

Capacity Prediction, Cost Reduction and Efficiency Improvement Optimization for Mold Enterprises Based on LSTM-RF Ensemble Learning

He Zhang

*Shenzhen College of International Education, Shenzhen, China
s23270.zhang@stu.scie.com.cn*

Abstract. To address challenges such as significant capacity fluctuations, low scheduling efficiency, and crude cost control in plastic mold manufacturing facilities, this study proposes an integrated forecasting framework that combines Long Short-Term Memory (LSTM) networks with Random Forest (RF) models, integrating greedy scheduling with an economic quantification model to construct a full-chain decision-making system comprising "forecasting–scheduling–evaluation". Based on the multi-dimensional production data of the past 365 days, lagged feature engineering and periodic feature engineering are used to build a dual-model system for rolling forecasts of equipment load rates, energy consumption and work order volumes; the scheduling process is optimised by a priority-weighted rule; and cost reduction benefits are quantified in four dimensions: labour, materials, energy consumption and maintenance. Based on the experiment, the integrated model has achieved a Mean Absolute Error (MAE) of 0.0324 and an R^2 value of 0.8916 for load rate forecasting; after scheduling optimisation, the idle rate has dropped by 17.82 percentage points, and a total monthly cost saving of about CNY 196,000 has been realised.

Keywords: Production capacity forecasting, ensemble learning, scheduling optimization, cost reduction and efficiency improvement, mold manufacturing plant

1. Introduction

Small and medium-sized enterprises in the manufacturing industry generally face challenges such as low Overall Equipment Effectiveness (OEE) and high production downtime caused by inefficient scheduling [1]. Data-driven capacity forecasting and intelligent scheduling constitute key pathways to achieving lean operations [2]. The LSTM network demonstrates excellent performance in time-series forecasting, with its gate mechanism effectively capturing long-range temporal dependencies [3]; whereas the random forest model offers advantages such as high robustness and insensitivity to hyperparameters in structured industrial data regression tasks [4]. However, single models struggle to simultaneously model both temporal dynamics and static characteristics. Zhang et al. [5] note that automatically designed dispatching rules and scheduling heuristics offer the advantage of low computational cost in real-time scheduling. The contributions of this paper are as follows: (1) constructing an integrated LSTM-RF dual-model framework for precise capacity forecasting; (2)

designing a priority-weighted greedy scheduling algorithm to optimize scheduling; and (3) establishing a four-dimensional economic benefit quantification model.

2. Literature review

Sampedro et al. [6] applied a multi-flow bidirectional LSTM model to fault mitigation in smart additive manufacturing, achieving an R^2 value of 0.9121, though their approach was limited to 3D printing equipment. Liang et al. [7] employed an adaptive genetic algorithm based on individual similarity to solve the multi-objective flexible job-shop scheduling problem, reducing the makespan, yet the computational complexity remains a challenge for large-scale real-time scheduling. Existing research lacks a comprehensive solution that deeply integrates production capacity forecasting, scheduling optimization, and the quantification of overall benefits; this paper addresses this gap.

3. Research methods

3.1. Data and features

The production dataset spans 365 days and contains operational records from 10 devices, 1,095 orders, 8 production processes, 6 types of raw materials, and 12 months of financial summaries. Feature engineering involves extracting 5 temporal periodic features: day of the week, month, quarter, whether it is a weekend, and ten-day period; additionally, lagged features of orders 1–3 are constructed based on equipment load averages, daily energy consumption, actual working hours, and order quantity, resulting in a 35-dimensional input matrix with a window width of 7 days. This paper proposes an LSTM-RF dual-model ensemble framework. The LSTM channel employs a single-layer architecture with 32 hidden units to capture temporal dependencies, generating a three-dimensional tensor input of shape [number of samples, 7, 35] using a 7-day sliding window; the RF channel is configured with `n\_estimators=200` and `max\_depth=12`. Both models are trained on the first 80% of the data and tested on the remaining 20%, while strictly preserving the chronological order along the time dimension. The forward propagation of the LSTM cell controls information transmission through three core gating structures: the forget gate, input gate, and output gate. The ensemble prediction result is obtained by averaging the outputs of the two models:

$$y_{ensemble} = \frac{y_{LSTM} + y_{RF}}{2} \quad (1)$$

The calculation formulas for the LSTM gating mechanism and cell state update are defined as follows:

$$f_t = \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \quad (2)$$

$$i_t = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(\mathbf{W}_c \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \quad (4)$$

$$o_t = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o), \quad \mathbf{h}_t = o_t \odot \tanh(c_t) \quad (5)$$

where σ refers to the sigmoid activation function; $\mathbf{W}_f, \mathbf{W}_i, \mathbf{W}_c, \mathbf{W}_o$ denote trainable weight matrices; $\mathbf{b}_f, \mathbf{b}_i, \mathbf{b}_c, \mathbf{b}_o$ are the corresponding bias vectors; $\mathbf{h}_{t-1}$ and $\mathbf{x}_t$ represent the previous hidden state and current input features, respectively; c_t is the cell state and $\mathbf{h}_t$ stands for the updated hidden state. In this work, $\cdot$ denotes matrix dot product, and $\odot$ denotes element-wise Hadamard product for rigorous mathematical distinction.

3.2. Scheduling optimization and economic model

To achieve multi-criteria adaptive job scheduling, a comprehensive priority scoring function is established, which synthetically integrates order urgency, order economic value, and overdue penalty factors. The priority score P_i is calculated as:

$$P_i = 1000U_i + \frac{A_i}{500} + 50 \cdot \max(0, D_i) \quad (6)$$

The weights are configured according to practical production characteristics of the mold workshop. Specifically, emergency order indicator U_i is assigned a dominant weight to avoid delivery delay risks. Order amount A_i and delayed days D_i are introduced to balance economic benefits and on-time delivery. All work orders are sorted in descending order of P_i and allocated to the earliest available machine. Orders with high urgency are forcibly scheduled in the initial time slot to guarantee delivery requirements.

A multi-dimensional cost-benefit evaluation model is constructed to quantitatively assess the optimization performance. The total monthly saved cost S_{total} is decomposed into savings on labor, raw materials, energy and equipment maintenance:

$$S_{\text{total}} = S_{\text{labor}} + S_{\text{material}} + S_{\text{energy}} + S_{\text{maintenance}} \quad (7)$$

The specific calculation rules are defined as follows:

$S_{\text{labor}} = C_{\text{labor}} \times 0.15$: labor cost savings originating from reduced overtime and idle manpower;

$S_{\text{material}} = \text{MaterialLoss} \times 0.3$: raw material waste reduction achieved via optimized production matching;

$S_{\text{energy}} = E_{\text{idle}} \times 0.7 \times 0.8 \text{ CNY/kWh}$: electricity savings by cutting idle equipment operation;

$S_{\text{maintenance}} = C_{\text{maintenance}} \times 0.12$: maintenance cost reduction benefited from balanced equipment load distribution.

4. Experimental results

4.1. Estimation performance

Table 1 quantitatively summarizes the prediction performance of the single LSTM model, single RF model and the proposed LSTM-RF ensemble model on three forecasting targets, including equipment load rate, energy consumption and work order quantity. For the equipment load rate prediction task, the ensemble method obtains an MAE of 0.0324 and an R^2 of 0.8916, which

outperforms the individual LSTM (MAE=0.0370, $R^2=0.8623$) and RF model (MAE=0.0401, $R^2=0.8547$). The ensemble framework can also be used to achieve a good-performing result in energy consumption and order quantity forecasting. Compared with a single model, the combined approach has a lower prediction error and better fit. The above experiments show that LSTM is suitable for learning the temporal structure of time series, and RF can extract non-linear relationships among static industrial features; thus, the two models are combined to improve the accuracy and generalisation ability of the prediction.

Table 1. Model performance comparison

Target	LSTM MAE	LSTM R^2	RF MAE	RF R^2	Ens. MAE	Ens. R^2
Load Rate	0.0370	0.8623	0.0401	0.8547	0.0324	0.8916
Energy (kWh)	48.21	0.8744	52.80	0.8612	42.56	0.9005
Order Qty	1.24	0.8012	1.31	0.7923	1.05	0.8260

Figure 1 shows the continuous comparison curve of the 30-day historical load rate (blue shaded area) and the following 30-day load forecast (orange dashed line). The historical load has a distinct 7-day cycle and is in the range of 78%-92%. The model has learned the weekly cycle and, therefore, the predicted load values are generally between 75% and 88%. The light-orange band around the forecast line is the prediction confidence interval, and the extent of this uncertainty varies. The horizontal red dashed line is the 80% alarm threshold; if the load rate exceeds this value, additional production capacity planning needs to be carried out to prevent resource oversupply.

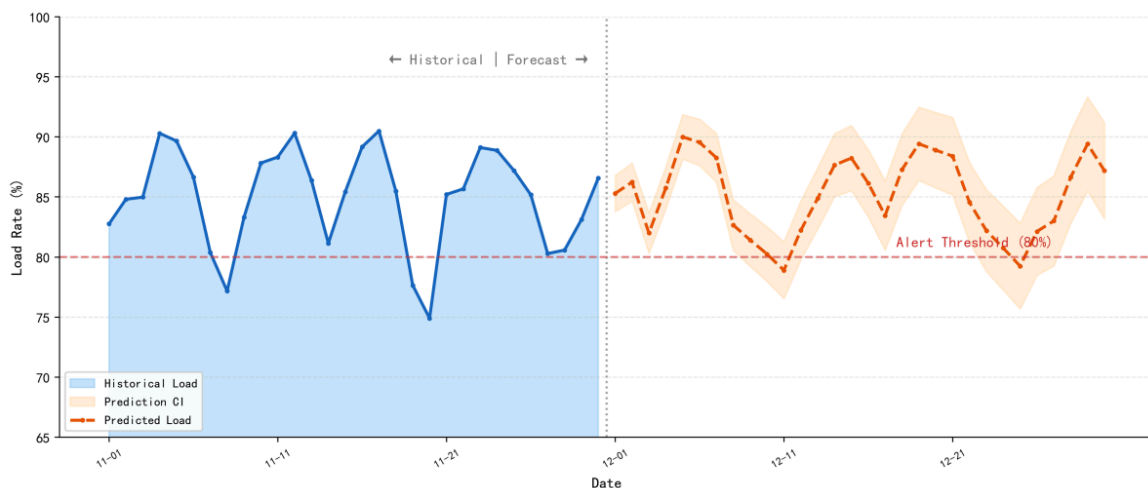


Figure 1. Load prediction: 30-day history+30-day forecast

Figure 2 shows a grouped bar chart of the average load and peak load rates for the four types of production equipment. Among all the workshops, the injection-moulding machine cluster is the main production capacity bottleneck, with an average load of 85.3% and a peak load of 97.2%; both exceed the critical congestion threshold. The CNC and grinding machine groups have a relatively low utilisation rate, at a peak load rate of 83.1% and 93.6%, respectively. On the other hand, the EDM equipment has the lowest resource utilisation at an average of 58.7%, indicating significant idle capacity of the equipment and uneven distribution of resources in the production line.

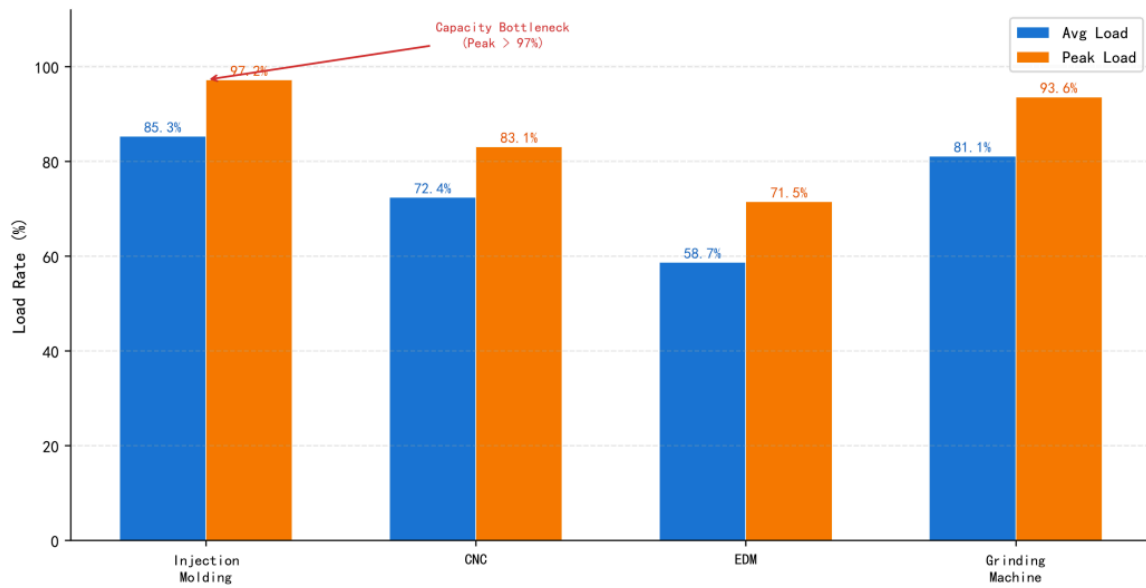


Figure 2. Equipment load by type: average vs. peak

4.2. Scheduling and economic benefits

Scheduling optimisation achieved good results: the total production cycle was reduced from 43,800 hours to 27,540 hours, a reduction rate of 37.12%; equipment idle time fell from 42.35% to 24.53%, a drop of 17.82 percentage points; and the proportion of delayed work orders decreased from 28.6% to 6.2%.

Table 2 shows the cost reduction benefits in four areas: a monthly total saving of CNY 196,000 and an annual estimated saving of CNY 2,352,000; a cost reduction rate of 18.76%; and an increase of about 2.87 percentage points in profit margin. The largest share of material loss savings (42.7%) was due to a reduction in the loss rate from 7.96% to 5.57%, and labour cost savings resulted from schedule optimisation, reduced overtime, and non-productive time for staff.

Table 2. Summary of economic benefits

Dimension	Monthly Saving (10k CNY)	Annual Saving (10k CNY)
Labor Cost	5.49	65.88
Material Loss	8.37	100.44
Energy Consumption	4.73	56.76
Equipment Maintenance	1.01	12.12
Total	19.60	235.20

5. Discussion

The LSTM-RF integration strategy can improve the R^2 value by approximately 0.03-0.04 with a relatively small increase in computational cost, and thus the combined strengths of temporal and static feature models have been confirmed. The R^2 value of the work order quantity prediction is relatively low (0.8260), which may be due to the inherent Poisson randomness of order arrivals;

purely history-driven predictions have a natural upper bound. Although greedy scheduling is not as good as metaheuristic methods globally, it has a very fast computation speed and supports dynamic reordering; therefore, it is more suitable for real-time decision-making at the field level. Research limitations: The improvement speed of the economic model varies by industry and requires customisation by enterprises.

6. Conclusion

This paper proposes a three-in-one integrated solution of "data-driven prediction - intelligent scheduling - benefit quantification". The main results are as follows: (1) The LSTM-RF hybrid model achieves an MAE of 0.0324 and an R^2 of 0.8916 in load rate forecasting, showing strong complementary effects of the two models; (2) The greedy scheduling algorithm reduces the cycle time by 37.12% and lowers the idle rate by 17.82 percentage points; (3) The multi-dimensional economic model estimates monthly cost savings of CNY 196,000 and annual savings of CNY 2,352,000, thus verifying the economic feasibility of the proposed solution. This paper offers a low-barrier path to digitalisation for small and medium-sized mould enterprises.

References

- [1] Liu, X., Li, J., Wang, H., Jia, W., Yang, J., & Guo, Z. (2023). Design of an optimal scheduling control system for smart manufacturing processes in tobacco industry. *IEEE Access*, 11, 33027–33036. <https://doi.org/10.1109/ACCESS.2023.3261883>
- [2] Kalir, A. A., et al. (2023). Leveraging machine learning for capacity and cost on a complex toolset: A case study. *IEEE Transactions on Semiconductor Manufacturing*, 36(4), 611–618. <https://doi.org/10.1109/TSM.2023.3314431>
- [3] Khan, A., Fouda, M. M., Do, D.-T., Almaleh, A., & Rahman, A. U. (2023). Short-term traffic prediction using deep learning long short-term memory: Taxonomy, applications, challenges, and future trends. *IEEE Access*, 11, 94371–94391. <https://doi.org/10.1109/ACCESS.2023.3309601>
- [4] Alfarizi, M. G., Tajiani, B., Vatn, J., & Yin, S. (2023). Optimized random forest model for remaining useful life prediction of experimental bearings. *IEEE Transactions on Industrial Informatics*, 19(6), 7771–7779. <https://doi.org/10.1109/TII.2022.3206339>
- [5] Zhang, F., Mei, Y., Nguyen, S., & Zhang, M. (2024). Survey on genetic programming and machine learning techniques for heuristic design in job shop scheduling. *IEEE Transactions on Evolutionary Computation*, 28(1), 147–167. <https://doi.org/10.1109/TEVC.2023.3255246>
- [6] Sampedro, G. A., Putra, M. A. P., Lee, J.-M., & Kim, D.-S. (2023). Industrial Internet of Things-based fault mitigation for smart additive manufacturing using multi-flow BiLSTM. *IEEE Access*, 11, 99130–99142. <https://doi.org/10.1109/ACCESS.2023.3312724>
- [7] Liang, X., Liu, Y., Gu, X., Huang, M., & Guo, F. (2022). Adaptive genetic algorithm based on individual similarity to solve multi-objective flexible job-shop scheduling problem. *IEEE Access*, 10, 45748–45758. <https://doi.org/10.1109/ACCESS.2022.3170032>