AI Powered Inventory and Stock Management System with Personalized AI Assistance

Komala D[1], Rohit Kumar[2], S Preetham[3], Shivam Kumar Singh[4], Sanjay Naik[5]
[1,2,3,4,5]Department of CSE, Don Bosco Institute of Technology,
Bengaluru, Karnataka, India
komala.dbit@dbit.co.in, rohit0908kr@gmail.com, preetham92.official@gmail.com,
shivamkrsingh1911@gmail.com,

Abstract
This review explores the integration of Artificial Intelligence (AI), with a focus on machine learning (ML) and Retrieval-Augmented Generation (RAG) techniques, in modern inventory management systems designed specifically for small and medium enterprises (SMEs). It synthesizes recent advancements in predictive analytics, deep learning models for demand forecasting, automated invoice data extraction, and real-time inventory monitoring through computer vision and edge AI technologies. Special attention is given to decision-support systems that leverage AI to enhance restocking precision, minimize stockouts and overstocks, and provide personalized business insights via natural language interfaces. The paper further examines the implementation of ML classifiers for inventory categorization, the use of optical character recognition (OCR) and natural language processing (NLP) for automated financial documentation, and the application of nature-inspired optimization algorithms for cost-effective inventory control. By integrating these methodologies within a RAG-based AI assistant framework, the review highlights potential solutions that empower small business owners with actionable, data-driven guidance to improve inventory turnover, optimize holding costs, and ultimately boost profitability without requiring advanced technical expertise.

Keywords
Inventory Management, Machine Learning, RAG, ABC Classification, Forecasting, Computer Vision, NLP, Invoice Automation

Introduction

Inventory management plays a crucial role in ensuring the smooth operation of supply chains and significantly impacts overall business profitability. Efficient inventory control helps businesses maintain the delicate balance between meeting customer demand and minimizing costs associated with holding excess stock or facing stockouts. For large enterprises, sophisticated Enterprise Resource Planning (ERP) systems provide integrated solutions that automate and optimize inventory processes using historical data and advanced analytics. However, small and medium-sized enterprises (SMEs) often face considerable challenges in adopting such complex systems due to limited budgets, technical expertise, and infrastructure.

With the rise of artificial intelligence (AI) and machine learning (ML), new opportunities have emerged to bridge this gap. AI-driven inventory management solutions can empower small business owners by automating forecasting, optimizing stock levels, and providing actionable insights without requiring deep technical knowledge. Among these advancements, Retrieval-Augmented Generation (RAG) models present a novel approach that combines the power of large language models with domain-specific data retrieval, enabling personalized and context-aware assistance tailored to each business's unique inventory and sales data.

This review aims to explore the state-of-the-art AI techniques applied to inventory management, focusing on how RAG models integrated with ML forecasting, computer vision, and natural language processing can revolutionize the inventory workflows of SMEs. By synthesizing research on demand forecasting, invoice automation, real-time stock monitoring, and optimization algorithms, the review highlights practical implementations that make intelligent decision support more accessible and actionable for small business owners. This democratization of AI tools promises to enhance operational efficiency, reduce inventory-related costs, and ultimately improve the profitability and competitiveness of smaller retail businesses in today's dynamic market environment.

Literature Review

This review synthesizes findings from seven peer-reviewed research papers that collectively form the foundation for TRIACT's AI-powered inventory management system. Each paper addresses a distinct yet complementary aspect of inventory control, demand forecasting, intelligent assistance, or retail automation, forming a comprehensive knowledge base for system design.

Demand Forecasting and Sales Prediction

Suresh and Suresh [1] provide a broad comparison of machine learning and deep learning models for retail sales forecasting. Their study highlights the superiority of LSTM and hybrid deep learning models over traditional approaches such as ARIMA. These findings support TRIACT's inclusion of time-series modelling for predictive restocking and trend forecasting in small retail environments.

Inventory Optimization Using ML and Metaheuristics

Ladva et al. [2] investigate both conventional ML classifiers (e.g., SVM, Decision Tree) and nature-inspired algorithms like PSO and NSGA-II for inventory classification and multi-objective optimization. Their study concludes that Random Forest achieves the highest classification accuracy (93%), and that metaheuristics can effectively balance trade-offs between inventory cost, space, and availability—core objectives in TRIACT's restocking logic.

Automated Invoice Processing

Alkhaled and Ng [3] present a deep learning approach to invoice processing using CNNs and Faster R-CNN paired with OCR tools. The system accurately extracts key metadata from scanned documents, which aligns directly with TRIACT's goal to automate invoice generation and archiving, improving back-office efficiency for small businesses.

Real-Time Shelf Management with Edge AI

Savit and Damor [4] propose a novel shelf management system using YOLOv5 and MobileNetV3 to detect voids and misplaced items in real time. Their use of edge-based computer vision proves practical for shop-floor deployment. TRIACT builds on this model to offer live shelf monitoring, which ensures visual confirmation of stock levels beyond POS data.

Classification and Forecasting with Lightweight Models

Bailkar et al. [5] present a comparative analysis of five machine learning models applied to inventory ABC classification. Their findings reinforce the use of Random Forest for dynamic categorization and the integration of forecasting models into intelligent retail systems. This supports TRIACT's approach to prioritizing inventory based on value, frequency, and profitability.

Speech Interface in Warehouse Systems

Song [6] explores the role of automatic speech recognition (ASR) in warehouse management. The study shows that voice-enabled systems improve human-computer interaction in noisy environments, suggesting future

integration potential for TRIACT's RAG assistant to support voice-based inventory queries—especially useful in busy shop environments.

RAG Models for Intelligent Assistants

Hammane et al. [7] introduce a self-evaluating RAG framework in a clinical reasoning context, showing how RAG models can deliver context-rich, accurate responses from domain-specific knowledge bases. Although applied to medicine, the underlying architecture validates TRIACT's conversational AI strategy—enabling small business owners to interact naturally with their sales and inventory data through a domain-trained assistant.

Comparative Analysis

This section provides a comparative summary of the machine learning techniques, AI components, and optimization strategies reviewed for the TRIACT system. The analysis focuses on their application in demand forecasting, invoice processing, inventory classification, and real-time stock monitoring. The table and chart below highlight key performance indicators that validate the proposed system architecture.

ML Classifier Performance in ABC Classification

Based on the literature, five machine learning algorithms were evaluated for their accuracy in classifying inventory items into ABC categories. These models were trained using features like item cost, shelf life, demand variability, and historical sales data.

Model	Accuracy (%)	Strengths	Limitations
Random Forest	**93**	High accuracy, handles non-linearity	Computationally heavier
Decision Tree	87	Simple, interpretable	Prone to over fitting
Naïve Bayes	74	Fast, handles noisy data	Assumes feature independence
K-Nearest Neighbors	54	Easy to implement	Poor with large dataset
Support Vector Machine	40	Effective in high-dimensional spaces	Sensitive to parameter tuning

Functional Comparison of Core AI Modules

Module	Technology	Purpose in TRIACT	Benefit
Demand Forecasting	LSTM, ARIMA, RF	Predict sales trends and restocking needs	Proactive ordering, reduced stockouts
ABC Classification	RF, DT, SVM	Prioritize inventory based on value and turnover	Customized stock policies
Invoice Automation	CNN, Faster R-CNN, OCR	Extract structured data from unstructured PDFs/images	Faster invoicing, searchable records
Optimization	PSO, NSGA-II, MOPSO	Minimize costs and balance inventory levels	Scenario modelling and cost-effective ordering
Conversational Assistant	RAG, NLP, Vector DB	Answer shop-specific questions in natural language	Personalized insights without dashboards

Key Findings Summary

- Accuracy & Flexibility: ML models like Random Forest offer high accuracy and adaptability in classifying inventory for smarter control.

- Automation Potential: CNN-based invoice processing and computer vision for shelves reduce manual workload.
- Smart Decision Support: RAG assistants enhance accessibility by allowing non-technical users to query and receive contextualized, data-driven answers.
- Optimization Impact: Nature-inspired algorithms help businesses make cost-effective decisions while considering multiple operational constraints.

Results/Discussion

The findings from this review demonstrate that integrating modern AI techniques—particularly machine learning (ML), deep learning, computer vision, and Retrieval-Augmented Generation (RAG)—offers substantial benefits for improving inventory management in small business environments.

Interpretation in the Context of Literature

Consistent with previous studies, Random Forest emerged as the most reliable ML classifier for inventory classification tasks, achieving the highest accuracy (93%) in ABC classification. This aligns with research emphasizing ensemble models' superiority in handling complex, multi-dimensional inventory data. Deep learning techniques such as LSTM also showed clear advantages over traditional statistical models (e.g., ARIMA) in forecasting variable retail demand, reinforcing the shift toward neural architectures in time-series prediction.

Computer vision and OCR models such as YOLOv5 and Faster R-CNN have shown great promise in real-time shelf monitoring and invoice data extraction, validating earlier research on their effectiveness in structured and unstructured retail tasks. The inclusion of RAG-based assistants marks a notable evolution beyond static dashboards, as they provide personalized, queryable, and context-aware insights by combining retrieval with generative language models—an approach not yet widely applied in SME inventory systems.

Moreover, nature-inspired algorithms like Particle Swarm Optimization (PSO) and NSGA-II allow TRIACT to support dynamic, multi-objective optimization. This capability helps balance trade-offs between cost, storage capacity, and demand fluctuation—paralleling successful implementations in larger supply chain systems.

Implications

The results strongly suggest that the TRIACT system can provide a practical, scalable, and user-friendly platform for inventory management in small businesses. By reducing reliance on manual calculations and enabling intelligent, data-driven decisions, the system supports:

- More accurate restocking schedules
- Lower operational and holding costs
- Higher inventory turnover
- Better responsiveness to changing market conditions

Despite these benefits, there are limitations to consider:

- Data Availability: Many small businesses may lack digitized historical data, limiting the initial performance of ML models.
- Model Cold Start: RAG and ML-based systems rely on adequate training data; in early deployment stages, suggestions may lack depth or accuracy.
- Hardware Constraints: Real-time computer vision using edge AI may be challenging for businesses with limited computing resources or outdated camera setups.
- User Trust: Business owners unfamiliar with AI may initially mistrust automated suggestions unless transparency and explainability are emphasized.

Future studies should explore:

- Lightweight, hybrid AI models suitable for low-resource environments
- Pre-trained RAG assistants with generic business logic to mitigate cold-start problems
- Integration of voice-command systems (as supported in one study) to improve accessibility
- Continuous learning systems that adapt as new inventory and sales data are recorded
- Human-in-the-loop feedback systems for refining AI recommendations over time

Conclusion

This review highlights the transformative potential of integrating AI-driven technologies—particularly machine learning, deep learning, RAG-based assistants, and nature-inspired optimization algorithms—into small business inventory management systems. The TRIACT framework demonstrates how these tools can be practically deployed to deliver personalized, intelligent decision support to non-technical users.

Key findings reveal that Random Forest classifiers offer highly accurate inventory classification, deep learning models such as LSTM improve forecasting precision, and RAG assistants enhance data accessibility through conversational interaction. Additionally, the use of OCR, computer vision, and metaheuristic optimization strategies significantly improves real-time stock visibility and cost efficiency.

Although the application is centred on retail and operational use cases, the implications of these advancements extend beyond commerce. Reliable inventory systems in small businesses—especially those in rural, underserved, or health-related retail sectors (e.g., pharmacies, clinics, or medical supply stores)—can lead to better resource availability, reduced waste, and improved service delivery. This contributes indirectly to public well-being and localized economic health, aligning with broader objectives in medical logistics and community care.

Recommendations & Applications:

- Deploy TRIACT in small medical supply stores or pharmacies to improve stock availability of critical health items.
- Integrate AI-assisted forecasting in community clinics to manage perishable or regulated medical inventory.
- Extend RAG-based assistants to support medical shop owners with alerts and compliance tracking.
- Develop low-resource variants of TRIACT for rural healthcare retailers with limited infrastructure.

References

[1] S. B. Suresh and M. Suresh, “A Comprehensive Analysis of Retail Sales Forecasting Using Machine Learning and Deep Learning Methods,” in *Proc. Int. Conf. on Data Science and Network Security (ICDSNS)*, 2023.

[2] V. Ladva, C. Vaghela, M. Shukla, T. Kshatriya, and N. H. Dholakia, “An Analysis on Various Machine Learning Algorithms & Nature Inspired Algorithms for Modern Inventory Management,” in *Proc. 14th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)*, Delhi, India, Jul. 2023.

[3] L. Alkhaled and Y. F. Ng, “Automated Invoice Processing System,” in *Proc. IEEE Int. Conf. on Industrial Engineering and Engineering Management (IEEM)*, 2023.

[4] A Savit and A. Damor, “Revolutionizing Retail Stores with Computer Vision and Edge AI: A Novel Shelf Management System,” in *Proc. 2nd Int. Conf. on Applied Artificial Intelligence and Computing (ICAAIC)*, 2023.

[5] Shreyas Bailkar, Kunal Shenoy, Amogh Bedekar, Sandip Bankar, Pranav More “Smart Inventory Optimization using Machine Learning Algorithms,” IDCIoT, IEEE, 2024.

[6] X. Song, “Optimizing the Human-Computer Interaction Interface of Warehouse Management Systems Using Automatic Speech Recognition Technology,” in *Proc. 4th Int. Conf. on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, 2024.

[7] Zakaria Hammane, Fatima-Ezzahraa Ben-Bouazza, Abdelhadi Fennan "SelfRewardRAG: Enhancing Medical Reasoning with Retrieval-Augmented Generation and Self-Evaluation in Large Language Models" in Intelligent Systems and Computer Vision (ISCV) IEEE, 2024

[8] C. Deng and Y. Liu, "A Deep Learning-Based Inventory Management and Demand Prediction optimization method for anomaly detection," Wireless Communications and Mobile Computing, vol. 2021, pp. 1–14, Oct. 2021, doi:10.1155/2021/9969357.

[9] S. Punia and S. Shankar, "Predictive analytics for demand forecasting: A deep learning-based decision support system," Knowledge-Based Systems, vol. 258, p. 109956, Oct. 2022, doi:10.1016/j.knosys.2022.109956.

[10] Preil, Deniz, and Michael Krapp. "Artificial intelligence-based inventory management: a Monte Carlo tree search approach." Annals of Operations Research (2022): 1-25.

[11] Teerasoponpong, Siravat, and ApichatSopadang. "Decision support system for adaptive sourcing and inventory management in small-and medium-sized enterprises." Robotics and Computer-Integrated Manufacturing 73 (2022): 102226.

[12] Mafalda Correia Piçarra Gonçalves "The possibility of implementing intelligent systems and the respective impact of Artificial Intelligence on inventory management and warehousing"iscteMarch, 2022

[13] Sun, Tienlan, Anaiy Somalwar, and Hinson Chan. "Multimodal Retrieval Augmented Generation Evaluation Benchmark." In 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring), pp.1-5. IEEE, 2024.

[14] Arslan, Muhammad, and Christophe Cruz. "Business-RAG: Information Extraction for Business Insights." In 21st International Conference on Smart Business Technologies, pp. 88-94. SCITEPRESS-Science and Technology Publications, 2024.

[15] Analytics, Data, and Dishant Sukhwal. "Retrieval Augmented Generation: An Evaluation of RAG-based Chatbot for Customer Support." Retrieval Augmented Generation: An Evaluation of RAG- based Chatbot for Customer Support (2024).

[16] Thiyagarajan, Kailash. "Multimodal RAG for enhanced information retrieval and generation in retail." Proceedings of the International Conference on Visual Analytics and Data Visualization (ICVADV-2025), IEEE (2025): 102–106.

Acknowledgments

We would also like to thank the authors of the referenced research papers, whose pioneering work laid the foundation for this analysis. Special acknowledgment goes to the contributors of IEEE Xplore and other academic databases for enabling access to critical scientific literature.

Copyright & License:

© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.