International Journal of Advances in
Soft Computing and Intelligent Systems (IJASCIS)
2025, Vol 04, Issue 02, 142-157
ISSN: 3048-4987

Science Transactions © 2025

Original Paper

REAL TIME SALES PRODUCTIVITY FORECASTING IN RETAIL ECOSYSTEMS THROUGH THE INTEGRATION OF CLOUD COMPUTING AND EDGE AI

Vishnu Dabhade [a*], , Bharat Naiknaware[b] , Akshay Deshpande[c],
Dr. G.Y. Pathrikar College of Computer Science & Information Technology,
MGM University, Chhatrapati Sambhajinagar, India
*Email ids: [a]adeshpande341@gmail.com, [b]bnaiknaware@mgmu.ac.in,
[c]vishnudabhade60@gmail.com

ABSTRACT

Today's retail environment is rapidly changing. How customers buy products has changed, as have methods of distribution. It is more competitive. Traditional ways of assessing predicted sales are becoming outdated. We argue that more advanced predictive tools and models that assess the productivity of different types of sales are needed. There is a need to be more efficient to satisfy corporate profit needs and to be more competitive. Advanced predictive models that look at sales productivity can help assess how effective sales are based on the cost of labor and the available sales inventory. Our research shows that the introduction of Artificial Intelligence to retail sales tools can help predict sales and avoid lost sales due to a lack of inventory. There are a number of areas that present new problems, including selling products that are customer specific and making products that customers want to buy. We present new ways of thinking that will allow the customer to secure their information and allow the sale to take place. This research presents the most advanced hybrid Cloud and Edge Artificial Intelligence system. The model is the most advanced and gives the most competitive edge to sales.

KEYWORDS

Edge AI, Cloud Computing, Hybrid Architecture, Real-Time Sales Productivity Forecasting, Retail Ecosystems, Machine Learning, Deep Learning, Transformer Models, 5G, Dynamic Staffing.

1. INTRODUCTION

The retail sector is being increasingly pressured to adapt to change like never before. The pandemic disrupted the decades long gradual shift to e-commerce and omnichannel commerce by leapfrogging this change by an entire decade. The pandemic highlighted the weaknesses in the traditional models for retail. The old models are gone for good. Retailers have been forced to adapt their business structures if they want to survive the pandemic. Retail structures will need to be repositioned for the constant changes in customer preferences, the shifting economic landscape, and the disruptions in the supply chain. The retail environment is constantly changing and has created the need to adapt faster than quarterly or weekly changes. The traditional view of the retail environment is a disconnected set of linear structures for the suppliers, distributors, retail partners, and customers. The retail environment has now changed to a highly integrated business environment with multiple structures. To succeed in this new environment, retailing will need to be backed by new technologies to support flexible structures, and integrated advanced systems to allow optimal and on demand business operations. Those retailers better positioned to continue to develop technology and systems through the pandemic did have some clear advantages, while those retailers that lagged on their development were left out.

To survive in a very competitive market, retailers need to implement new technologies that can drive sales and improve operational performance, but that also provide advanced analytical insights into customer behavior and improve customer experience at a lower cost.

A fundamental component of this challenge is forecasting. For decades, retail forecasting has been foundational for managing inventory, planning retail supply, and forecasting retail sales and finances. Unfortunately, methods that have been historically used to address forecasting have, for the most part, been incompatible with modern retail. Those historical methods, which are based on simple statistical methodologies such as linear regression or time-series analysis, have been based on an unreliable foundation. The modern retail environment is characterized by unanticipated disruptions, which have made the assumption that past behaviors will always predict future happenings obsolete and invalid. Coupled with that, many forecasting techniques are traditionally dependent on the historical sales and, therefore, tend to be structurally and functionally blind to the current situation. As a result, there is an inability to integrate the numerous streams of unstructured, real-time, and external data concerning the economy, the weather, and even the crime in the retail environment, that are optimal indicators of demand. Traditional forecasting even struggles to address simple concerns, such as the impact of a marketing or pricing decision to offer a demand-elastic service. Traditionally, retailers have been forced to make multi-million-dollar decisions on their retail activities, even when there are substantial and unacceptable variances in the underlying predictions.

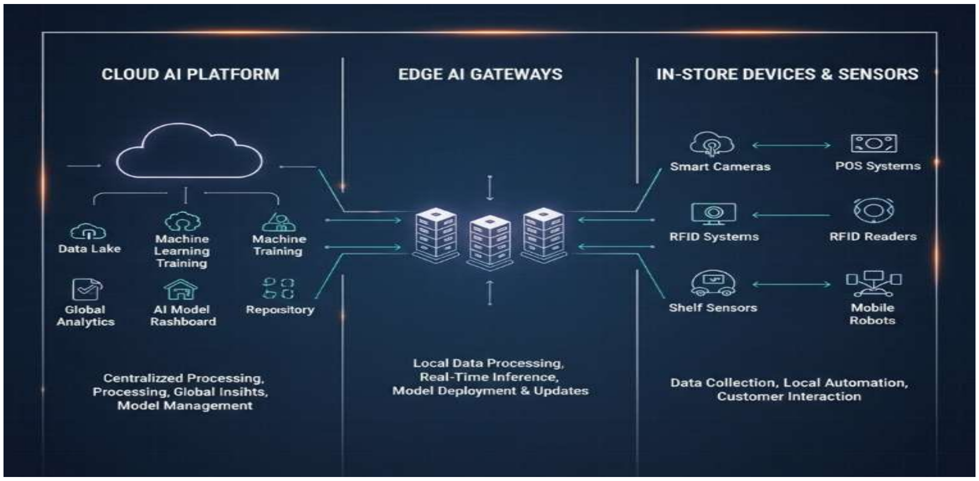

Fig. 1. The Hybrid Cloud-Edge AI Architecture for Retail.

This paper claims that forecasting goals in today's retail systems should change. Goals should no longer focus on predicting future revenue or sales volume. Instead, forecasting sales productivity is imperative. Sales productivity is a more advanced measurement of sales efficiency and effectiveness and is more strategically important. Forecasting sales productivity requires the ability to analyze operational components, such as labor and inventory, and sales as they happen. The main idea of this research is that the combination of centralized Cloud Computing with decentralized Edge AI offers the technology needed to meet the demand for advanced, real-time sales productivity forecasting.

2. THE HYBRID CLOUD-EDGE AI PARADIGM: A CONCEPTUAL AND TECHNICAL FRAMEWORK

Cloud Computing and Edge AI create a continuous learning and optimizing system for the retail sector. A hybrid system integrates both Edge AI and Cloud computing to create a self-learning system. This hybrid system integrates both Edge AI and Cloud computing to create a self-learning system. It creates a self-learning system by combining both Cloud computing and Edge AI. It provides the means to gather the collective knowledge of the whole system and act in real time at the local level of each store. The following explains the components of this system, the life cycle of data and models, various forms of machine learning algorithms, and finally the role of 5G technology as a critical enabler.

2.1. A Symbiotic Architecture: The Data and Model Lifecycle

The hybrid Cloud-Edge AI architecture integrates cloud and edge processing in a closed-loop system. In this black box system, the edge houses the cloud’s intelligence to carry out the functions. The cloud houses the edge’s data to learn and perform functions. The cycles consist of the following six stages:

- Data Ingestion & Pre-Processing (at the Edge): This stage of the system engages the physical retail setting. Different types of devices (e.g. smart cameras, RFID tags, POS terminals, etc.) are all examples of the sensors used to illuminate customer behavior. The data discovered and collected is very large and very noisy. At the Edge, data processing is done by Edge Gateways or on-premises servers. The first step in the processing is to remove the noise, then aggregate the data (e.g. transforming all touch points on a display to a single store traffic count). This preliminary data processing occurs at the Edge, substantially decreasing the data and increasing the quality of the data before it reaches the cloud. [15]
- Local Inference and Action (at the Edge): Edge devices and local servers have lightweight, pre-trained machine learning models deployed and run at their locations. Inference happens in real time on the processed data to implement in-store actions immediately. Examples of this include computer vision models on cameras that populate a queue at checkout and trigger a cashier alert, smart shelves that identify a stockout and alert an associate, and POS systems that use a fraud detection model to identify a suspicious transaction in milliseconds. These actions are low latency and can occur when a store's internet connection is down.
- Selective Data Synchronization (from the Edge to the Cloud): One of the key concepts of this hybrid architecture is to provide the ability to not send all raw data to the cloud. This design makes it possible to transmit only significant aggregated pieces of information and to send only those samples of raw data designed to support the retraining of models under cloud conditions. This methodology, referred to as selective synchronization, provides a significant reduction in bandwidth costs and provides the central system the means to learn and improve.
- Centralized Storage & Global Analytics (in the Cloud): A centralized cloud data lake or data warehouse compiles the structured data streams from thousands of edge locations. This results in an enormous company-wide data set that provides the "single source of truth." With data positioned in the cloud, data scientists and business analysts can run analytical queries of this enterprise data to identify global patterns and regional or store type comparative analyses as well as discover patterns of consumer behavior that would typically be obscured in a single store.
- Model Training & Retraining (in the Cloud): The training and ongoing retraining of sophisticated, complex global forecasting models is conducted in the Cloud and utilizes its massive computational resources. The models use the data of the entire retail network to discover the relationships that sales, promotions, and other things that affect the models may have on the models. This consolidated approach to training the models makes use of the expertise of all the stores and enables the models to be more accurate and have a better coverage and applicability compared to the other models that have been created using the data of only one store.
- Model Optimization & Deployment (from Cloud to Edge): Once a model is newly or better trained in the cloud, you cannot simply deploy the model to edge devices that have limited resources. Several model optimization techniques must take place first. These include quantization, meaning that the

precision of the model's weights is lessened, and pruning, which refers to removing unnecessary connections. The model's size and the model's computational needs are reduced this way. After this process, the model is packaged, typically in a container, and finally deployed to all the edge devices. This process finalizes the learning loop. This process also improves the model's local inference and actions. The model does this by being even more accurate and intelligent. This work is important as the increased accuracy of the models deployed on edge devices is of extremely high importance [17].

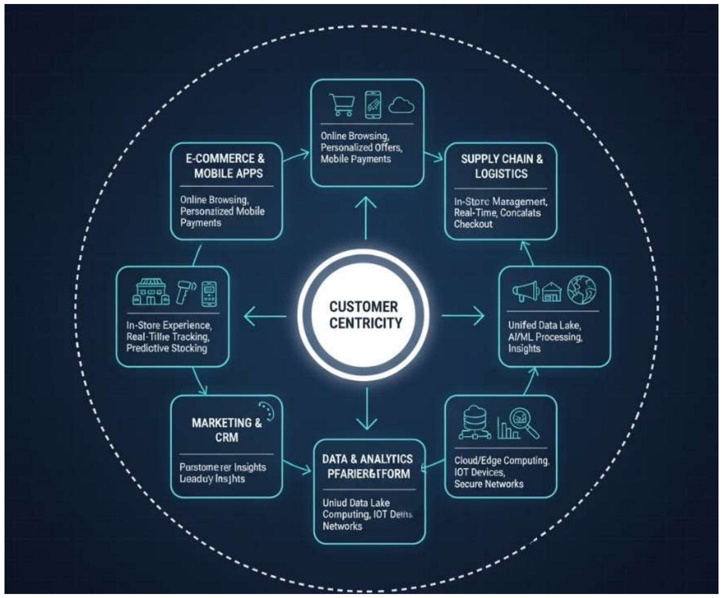

Fig. 2. The Modern Interconnected Retail Ecosystem

The cycle creates a "data network effect." The value to the system increases with the number of stores added, each one acting as an edge node. The edge system of an individual store optimizes that store's operations, and therefore provides a linear value. However, once thousands of stores' data are collected and placed in the cloud, the central AI is able to find subtle, large-scale patterns, and create insights that individual stores cannot. As an example, the central AI could find that a specific marketing campaign is effective for urban stores in one region, but a campaign for rural stores is ineffective. The central AI updates the forecasting in cloud ML and publishes the updated campaigns back to the stores. The circular cycle provides each store greater value than the previous value, and therefore the network is a competitive advantage, given that the raw data is proprietary and the ML operational infrastructure is self-improving.

2.2. Machine Learning Models Across the Continuum

The effectiveness of hybrid architecture relies on the correct selection of machine learning models deployed at each tier. Models trained on the cloud are made to balance the tradeoff between accuracy and complexity. On the other hand, models deployed on the edge are optimized for latency and throughput.

Table 1. Literature review.

Function / Task	Optimal Location	Rationale (Why it belongs there)
Real-Time Stockout Detection on Shelf	Edge	Requires millisecond-level response from video/sensor data to trigger immediate restocking alerts. Cloud latency is too high [22].
Frictionless Checkout / Theft Detection	Edge	Needs instantaneous processing of video feeds to ensure a seamless customer experience and immediate fraud prevention [8].
In-Store Personalized Promotion Delivery	Edge	Triggering a location-based offer requires real-time processing of a customer's presence and movement. The inference happens at the edge [12].
Global Demand Forecasting Model Training	Cloud	Requires massive historical datasets from all stores and significant computational power (GPUs). Not latency-sensitive [9].
Enterprise-Wide Inventory & Supply Chain Analytics	Cloud	Involves aggregating and analyzing data from all stores, warehouses, and suppliers to make strategic, long-term decisions [23].
Long-Term Customer Behavior Analysis (CLV)	Cloud	Requires analyzing years of transactional and behavioral data across all channels to build comprehensive customer profiles [7].
Archival of Transactional Data	Cloud	Offers cost-effective, durable, and scalable long-term storage for compliance, auditing, and future large-scale analysis [11].

- **Cloud-Based Training Models**

Retailers can access the models that provide them the best possible accuracy for their forecasts thanks to the virtually unlimited computational resources of the cloud [24-27]:

Modern Time-Series Models: Deep learning models likely to meet the complex challenges of the retail context are in most situations a better option than traditional statistical models, like ARIMA, which are useful in stable contexts. When it comes to retail, sales data often is inherently sequential. Recurrent Neural Networks (RNN) of which Long Short-Term Memory (LSTM) networks are an advanced variant and are capable of capturing long-term dependencies in data. CNN-LSTM is a hybrid model that combines CNN with LSTM. CNNs are excellent at spatial data and feature extraction. CNN-LSTM models are excellent at capturing dependencies with relation to time (temporal). CNN-LSTM models also serve to lower the error rates of forecasting compared to standard forecasting models.

More recently, transformer-based architectures have advanced models beyond their initial use in natural language processing, achieving even better performance for time series forecasting.

Due to the transformer's self-attention mechanism, models can process a full data sequence with equal consideration to the range of all preceding data points. For this reason, transformer models can identify peak hour demand, sales, and complex relationships between a variety of different streams such as price and promotions. Advanced variants of the architecture, like the temporal fusion transformer (TFT), outperform classical models and methods used in time series forecasting retail demand.

These methods aggregate the forecasts of several distinct models to achieve a more accurate and resilient prediction. Gradient boosting methods, particularly those modeled as XGBoost and LightGBM, have demonstrated high performance in retail sales forecasting and are extensively applied in competitions. Among ensemble methods, Random Forests achieve a good balance of accuracy with high dimensional data due to its structure which prevents overfitting.

- **Big Models, Small Devices**

Edge devices have a limited budget for hardware resources like memory and energy consumption. Because of this, large models trained in the cloud must be adapted to run efficient in the edge. This relies on the use of smart architectures. Some neural architectures that are designed from the ground up for efficiency. MobileNet for instance, employs depthwise separable convolutions and SqueezeNet features are formulated to retain their accuracy with a reduced number of parameters. These small, efficient models fit the budget of mobile phones and smart cameras.

- **Model Compression Techniques**

There are a number of methods that can shrink models for better efficiency on edge computing devices. For example:

- Pruning: This involves removing weights from the neural network that are unimportant.
- Knowledge Distillation: This involves teaching a smaller neural network, called the student model, to mimic the neural network, called the teacher model. The teacher model has undergone a lot of training. Using this method allows for the knowledge of a more advanced and complex model to be transferred to a form that is far more efficient and doesn't rely on the original training data.
- Quantization: This changes the numerical precision of the weights in the model. An example would be changing the 32-bit floating point numbers to 8-bit integers. This method is highly compressive and can be used to speed up calculations for certain types

of hardware.

- Finally, TinyML focuses on using less powerful microcontrollers to run machine learning models. Compression frameworks like TensorFlow Lite for Microcontrollers, provide the ability to run machine learning models of very small sizes and occupy as little as kilobytes of memory to efficiently be used in the small Internet of Things.

2.3. The Role of 5G as a Critical Enabler

The introduction of 5G wireless technology worldwide boosts many elements of the hybrid Cloud-Edge AI architecture. It solves several communication problems that limit the full potential of this architecture. Compared to 4G/LTE, 5G is a much more distinct technology for retail in several ways:

- 5G has considerably higher bandwidth than previous generations of wireless technology. Therefore, it can more easily manage and transmit large volumes of Edge data to the Cloud. This is extremely useful for retail scenarios where the Cloud needs to process high-definition video, or where large data batches need to be transferred to the Cloud for the initial data set in order to build or retrain AI models.
- 5G drastically reduces the time it takes for a signal to be sent and received (latency). This enables near-simultaneous interaction between multiple Edge devices in retail and the Cloud (local Edge Cloud, public Cloud, etc.). This is particularly important for time-sensitive retail use cases, such as coordinating multiple autonomous robots operating in the retail space, or AR shopping applications.
- 5G was designed to allow a larger number of devices to connect to a single cellular tower than in previous wireless generations. This allows retail to create Smart Stores that use thousands of IoT devices (sensors, cameras, etc.) to capture real-time data to create a comprehensive digital representation of the Smart Store.

3. QUANTITATIVE ANALYSIS: MEASURING THE LEAP IN ACCURACY AND ROI

The benefits of a hybrid Cloud-Edge AI architecture are becoming increasingly clear and go beyond theory, thanks to the increased quantity of empirical evidence provided by industry and academia. This section consolidates empirical evidence for a clear assessment of the technology’s impact. This analysis has two components. The first of these components assesses the accuracy improvements in forecasts by AI-enhanced models versus non-AI models. The second of these components addresses the returns related to the improvements in the accuracy of forecasts across main business operations [29].

3.1. Improvements in Forecast Accuracy

An important, positive effect gained through the use of AI and ML driven forecasting is a measurable increase in accuracy. Many classical models are often unable to capture accurate results in the face of instability and complexity of real-world data. AI is naturally a better candidate to process large and structurally complicated data sets, as it is better at locating patterns and structures to make useful, accurate predictions.

The extent of this advancement is significant. The magnitude of this improvement is noteworthy. A McKinsey study reveals that utilizing AI-driven forecasting for supply chain management can lead to a 20%-50% error reduction from traditional forecasting methods that rely on spreadsheets. More detailed studies involving the retail sector support this finding. One of these studies was a quantitative experimental one that compared the error rates of retail MSMEs that used AI-based forecasting (Random Forest) with those that used traditional forecasting (Moving Average, Exponential Smoothing). The Random Forest Model yielded a Mean Absolute Percentage Error (MAPE) of 8.5%, while the traditional forecasting methods were reported to have a MAPE of 15.2%, Therefore, error rates of the AI-based forecasting model were almost two times better than those of the traditional methods. This advantage of AI-based forecasting is also seen in retail prediction models of large-scale retail operations. One of the published case studies for one of the large retail chains, that is present in multiple countries, showed that the AI-based Demand Forecasting System that was implemented by them improved the forecast error rates (when measured by Mean Absolute Error) from 37% to 25.6% [30]. The improvement of forecast accuracy was 11.6% in an industry that has a low margin for many concurrent transactions. Also, the retailers that use Amazon Forecast, a managed ML service, have shown an improvement of 10% in the Weighted Absolute Percentage Error (WAPE), compared to other manual methods of forecasting that they used. AI models perform best in forecasting, when the challenge of forecasting is the greatest.

Research demonstrates that machine learning methods are able to model multi-dimensional cycles (e.g., holidays, back-to-school) much better than traditional techniques. For products that exhibit intricate seasonality, machine learning methods can forecast demand 25% to 40% better than traditional methods. When the market is extremely volatile, as it was during the COVID-19 global pandemic, AI-based approaches produced results that were 38.6% better than traditional approaches based on statistical methods. ML techniques are also preferred even for short-term forecasting in the range of 1-3 months, where traditional statistical methods are considered better. In the conventional forecasting paradigm, MAPE in the range of 10-25% is accepted for a 1–3-month forecast, while ML techniques are able to provide MAPE in the range of 5-15%. The breadth of the operational gains and the resultant financial benefits is a result of improvement in forecast accuracy.

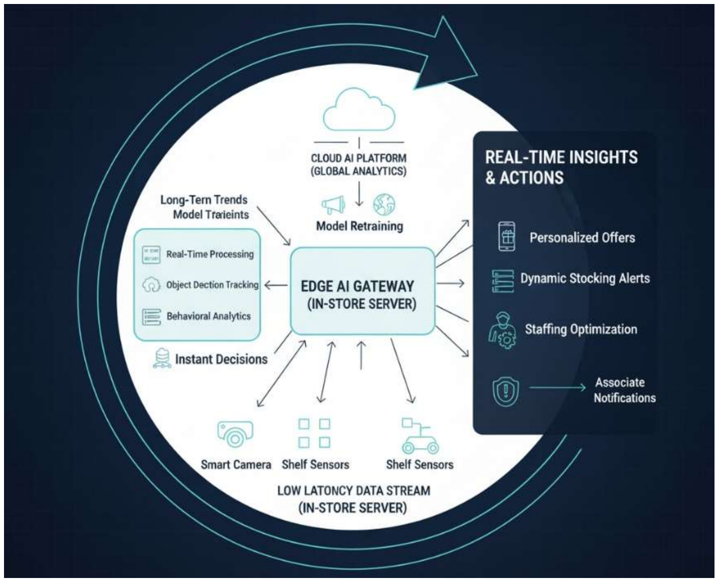

Fig. 3. Real-Time In-Store Analytics via Edge AI

3.2. Tangible Return on Investment (ROI)

Accurate forecasting is useful as a tool for better business decision making and getting a positive return on investment, but it is not valuable in and of itself. The returns on investment from implementing a hybrid Cloud-Edge AI system are from a combination of many interconnected positive effects across the full retail value chain stretching from the many operational efficiencies it creates to increased revenue directly resulting from its implementation.

- **Improved Forecasting, Enhanced Efficiency, Lower Costs**

Perhaps the largest impact accurate forecasting has is in inventory management. Demand forecasting helps retailers achieve optimal inventory levels and thus significant cost savings. AI forecasts reduce inventories 20% to 30% across the industry, and some organizations achieve reductions of 50% for certain categories. Less inventory is easier and cheaper to manage, reducing costs, waste, and improving cash flow. An excellent example of this is AI implementation in Walmart inventory. Inventory surplus was reduced by 15%, a significant savings for a retailer of Walmart's size. AI accomplishes this and inventory savings by reducing costs with improved efficiency and reduced need for labor. In a notable case, implementing

Amazon Forecast resulted in a direct cost savings of 16 hours of labor that was previously used for manual forecasting tasks. Dynamic staffing uses cost savings from reduced overstaffing to pay for the cost savings from reduced understaffing.

- **Revenue Growth and Profitability**

The complement of the reduction of overstock is prevention of stockouts. Stockouts generate long-term customer dissatisfaction and the lost sales are directly attributable to stockouts. A McKinsey analysis showed that a 20-50% reduction in forecasted demand led to a 65% reduction in lost sales and the stockouts of unaffordable products. This also means a recapture of sales that would be lost to other companies. One estimate in line with the analysis forecasted that sales would increase by 11.8% if a retailer provided the correct amount of stock, as determined by a more accurate demand forecast. A reduction of out-of-stock events by 30% seen by Walmart illustrates that the prevention of lost sales is real. In addition to that, a hybrid architecture unlocks new revenue streams and improves the profitability of other existing streams. Retailers can generate 20% more revenue and 22% more profit through the implementation of a dynamic pricing system compared to the traditional pricing system. Furthermore, the use of AI in designing and implementing personalized recommendations can result in a 288% higher sales conversion rate in e-commerce and 25% more revenue, with the return from these investments occurring almost immediately.

The 2025 Google Cloud study states that 74% of companies see positive ROI from generative AI within the first year [33]. 68% of companies in the retail and consumer packaged goods sector report seeing value with generative AI from customer experience improvements, up from 57% in the previous year. Dell Technologies says edge deployments in retail save 71% time achieving ROI. While the business case is strong, a complete picture must be seen. ROI comes from a multitude of interconnected improvements. The more precise forecasting and prediction is, the greater the first-order effects. From the forecasting and predictions, costs of holding inventory decrease while revenues increase from eliminated stockouts. These also create second order effects, like customer satisfaction which improves from having better inventory, which in turn improves customer loyalty and customer lifetime value. At the same time, the automated forecasting and inventory processes create an opportunity for employees to focus on customer service instead, which in turn, creates an increase in conversion. Focusing on a single benefit, like the reduction of carrying costs, from a narrow perspective, gets a much smaller value of the hybrid cloud-edge AI implementation. The improvements in the table 2 are real-world examples taken from the retail industry [35].

Table 2. Real-world examples taken from the retail industry.

Retailer / Case Study	Technology Implemented	Key Metric	Quantitative Improvement
Walmart	AI-driven Inventory Management (Edge AI with Computer Vision & Sensors)	Out-of-Stock Events	30% Reduction
Walmart	Advanced AI Systems for Demand Forecasting	Surplus Stock	15% Reduction
Large International Retail Chain	AI Demand Forecasting System (Ensemble ML Models)	Forecast Error Rate (Mean Absolute Error)	Reduced from 37% to 25.6% (11.6 percentage point improvement)
Amazon Forecast Customer	Managed ML Service for Demand Forecasting	Forecast Accuracy (Weighted Absolute Percentage Error)	10% Average Improvement over manual methods
General AI Adoption (McKinsey Study)	AI-driven Supply Chain Forecasting	Reduction in Lost Sales due to Stockouts	Up to 65%
Retail MSMEs Study	AI Forecasting (Random Forest) vs. Traditional	Forecast Accuracy (Mean Absolute Percentage Error)	8.5% (AI) vs. 15.2% (Traditional)
General AI Adoption (McKinsey Study)	AI-driven Supply Chain Forecasting	Forecast Error Reduction	20% - 50%

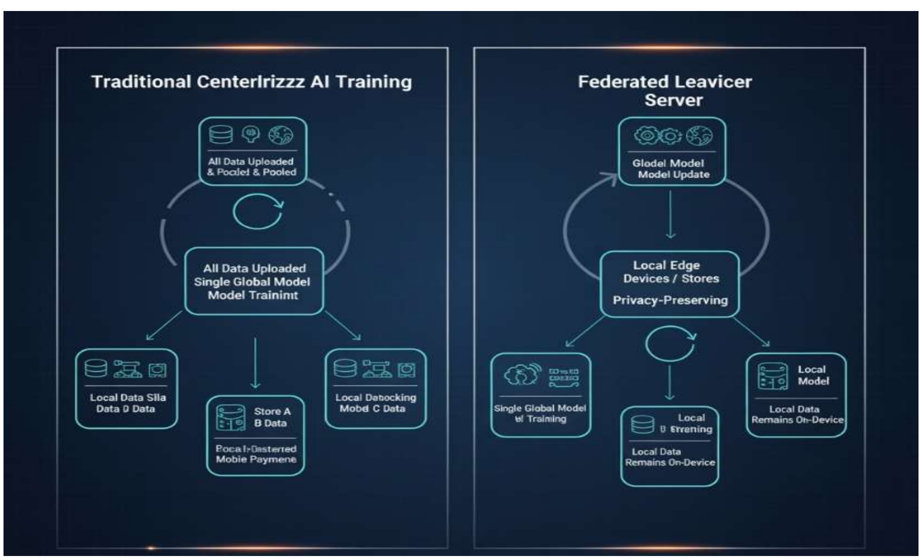

Fig. 4. Federated Learning vs. Traditional Centralized AI Training.

4. Conclusions

The complexity and volatility of today's retail environment have made many standard forecasting tools inadequate and even strategically harmful. This paper addresses this challenge and advocates for an evolutionary shift, not an incremental one, from forecasting sales volume to forecasting sales productivity. This shift would more effectively incorporate forecasting and predictive analytics with the core business strategies of operational efficiency and profitability. This research identifies the integration of the centralized authority of Cloud Computing with the distributed intelligence of Edge AI as the primary technological enabler of this shift. From the analysis and evidence this paper has provided, a specific and strong recommendation can be made. For retail business' in the 21st century, adopting a Cloud-Edge AI hybrid model in order to forecast sales productivity in a real-time context should not be considered an additional and progressive technology investment, but the most essential and strategic technology investment in order to survive and have a competitive advantage in the industry. The competitive barriers of the past, including physical presence, a robust supply chain, and customer loyalty, have all been undermined by new competitors who leverage the data to refine their business models. Retailers have a great opportunity to create a sustainable advantage by how sophisticated their data feedback loops are and how fast they can act on data. Now the question is not why the retail sector should undergo this transformation, but how rapidly the transformation should take place.

Authors Contributions:

All authors have equally contributed.

References

1. Alicke, K., Dilda, V., Görner, S., Mori, L., Rebuffel, P., Reiter, S., & Samek, R. (2021). Succeeding in the AI supply-chain revolution. McKinsey & Company. https://www.mckinsey.com/capabilities/operations/our-insights/succeeding-in-the-ai-supply-chain-revolution
2. Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, B., Maddix, D., Turkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., & Januschowski, T. (2022). Deep learning for time series forecasting: Tutorial and literature survey. ACM Computing Surveys, 55(6), Article 109. https://doi.org/10.1145/3533382
3. Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012). Fog computing and its role in the internet of things. Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, 13–16. https://doi.org/10.1145/2342509.2342513
4. Byers, C. C. (2017). Architectural imperatives for fog computing: Use cases, requirements, and architectural techniques for fog-enabled IoT networks. IEEE Communications Magazine, 55(8), 14–20. https://doi.org/10.1109/MCOM.2017.1600885
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
6. Damacharla, P. (2025). Temporal fusion transformer for multi-horizon probabilistic forecasting of weekly retail sales. arXiv preprint. https://arxiv.org/abs/2511.00552
7. Dwork, C. (2008). Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), Theory and Applications of Models of Computation (pp. 1–19). Springer. https://doi.org/10.1007/978-3-540-79228-4_1

8. Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. International Journal of Forecasting, 35(1), 1–12. https://doi.org/10.1016/j.ijforecast.2018.10.002
9. Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media. https://doi.org/10.1007/978-1-4842-8843-2
10. Han, S., Mao, H., & Dally, W. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1510.00149
11. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90
12. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1314–1324. https://doi.org/10.1109/ICCV.2019.00140
13. Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv preprint. https://doi.org/10.48550/arXiv.1602.07360
14. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2704–2713. https://doi.org/10.1109/CVPR.2018.00286
15. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.08357
16. Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012
17. Lin, J., Chen, W.-M., Lin, Y., Cohn, J., Gan, C., & Han, S. (2020). MCUNet: Tiny deep learning on IoT devices. Advances in Neural Information Processing Systems, 33, 11711–11722. https://doi.org/10.48550/arXiv.2007.10319
18. Mach, P., & Becvar, Z. (2017). Mobile edge computing: A survey on architecture and computation offloading. IEEE Communications Surveys & Tutorials, 19(3), 1628–1656. https://doi.org/10.1109/COMST.2017.2682318
19. McKinsey & Company. (2021). Artificial intelligence in supply chain management. McKinsey Global Institute Report. https://www.mckinsey.com/capabilities/operations/our-insights/ai-powered-supply-chain-management
20. Pan, J., & McElhannon, J. (2017). Future edge cloud and edge computing for internet of things applications. IEEE Internet of Things Journal, 5(1), 439–449. https://doi.org/10.1109/JIOT.2017.2767608
21. Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), 2627–2633. https://doi.org/10.24963/ijcai.2017/366
22. Rost, P., Mannweiler, C., Michalopoulos, D. S., Sartori, C., Sciancalepore, V., Sastry, N., Holland, O., Tayade, A. A., Maeder, A., & Doppler, K. (2017). Network slicing to enable scalability and flexibility in 5G mobile networks. IEEE Communications Magazine, 55(5), 72–79. https://doi.org/10.1109/MCOM.2017.1600945
23. Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting, 36(3), 1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001
24. Santos, J., Wauters, T., Volckaert, B., & De Turck, F. (2018). Fog computing: Enabling the management and orchestration of smart city applications in 5G networks. Entropy, 20(1), 4. https://doi.org/10.3390/e20010004
25. Schulz, P., Matthe, M., Klessig, H., Simsek, M., Fettweis, G., Ansari, J., Ashraf, S. A., Almeroth, B., Voigt, J., Riedel, I., Puschmann, A., & Wunder, G. (2017). Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture. IEEE Communications Magazine, 55(2), 70–78. https://doi.org/10.1109/MCOM.2017.1600435CM

26. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198
27. Shi, Y., Yang, Q., & Tang, K. (2022). Evaluating the effectiveness of time series transformers for demand forecasting in retail. Mathematics, 12(17), 2728. https://doi.org/10.3390/math12172728
28. Theodoridis, G., & Tsadiras, A. (2025). Retail demand forecasting: A comparative analysis of deep neural networks and the proposal of LSTMixer, a linear model extension. Information, 16(7), 596. https://doi.org/10.3390/info16070596
29. Vanjare, G. D. (2025). Edge computing and 5G networks: Catalysts for next-generation smart retail transformation. Journal of Advanced Artificial Intelligence, 2(1), 18–23. https://jaaionline.phdfocus.com/archives/volume2/number1/edge-computing-and-5g-networks-catalysts-for-next-generation-smart-retail-transformation/
30. Wanchoo, K. (2019). Retail demand forecasting: A comparison between deep neural network and gradient boosting method for univariate time series. 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 1–5. https://doi.org/10.1109/I2CT45611.2019.9033651
31. Wang, S., Cao, J., & Yu, P. (2020). Deep learning for spatio-temporal data mining: A survey. IEEE Transactions on Knowledge and Data Engineering, 34(8), 3681–3700. https://doi.org/10.1109/TKDE.2020.3025580
32. Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2020). Detectron2: A PyTorch-based modular object detection library. arXiv preprint. https://doi.org/10.48550/arXiv.2006.06683
33. Xiao, Y., & Krunz, M. (2017). QoE and power efficiency tradeoff for fog computing networks with fog node cooperation. 2017 IEEE Conference on Computer Communications (INFOCOM), 1–9. https://doi.org/10.1109/INFOCOM.2017.8057196
34. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), Article 12. https://doi.org/10.1145/3298981
35. Zreikat, A. I., AlArnaout, Z., Abadleh, A., Elbasi, E., & Mostafa, N. (2025). The integration of the Internet of Things (IoT) applications into 5G networks: A review and analysis. Computers, 14(7), 250. https://doi.org/10.3390/computers14070250