Data Science and Engineering (DSE) Record, Volume 6, issue 1.

Forecasting Method Evaluation Framework for Production Planning under Uncertain Customer Demand

Anongphorn Janboonpeng [1] and Chompoonoot Kasemset [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Industrial Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
anongphorn_J@cmu.ac.th,chompoonoot.kasemset@cmu.ac.th

Abstract. Customer demand volatility in rapidly changing market environments poses significant challenges to production planning, particularly for short life-cycle products that exhibit pronounced seasonal patterns. Exploratory Data Analysis (EDA) conducted in this study revealed clear annual cycles and distinct seasonal peak periods in customer demand, highlighting the need for forecasting methods capable of effectively and reliably capturing such seasonality. Therefore, this research aims to develop a Forecasting Method Evaluation Framework to guide the selection of appropriate forecasting models for production planning under uncertain demand conditions. The proposed framework consists of five key components: (1) data preparation and seasonal feature engineering informed by EDA findings, (2) development of statistical and deep learning forecasting models, including SARIMA, Holt-Winters, LSTM, GRU, and a hybrid SARIMA–LSTM model, (3) performance evaluation using MAE, RMSE, MAPE, and R^2, and (4) integration of forecast outputs into production planning processes to effectively accommodate demand variability. The framework supports improved production stability by reducing the frequency of production plan adjustments and enhances operational readiness across manpower, machinery, materials, and production methods (Man–Machine–Material–Method), ensuring alignment with future demand conditions.

Keywords: Exploratory Data Analysis (EDA), SARIMA, Holt-winters, GRU, LSTM.

1 Introduction

In today's fast-paced and volatile market environment, customer demand has become increasingly unpredictable, often fluctuating significantly within short periods. This instability presents a major challenge for production planning, as manufacturers are frequently required to revise their production schedules in response to updated demand information. For instance, an initial forecast of 200 units may be revised to 300 units, reduced to 250 units, and ultimately updated to 350 units in the final version. Such repeated adjustments disrupt scheduling, complicate inventory management, and reduce overall operational efficiency.

Short life-cycle products particularly those exhibiting strong seasonal patterns further amplify these challenges due to the limited amount of historical data available and the pronounced variability in customer demand. Traditional statistical forecasting models such as Seasonal Autoregressive Integrated Moving Average (SARIMA) and Holt-Winters have been widely used to capture linear trends and seasonal structures. However, their performance often diminishes when confronted with nonlinear fluctuations or irregular patterns that frequently occur in real-world environments. In contrast, deep learning models, especially Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), are capable of modeling complex temporal relationships but may require careful feature engineering and data preparation to perform effectively with small datasets.

Despite the availability of various forecasting techniques, manufacturers still lack a systematic approach for evaluating and selecting the most suitable forecasting method under conditions of uncertain and seasonally driven demand. Existing studies tend to focus on comparing individual models or improving specific algorithms, while relatively few offer a structured framework that integrates data analysis, model development, performance evaluation, and production planning considerations into a unified process.

To address this gap, this study proposes a Forecasting Method Evaluation Framework designed to guide the selection and assessment of forecasting methods for short life-cycle products with seasonal demand characteristics. Rather than emphasizing model performance outcomes, the framework emphasizes the process from data preparation and seasonal feature engineering to statistical and deep learning model development, hybrid model integration, evaluation criteria, and practical application to production planning.

From a theoretical perspective, this research contributes a structured methodology for forecasting under data-scarce and seasonally fluctuating environments an area that remains underexplored. From a practical perspective, the framework aims to enhance forecast stability, reduce the frequency of production plan revisions, and improve overall supply chain responsiveness in uncertain demand environments.

2 Related Works

2.1 Traditional Statistical Forecasting

Forecasting: Principles and Practice by Rob J. Hyndman and George Athanasopoulos [1] offers a widely adopted framework for time series forecasting that focusses both data exploration and model construction. Their work highlights the use of key exploratory tools such as time series plots, STL decomposition, ACF/PACF analysis, and the Augmented Dickey-Fuller (ADF) test to understand underlying patterns and assess stationarity. For model development, the authors recommend classical approaches like SARIMA and Holt-Winters, which are well-suited for handling trend and seasonality. SARIMA offers flexible seasonal modeling, while Holt-Winters focuses on smoothing seasonal and trend components. These methods remain fundamental due to their interpretability and effectiveness across various real-world forecasting applications.

2.2 Deep Learning Model Forecasting

Gated Recurrent Unit (GRU) is a variant of recurrent neural networks (RNNs) developed to address the limitations of traditional RNNs in learning long-term dependencies, particularly the vanishing gradient problem. GRU simplifies the LSTM architecture by using only two gates reset and update allowing for faster training and fewer parameters while still retaining effective memory over time (Bousnguar et al., 2023) [2]. Recent studies have shown that GRU performs competitively with, and sometimes outperforms, LSTM in time series forecasting tasks, especially when the dataset is small or the sequence length is short. In the context of higher education, GRU was successfully applied to predict student enrollment data, achieving lower RMSE and MAE than LSTM. These findings highlight GRU's capability to model non-linear and non-stationary time series data efficiently and accurately.

Taslim and Murwantara [3] conducted a comparative study between ARIMA and LSTM models using time series data with varying lengths (60, 120, and 228 data points) and different levels of missing values. The results showed that LSTM outperformed ARIMA in terms of accuracy when applied to short datasets (60 data points) and also demonstrated better stability in the presence of missing data. The findings highlight LSTM's effectiveness in modeling volatile or incomplete time series, especially when data quantity is limited, making it a suitable choice over traditional statistical methods in such contexts.

2.3 Hybrid Model Forecasting

The Hybrid SARIMA-LSTM model integrates SARIMA for capturing linear and seasonal patterns, and LSTM for modeling non-linear dependencies. By using SARIMA to forecast the trend and seasonality and feeding its residuals into LSTM, the model effectively addresses both components of complex time series. According to Panicker and Valarmathi (2024) [4], this hybrid approach consistently outperforms individual SARIMA and LSTM models in accuracy, making it suitable for non-stationary and seasonal datasets.

2.4 Feature Engineering for Forecasting

Min-Max Normalization is one of the most widely used data preprocessing techniques for transforming numerical features into a fixed range, typically [0, 1]. This approach ensures that variables with different units and magnitudes become directly comparable, especially in the context of time series aggregation or machine learning algorithms sensitive to scale, such as neural networks. According to Mazziotta and Pareto (2021) [5], the Min-Max method has been traditionally employed in the construction of composite indices, such as the Human Development Index (HDI), where indicators are rescaled to a common range. This method is advantageous due to its simplicity and its ability to preserve the relative distances between values. However, it lacks a central reference point such as the mean making it difficult to interpret the normalized values in terms of central tendency. For instance, a normalized value of 0.3 does not directly indicate whether the original value was below or above average.

Jaén-Vargas et al. (2022) [6] investigated how sliding window size influences the performance of deep learning models for human activity recognition and found that selecting an appropriate window length significantly affects accuracy, inference time, and computational efficiency. Their evaluation of multiple architectures including DNN, CNN, LSTM, and CNN-LSTM showed that a fixed window size of 20–25 frames provided the best balance between accuracy and processing cost. This finding supports the use of a fixed lookback window in time-series forecasting, where a defined sequence of past observations (e.g., 12 months) enables LSTM and GRU models to effectively capture temporal dependencies for improved demand prediction.

Cyclical features such as hour-of-day or month-of-year can lead to misleading representations when encoded as integers, since values like 0 and 23 (midnight and 11 PM) appear distant despite being adjacent in time. Mahajan et al. (2021) [7] recommend using sine and cosine transformations to encode cyclical features, effectively mapping them onto a unit circle. This ensures that temporally adjacent values are also geometrically close in feature space.

3 Data and Methodology

The overall research framework is illustrated in Fig.1., which consists of four sequential components: data preparation, exploratory data analysis and parameter identification, model development (statistical, deep learning, and hybrid approaches), and model evaluation using forecasting accuracy metrics.

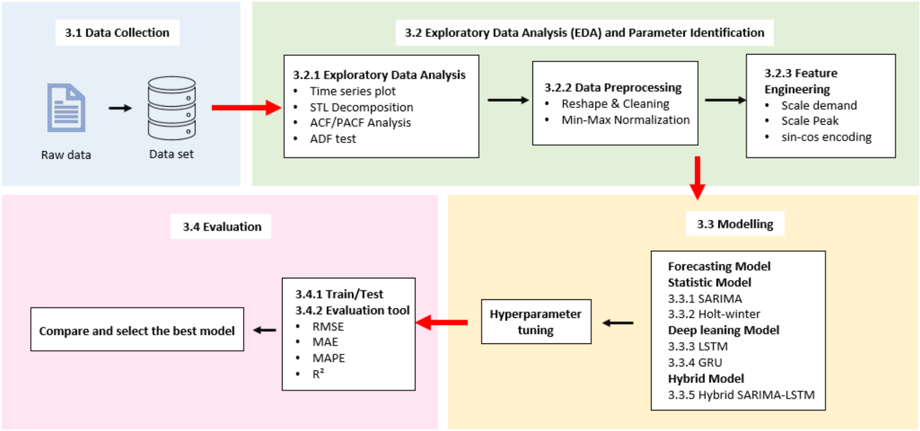

Fig.1. The research framework overall design

3.1 Data Collection

In demand planning environments where customer forecasts are updated frequently, it is common for each month to contain multiple revisions. This reflects the dynamic nature of short life-cycle products and introduces variability into the raw dataset.

To maintain consistency within the forecasting framework, only the final confirmed revision for each month is retained as the historical demand input.

3.2 EDA and Parameter Identification

3.2.1 Exploratory Data Analysis

Before applying forecasting models, exploratory data analysis (EDA) is conducted to understand the dataset's structure, reveal patterns, and detect anomalies. The process begins with a time series plot to observe overall trends and seasonality, followed by STL decomposition to separate trend, seasonal, and residual components. Autocorrelation (ACF) and partial autocorrelation (PACF) plots are then used to examine temporal dependencies, and the Augmented Dickey–Fuller (ADF) test assesses stationarity. These steps provide essential insights for selecting suitable forecasting models.

3.2.2 Data Preprocessing

1) Reshape & Cleaning

This step ensures the dataset is consistent, well-structured, and ready for time series modeling. Because the original data contained multiple revisions in wide format, it was reshaped into long format so each row represents a single product month observation. The records were then cleaned by sorting chronologically, extracting the month, and selecting only the latest revision per month to retain the final confirmed demand. The dataset was ultimately simplified into key fields such as Month, Demand, and Product, providing a clean, up-to-date structure suitable for forecasting.

2) Min-Max Normalization

Min–Max Normalization [5] was applied to scale numerical features into a fixed range of [0, 1], ensuring consistency across inputs and improving learning efficiency in deep learning models. The normalization was performed using the MinMaxScaler from the sklearn.preprocessing library.

$$\mathrm{x}' = \frac{\mathrm{x} - \min(x)}{\max(x) - \min(x)} \quad (1)$$

3.2.3 Feature Engineering

1) Scaled Demand

Scaled demand representing the normalized monthly demand and serving as the primary predictor.

2) Peak Season Indicator (Scale peak)

Identifying periods of sustained seasonal demand increases through a two-stage assessment:

- pattern detection using exploratory data analysis to observe recurring surges across annual cycles, and
- domain validation based on organizational knowledge from historical production and sales activities.

Within the proposed framework, these complementary sources of evidence are used to formally specify peak-season intervals, which are subsequently encoded into a binary variable.

This variable is incorporated into deep learning and hybrid model components of the framework, serving as an explicit seasonal signal that supports model stability and enhances the interpretability of forecasting behavior under recurring high-demand conditions.

3) Cyclical Encoding (Sin-Cos encoding)

To capture the cyclical nature of monthly seasonality, the month index was transformed using sine and cosine functions. This approach preserves temporal continuity, particularly the natural transition between December and January, which is not possible with conventional ordinal encoding. As suggested by Mahajan et al. (2021) [7], projecting periodic variables onto a unit circle enables the model to learn seasonal patterns more effectively.

In this study, the month of each observation (extracted from the Date column) was encoded using the following equations (2).

$$Scaled\_\sin\left(\frac{2\pi \cdot month}{12}\right), Scaled\_\cos\left(\frac{2\pi \cdot month}{12}\right) \qquad (2)$$

3.3 Modeling

In this study, the focus is on evaluating and comparing multiple forecasting models, including both statistical and deep learning approaches. The models considered are: SARIMA, Holt-Winters, LSTM, GRU, and a Hybrid SARIMA-LSTM model. Each of these models is employed to analyze the historical demand data, which has been processed through feature engineering and normalization.

3.3.1 SARIMA

The seasonal structure of the time series, a Seasonal ARIMA (SARIMA) model was employed. SARIMA extends the classical ARIMA model by incorporating both non-seasonal (p, d, q) and seasonal (P, D, Q, s) components as the table 1, making it suitable for data with recurring seasonal patterns.

The general form of the SARIMA model is expressed as the below.

$$SARIMA = (p, d, q)\ (P, D, Q)s \tag{3}$$

Table 1. Model Final Parameter Selection.

Parameter	Selection Basic	Explanation
p	PACF	Based on PACF spikes at early lags.
d	ADF test & TS Plot	Non-stationarity indicated by ADF.
q	ACF	From ACF spikes at early lags.
P	Seasonal PACF	Showing weak or absent seasonal AR effects.
D	STL decomposition	Clear seasonal patterns observed in STL.
Q	Seasonal ACF	Seasonal ACF spikes at seasonal lags.
s	Data frequency	Data frequency defining seasonal cycle length.

3.3.2 Holt-Winters

The time series data with trend and seasonality, the Holt-Winters exponential smoothing method was applied. This method extends simple exponential smoothing by incorporating both a trend component and a seasonal component, making it suitable for series exhibiting regular seasonal variation and trend over time.

There are two versions of the Holt-Winters method.

$$\begin{aligned} \hat{y}_{t+h|t} &= l_t + hb_t + s_{t+h-m(k+1)} \\ l_t &= \alpha(y_t - s_{t-m}) + (1\text{-}\alpha)(l_{t-1} + b_{t-1}) \\ b_t &= \beta^*(l_t\text{-}l_{t-1}) + (1\text{-}\beta^*)b_{t-1} \\ s_t &= \gamma(y_t\text{-}l_{t-1}\text{-}b_{t-1}) + (1\text{-}\gamma)s_{t-m} \end{aligned} \tag{4}$$

Level: $l_t = \alpha(y_t - s_{t-m}) + (1\text{-}\alpha)(l_{t-1} + b_{t-1})$
Trend: $b_t = \beta^*(l_t\text{-}l_{t-1}) + (1\text{-}\beta^*)b_{t-1}$
Seasonal: $s_t = \gamma(y_t\text{-}l_{t-1}\text{-}b_{t-1}) + (1\text{-}\gamma)s_{t-m}$
Forecast: $\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)}$

α = smoothing level
β = smoothing trend
γ = smoothing seasonal
m = The length of the seasonal cycle (e.g., 12 for monthly data).

Within the framework, Holt-Winters parameters (level, trend, and seasonality smoothing factors) are determined through iterative tuning rather than predetermined fixed values. The selection process emphasizes achieving stable convergence and capturing the observed seasonal structure.

3.3.3 Long Short-Term Memory (LSTM)

This study employed LSTM neural network to forecast monthly product demand. LSTM, introduced by Hochreiter and Schmidhuber (1997) [8], addresses vanishing gradient issues through memory cells and gating mechanisms, enabling effective modeling of sequential dependencies as shown in Fig.2.

LSTM works by using three main gates to control the flow of information:

- Forget Gate: Decides what past information to forget.
- Input Gate: Selects what new information to store.
- Output Gate: Determines what information to output at the current time step.

The forget gate was introduced later by Gers et al. (1999), but is part of all modern LSTM implementations.

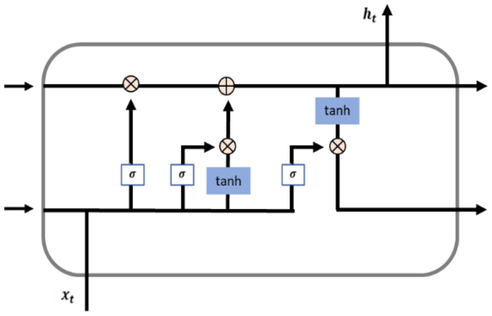

Fig.2. Architecture of a Long Short-Term Memory (LSTM) Cell

LSTM networks were utilized to model nonlinear trends and seasonal demand patterns, leveraging their strength in capturing long-term temporal dependencies.

3.3.4 Gated Recurrent Unit (GRU)

In this study, GRU model, a type of RNN, was employed to forecast student enrollment data. GRU is particularly suitable for time-series tasks due to its ability to capture temporal dependencies without the vanishing gradient problem faced by traditional RNNs. Compared to LSTM, GRU is computationally more efficient, using fewer parameters and a simpler architecture as shown in Fig.3.

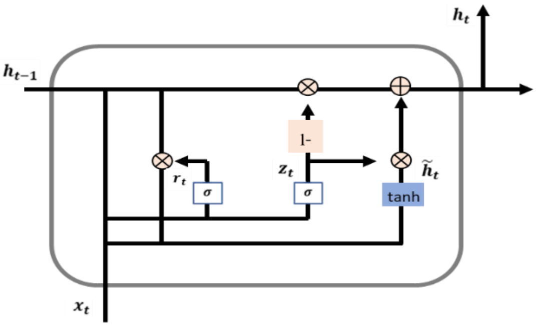

Fig.3. Architecture of a Gated Recurrent Unit (GRU) Cell.

3.3.5 Hybrid SARIMA-LSTM

This study adopts a hybrid SARIMA–LSTM model (Fig.4.) to enhance time series forecasting by leveraging both linear and nonlinear modeling capabilities. Initially, a SARIMA model captures the linear trend and seasonal patterns, serving as the baseline forecast. The residuals defined as the differences between actual values and SARIMA predictions were then modeled using LSTM network to capture nonlinear components.

The final forecast is computed as:

$$\hat{A}_t = \widehat{L_t} + \widehat{N_t} \tag{5}$$

Where $\widehat{L_t}$ is the SARIMA forecast and $\widehat{N_t}$ is the LSTM prediction of residuals. Modeling Procedure.

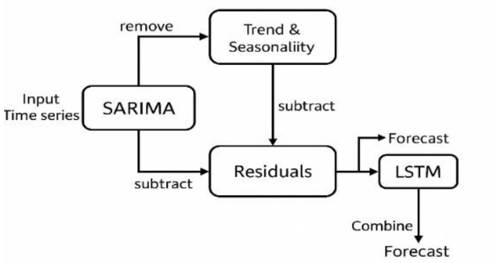

Fig.4. Architecture of a Hybrid SARIMA–LSTM Forecasting Model.

This residual-based hybrid structure enables the model to address both trend-seasonal and complex temporal dynamics more effectively than either model alone, following the methodology proposed by Panicker & Valarmathi (2024) [4].

3.4 Performance Evaluation

3.4.1 Train/Test Dataset

In this framework, a temporal train–test split was adopted to preserve the sequential structure of the time-series data. Given that the product exhibits a short life cycle with limited historical observations, the training portion was maximized to enable the models to capture the broadest possible seasonal patterns before proceeding to evaluation. For the deep learning models, a sliding-window approach [5] was applied to transform the continuous series into supervised sequences while maintaining chronological order.

3.4.2 Measurement

In time series forecasting, evaluating model performance is essential for selecting an appropriate method. Common metrics such as MAE, RMSE, MAPE, and R² assess different aspects of prediction accuracy. Mehdiyev et al. (2016) [9] emphasized that no single accuracy measure is universally reliable, as different metrics may produce different evaluations of model performance. Therefore, multiple accuracy measures should be used to ensure a more robust and comprehensive assessment of forecasting methods. Therefore, this study adopts the following evaluation.

1) Mean Absolute Error (MAE)
The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It is the mean of the absolute differences between the actual and forecasted values:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t \,| \qquad (6)$$

2) Root Mean Squared Error (RMSE)
The Root Mean Squared Error (RMSE) is defined as the square root of the average of the squared forecast errors:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \qquad (7)$$

3) Mean Absolute Percentage Error (MAPE)
The Mean Absolute Percentage Error (MAPE) expresses forecast accuracy as a percentage by comparing the absolute forecast error relative to the actual observed values:

$$\text{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \qquad (8)$$

4) Coefficient of Determination (R²)
The coefficient of determination, denoted as R², is a statistical measure that indicates how well a forecasting model explains the variability of the actual data. It is defined as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad (9)$$

4 Conclusion

This research presents a Forecasting Evaluation Framework designed for short life-cycle, small-sample, and seasonally driven demand data. The framework provides a systematic basis for selecting appropriate forecasting methods by considering data characteristics, structural suitability, and model robustness, rather than relying solely on numerical accuracy.

Within the proposed forecasting framework, each model demonstrates distinct strengths and limitations that determine its suitability for different short life-cycle and small-data conditions. SARIMA performs well when seasonal patterns are stable and easily identifiable, owing to its transparent structure, but its capability is limited when nonlinear behavior is present. Holt-Winters offers a simple and fast approach suitable for very small datasets with clear seasonality, although it struggles to adapt to abrupt demand shifts. LSTM is effective in capturing complex nonlinear patterns but tends to overfit when historical data are limited, making it more appropriate for larger datasets or those enriched with multiple explanatory features. GRU, by contrast, is lightweight and more reliable than LSTM under small-sample constraints but may underfit highly complex demand structures. The Hybrid SARIMA–LSTM model provides the most balanced performance, as it integrates both linear and nonlinear components and remains robust when only limited data are available, making it particularly suitable for short life-cycle products exhibiting nonlinear seasonal fluctuations.

To enhance forecasting reliability under short life-cycle and small-sample constraints, the framework recommends incorporating explicit seasonal features, such as a Peak Season Indicator, to better capture cyclical demand structures. The adoption of residual-based hybrid modeling is encouraged to reduce overfitting and improve the handling of nonlinear variations, particularly when data availability is limited. Furthermore, conducting real-field evaluation is essential for assessing alignment with actual demand, verifying reductions in production plan revisions, and ensuring the model’s stability in operational environments characterized by uncertainty.

The proposed framework offers a structured pathway for selecting forecasting methods tailored to small-sample, seasonal, and short life-cycle demand conditions. The findings highlight Hybrid SARIMA–LSTM as the most balanced and reliable approach, while other models remain applicable depending on data characteristics and operational requirements.

References

1. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts. Retrieved from https://otexts.com/fpp2/index.html
2. Bousnguar, H., Battou, A., & Najdi, L. (2023). Gated recurrent units (GRU) for time series forecasting in higher education. International Journal of Engineering Research & Technology, 12(3), 152–155.
3. Taslim, D. G., & Murwantara, I. M. (2024). Comparative analysis of ARIMA and LSTM for predicting fluctuating time series data. Bulletin of Electrical Engineering and Informatics, 13(3), 1943–1951.

4. Panicker, N. K. K., & Valarmathi, J. (2024). A hybrid SARIMA-LSTM approach for improved time series prediction of aerosol optical depth across Delhi, India. Journal of Theoretical and Applied Information Technology, 102(11), 4836–4851.
5. Mazziotta, M., & Pareto, A. (2021). Data normalization for aggregating time series: The constrained Min-Max method. Rivista Italiana di Economia Demografia e Statistica, 75(4), 101–108.
6. Jaén-Vargas, M., Reyes Leiva, K. M., Fernandes, F., Gonçalves, S. B., Silva, M. T., Lopes, D. S., & Serrano Olmedo, J. J. (2022). Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models. PeerJ Computer Science, 8: e1052.
7. Mahajan, K., Singh, M., & Bruns, R. (2021). An experimental assessment of treatments for cyclical data. Proceedings of the CSCSU 2021 Conference, 1–6.
8. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
9. Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating Forecasting Methods by Considering Different Accuracy Measures. Procedia Computer Science, 95, 264–271.