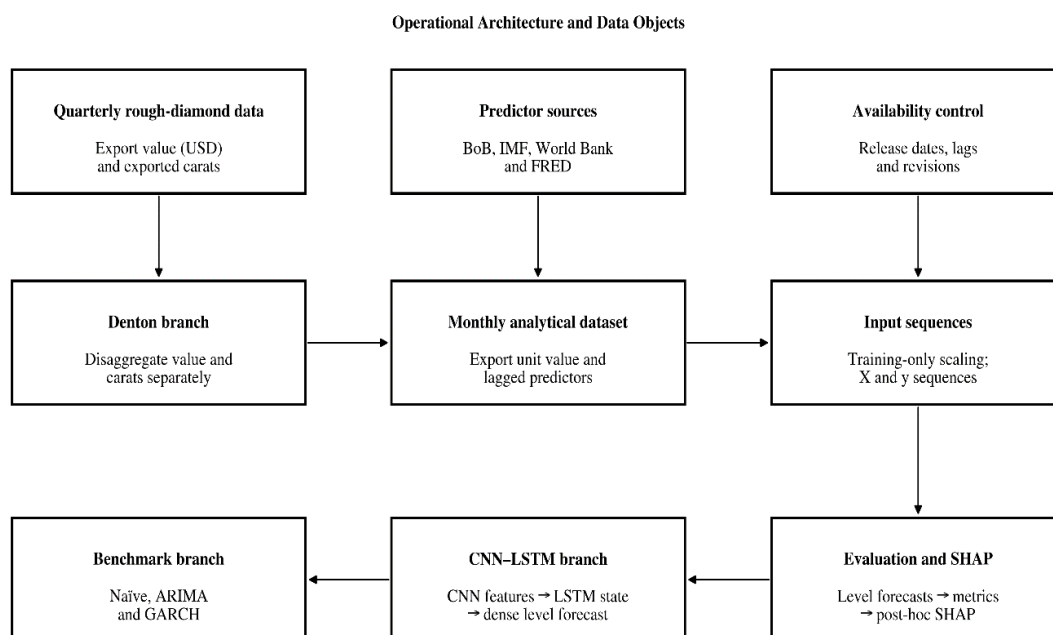


Figure 3.2: Operational architecture of the proposed CNN–LSTM forecasting framework

Data Description and Variables

The proposed study will use secondary time-series data covering January 2004 to December 2025, subject to final verification of coverage, definitions and release histories. The principal sources will include the Kimberley Process, Statistics Botswana, Bank of Botswana, International Monetary Fund, World Bank and Federal Reserve Economic Data. Kimberley Process statistics report rough-diamond trade and production by value and volume, but their explanatory notes, release delays and revisions must be considered (Kimberley Process, n.d.). The target is Botswana’s rough-diamond export unit value in USD per carat, derived from export value and exported carats. It is an average, composition-sensitive export measure rather than a homogeneous market price (Silver, 2009).

Predictors are grouped as macroeconomic, diamond-market, country-level and global-supply variables. Table 3.3 provides a provisional data dictionary. Before empirical implementation, each entry will be finalised with an exact provider series name, code or ticker, unit, native frequency, coverage, transformation, publication lag and availability rule. Undefined proxy labels will not be retained in the final specification.

Table 3.3: Provisional Data Description and Variable Specification

Class	Variable	Unit	Source	Native frequency	Transformation and leakage control
Target	Rough-diamond export unit value	USD/carat	KP value ÷ KP carats	Monthly estimate	Disaggregate value and carats separately; never use contemporaneous value as predictor
Macroeconomic	Global activity indicator	Index or growth	World Bank/IMF exact series	Monthly/quarterly	Lag to publication availability

Macroeconomic	Botswana real GDP growth	Percent	Statistics Botswana/BoB	Quarterly	Denton only with stated indicator; apply release lag
Macroeconomic	Botswana CPI inflation	Percent	Statistics Botswana/BoB	Monthly	Use publication-available lag
Macroeconomic	BWP per USD exchange rate	BWP/USD	Bank of Botswana	Monthly/daily	Monthly aggregate known by forecast origin
Macroeconomic	Policy interest rate	Percent	Bank of Botswana	Event/monthly	Carry forward only from effective date
Macroeconomic	Brent crude oil price	USD/barrel	FRED series exact	Daily/monthly	Monthly mean or end-period, pre-specified
Diamond market	Gold price	USD/troy ounce	FRED series exact	Daily/monthly	Pre-specified monthly aggregation and lag
Diamond market	Global liquidity indicator	Exact unit	IMF/FRED exact series	Monthly	Code and transformation to be fixed
Diamond market	Luxury-demand indicator	Exact index	Named provider/ticker required	Monthly/quarterly	No vague proxy; justify coverage and transformation
Country	Foreign-exchange reserves	USD months or of imports	Bank of Botswana/IMF	Monthly	Lag to first public release
Country	Lagged diamond export value	USD	Kimberley Process	Quarterly	Optional; strictly lagged and leakage-tested
Global supply	Competitor rough production	Carats	Kimberley Process	Semi-annual/annual	Release-aware lag; no contemporaneous unavailable values
Diamond market	Rough-diamond market indicator	Exact index/unit	Online	Monthly	Pre-specify licensed/public proxy and sensitivity test

Expected directional signs will not be presented as causal effects. If retained in the final data dictionary, the column will be titled “Expected predictive association,” and every sign will be justified from theory or prior evidence. Predictor selection will be conducted within the training and validation periods only.

Contemporaneous export value will not be used as a predictor because it forms the numerator of the target and would create mathematical target leakage. Any retained export-value feature must be strictly lagged, observable at the forecast origin and evaluated through sensitivity analysis. The same real-time availability rule will apply to every predictor, including revised macroeconomic series.

The final harmonised data matrix will be subjected to checks for missingness, duplicates, implausible values, structural breaks and revision inconsistencies. A reproducibility register will preserve original downloads, extraction dates, units, transformations, lags and exclusions. These procedures define the proposed data design; empirical sample counts and final predictor specifications will be reported only after data verification and implementation.

Data Preprocessing and Temporal Disaggregation

All variables will be harmonised to monthly frequency before estimation. Series already reported monthly or more frequently will be aggregated according to a pre-specified rule. End-of-period values will be used for stock variables where conceptually appropriate, while monthly means or sums will be used for rate and flow variables according to their definitions. Lower-frequency series will be temporally disaggregated only when this is required by the framework and when the resulting assumptions can be stated transparently (International Monetary Fund, 2017).

Quarterly rough-diamond export value and quarterly exported carats will be disaggregated separately. For export value, the preferred high-frequency indicator will be Statistics Botswana's monthly diamond-export value series, provided that its commodity coverage, valuation basis and currency definition can be reconciled with the Kimberley Process series. For exported carats, a verified monthly rough-diamond export-volume series will be used if one is available from Statistics Botswana, the relevant national authority or an auditable industry source. Production or sales carats will not be substituted automatically for exported carats because those concepts may differ through inventories, timing and channel composition.

If no defensible monthly indicator exists for exported carats, the proportional Denton–Cholette procedure will be applied to a flat preliminary monthly indicator, subject to the quarterly carat-volume constraint. In that case, the monthly pattern will be identified principally by the benchmarking restriction rather than observed within-quarter export movements. This limitation will be reported explicitly, and quarterly-frequency estimation will be retained as a robustness specification.

For a positive quarterly flow Y_q and a preliminary monthly indicator x_t , the proportional first-difference Denton problem is represented as follows (Denton, 1971; International Monetary Fund, 2017):

$$\min_{\{y_t\}_{t=1}^T} \sum_{t=2}^T \left(\frac{y_t}{x_t} - \frac{y_{t-1}}{x_{t-1}} \right)^2$$

subject to the quarterly aggregation constraint:

$$\sum_{m=1}^3 y_{q,m} = Y_q$$

This optimisation will be performed separately for export value and exported carats. Monthly export unit value will then be computed as:

$$U_q = \frac{\sum_{m=1}^3 V_{q,m}}{\sum_{m=1}^3 C_{q,m}}$$

Equivalent:

$$U_q = \frac{V_{q,1} + V_{q,2} + V_{q,3}}{C_{q,1} + C_{q,2} + C_{q,3}}$$

where $V_{q,m}$ is the export value and $C_{q,m}$ is the exported carat volume for month m within quarter q . Denton estimates will not be described as newly observed data or as additional independent information.

Preprocessing will include definition checks, duplicate detection, missingness assessment, outlier diagnostics and transformation of skewed positive series where justified. Missing observations will not be filled using future information. Any interpolation or imputation will be fitted within each training window and documented. Potential outliers will be retained unless a demonstrable data error exists; sensitivity analysis will examine winsorisation and robust scaling.

All scalars, imputation rules, feature-selection decisions and transformation parameters will be estimated on the training data and applied unchanged to validation and test observations. Predictor lags will incorporate both economic timing and actual publication delays. Each forecast-origin dataset will therefore include only information that would have been observable when the forecast was generated.

Hybrid CNN–LSTM Model Formulation

The input to the hybrid model will be a three-dimensional tensor X with dimensions [number of samples, sequence length, number of features]. Each sample will contain the lagged monthly rough-diamond export unit value and the admissible lagged predictors over the preceding sequence window. The output will be a one-month-ahead forecast of Botswana's rough-diamond export unit value in USD per carat. This sequence-to-one construction is consistent with established deep-learning approaches to time-series forecasting (Goodfellow et al., 2016; Sezer et al., 2020).

For convolutional filter k , the one-dimensional feature map at time t is defined as follows (Goodfellow et al., 2016):

$$a_t^{(k)} = \text{ReLU}\left(\sum_{j=0}^{K-1} W_j^{(k)} X_{t-j} + b^{(k)}\right)$$

where K is the kernel size, W denotes convolutional weights, b is the bias term and ReLU is the activation function. The resulting feature sequence is passed to the LSTM. Following the standard gated recurrent formulation, the complete LSTM updates are (Hochreiter and Schmidhuber, 1997; Goodfellow et al., 2016):

$$f_t = \sigma(W_f a_t + U_f h_{t-1} + b_f)$$

$$i_t = \sigma(W_i a_t + U_i h_{t-1} + b_i)$$

$$g_t = \tanh(W_g a_t + U_g h_{t-1} + b_g)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$o_t = \sigma(W_o a_t + U_o h_{t-1} + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$

Here f_t , i_t and o_t are the forget, input and output gates; g_t is the candidate cell state; c_t is the updated cell state; h_t is the hidden state; σ is the logistic sigmoid function; and $\odot$ denotes element-wise multiplication. The one-month-ahead level forecast is generated by the dense output layer:

$$\hat{U}_{t+1} = W_y h_t + b_y$$

where $\hat{U}_{t+1}$ denotes the predicted rough-diamond export unit value in USD per carat for the next month, and W_y and b_y are the output-layer parameters. Training will minimise mean squared error:

$$MSE = \frac{1}{N} \sum_{t=1}^N (U_t - \hat{U}_t)^2$$

where N is the number of evaluated forecasts, U_t is the observed rough-diamond export unit value at time t and $\hat{U}_t$ is the corresponding forecast. The initial search space is specified in Table 3.4. The final configuration will be selected using validation performance only and then fixed before the test period is evaluated.

Table 3.4: Proposed CNN–LSTM Architecture and Hyperparameter Search Space

Component	Candidate specification	Purpose
Sequence length	6, 12, 18 or 24 months	Captures alternative short- and medium-term histories
CNN filters	16, 32 or 64	Controls local feature-map capacity
Kernel size	2, 3 or 4	Controls local temporal receptive field
Padding	Causal	Prevents future positions entering convolution
Activation	ReLU	Nonlinear feature transformation
Pooling	None or max-pooling size 2	Selected on validation data
LSTM units	16, 32 or 64	Controls recurrent representation
Dropout	0.0, 0.1, 0.2 or 0.3	Regularisation
Dense output	One linear unit	One-month-ahead level forecast
Optimiser	Adam	Gradient-based optimisation
Learning rate	0.001, 0.0005 or 0.0001	Validation-selected step size
Batch size	8, 16 or 32	Appropriate to the limited time-series sample
Epoch limit	Maximum 300	Subject to early stopping
Early stopping	Patience 20; restore best weights	Uses validation loss only

Benchmark Models and Forecast Evaluation

The benchmark suite will comprise naïve, ARIMA and GARCH models. These benchmarks represent progressively more structured approaches to persistence, conditional-mean forecasting and volatility modelling. Their performance will be evaluated without presuming that any benchmark or the proposed CNN–LSTM model will be superior.

The naïve benchmark will set the next month's rough-diamond export unit value equal to the most recently observed value:

$$\hat{U}_{t+1|t} = U_t$$

where U_t is the observed rough-diamond export unit value at time t , and $\hat{U}_{t+1|t}$ is its one-month-ahead forecast produced using information available at time t .

The ARIMA model will provide a univariate linear benchmark for the conditional mean of the export unit-value series. Its general specification is:

$$\Phi(B)(1 - B)^d U_t = c + \Theta(B)\varepsilon_t$$

where B is the backshift operator, d is the order of differencing, $\Phi(B)$ is the autoregressive polynomial, $\Theta(B)$ is the moving-average polynomial, c is a constant or drift term where admissible, and ε_t is the innovation. Candidate ARIMA orders will be selected using the corrected Akaike Information Criterion within pre-specified bounds and using only observations available at each forecast origin (Box et al., 2015; Hyndman and Athanasopoulos, 2021).

The GARCH benchmark will distinguish conditional-mean forecasting from conditional-variance modelling. Standard GARCH primarily models conditional heteroskedasticity and does not independently generate a rough-diamond export unit-value level forecast comparable with the naïve, ARIMA and CNN–LSTM forecasts (Engle, 1982; Bollerslev, 1986). Accordingly, the GARCH benchmark will be estimated with an ARMA conditional-mean equation:

$$U_t = \mu + \sum_{i=1}^p \phi_i U_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t$$

$$\varepsilon_t = \sigma_t z_t, \quad \sigma_t^2 = \omega + \sum_{i=1}^P \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^Q \beta_j \sigma_{t-j}^2$$

where U_t is the rough-diamond export unit value; μ is the conditional-mean constant; p and q are the AR and MA orders; ϕ_i and θ_j are their respective coefficients; ε_t is the innovation; σ_t^2 is the conditional variance; P and Q are the ARCH and GARCH orders; and z_t is a standardised innovation satisfying $E(z_t) = 0$ and $\text{Var}(z_t) = 1$.

The conditional-mean component will generate the GARCH benchmark's one-month-ahead unit-value forecast. This forecast will enter the level-accuracy comparison. The corresponding conditional-variance forecast will be reported separately as a measure of forecast uncertainty and will not be compared with unit-value forecasts using level-error metrics.

For H one-month-ahead test forecasts, the forecast error is:

$$e_h = U_h - \hat{U}_h$$

The level-forecast accuracy measures are defined as:

$$RMSE = \sqrt{\frac{1}{H} \sum_{h=1}^H (U_h - \hat{U}_h)^2}$$

$$MAE = \frac{1}{H} \sum_{h=1}^H |U_h - \hat{U}_h|$$

$$MAPE = \frac{100}{H} \sum_{h=1}^H \left| \frac{U_h - \hat{U}_h}{U_h} \right|$$

MAPE will be reported because rough-diamond export unit values are expected to remain strictly positive. Nevertheless, MAPE can become unstable when observed values approach zero. A scale-free alternative, such as Mean Absolute Scaled Error, will therefore be reported as a robustness measure if near-zero or unstable denominators are encountered (Hyndman and Athanasopoulos, 2021).

The Diebold–Mariano test will be applied to pairwise forecast-loss differentials rather than attributed to an individual model (Diebold and Mariano, 1995). The primary comparisons will be: CNN–LSTM versus naïve; CNN–LSTM versus ARIMA; and CNN–LSTM versus GARCH.

Additional pairwise comparisons among the three benchmarks may be reported for completeness. Squared-error loss will support comparisons corresponding to RMSE, while absolute-error loss will support comparisons corresponding to MAE. A heteroskedasticity-and-autocorrelation-consistent variance estimator will be employed. For the one-month-ahead forecast horizon, the baseline truncation lag will be zero, and the Harvey–Leybourne–Newbold small-sample correction will be reported as a sensitivity analysis (Harvey et al., 1997).

This version is consistent with your decision to retain only naïve, ARIMA and GARCH as benchmarks while ensuring that the GARCH branch produces a level forecast that can legitimately be compared with the CNN–LSTM forecast.

Model Training, Validation and Forecasting Design

The intended monthly period January 2004 to December 2025 contains 264 observations before sequence construction and missing-data losses. The fixed chronological split will comprise January 2004–December 2018 for training (180 months), January 2019–December 2021 for validation (36 months), and January 2022–December 2025 for final testing (48 months). Effective sequence counts will be lower by the selected look-back length and will be reported after harmonisation. Disaggregated monthly estimates will be included in all three periods, and quarterly-frequency models will provide an explicit robustness check.

The primary forecasting task will be one-month-ahead prediction using an expanding window, thereby preserving temporal order and approximating real-time forecast generation (Cerqueira et al., 2020; Hyndman and Athanasopoulos, 2021). Hyperparameters will be selected once from the training and validation periods and fixed before test evaluation. During the test period, the CNN–LSTM weights will be re-estimated at each monthly origin using all observations available up to that origin, but its architecture and hyperparameters will not be re-tuned. A computational robustness check will compare less frequent quarterly re-estimation.

At each forecast origin, naïve forecasts will be updated mechanically. ARIMA orders will be reselected using the corrected corrected Akaike Information Criterion (AICc) within bounded autoregressive and moving-average orders, pp and qq, ranging from 0 to 5, and admissible differencing orders, dd, ranging from 0 to 2. GARCH specifications will be selected from GARCH(1,1), GARCH(1,2), GARCH(2,1) and GARCH(2,2) using the Bayesian Information Criterion (BIC) and residual diagnostics. All model selection will use only the

data available at the relevant forecast origin. No test-period forecast error will be used to select a model order or hyperparameter.

Neural-network results will be reproduced across 30 random seeds. At each forecast origin, the primary CNN–LSTM forecast will be the median prediction across the validated seed runs, while the distribution of forecast errors will also be reported. Early stopping will monitor validation loss only during architecture selection. During expanding-window re-estimation over the test period, the stopping epoch will be fixed using evidence from the validation stage or determined using an internal training-only tail that excludes the observation being forecast.

Robustness analysis will examine quarterly rather than monthly target estimation; alternative Denton indicators and the no-indicator Denton–Cholette specification; predictor subsets organised by conceptual group; predictor lags of 1, 3, 6 and 12 months; sequence lengths of 6, 12, 18 and 24 months; 30 random-seed repetitions; alternative ARIMA orders and GARCH specifications; robust versus standard scaling; and the exclusion of lagged export unit value. Conclusions will be described as robust only when the relative model ranking, statistical comparisons and principal SHAP patterns remain substantively stable.

Important methodological note: standard GARCH primarily forecasts conditional variance. If GARCH is compared directly with ARIMA and CNN–LSTM unit-value forecasts, its conditional-mean equation and the procedure used to derive the level forecast must be stated explicitly.

Explainability Framework: SHAP-Based Interpretation

SHAP will be implemented only after the CNN–LSTM has been developed, fitted and evaluated (Lundberg and Lee, 2017). GradientExplainer is proposed as the primary implementation because its expected-gradients formulation provides gradient-based attribution for differentiable models (Erion et al., 2021). DeepExplainer will be used as a sensitivity implementation where it is compatible with the final network layers and software environment. The exact software versions, model wrapper and output dimensions will be recorded.

The SHAP background distribution will comprise 100 training sequences selected by k-means summarisation of the scaled training tensor. If fewer than 100 valid training sequences remain, all available training sequences will be used. Validation and test observations will not be included in the background dataset. SHAP values will be computed for each predictor at each lag within the selected sequence window, producing an array indexed by forecast, time step and feature.

Local explanations will display signed contributions for an individual forecast. Global importance for feature j will be calculated as the mean absolute SHAP value across forecasts and sequence positions:

$$I_j = \frac{1}{HL} \sum_{h=1}^H \sum_{l=1}^L |\phi_{h,l,j}|$$

where I_j is the global SHAP importance of predictor j , H is the total number of evaluated forecasts, L is the number of time steps or lags in each input sequence and $\phi_{h,l,j}$ is the SHAP value assigned for predictor j , at sequence position l , for forecast h ; and $|\phi_{h,l,j}|$ is its absolute predictive contribution. Temporal-lag importance will retain the lag dimension, while calendar-time analysis will aggregate signed and absolute contributions by forecast origin. This will allow examination of whether predictive contributions vary across the evaluation period without describing those patterns as causal effects.

Because correlated predictors can share or redistribute attribution, SHAP rankings will not be interpreted as unique structural effects (Aas et al., 2021). Correlation matrices and hierarchical clustering will identify strongly related predictors. Results will be reported both at individual-variable level and for pre-defined correlated groups, supported by sensitivity analyses that remove one variable from highly correlated pairs. Directional language will remain predictive: positive or negative SHAP values indicate contributions relative

to the background expectation, not that a predictor causes the export unit value to change (Barredo Arrieta et al., 2020).

The SHAP stage therefore complements, but does not replace, forecast-accuracy evaluation. It is pre-specified within the proposed analytical pipeline while remaining post-hoc in its application to the fitted CNN–LSTM forecasts.

DISCUSSION AND METHODOLOGICAL IMPLICATIONS

This paper proposes an integrated framework for forecasting Botswana’s rough-diamond export unit values using a hybrid CNN–LSTM model, conventional benchmarks and post-hoc explainability. Because the framework has not yet been implemented empirically, this section discusses its methodological rationale and anticipated implications rather than presenting realised forecasting results. No claim is therefore made that the proposed CNN–LSTM will outperform the benchmark models.

The benchmark suite comprises naïve, ARIMA and GARCH models. The naïve model provides a minimum persistence benchmark by setting the next month’s export unit value equal to the most recently observed value. ARIMA provides a conventional univariate benchmark for modelling linear dependence in the conditional mean of the export unit-value series through autoregressive, differencing and moving-average components (Box et al., 2015; Hyndman and Athanasopoulos, 2021). These models establish whether the additional complexity of the CNN–LSTM produces forecast improvements beyond simple persistence and established linear time-series dynamics.

The role of GARCH requires a distinction between level forecasting and volatility modelling. Standard GARCH specifications principally model conditional variance and volatility clustering rather than the level of the forecast variable (Engle, 1982; Bollerslev, 1986). Accordingly, the GARCH benchmark will include an ARMA conditional-mean equation for generating rough-diamond export unit-value forecasts and a GARCH conditional-variance equation for modelling time-varying forecast uncertainty. Only the conditional-mean forecasts will be evaluated directly against the naïve, ARIMA and CNN–LSTM level forecasts using RMSE, MAE and MAPE. Conditional-variance forecasts will be reported separately and interpreted as measures of volatility rather than unit-value forecasts.

The proposed CNN–LSTM architecture is motivated by the complementary capabilities of its components. The convolutional layers are designed to extract local patterns from multivariate input sequences, whereas the LSTM layers use gated recurrent mechanisms to represent longer-term temporal dependencies (Hochreiter and Schmidhuber, 1997; Goodfellow et al., 2016; Sezer et al., 2020). This architectural rationale provides a theoretical basis for evaluating the hybrid model but does not establish its empirical superiority. Forecasting performance will therefore be determined through direct out-of-sample comparison with the naïve, ARIMA and GARCH benchmarks.

All models will be evaluated using an expanding-window walk-forward procedure that preserves the chronological ordering of observations and approximates real forecasting conditions. At each forecast origin, data preprocessing, model estimation, order selection and forecast generation will use only information available at that time. This approach reduces the risk of look-ahead bias and information leakage that may arise when conventional random cross-validation is applied to time-series data (Bergmeir et al., 2018; Cerqueira et al., 2020).

The comparisons must nevertheless be interpreted in relation to the information available to each model. The naïve and ARIMA models use historical values of the dependent variable, while the CNN–LSTM uses multivariate sequences containing lagged macroeconomic, diamond-market, country-level and global-supply indicators. The GARCH benchmark uses the historical export unit-value series through its conditional-mean and conditional-variance components. Differences in forecasting performance may consequently reflect both model architecture and differences in the information sets employed. The results will therefore be interpreted

as comparisons between complete forecasting approaches rather than as isolated tests of algorithmic superiority.

Forecast accuracy will be evaluated using RMSE, unordered MAE and MAPE, which provide complementary assessments of the magnitude and distribution of forecast errors. Pair CC wise Diebold–Mariano tests will be used to assess whether differences in forecast loss between the CNN–LSTM and each benchmark are statistically significant (Diebold and Mariano, 1995). The principal comparisons will be CNN–LSTM versus naïve, CNN–LSTM versus ARIMA and CNN–LSTM versus GARCH. The Harvey–Leybourne–Newbold small-sample correction will be reported as a sensitivity analysis (Harvey et al., 1997). Additional comparisons among the three benchmark models may be included where they contribute to the interpretation of relative forecasting performance.

The proposed data architecture addresses the mixed-frequency structure of the candidate variables. Quarterly rough-diamond export value and exported carats will be temporally disaggregated separately using the specified Denton procedure, subject to quarterly aggregation constraints. Monthly export unit values will subsequently be constructed by dividing the estimated monthly export value by the corresponding estimated monthly carats (Denton, 1971; International Monetary Fund, 2017). The export unit value itself will not be disaggregated directly because it is a non-additive ratio.

This treatment is methodologically important because Botswana’s rough-diamond export unit value is not equivalent to the market price of a homogeneous commodity. The measure represents the average export value per carat and may change because of underlying market conditions, variation in the size and quality of exported goods, or broader changes in assortment composition (Silver, 2009). Consequently, an increase in export unit value does not necessarily indicate that the prices of all rough-diamond categories increased proportionately. The dependent variable must therefore be described consistently as a composition-sensitive rough-diamond export unit value rather than simply as a diamond price.

Variables obtained from different sources will be aligned to a common monthly calendar using their publication frequencies, release dates, revision histories and availability at each forecast origin. These controls are necessary to prevent subsequently released or revised information from entering earlier forecasts. The harmonised predictors will be evaluated according to whether their historically available values provide useful information for forecasting subsequent rough-diamond export unit values.

The relationships identified by the forecasting models will be interpreted as predictive rather than causal. Neither the multivariate CNN–LSTM specification nor SHAP attribution provides an identification strategy for estimating causal effects. The study will therefore avoid statements that individual predictors determine or cause changes in rough-diamond export unit values. Establishing causal relationships would require a separate research design and additional identifying assumptions beyond the scope of the proposed framework.

Explainability is incorporated into the research design while remaining post-hoc in its analytical application. SHAP will be applied only after the CNN–LSTM has been developed, fitted, used to generate forecasts and evaluated against the benchmarks (Lundberg and Lee, 2017). The analysis will provide local explanations for selected forecasts, global estimates of predictor importance and temporal summaries showing whether predictive contributions vary across forecast origins.

SHAP values will be interpreted as model-dependent contributions relative to the selected background expectation. Positive and negative SHAP values will indicate whether a predictor increases or decreases a particular forecast relative to that expectation; they will not demonstrate that the predictor causes the export unit value to change. Because correlated predictors may share or redistribute attribution, individual feature rankings may be sensitive to the predictor set and background distribution. Correlation diagnostics, grouped-predictor reporting and sensitivity analyses will therefore accompany the SHAP results (Aas et al., 2021; Barredo Arrieta et al., 2020).

The framework’s principal methodological contribution lies in connecting target construction, temporal harmonisation, benchmark forecasting, hybrid deep learning, temporally ordered evaluation and post-hoc

explanation within a single reproducible protocol. It does not presume that greater model complexity necessarily produces greater forecasting accuracy. Instead, it provides a structured procedure for determining whether the CNN–LSTM delivers incremental predictive value and whether its forecasts can be interpreted sufficiently for policy-related analysis.

As this is a proposed-framework paper, no empirical findings are reported at this stage. Subsequent implementation should include a distinct Results section presenting the final sample and descriptive statistics; temporal-disaggregation validation; selected ARIMA and GARCH specifications; CNN–LSTM architecture and hyperparameters; actual-versus-forecast plots; RMSE, MAE and MAPE results; pairwise Diebold–Mariano tests; robustness analyses; and local, global and temporal SHAP outputs. Only after these results have been reported should conclusions be drawn regarding comparative accuracy, statistical significance, attribution stability and policy relevance.

Expected Contribution to Knowledge

This study is expected to contribute to forecasting research by specifying a transparent and reproducible protocol for forecasting Botswana’s rough-diamond export unit values. Rather than presuming that the proposed hybrid model is universally superior, the framework integrates target construction, mixed-frequency data harmonisation, release-aware predictor alignment, benchmark specification, walk-forward evaluation and post-hoc explanation within a pre-specified analytical sequence. Documenting these stages makes the underlying assumptions, modelling decisions and evaluation procedures available for replication and methodological scrutiny.

The second expected contribution is a structured comparison of the hybrid CNN–LSTM with three clearly defined benchmarks: naïve, ARIMA and GARCH. The naïve model represents simple persistence, ARIMA provides a conventional univariate conditional-mean benchmark, and the GARCH benchmark combines an ARMA conditional-mean equation with a GARCH conditional-variance specification. This design permits direct comparison of the models’ unit-value forecasts while retaining the GARCH model’s separate representation of time-varying volatility.

The benchmark comparison will also clarify the importance of the information available to each forecasting approach. The naïve and ARIMA models rely on historical values of the rough-diamond export unit-value series, while the CNN–LSTM uses multivariate sequences containing lagged macroeconomic, diamond-market, country-level and global-supply indicators. The GARCH benchmark uses the historical export unit-value series through its conditional-mean and conditional-variance components. Differences in predictive performance will therefore be interpreted in relation to both model structure and information set, rather than as an unconditional comparison of modelling classes.

The third expected contribution concerns the integration of forecasting and explainability. SHAP is pre-specified within the analytical protocol but will be applied post-hoc only after the CNN–LSTM has been fitted, used to generate forecasts and evaluated against the benchmarks (Lundberg and Lee, 2017). The proposed analysis is designed to produce local explanations for individual forecasts, global summaries of predictor importance and temporal profiles showing how predictive contributions may vary across input lags and forecast origins.

This treatment extends the interpretation of the fitted CNN–LSTM beyond a single static ranking of predictors. It may reveal whether particular macroeconomic, diamond-market, country-level or global-supply indicators contribute more strongly to forecasts during specific periods. However, the resulting SHAP values will remain conditional on the fitted model, selected predictors and reference distribution. They will therefore be interpreted as predictive contributions rather than causal effects. Potential attribution instability arising from correlated predictors will be examined through correlation diagnostics, grouped reporting and sensitivity analyses (Aas et al., 2021; Barredo Arrieta et al., 2020).

The fourth expected contribution is the study’s domain-specific treatment of Botswana’s rough-diamond export unit value. The target will be constructed by dividing rough-diamond export value by the corresponding

volume of exported carats. Because export value and carats are additive but their ratio is not, the two components will be temporally disaggregated separately before calculating the monthly export unit value. This procedure provides a methodologically coherent approach to constructing the monthly forecasting target from the quarterly Kimberley Process data.

The study further recognises that rough diamonds are heterogeneous and that the resulting export unit value is a composition-sensitive measure rather than the directly observed market price of a homogeneous product. Changes in the measure may reflect market conditions, variation in the size and quality of exported diamonds, changes in assortment composition, or a combination of these factors (Kimberley Process, n.d.; Silver, 2009). Maintaining this distinction throughout the analysis is expected to improve the conceptual precision of research using aggregated diamond trade data.

Finally, the framework is expected to provide a foundation for policy-relevant forecasting research in commodity-dependent economies. By integrating transparent benchmark comparison, nonlinear predictive modelling, temporally valid evaluation and post-hoc explanation, the protocol is designed to support forecasts whose accuracy and predictive drivers can be examined systematically. Nevertheless, the practical contribution will depend on the subsequent empirical results. Claims concerning comparative accuracy, statistical significance, robustness and policy usefulness will therefore be made only after the proposed framework has been implemented and evaluated.

LIMITATIONS

This study has several limitations. First, it presents a proposed methodological framework and does not report completed empirical forecasting or SHAP results. No conclusions can therefore be drawn at this stage regarding the comparative accuracy of the CNN–LSTM, naïve, ARIMA and GARCH models, the statistical significance of forecast differences, or the framework’s practical usefulness. These conclusions must await implementation and out-of-sample evaluation.

Second, the proposed monthly period of 2004–2025 provides approximately 264 observations before accounting for missing values, input lags, sequence construction and the division of the data into training, validation and test periods. This is a relatively modest sample for a multivariate CNN–LSTM and may increase the risk of overfitting. The study will mitigate this risk through parsimonious predictor selection, regularisation, early stopping, temporally ordered validation and robustness analysis across alternative specifications and random seeds.

Third, the models use different information sets. The naïve and ARIMA benchmarks rely on historical export unit values, whereas the CNN–LSTM uses multivariate predictor sequences. The GARCH benchmark models conditional-mean and conditional-variance dynamics using the historical export unit-value series. Differences in forecasting performance may therefore reflect both model structure and the information supplied to each model. In addition, the GARCH level forecast and conditional-variance forecast answer different questions and will be reported separately.

Fourth, the study relies on secondary data that may contain missing observations, revisions, reporting delays, definitional changes and measurement errors. Kimberley Process export values and carat volumes aggregate heterogeneous rough diamonds; consequently, the derived unit value may reflect changes in assortment composition as well as market conditions (Kimberley Process, n.d.; Silver, 2009). It must therefore be interpreted as a composition-sensitive average export value per carat rather than the price of a homogeneous product.

Fifth, temporal disaggregation introduces uncertainty because the within-quarter distribution of exports is estimated rather than directly observed. Although the Denton procedure preserves quarterly aggregation constraints, the estimated monthly series depends on the selected indicator, disaggregation specification and associated assumptions (Denton, 1971; International Monetary Fund, 2017). Alternative specifications will therefore be examined through sensitivity analysis.

Finally, SHAP explanations are conditional on the fitted model, selected predictors and background distribution. Correlated predictors may share or redistribute their attributed contributions, potentially affecting feature-importance rankings (Aas et al., 2021). SHAP values will consequently be interpreted as model-dependent predictive contributions rather than causal effects. Furthermore, because Botswana's diamond sector has distinctive institutional and economic characteristics, future empirical findings may not generalise directly to other countries, commodities, forecasting horizons or data periods.

CONCLUSION

This paper proposes a unified and reproducible framework for forecasting Botswana's rough-diamond export unit values using a hybrid CNN–LSTM model evaluated against naïve, ARIMA and GARCH benchmarks. The framework is designed to use monthly information covering 2004–2025, subject to data availability and release-aware alignment. Quarterly rough-diamond export value and exported carats will be temporally disaggregated separately before the monthly export unit value is calculated, thereby preserving the distinction between additive export components and their non-additive ratio. Forecasting performance will be assessed through walk-forward validation using RMSE, MAE, MAPE and pairwise Diebold–Mariano tests. The GARCH benchmark will use an ARMA conditional-mean equation to generate comparable unit-value forecasts while modelling conditional volatility separately.

SHAP is pre-specified as a post-hoc interpretation stage and will be applied only after the CNN–LSTM has been fitted, forecasts have been generated and predictive performance has been evaluated. Its values will describe model-dependent predictive contributions without implying causation. Because this is a proposed-framework paper, no completed forecasting or SHAP results are reported, and no claim is made that the CNN–LSTM will outperform the benchmarks. The contribution of the paper is therefore a coherent protocol integrating composition-aware target construction, temporal harmonisation, conventional benchmarking, hybrid deep learning, temporally valid evaluation and explainability to support subsequent empirical investigation in Botswana's diamond-dependent economy.

REFERENCES

1. Ameer, H. B., Boubaker, S., Ftiti, Z., Louhichi, W., & Tissaoui, K. (2023). Forecasting commodity prices: Empirical evidence using deep learning tools. *Annals of Operations Research*, 339(1), 349–367. <https://doi.org/10.1007/s10479-022-05076-6>
2. Aas, K., Jullum, M. and Løland, A. (2021) “Explaining individual predictions when features are dependent: More accurate approximations to Shapley values”, *Artificial Intelligence*, 298, 103502. <https://doi.org/10.1016/j.artint.2021.103502>
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82–115.
4. Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, Y., Maddix, D., Turkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., Callot, L., & Januschowski, T. (2022). Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6), Article 121. <https://doi.org/10.1145/3533382>
5. Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83
6. Bollerslev, T. (1986) “Generalized autoregressive conditional heteroskedasticity”, *Journal of Econometrics*, 31(3), pp. 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
7. Borsky, S. and Leiter, A.M. (2022) ‘International trade in rough diamonds and the Kimberley Process Certification Scheme’, *World Development*, 152, 105786. <https://doi.org/10.1016/j.worlddev.2021.105786>
8. Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015) *Time Series Analysis: Forecasting and Control*. 5th edn. Hoboken, NJ: Wiley.

9. Çelik, T. B., İcan, Ö., & Bulut, E. (2023). Extending machine learning prediction capabilities by explainable AI in financial time series prediction. *Applied Soft Computing*, 132, 109876. <https://doi.org/10.1016/j.asoc.2022.109876>
10. Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11), 1997–2028. <https://doi.org/10.1007/s10994-020-05910-7>
11. Denton, F. T. (1971). Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization. *Journal of the American Statistical Association*, 66(333), 99–102. <https://doi.org/10.1080/01621459.1971.10482227>
12. Diebold, F.X. and Mariano, R.S. (1995) “Comparing predictive accuracy”, *Journal of Business & Economic Statistics*, 13(3), pp. 253–263. <https://doi.org/10.1080/07350015.1995.10524599>
13. Engle, R.F. (1982) ‘Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation’, *Econometrica*, 50(4), pp. 987–1007. <https://doi.org/10.2307/1912773>
14. Erion, G. et al. (2021) “Improving performance of deep learning models with axiomatic attribution priors and expected gradients”, *Nature Machine Intelligence*, 3, pp. 620–631. <https://doi.org/10.1038/s42256-021-00343-w>
15. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
16. Harvey, D., Leybourne, S. and Newbold, P. (1997) “Testing the equality of prediction mean squared errors”, *International Journal of Forecasting*, 13(2), pp. 281–291. [https://doi.org/10.1016/S0169-2070\(96\)00719-4](https://doi.org/10.1016/S0169-2070(96)00719-4)
17. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
18. Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice* (Vol. 2). Melbourne, Australia: OTexts.
19. International Monetary Fund (2017) *Quarterly National Accounts Manual: 2017 Edition*. Washington, DC: International Monetary Fund.
20. International Monetary Fund. (2024). Botswana: 2024 Article IV Consultation—Press Release; Staff Report; and Statement by the Executive Director for Botswana (Country Report No. 24/286). <https://doi.org/10.5089/9798400288135.002>
21. Kigo, S.N., Omondi, E.O. and Omolo, B.O. (2023) ‘Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model’, *Scientific Reports*, 13, 17315. <https://doi.org/10.1038/s41598-023-44326-w>
22. Kimberley Process (n.d.) *Rough Diamond Statistics: Public Statistics Area*. Available at: https://kimberleyprocessstatistics.org/public_statistics
23. Lundberg, S.M. and Lee, S.-I. (2017) “A unified approach to interpreting model predictions”, *Advances in Neural Information Processing Systems*, 30, pp. 4765–4774.
24. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74. <https://doi.org/10.1016/j.ijforecast.2019.04.014>
25. Petropoulos, F. et al. (2022). *Forecasting: Theory and practice*. *International Journal of Forecasting*, 38(3), 705–871. <https://doi.org/10.1016/j.ijforecast.2021.11.001>
26. Sezer, O.B., Gudelek, M.U. and Ozbayoglu, A.M. (2020) “Financial time series forecasting with deep learning: A systematic literature review: 2005–2019”, *Applied Soft Computing*, 90, 106181. <https://doi.org/10.1016/j.asoc.2020.106181>
27. Shah, J., Vaidya, D., & Shah, M. (2022). A comprehensive review on multiple hybrid deep learning approaches for stock prediction. *Intelligent Systems with Applications*, 16, 200111.
28. Silver, M. (2009) ‘Do unit value export, import, and terms-of-trade indices misrepresent price indices?’, *IMF Staff Papers*, 56(2), pp. 297–322. <https://doi.org/10.1057/imfsp.2008.24>