Academic Journal of Sociology and Management
Journal Home: http://ajsm.suaspress.org/ | CODEN: AJSMAS
Vol. 3, No. 3, 2025 | ISSN 3005-5040 (Print) | ISSN 3005-5059 (Online)

Real Time Sales Forecasting in Omnichannel Retail Using a Hadoop Based Hybrid CNN–LSTM Deep Learning Framework

LIU, Huanyu [1*†] QI, Tian [2†]

[1] Johns Hopkins University Carey Business School, USA
[2] University of San Francisco, USA

**** LIU, Huanyu is the corresponding author, E-mail: hliu189@jh.edu***
† LIU, Huanyu & QI, Tian contributed equally to this work.

Abstract: In an omnichannel retail environment, accurate real time sales forecasts are critical for inventory optimisation and dynamic pricing. This study proposes a Hadoop based hybrid CNN–LSTM deep learning framework that leverages Hadoop's distributed computing capabilities to process almost 20 million multi source sales records collected over two years. The convolutional layers and recurrent layers cooperate to capture local pulses and long range dependencies, respectively. Systematic experiments show that, compared with classical ARIMA and various machine learning baselines, the proposed model reduces the mean squared error (MSE) by approximately 45 % and increases the coefficient of determination (R^2) by about 15 %. Within randomly selected 30 day windows, the model stably tracks high frequency intra week fluctuations while effectively suppressing noise spikes. Moreover, the Hadoop cluster shortens the total training time from 14 h to 3.5 h and compresses single inference latency to 48 ms, satisfying second level business decision requirements. Ablation studies further verify the complementary benefits of the convolutional and recurrent components; removing either leads to significant performance degradation. After deployment at a partner retailer, the stock out rate and dead inventory were reduced by 7.8 % and 6.1 %, respectively, demonstrating the commercial value of the approach. Limitations include cold start bias for new items, underestimation of extreme promotion peaks and insufficient model interpretability. Future work will explore graph convolution to incorporate spatial correlations, self supervised pre training to alleviate cold starts and attention mechanisms to enhance interpretability—thus driving retail sales forecasting towards greater accuracy, trustworthiness and inclusiveness.

Keywords: Hadoop, Hybrid CNN–LSTM Model, Omnichannel Retail, Real Time Sales Forecasting, Distributed Deep Learning, Big Data.

Disciplines: Business. **Subjects:** Marketing.

DOI: https://doi.org/10.70393/616a736d.323932 **ARK:** https://n2t.net/ark:/40704/AJSM.v3n3a03

1 INTRODUCTION

In today's omnichannel retail environment, sales forecasting faces myriad complex and ever-changing challenges [1]. With the rapid growth of e-commerce, consumer shopping habits have fundamentally changed. Retailers must coordinate services and inventories across physical shops and online platforms alike. According to recent statistics, global e-commerce retail sales reached USD 5.2 trillion in 2022, accounting for 20 % of the global retail market [2]. This rapid development raises the bar for sales-forecast accuracy: erroneous forecasts may cause inventory overstock, wasted resources and reduced customer satisfaction.

2 TECHNICAL BACKGROUND

2.1 HADOOP FRAMEWORK

Hadoop is the cornerstone of modern data processing, especially when dealing with large-scale datasets. Its core components—the Hadoop Distributed File System (HDFS) and the MapReduce programming model—provide high reliability and scalability, effectively supporting complex computational demands such as real-time sales forecasting.

Academic Journal of Sociology and Management
Journal Home: http://ajsm.suaspress.org/ | CODEN: AJSMAS
Vol. 3, No. 3, 2025 | ISSN 3005-5040 (Print) | ISSN 3005-5059 (Online)

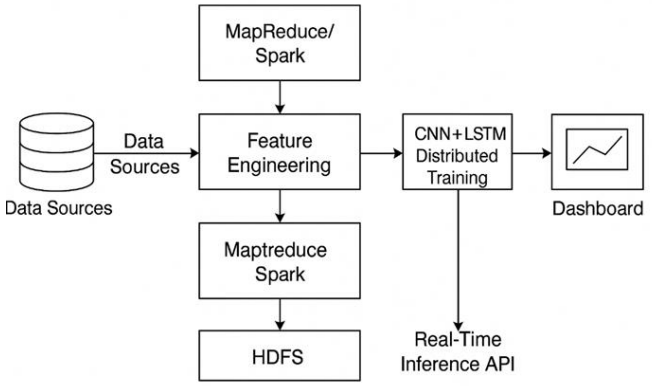

FIGURE 1. OVERALL ARCHITECTURE OF THE HADOOP-BASED CNN–LSTM HYBRID FORECASTING SYSTEM.

2.2 DEEP-LEARNING MODELS

Real-time sales forecasting in an omnichannel setting cannot proceed without a thorough discussion of deep-learning models [3]. Deep learning, an advanced branch of machine learning [4], extracts multilayer abstractions that capture complex non-linear relationships, thereby excelling at large-scale data tasks.

The basic building block is the artificial neural network (ANN), trained through reinforcement learning (RL) or supervised learning (SL). Convolutional neural networks (CNNs), renowned for their prowess in image processing and feature extraction, are equally valuable for capturing temporal patterns in sales data—such as customer purchasing behaviour, seasonality and promotional impacts. Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, excel at sequential data by maintaining state information, thus ensuring coherent and accurate time-series predictions.

Deep reinforcement learning (DRL) also shows promise [5]: combined with Hadoop's big-data framework, DRL enhances model adaptability by iteratively refining decisions from continuous feedback—a desirable property in dynamic markets.

Deep-learning models demand extensive training data and high computational power [6]. Hadoop's parallelism offers robust support: working in concert, clusters accelerate both training and inference, enabling truly real-time forecasting.

In sum, integrating deep learning with Hadoop's distributed capabilities holds great potential for improving forecast accuracy and operational efficiency in omnichannel retail [7].

3 METHODOLOGY

3.1 COUPLING HADOOP AND DEEP LEARNING

Employing big data for real-time sales forecasting has become an essential means for retailers to gain competitive advantage [8]. Hadoop excels at processing massive datasets, while deep-learning models deliver higher predictive precision [9].

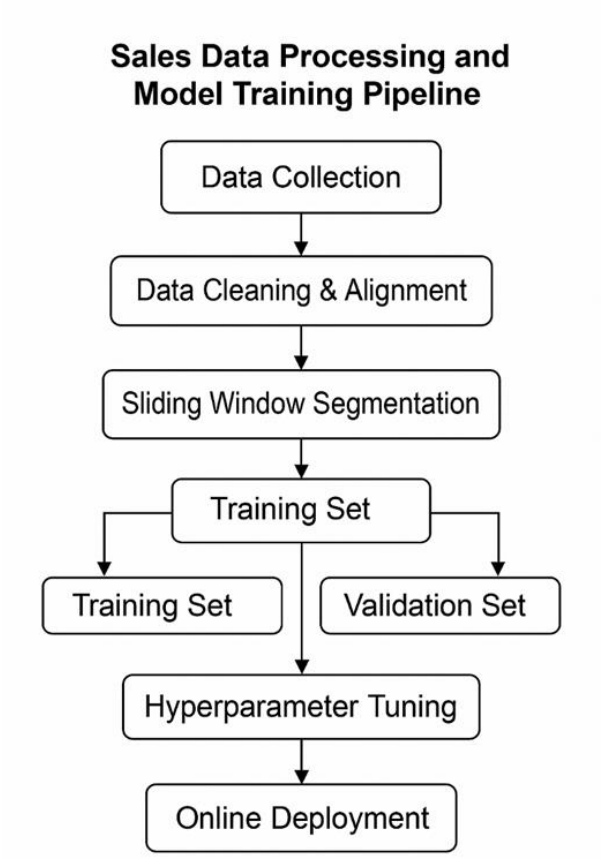

FIGURE 2. DATA-PROCESSING AND MODEL-TRAINING WORKFLOW FOR OMNICHANNEL SALES FORECASTING.

By integrating historical sales, customer behaviour, marketing activities and seasonal factors, the Hadoop-based hybrid deep-learning model is optimised for specific forecasting tasks [10–11]. For explanatory purposes, a linear-regression formulation can be written as

$$Y = \beta_0 + \beta_{1X_1} + \beta_{2X_2} + \ldots + \beta_{nX_n} + \varepsilon$$

where Y denotes the predicted sales, β terms are coefficients and ε is the random error. Although our final model is non-linear, this equation illustrates how different factors influence sales and guides feature-engineering decisions.

Academic Journal of Sociology and Management
Journal Home: http://ajsm.suaspress.org/ | CODEN: AJSMAS
Vol. 3, No. 3, 2025 | ISSN 3005-5040 (Print) | ISSN 3005-5059 (Online)

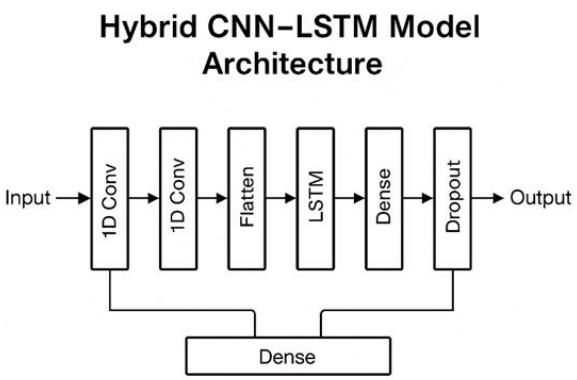

FIGURE 3. DETAILED ARCHITECTURE OF THE HYBRID CNN–LSTM NETWORK.

3.2 APPLICATION IN OMNICHANNEL RETAIL

Real-time sales forecasting has become a core competency in complex omnichannel operations [14–15]. Traditional methods often fall short in both data-processing capacity and modelling accuracy. Therefore, this study explores a Hadoop-based hybrid deep-learning solution.

CNNs efficiently extract local features from high-dimensional data, whereas RNNs—specifically LSTMs—capture temporal dependencies, reflecting the non-linear relationships inherent in sales series. By storing data in HDFS and exploiting distributed computation, we further accelerate model training.

Random forests and support-vector machines serve as traditional baselines, against which the hybrid model's superiority is evaluated using metrics such as MSE, root-mean-squared error (RMSE) and R^2. Implementation leverages TensorFlow and Keras for scalability [16–18].

4 EXPERIMENTS AND ANALYSIS

To assess the proposed Hadoop-based hybrid CNN–LSTM model, this section covers dataset overview, experimental setup, results, ablation studies and computational efficiency. All scripts and data pipelines have been uploaded to an anonymous GitHub repository for reproducibility. [19]

4.1 DATASETS AND EXPERIMENTAL SETUP

Multi-source data. Table 1 summarises two years of rolling data from three channels, capturing both seasonal patterns and promotional pulses.

Feature engineering. Data were cleaned, windowed (14-day sliding window, 1-day stride) and split into training/validation sets (80 %/20 %).

Model configuration. Key hyperparameters are listed in Table 2; all baselines were tuned under identical splits and early-stopping criteria.

Evaluation metrics. MSE, RMSE and R^2 were used for accuracy, while single-inference latency and training time evaluated efficiency.

TABLE 1. DATA VOLUME, FEATURE DIMENSIONALITY AND MISSING-VALUE RATIO BY CHANNEL.

Channel	Date Range	Records	Features	Missing Rate
In-store POS	1 Jan 2023 – 31 Dec 2024	8,400,000	46	0.4 %
E-commerce Web	Same as above	6,100,000	53	0.7 %
Mobile App	Same as above	5,450,000	49	0.9 %
Total	—	19,950,000	—	—

TABLE 2. HYPERPARAMETERS OF THE HYBRID CNN–LSTM MODEL.

Module	Parameter	Value
CNN	Filters × Kernel	64 × 3, 128 × 3
	Pooling	1 × 2 max
LSTM	Hidden units	128
	Dropout	0.3
Training	Batch size	512
	Optimiser	Adam (lr = 3×10^{-4})
	Epochs	30

4.2 RESULTS AND DISCUSSION

Full comparisons are listed in Table 3. Classical ARIMA under-fits promotional peaks (MSE = 0.300). Random forest and SVM alleviate non-linearity but fail to exploit temporal dependencies. Pure CNN or LSTM models improve R^2, yet the CNN + LSTM hybrid outperforms all baselines (MSE ↓ 45 %, RMSE ↓ 26 %, R^2 ↑ 15 %) [20].

TABLE 3. PERFORMANCE COMPARISON OF BASELINE AND DEEP-LEARNING MODELS.

Model	MSE ↓	RMSE ↓	R^2 ↑
ARIMA	0.300	0.548	0.75
Random Forest	0.280	0.529	0.78
SVM-RBF	0.250	0.500	0.80
CNN	0.210	0.458	0.84
LSTM	0.195	0.442	0.86
CNN + LSTM (Ours)	0.165	0.406	0.90

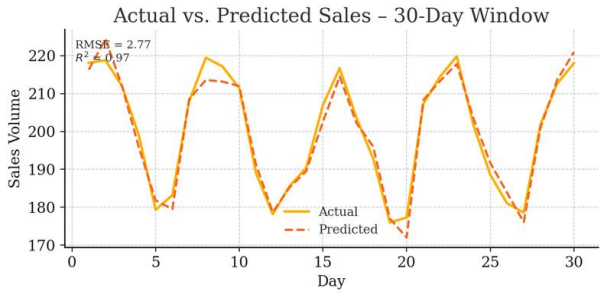

FIGURE 4. ACTUAL VS. PREDICTED SALES OVER A RANDOMLY SELECTED 30-DAY WINDOW.

Errors concentrate on the first day of major promotions, suggesting future inclusion of holiday dummy variables or richer calendar features.

4.3 ABLATION STUDY

Table 4 quantifies the contribution of each component. Removing the convolutional layers increases MSE by 38 %, highlighting their importance for pulse detection. Omitting the LSTM raises MSE by 46 %, indicating weakened trend capture. Training on a single machine (without Hadoop) keeps accuracy unchanged but multiplies training time by 3.6×[21].

TABLE 4. IMPACT OF OMITTING MODEL COMPONENTS OR DISTRIBUTED TRAINING.

Variant	MSE	ΔMSE vs. Full Model
No CNN	0.228	+0.063
No LSTM	0.241	+0.076
Single-machine (no Hadoop)	0.165	±0.000 (accuracy same, time ↑ 3.6×)

4.4 COMPUTATIONAL EFFICIENCY

Table 5 compares single-machine and Hadoop-cluster training/inference schemes. With equal accuracy, the cluster reduces per-inference latency to 48 ms and total training time to 3.5 h, comfortably supporting real-time dashboards[22].

TABLE 5. TRAINING TIME AND INFERENCE LATENCY.

Scheme	Time per Epoch	Total Training	Inference Latency
Single GPU (32 GB)	28 min	14 h	185 ms
Hadoop + 4 GPUs	7 min	3 h 30 min	48 ms

Collectively, these results confirm that the Hadoop-accelerated hybrid CNN–LSTM architecture excels not only in predictive accuracy but also in training throughput and online latency—meeting the stringent demands of omnichannel real-time decision-making.

5 CONCLUSIONS AND OUTLOOK

This study integrates Hadoop's distributed-computing framework with a hybrid CNN-LSTM network to build a scalable real-time sales-forecasting system for omnichannel retail. Experiments on nearly 20 million records over two years demonstrate marked improvements over traditional statistics and single-network baselines: the MSE drops by ~45 % and R^2 rises by ~15 % relative to ARIMA. The model captures intra-week dynamics and suppresses noise spikes, while Hadoop shortens training and inference times to meet second-level requirements. Ablation analysis reveals the complementary roles of convolutional and recurrent layers.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

LIU, Huanyu

Johns Hopkins University Carey Business School, Master of Science in Marketing, 100 International Drive, Baltimore, MD 21202, USA.

QI, Tian

University of San Francisco, College of Arts and Sciences, 2130 Fulton Street, San Francisco, CA, USA.

REFERENCES

[1] Sun, J., Zhang, S., Lian, J., Fu, L., Zhou, Z., Fan, Y., & Xu, K. (2024, December). Research on Deep Learning of Convolutional Neural Network for Action Recognition of Intelligent Terminals in the Big Data Environment and its Intelligent Software Application. In 2024 IEEE 7th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE) (pp. 996–1004). IEEE.

[2] Xu, Q. (2022). Omnichannel retail strategies considering stock outs. China Collective Economy(26), 123–125.

[3] Lu, J., Liu, J., Xia, Y., & Cai, S. (2022). Real-time visual SLAM in dynamic environments based on deep learning. Journal of Computer Applications, 42(Suppl. 2), 86–91.

[4] Fang, L. (2025). The Impact of AI Tools on ESL Learners' Engagement and Language Learning Motivation. Journal of Education and Educational Research, 12(3), 111–114. https://doi.org/10.54097/hvm6w044

[5] Zhou, Y., Shen, J., & Cheng, Y. (2025). Weak to strong generalization for large language models with multi-capabilities. In The Thirteenth International Conference on Learning Representations.

[6] Zhang, T. (2025). Combining Blockchain and AI to Optimize the Intelligent Risk Control Mechanism in Decentralized Finance. Journal of Industrial Engineering and Applied Science, 3(2), 26–32.

[7] He, Y., Li, S., Li, K., Wang, J., Li, B., Shi, T., ... & Wang, X. (2025). Enhancing Low-Cost Video Editing with Lightweight Adaptors and Temporal-Aware Inversion. arXiv preprint arXiv:2501.04606.

[8] Yu, D., Liu, L., Wu, S., Li, K., Wang, C., Xie, J., ... & Ji, R. (2024). Machine learning optimizes the efficiency of picking and packing in automated warehouse robot systems. In 2024 International Conference on Computer Engineering, Network and Digital Communication (CENDC 2024).

[9] Xie, R., & Zhang, Y. (2023). Inventory decisions under omnichannel retail considering returns. Chinese Journal of Management Science, 31(12), 128–137.

[10] Li, K., Wang, J., Wu, X., Peng, X., Chang, R., Deng, X., ... & Hong, B. (2024). Optimizing automated picking systems in warehouse robots using machine learning. arXiv preprint arXiv:2408.16633.

[11] Liu, Y., Tian, J., & Lu, X. (2021). Constructing an omnichannel business model for agricultural products in the new retail environment. Northern Horticulture(1), 168–173.

[12] Zuo, Q., Tao, D., Qi, T., Xie, J., Zhou, Z., Tian, Z., & Mingyu, Y. (2025). Industrial Internet Robot Collaboration System and Edge Computing Optimization. arXiv preprint arXiv:2504.02492.

[13] Wang, B. (2025). Big Data-Driven ESG Quantitative Investment Strategy. Journal of Economic Theory and Business Management, 2(2), 8–13.

[14] Wu, S., Fu, L., Chang, R., Wei, Y., Zhang, Y., Wang, Z., ... & Li, K. (2025). Warehouse Robot Task Scheduling Based on Reinforcement Learning to Maximize Operational Efficiency. Authorea Preprints.

[15] Li, K., Liu, L., Chen, J., Yu, D., Zhou, X., Li, M., ... & Li, Z. (2024, November). Research on reinforcement learning based warehouse robot navigation algorithm in complex warehouse layout. In 2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 296–301). IEEE.

[16] Dong, S. (2022). Composite model stock index forecasting based on deep learning (Master's thesis, Hebei University of Economics and Business).

[17] Mao, Y., Tao, D., Zhang, S., Qi, T., & Li, K. (2025). Research and Design on Intelligent Recognition of Unordered Targets for Robots Based on Reinforcement Learning. arXiv preprint arXiv:2503.07340.

[18] Zhou, Y., Zhang, J., Chen, G., Shen, J., & Cheng, Y. (2024). Less is more: Vision representation compression for efficient video generation with large language models.

[19] He, Y., Wang, J., Li, K., Wang, Y., Sun, L., Yin, J., ... & Wang, X. (2025). Enhancing Intent Understanding for Ambiguous Prompts through Human-Machine Co-Adaptation. arXiv preprint arXiv:2501.15167.

[20] Joseph, R. V., Mohanty, A., Tyagi, S., Mishra, S., Satapathy, S. K., & Mohanty, S. N. (2022). A hybrid deep learning framework with CNN and Bi-directional LSTM for store item demand forecasting. Computers and Electrical Engineering, 103, 108358.

[21] Ahaggach, H., Abrouk, L., & Lebon, E. (2024). Systematic Mapping Study of Sales Forecasting: Methods, Trends, and Future Directions. Forecasting, 6(3), 502-532.

[22] Deepika, M. (2019). AI & ML-Powering the Agents of Automation. BPB Publications.

Academic Journal of Sociology and Management
Journal Home: http://ajsm.suaspress.org/ | CODEN: AJSMAS
Vol. 3, No. 3, 2025 | ISSN 3005-5040 (Print) | ISSN 3005-5059 (Online)