DOI: https://doi.org/10.59294/HIUJS2025031

Effective forecasting method for determining production demand: A case study

Nguyen Thi Thu Hao[1,2], Ha Trong Khoi [1,2], Le Thi Diem Chau[1,2,*]
[1]*Ho Chi Minh City University of Technology*
[2]*Vietnam National University Ho Chi Minh City*

ABSTRACT

Forecasting has long been essential for improving enterprise performance, especially in inventory management, production planning, and overall economic efficiency. This study aims to predict production demand for a local stationery manufacturer with the aim of comparing and finding effective methods between three classical forecasting methods - Moving Average, Exponential Smoothing, and ARIMA - with a modern deep learning approach, Long Short-Term Memory (LSTM), to optimize production or inventory planning and management, and support enterprises in reducing waste and saving many expenses. Despite the ease of implement, the classical forecasting methods have difficulties in capturing and remembering economic trends, while machine learning gives more effective results in this circumstance. The data set includes quantities of demand and production collected and pre-processed by normalizing and stationarity testing. Classical methods are applied to capture the trends and variations in the data. At the same time, the LSTM is built to learn models with more complex patterns or some essential elements that traditional methods might overlook. The efficiency of the model is evaluated through forecasting errors such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). According to experimental results, classical methods provide a sustainable baseline. Still, the LSTM model shows high accuracy, especially in catching up with subtle fluctuations and unusual patterns in the dataset. It is ensured that using advanced deep learning techniques, such as LSTM, could improve forecasting error and reliability and significantly benefit manufacturing planning and inventory control. This study contributes to this field by providing a comprehensive comparative analysis of forecasting methods and practical insights for industrial companies to upgrade their capabilities in forecasting production demand effectively.

Keywords: *forecasting, moving average, exponential smoothing, ARIMA, LSTM, time series*

1. INTRODUCTION

In a competitive manufacturing environment, precisely forecasting demand is essential for optimizing production planning, inventory management, and overall operational efficiency to help the company gain many economic benefits. For companies in dynamic markets, such as pen manufacturers, customer demand fluctuates over time. At certain periods, production output falls short of actual demand, leading to shortages and supply disruptions, while at other times, output exceeds demand, resulting in excess inventory and increased storage costs. As a consequence, these fluctuations make cost control challenging and inefficient. Therefore, accurately forecasting demand is essential for businesses to reduce costs, minimize waste, and enhance customer satis-faction. Despite various forecasting techniques, choosing the most suitable model to balance accuracy, interpretability, and compu-tational efficiency is still tricky. Besides that, comparing different forecasting methods provides a more comprehensive and objective perspective for the study and helps enterprises to identify the most suitable approach for practical application.

Traditional forecasting methods - Moving Average, Exponential Smoothing, ARIMA, for instance - have been popular for predicting demand in an industrial background because of their simplicity and easy implementation. These methods effectively capture trends and patterns in a stable production environment. However, they also have limitations when applied to more complex models and fluctuations.

Recently, deep learning has advanced, particularly

Corresponding author: Le Thi Diem Chau
Email: lechau@hcmut.edu.vn

with the innovation of the Long-Short Term Memory (LSTM) network; therefore, new avenues have been opened for modeling time series data. Capturing long-term or short-term dependencies and nonlinear relationships can help LSTM become a promising technique for demand forecasting in industries with rapid market fluctuations. Nevertheless, deep learning models also have drawbacks; they can be data-intensive and sometimes operate as "black-boxes", which can be challenging for researchers when practicing in environments requiring interpretability.

This study comprehensively compares three forecasting methods - Moving Average, Exponential Smoothing, and ARIMA - to the LSTM approach for predicting production demand in a local pen manufacturing company. By taking advantage of historical data, this study seeks to answer key questions: How do classical methods capture demand patterns in a volatile market? Can LSTM provide better forecasting accuracy despite its complexity? And what are the critical factors influencing the performance of these models?

The study is confined to the production data of a local pen manufacturer, offering a focused investigation of the challenges and nuances inherent in this industry. Through implementing and evaluating models using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), the research efforts provide possible insights and recommendations for both academic researchers and industry practitioners.

By bridging the gap between traditional forecasting and modern deep learning approaches, this study contributes to academic literature. It gives practical implications for manufacturers seeking to enhance their production planning processes in a rapidly evolving market.

Section 1 provides an overview of the current situation, the problem, the objective, and the methodology used in the study. The Literature Review briefly summarizes previous research and compares those studies with the current one. Section 3, Methodology, presents the methods and types of errors used in the analysis. Section 4 presents the results of the forecasting models, along with a comparison between them to provide a more comprehensive view. The final section is the Conclusion, which summarizes the study's main findings and suggests directions for future research.

2. LITERATURE REVIEW

Classical statistical methods have been used widely in early explorations about predicting production demand due to their simplicity and ease of interpretation, such as Moving Average and Exponential Smoothing. For instance, Fattah et al. (2018) compares the performance of Moving Average and Exponential Smoothing through predicting quantities of sales in the future. The result demonstrates that while these methods are computationally efficient and easy to implement, they tend to struggle with highly uncertain data and seasonality effects [1]. Basri et al. (2019) implemented Exponential Smoothing to predict production demands, emphasizing that parameter tuning is important to achieve higher accuracy [2]. Additionally, Kahraman and Akay (2023) conducted a study regarding to compare Exponential Smoothing methods in forecasting global prices of main metals [3]. Exponential smoothing seemed to be effective for short-term forecasting, but there are some limitations in capturing sudden market fluctuations and external economic factors. The primary advantage of these methods is their simplicity, but they lack adaptability to complex nonlinear relationships.

The Autoregressive Integrated Moving Average (ARIMA) model is a popular forecasting method for time series data. It is known for its ability to capture linear patterns in data. Zhang (2003) conducted a comprehensive study comparing ARIMA with artificial neural networks (ANNs). This study concludes that ARIMA performs well for short-term forecasting with stable data than capture nonlinear patterns effectively [4]. Another study using ARIMA to predict the number of COVID-19 cases by Munarsih et al. (2020), demonstrating that this method can provide reliable short-term forecasts, but its performance falls with complex datasets [5]. Chodakowska et al. (2023) evaluate the effectiveness of ARIMA models for forecasting solar radiation in different climatic conditions, demonstrating their practical application and highlighting the need for location-specific model development [6]. However, the limitation of ARIMA is its assumption of stationarity, requiring extensive preprocessing to transform non-stationary data into a form suitable for forecasting.

With the rise of deep learning, LSTM networks have emerged as a powerful forecasting technique capable of handling complex time-series data. The

study by Fattah et al. (2018) applied LSTM for demand forecasting and found that it significantly outperformed traditional statistical models in capturing long-term dependencies [1]. Similarly, Munarsih et al. (2020) and Basri et al. (2019) demonstrated the superiority of LSTM over ARIMA in forecasting COVID-19 cases and production demand, respectively [2, 5]. The primary advantage of LSTM lies in its ability to recognize intricate patterns in sequential data, making it highly effective for non-linear time series. However, these models require extensive computational resources and large datasets for training, making them less accessible for small-scale applications.

A growing body of research suggests that hybrid models combining statistical and deep learning approaches yield superior forecasting accuracy. Zhang (2003) explored the integration of ARIMA and artificial neural networks, highlighting that such models can leverage the strengths of both techniques [4]. Basri et al. (2019) and Munarsih et al. (2020) also experimented with hybrid models, demonstrating improved performance over standalone ARIMA or LSTM methods [2]. The main challenge associated with hybrid models is their complexity, requiring careful design and optimization to ensure balanced contributions from both statistical and machine learning components.

In this study, the production demand of a pen manufacturing company is forecasted by using five previously mentioned forecasting methods and evaluated their performance based on the value of MAD, MSE, MAPE, and RMSE. This approach allows for an objective assessment of the results and the selection of the most suitable method for the dataset, rather than limiting the comparison to only two or three methods. Additionally, this study applies a novel approach to Artificial Neural Networks for medium-term datasets by reducing the number of neural layers during data processing.

3. METHODOLOGY

3.1. Problem

From the data for this study, that were collected from a local company specializing in office supplies manufacturing, which is experiencing difficulties in meeting customer demand. As shown in Figure 1, the production volume fails to meet more than 90% of the required supply for customers from 2023 to 2024. The blue line and orange line demonstrate quantities that were forecasted and actual customer demand, respectively, from January 2023 to June 2024.

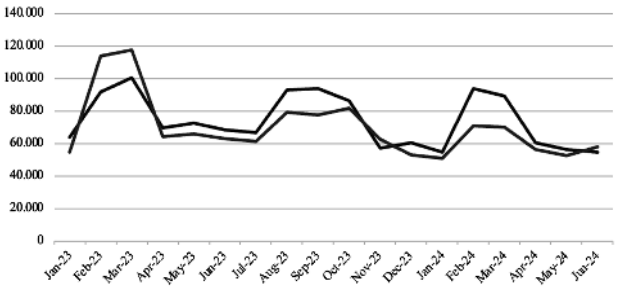

Figure 1. Quantities of forecasting and demand from January 2023 to June 2024

According to available data and conducting to find out the key reasons for this problem, Fishbone Diagram points out that the company has not had any systematic and effective method yet. Staff members primarily rely on personal experience to make predictions, while there are a lot of significant changes in the economy.

3.2. Forecasting technique

3.2.1. Moving average

A fixed number of data points is selected, and the average is calculated based on the most recent observations [7]. As new data arrives, the oldest value is replaced, and the updated average is used for forecasting. This method is known as Moving Average (k-order MA(k)), is calculated as follows Equation 1:

$$\hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-k+1}}{k} \quad (1)$$

Where:

$\hat{Y}_{t+1}$: Forecasted value for the next period.

Y_t: Actual observed value of the current period.

k: Number of periods.

3.2.2. Exponential smoothing

The Simple Exponential Smoothing method provides a weighted moving average of all previously observed values, with exponentially decreasing weights [7]. It continuously updates estimates based on recent observations, smoothing past values in a way that declines exponentially. The Exponential Smoothing function follows Equation 2:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t \quad (2)$$

Where:

$\hat{Y}_{t+1}$: The new smoothed value or the forecast value for the next period.

Y_t: The new observation or the actual value of the series in period t.
$\hat{Y}_t$: The old smoothed value or the forecast for period t.
α: The smoothing constant $(0 < \alpha < 1)$.

3.2.3. Autoregressive integrated moving average (ARIMA)

This method combines three components: the AutoRegressive (AR) model, the Integrated (I) property of the data series, and the Moving Average (MA) model. AutoRegressive (AR) Model: This component represents the autoregressive process, incorporating lagged values of the current variable [7]. Specifically, the AR(p) model (pth-order autoregressive model) is expressed as follows Equation 3:

$$Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p} + \varepsilon_t \quad (3)$$

Where:
Y_t: The response variable at time t.
$Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$: The response variable at time lags $t-1, t-2, \ldots, t-p$ respectively.
$\Phi_0, \Phi_1, \Phi_2, \ldots, \Phi_p$: The coefficients to be estimated.
ε_t: The error term at time t.

Moving Average (MA) model: The moving average process of order qqq, denoted as MA(q), is defined as follows Equation 4:

$$Y_t = \mu + \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \cdots - \omega_q \varepsilon_{t-q} \quad (4)$$

Where:
Y_t: The response variable at time t.
ε_t: The error term at time t.
μ: White noise error term at time t.
$\omega_1, \omega_2, \ldots, \omega_q$: Coefficients of the moving average terms.

Combine MA and AR, the ARMA (p,q) equation is Equation 5:

$$Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p} + \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \ldots - \omega_q \varepsilon_{t-q} \quad (5)$$

3.2.4. Long-short term memory (LSTM)

A recurrent neural network (RNN), first defined by Rumelhart, Hinton, and Williams (1986) [8], is a deep learning model that is trained to process and transform sequential data inputs into specific sequential data outputs. Sequential data is data, such as words, sentences, or time series data, in which the sequential components are correlated based on complex semantics and syntactic rules. An RNN is a software system composed of multiple components that are interconnected in the same way that humans perform sequential data transformations, such as translating text from one language to another. RNNs are largely being replaced by artificial intelligence (AI) based on transformation engines and large language models, which are much more efficient at pro-cessing sequential data.

Long-short-term memory (LSTM), developed by Hochreiter and Schmidhuber (1997), is a variant of RNN that allows the model to expand its memory capacity to accommodate longer time horizons. LSTM is designed to avoid the long-term dependency problem. Remembering information over a long period of time is their default property, and we do not need to train them to remember it. That is, they can remember it by nature without any intervention. RNNs can only remember the immediate recent input, but RNNs cannot use input from some previous sequence to improve predictions.

The LSTM model consists of 3 gates, stated in turn through Equations 6, 7, 8:
- Input Gate:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (6)$$

- Forget Gate:

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \quad (7)$$

- Output Gate:

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \quad (8)$$

Where:
σ: Sigmoid function.
w_x: Weight for the respective gate (x) neurons.
h_{t-1}: Output of the previous LSTM block (at timestamp $t-1$).
x_t: Input at current timestamp.
b_o: Biases for the respective gates (x).

3.2.5. Forecasting error

We compare the effectiveness of forecasting methods using the following four forecasting error metrics (Equations 9, 10, 11, 12):
- Mean absolute deviation:

$$MAD = \frac{1}{n}\sum_{t=1}^{n} |Y_t - \hat{Y}_t| \quad (9)$$

- Mean squared error:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 \quad (10)$$

- Root mean squared error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2} \quad (11)$$

- Mean absolute percentage error:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \hat{Y}_t|}{|Y_t|} \quad (12)$$

4. RESULT

The study conducts forecasting and comparison for the last 20 periods of the dataset, with the results presented in Figures 2, 3, 4, and 5. Comparing the Autocorrelation Function (ACF) of three types of pens gives a brief overview for total products in the company, illustrating the appropriation of each method to the data set.

4.1. Moving average

When fitting the autocorrelation function to the residuals of the Moving Average model for pen A, B, and C, most of the lags are within significant limits. Although some of them are outside the limitation, they do not strongly affect to effect of this method (Figure 2).

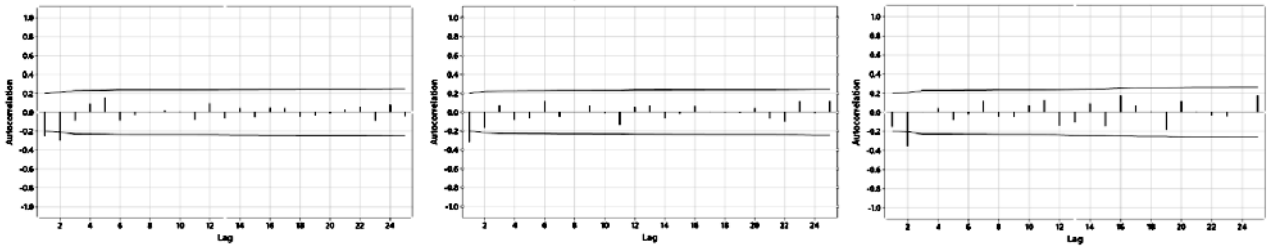

Figure 2. Moving average autocorrelation function plot of pen A, B, and C

4.2. Exponential smoothing

The residuals are all within the limitations of the ACF plot, which means that ARIMA works well and is more appropriate for the dataset (Figure 3).

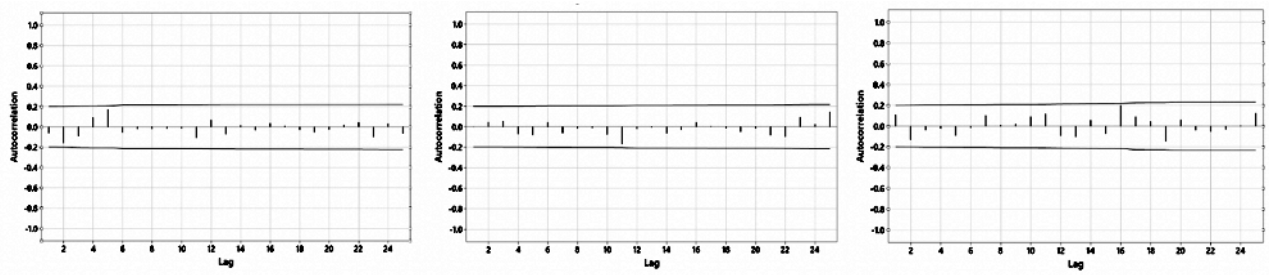

Figure 3. Exponential smoothing autocorrelation function plot of Pen A, B, C

4.3. Autoregressive integrated moving average

This method gives the ACF Plot with quite small residuals, and all of them are in the limitation. Therefore, its forecasting errors for the 3 types of pens are lower than 2 methods before (Figure 4).

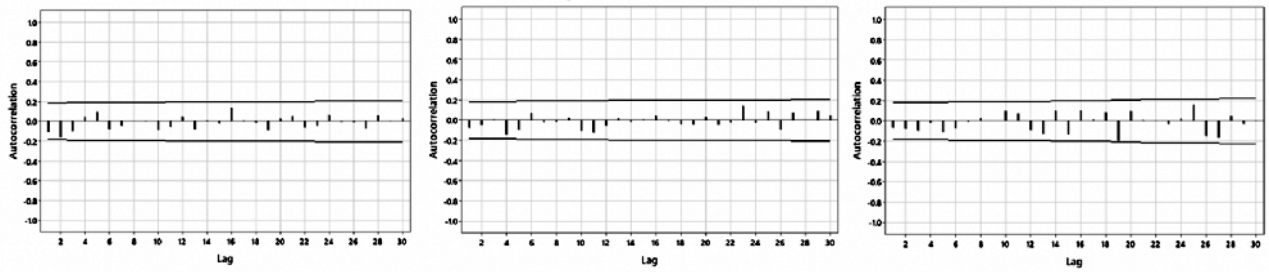

Figure 4. ARIMA autocorrelation function plot of pen A, B, C

4.4. Long-short term memory

All three plots show a decreasing trend in loss over time, indicating that the model is learning and improving as training progresses. The loss starts at a relatively high value and gradually reduces, which is expected in most machine learning training processes. Due to the low final loss, these plots show that the models are learning well by detecting non-linear patterns or hidden elements in sequential data that traditional methods often miss.

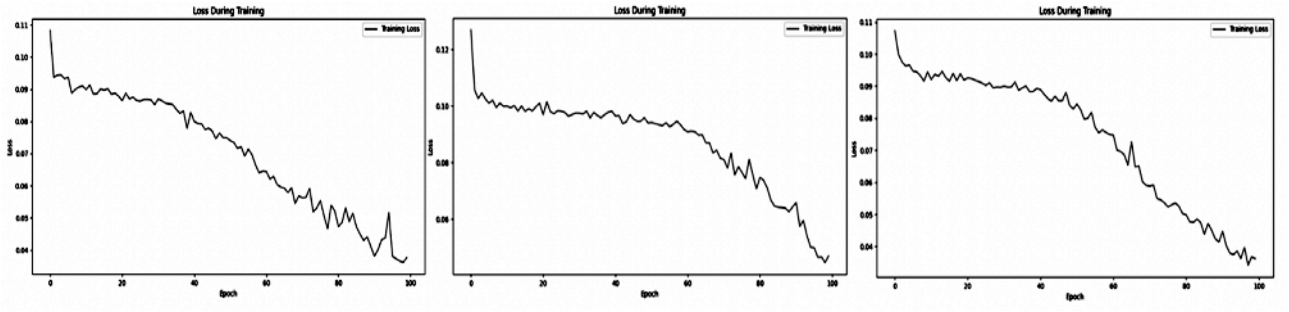

Figure 5. Loss function plot of pens A, B, and C

For a more obvious observation, the results can be summarized in Table 1, which compares the forecasting error metrics of the five methods. It can be observed that the LSTM method yields the most favorable error results compared to the other methods. An ANN network is implemented in Python to achieve the optimal results. The forecasting process relies on the previous 16 weeks of data (e.g., to predict sales for week 17, data from weeks 1 to 16 is used; for week 18, data from weeks 2 to 17 is utilized). The core of the model lies in the LSTM layer with 80 units. The model is designed to predict a single output. The Mean Squared Error (MSE) is used as the loss function during training.

Assuming the production cost of one box of pens is 100,000 VND, the average savings for types A, B, and C are 15,481,070 VND, 15,693,820 VND, and 11,642,130 VND, respectively. This appears to be quite feasible, and if applied to all of the company's products, it would result in significant cost savings.

Table 1. Comparative table of three types of pens in terms of forecasting metrics

PEN A				
	Moving Average	**Exponential Smoothing**	**ARIMA**	**LSTM**
MAD	191.775	173.8624	170.92	139.3938
MSE	51,950.7125	40,195.502	42,225.1	23,966.3428
MAPE	9.94%	9.41%	9.08%	7.41%
RMSE	227.927	200.4882	205.49	154.8107
PEN B				
	Moving Average	**Exponential Smoothing**	**ARIMA**	**LSTM**
MAD	170.625	145.3838	147.71	122.2053
MSE	39,857.4125	28,739.9199	33,841.59	24,629.6053
MAPE	20.54%	16.49%	17.74%	13.88%
RMSE	199.6432	169.5285	183.96	156.9382
PEN C				
	Moving Average	**Exponential Smoothing**	**ARIMA**	**LSTM**
MAD	136.55	117.0827	119.51	97.093
MSE	26,873.575	16,900.8302	18,600.27	13,553.9158
MAPE	9.86%	8.38%	8.69%	7.14%
RMSE	163.931617	130.0032	136.38	116.4213

5. CONCLUSION

This research has provided a comprehensive comparison of traditional forecasting methods, including Moving Average, Exponential Smoothing, ARIMA, with a deep learning approach - Long-Short Term Memory (LSTM) networks to predict the production demand of a local pen manufacturing company. The study describes strengths of classical techniques such as simplicity, transparency, and ease of implementation in stable conditions and demonstrates the difficulty these methods face in some circumstances. However, although LSTM gives strong effectiveness to research in unusual backgrounds, it also comes with trade-offs for SMEs.

This study contributes a new approach to resolve

with this problem by using LSTM, a strong technique to forecast, combining with reducing the layers of this network to bring a possible result and then offering a feasible solution for resource-constrained manufacturers. By emphasizing the importance of selecting forecasting models that align with specific production environments and economic contexts, this research offers actionable insights for SMEs seeking to improve operational efficiency and responsiveness.

In the future, research will concentrate on developing more interpretable deep learning models. This recommendation is crucial for SMEs to respond effectively to demand fluctuations and optimize their production planning and cost savings in an ever-changing economic landscape.

ACKNOWLEDGEMENT

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

REFERENCES

[1] M. S. Haque, M. S. Amin, and J. Miah, "Retail demand forecasting: a comparative study for multivariate time series," *arXiv preprint arXiv:.11939,* 2023.

[2] I. Sumitra, "Comparison of forecasting the number of outpatients visitors based on naïve method and exponential smoothing," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 662, no. 4, p. 042002: IOP Publishing.

[3] E. Kahraman and O. Akay, "Comparison of exponential smoothing methods in forecasting global prices of main metals," *Mineral Economics,* vol. 36, no. 3, pp. 427-435, 2023.

[4] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing,* vol. 50, pp. 159-175, 2003.

[5] E. Munarsih and I. Saluza, "Comparison of exponential smoothing method and autoregressive integrated moving average (ARIMA) method in predicting dengue fever cases in the city of Palembang," in *Journal of Physics: conference series*, 2020, vol. 1521, no. 3, p. 032100: IOP Publishing.

[6] E. Chodakowska, J. Nazarko, Ł. Nazarko, H. S. Rabayah, R. M. Abendeh, and R. Alawneh, "ARIMA models in solar radiation forecasting in different geographic locations," *Energies,* vol. 16, no. 13, p. 5029, 2023.

[7] J. E. Hanke and D. W. Wichern, *Business forecasting*. Pearson Educación, 2005.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

Phương pháp dự báo hiệu quả trong xác định nhu cầu sản xuất: Một nghiên cứu điển hình

Nguyễn Thị Thu Hảo, Hà Trọng Khôi, Lê Thị Diễm Châu

TÓM TẮT

Dự báo từ lâu đã là yếu tố thiết yếu giúp doanh nghiệp cải thiện hiệu suất và tăng hiệu quả về kinh tế. Các phương pháp dự báo kinh điển như trung bình động (Moving Average), làm mịn theo hàm mũ (Exponential Smoothing), ARIMA cùng với phương pháp bộ nhớ dài hạn ngắn hạn (LSTM) đã được sử dụng trong nghiên cứu này với mục đích so sánh và tìm ra phương pháp hiệu quả nhằm giúp tối ưu hoá kế hoạch và quản lý sản xuất hoặc quản lý tồn kho, đồng thời giúp doanh nghiệp tiết kiệm chi phí. Các phương pháp cổ điển tuy dễ thực hiện nhưng khó nắm bắt và khó ghi nhớ xu hướng thị trường, trong khi phương pháp học máy cho ra kết quả rất tốt đối với tình huống này. Bộ dữ liệu được sử dụng được lấy từ một công ty sản xuất bút địa phương, được xử lý bằng cách chuẩn hoá và kiểm tra tính dừng. Hiệu quả của mô hình được đánh giá thông qua các chỉ số như sai số bình phương trung bình (MSE), căn bậc hai sai số bình phương trung bình (RMSE) và sai số phần trăm tuyệt đối trung bình (MAPE). Theo kết quả thử nghiệm, các phương pháp cổ điển cung

cấp đường cơ sở bền vững. Tuy nhiên, LSTM cho thấy dự báo có độ chính xác cao hơn, có thể cải thiện sai số dự báo và độ tin cậy, mang lại lợi ích về mặt kỹ thuật và kinh tế cho doanh nghiệp. Nghiên cứu này đóng góp vào lĩnh vực sản xuất bằng cách cung cấp phân tích so sánh toàn diện về các phương pháp dự báo và hiểu biết thực tế cho các công ty công nghiệp để nâng cao năng lực dự báo nhu cầu sản xuất hiệu quả.

Keywords: *dự báo, trung bình động, làm mịn theo hàm mũ, arima, bộ nhớ dài hạn ngắn hạn, chuỗi thời gian*

Received: 27/3/2025
Revised: 09/6/2025
Accepted for publication: 12/6/2025