# Optimal Inventory Management for New Tools Using Model-Based Deep Reinforcement Learning: A Solution for Short Product Life Cycles

**Sonal Modh Bhardwaj[1], Shimpy Harbhajanka Goyal[2], Anusha Jain[3],Priyanka Dhasal[4]**

[1] Assistant Professor,Department of Computer Science and Engineering, Medi-Caps University, Indore, Madhya Pradesh, India , Email: sonalmodh@gmail.com
[2]Assistant Professor, Department of Computer Science and Engineering, Medi-Caps University, Indore, Madhya Pradesh, India, Email: shimpy.h@gmail.com
[3]Assistant Professor, Department of Computer Science and Engineering, Medi-Caps University, Indore, Madhya Pradesh, India, Email: anusha.jain9@gmail.com
[4]Assistant Professor, Department of Computer Science and Engineering, Medi-Caps University, Indore, Madhya Pradesh, India, Email: priyanka.dhasal07@gmail.com

Received: 13.07.2024 Revised: 16.08.2024 Accepted: 24.09.2024

**ABSTRACT**
This research examines the optimum inventory management issue for new instruments, serving as a pertinent illustration of a supply chain characterised by a short product life cycle. Establishing the correct inventory level minimises wasted opportunities and faulty stock, which is crucial for enhancing profitability. Mathematical optimisation and reinforcement learning methodologies have been suggested for inventory management; nonetheless, the majority of these methodologies concentrate on things that are consistently offered over an extended duration. Consequently, when the objective is a new product, optimising inventory management from the date of its launch is challenging owing to an insufficiency of data for analysis. We address this issue by concentrating on model-based deep reinforcement learning characterised by high sample efficiency and present an inventory management strategy for new items that integrates model learning in an offline setting with planning in an online context. Simulations using authentic historical sales datasets indicate that the suggested strategy surpasses current methodologies for profitability, efficiency, and customer happiness. The suggested strategy enhances overall incentives and inventory turnover by under 5% compared to the trust area policy optimisation method, while preserving the same stock-out rate. Furthermore, the findings indicate that the suggested strategy can sustain consistent inventory management for multiproduct and multistore supply chains.

**Keywords:** Optimization, Machine learning, Heuristic, Bayesian

## 1. INTRODUCTION

Inventory management is a crucial procedure in the supply chain that ensures adequate inventory levels by calculating the optimal order amounts to meet demand throughout a product's sales cycle. Optimising daily orders helps prevent missed sales opportunities due to stockouts. Moreover, an escalation in inventory costs may be avoided by avoiding over-ordering. The supply chain is susceptible to several variables, such as abrupt demand swings and supply delays. Moreover, when a product has a brief life cycle and several variants, it is essential to manage swift environmental fluctuations and elements associated with demand variety. In this research, we concentrate on manufacturing tools, a product inside the supply chain, to tackle the inventory management issue of optimising order quantity while meeting the demand for new goods in a multiproduct and multistore context. Manufacturing tool stores often manage a diverse array of items from several vendors, with each season introducing a new assortment from these providers. In a diversified and rapidly evolving environment, the objective is to enhance inventory management from the moment new items are introduced. Effective optimisation enhances revenues at each shop by decreasing faulty inventory while minimising the stock-out rate of new goods. Moreover, efficient optimisationminimises ordering tasks at each shop and guarantees the continuity of sales operations by autonomously ascertaining suitable order volumes. This multiproduct, multistore inventory management issue involves daily determination of inventory levels for new goods across each product and store. A variety of inventory management techniques using mathematical optimisation and

reinforcement learning (RL) have been examined. For instance, methodologies conceptualised as dynamic programming issues (Van Roy et al., 1997) or stochastic programming problems (Dillon et al., 2017) have been suggested, and these techniques have shown notable efficacy. In recent years, deep reinforcement learning (DRL), which integrates deep learning, has been advanced across multiple domains, with its application to inventory management documented by Boute et al. (2021). Nonetheless, these prior research concentrated on the inventory management of established items that are sold consistently over extended durations; hence, the applicable methodologies are not suitable for newly launched products. This work proposes a technique to optimise inventory management for newly launched items by integrating demand forecasting model development in an offline setting with online planning via model-based deep reinforcement learning, characterised by high sample efficiency. Initially, during the learning phase, the environment, including product demand, is simulated using a historical sales record of previous items. In this demand forecasting model, we utilise a Bayesian neural network (BNN) (Gal & Ghahramani, 2016) capable of probabilistic predictions, alongside a meta-learning approach known as model-agnostic meta-learning (MAML) (Arango et al., 2021; Finn et al., 2017) to accurately forecast demand for new products. During the planning phase, the best daily order amount for each product and shop is established by random shooting (RS) (Rao, 2009; Richards, 2005), including a buffer to the demand quantity based on statistical estimations. Numerical simulations using real manufacturing tool sales datasets show that the suggested strategy effectively manages high profitability and inventory turnover while decreasing the stock-out rate for new goods.

## 2. Mathematical model-based analysis

### 2.1. Reinforcement learning (RL)

Reinforcement Learning (RL) is a machine learning technique whereby an agent develops an optimum policy via trial and error within a specified environment (Sutton & Barto, 2011). Reinforcement learning progresses by monitoring the state $st$ and reward $rt$, then executing action $at$ inside a Markov decision process (MDP). The agent then acquires the policy $\pi(at|st)$ from the environment, which maximises the reward, together with the action value function $Q(at,st)$ and the state value function $V(st)$. Reinforcement learning has been implemented in several domains. For instance, it has been used in robotics (Kober et al., 2013), Mobility as a Service (Xu et al., 2018), healthcare (Asoh et al., 2013; Wang et al., 2018), marketing (Halperin, 2017; Theocharous et al., 2015), and wireless communication (Yajna Narayana et al., 2020). An illustrative instance in the MaaS sector is the application for order dispatching in extensive on-demand ride-hailing systems. Xu et al. (2018) introduced a methodology in which an agent corresponding to each vehicle acquires the spatiotemporal state value function $V(st)$ from historical driving data using the temporal-difference update rule, a reinforcement learning technique.

### 2.2. Model-based RL

Reinforcement learning algorithms may often be categorised into model-free reinforcement learning and model-based reinforcement learning (Li, 2017). The model denotes the state transition function and the reward function of the environment, whereas the category of reinforcement learning is defined by the agent's utilisation of the dynamics model. In model-free reinforcement learning, the agent acquires a policy independently of the dynamics model; hence, this prevalent method is applicable in contexts where the transition and reward functions are unspecified. Conversely, in model-based reinforcement learning, the agent can foresee scenarios and make suitable decisions within the action space by using the dynamics model to forecast state transitions and rewards; hence, model-based reinforcement learning exhibits very efficient learning. Historically, the capacity to model a specific environment has been constrained; however, this is becoming more attainable via the use of highly expressive deep learning, with several model-based methodologies presented in recent years (Luo et al., 2022; Moerland et al., 2020; Wang et al., 2019). One method involves using the Dyna algorithm (Sutton, 1991), which executes the following two phases for the purpose of learning. (1) In model-free reinforcement learning, actions are determined by the existing policy, from which data is gathered from the environment to subsequently learn the dynamics model. The policy is then revised using simulated data produced by the learnt dynamics model. The Model Ensemble Trust-Region Policy Optimisation (ME-TRPO) method, developed by Kurutach et al. in 2018, enhances the Dyna algorithm. The ME-TRPO methodology employs model learning using an ensemble of neural networks and updates policies using TRPO (Schulman et al., 2015) with nonlinear optimisation techniques to enhance sampling efficiency and mitigate model bias. A different model-based method is the Shooting algorithm used in model predictive control (Camacho & Alba, 2013; Wang et al., 2019). In RS (Rao, 2009; Richards, 2005), the agent produces several random action sequences from a uniform distribution and assesses the rewards of each sequence using the acquired dynamics model. The agent conducts just the first action in the sequence deemed best, with

preplanning conducted at each stage. Wang et al. (2019) evaluated over 18 benchmarking conditions and found the Shooting method, including RS, to be successful and resilient across many settings. Conversely, a disadvantage may arise from the challenges of conducting enough investigation in the presence of an extensive action or state space.

### 2.3. Applying RL to inventory management

Numerous reinforcement learning methodologies have been suggested for inventory management throughout the supply chain. Giannoccaro and Pontrandolfo (2002) introduced a Markov Decision Process (MDP) formulation and reinforcement learning (RL) methodology for three-stage inventory management including Supplier, Manufacturer, and Distributor, applicable to more extensive issues than dynamic programming. Kara and Dogan (2018) used Q-learning (Watkins, 1989) and the state–action–reward–state–action algorithm (SARSA) (Singh & Sutton, 1996) for inventory management in a perishable inventory system, demonstrating the efficacy of this approach. Recently, deep reinforcement learning has been used in inventory management by Boute et al. (2021). Moreover, Meisheri et al. (2020) used the advantage actor-critic (A2C) algorithm (Konda &Tsitsiklis, 1999) and a deep Q-network (Tavakoli et al., 2018) in a multiproduct inventory management system for extensive challenges including the management of 100 to 200 goods. Gijsbrechts et al. (2021) implemented the asynchronous advantage actor-critic (A3C) algorithm (Mnih et al., 2016) in a dual-sourcing and multi-echelon inventory management framework. Despite advancements in the application of Deep Reinforcement Learning (DRL) to inventory management, the associated methods have mostly been confined to model-free methodologies. A benefit of model-free deep reinforcement learning is its adaptability in application, enabling its usage in unfamiliar situations. Conversely, a drawback is the inefficient learning process, requiring substantial data and several training events. Recent years have identified inefficiencies in the learning process of model-free deep reinforcement learning (DRL) and some instances of model-based DRL applied to inventory management. Malik et al. (2019) used a model-based deep reinforcement learning approach to a perishable inventory system. A demand forecasting model for perishable commodities in a supermarket chain is developed and used to estimate rewards in deep reinforcement learning (DRL). By adjusting the forecasting model to address uncertainty, they attained substantial performance enhancements relative to heuristic approaches.

### 2.4. Application issues for new products

Previous research on inventory management techniques using model-free and model-based deep reinforcement learning mostly focus on supply chains managing established items that are consistently sold over extended durations. Consequently, the agent's policy learning is conducted using an extensive dataset of previous sales for the items under management. Consequently, implementing a strategy for the inventory management of newly released items is challenging. The statistical technique for inventory management of new items is established by Wanke et al. (2016). This strategy employs a triangle distribution to reflect product demand, while the supply amount is dictated by the $(Q,r)$ model, a traditional inventory management policy. In the $(Q,r)$ model, the policy dictates that inventory should be replenished to the level $Q$ when it goes below the predetermined reorder point $r$. Rojas (2017) offered an inventory management technique using the autoregressive moving average model as a temporal extension of the approach established by Wanke et al. (2016). The benefits of these established systems are their straightforwardness and ease of use in practical inventory management. Conversely, the drawbacks are as follows. From a demand forecasting perspective, the accuracy of forecasts may be poorer than that of machine learning models due to the simplicity of statistical models, which fail to account for exogenous factors such as product attributes and area features. Moreover, from an inventory management perspective, the heuristic approach, shown by the $(Q,r)$ model, may result in stock-outs or excessive ordering in the context of extremely variable demand. Figure 1. Target supply chain for inventory management of new manufacturing tool goods in a multiproduct and multistore context.

## 3. Problem settings

In this study, we take manufacturing tools as an example and address the problem of determining the order quantities for new products to maximize the total rewards and inventory turnover while minimizing the stock-out rate.

### 3.1. Inventory management of new products

As shown in Fig. 1, the target supply chain comprises a logistics center, multiple retail stores, multiple new products with stochastic demand, and end customers.

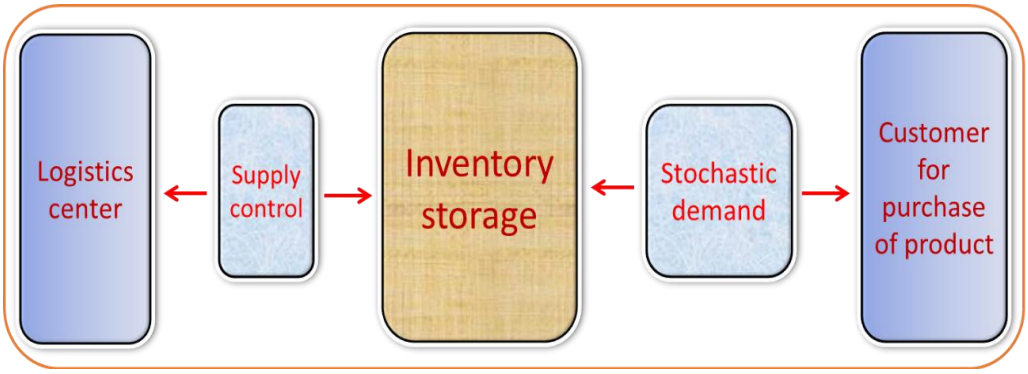


**Fig 1.** Supply chain targets

The agent monitors the inventory levels of each product at every shop, calculates suitable order amounts to ensure enough supply, and then distributes each product to the respective stores. Inventory management begins on the day each product is launched, and an arbitrary beginning value is set to the inventory level for each shop and product on the release date. The agent assesses the supply order daily by monitoring inventory variations, with items delivered the next day. Each product is sourced from a distinct provider, and demand patterns vary per product. Moreover, the product life cycle of a production tool is brief, with each supplier introducing new goods each season. It is typical for each product to possess many colour variants; nevertheless, for the sake of simplicity, this research does not differentiate between colours and concentrates on the inventory levels of each product series. The demand pattern for each product varies based on the geographical location of each retailer. In metropolitan regions, the newest flagship items see immediate high demand upon release. Conversely, the desire for older items persists robustly in rural regions. To meet the sales demand for new manufacturing tool goods with these attributes, the agent makes supply choices on a daily, product, and store basis.

### 3.2. Markov decision process

This document outlines the formulation of the relevant Markov Decision Process (MDP) for the issue of ascertaining the supply of new items. Figure 2 illustrates the interactions between the agent and the environment. The environment constitutes a supply chain characterised by stochastic demand for various items across many retail locations.


**Fig 2.** Interactions between agents and customers

The agent controls the inventory by determining the next quantity supplied $At$ while observing the state $St$ including the inventory level at each step.In this problem set, $t$ is assumed to be discrete-time, and the time step is one day. In other words, the agent maintains an appropriate inventory level through daily supply decisions. The reward $r'$ at time $t$ + 1 for taking action $At$ in state $St$ can be expressed as follows with unknown demand $d$ as in transition probabilities: $r' = R(St = s, At = a) = g(\min\{\min\{q + a, M\}, d\}) - c1(q) - c2(a)$ (5) where $g(\cdot)$ is a function to compute the revenue for the number of units that could be sold at $t$ +1, $c1(\cdot)$ is a function to compute the inventory cost for the number of units held at $t$, and $c2(\cdot)$ is a function to compute the ordering cost for the number of units ordered at $t$. For simplicity, in this study,

functions $g$,$c1$, and $c2$ are known and assumed to be linear.The inventory level of product $i$ and store $j$ at time $t$ = 0, i.e., the starting position of the inventory operation, is given as the initial value $q$.

### 3.3. Inventory management metrics

The agent uses the following metrics to measure inventory management performance over the entire period $T$: total reward from a profitability perspective, inventory turnover from an efficiency perspective, and stock-out rate from a customer satisfaction perspective.

The total reward is the sum of the rewards at each time step over the entire period $T$. The total reward in one episode for product $i$ and store $j$ is calculated as follows. $T \sum t=1\ rt$ (6) Inventory turnover Inventory turnover is a measure of how many times inventory is replaced in a given time period $T$, which is obtained as follows: $\sum Tt=1\ d'\ t\ (q0\ +\ qT)2$ (7) where $d'\ t$ represents the daily sales. Here, $d'\ t$ is calculated according to the inventory level $qt-1$, the order $at-1$, and the next step demand $dt$ as follows. $d'\ t$ = min{min{$qt-1\ +at-1,M$},$dt$ }. In this study, to evaluate inventory management in a single period $T$, average inventory is by averaging beginning inventory and ending inventory according to Ali (2011), Unleashed (2022), and Amazon (2022). Stock-out rate The stock-out rate is the ratio of the number of days during which shortages min{$qt\ +at,M$} <$dt+1$ occur over $T$. The stock-out rate is computed as follows: $T \sum t=1\ utT$ (8) where $ut+1$ denotes the occurrence of missing product $i$ at store $j$ in each time step. Here, $ut+1$ is obtained as follows. { $ut+1$ = } 1, 0, min{$qt\ +at,M$} <$dt+1\ otherwise$.

4. Proposed method

In this section, we describe the proposed model-based DRL inven tory management method for new products. The proposed method is illustrated in Fig. 3. As shown, the proposed method comprises two main phases, i.e., a learning phase for the demand forecasting model in an offline environment and a planning phase for supply decisions in an online environment.

### 4.1. Demand forecasting

First, we describe the demand forecasting model used in the proposed and learning methods as well as the features used in the learning method.

#### 4.1.1. Model definition

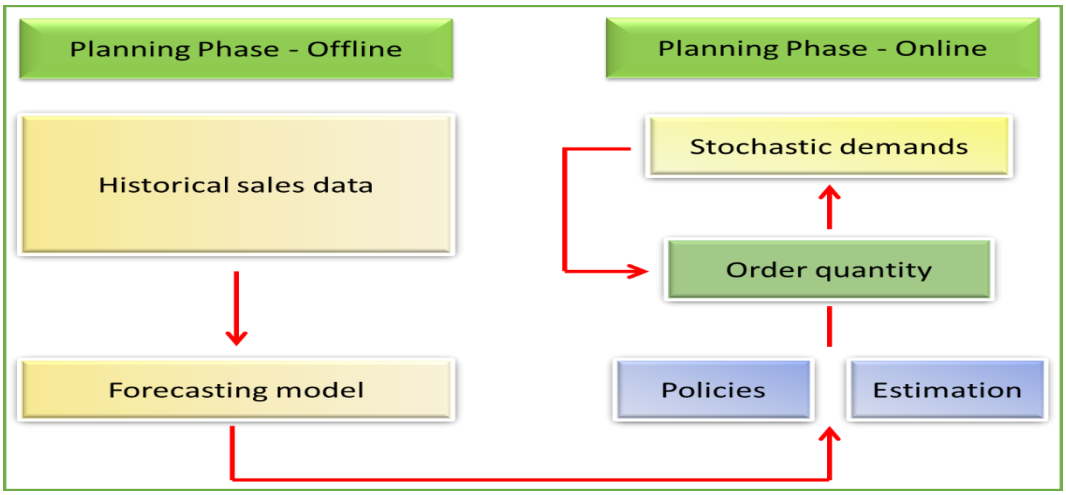


**Fig 3.** Model-based management system of inventory

#### 4.1.2. Model learning

In the target problem, the inventory management of new products begins at the sales start date $t$ = 0. At this time, the inventory level $q'$ and reward $r'$ at time $t+1$ can be estimated using$\hat{}$ $d$. However, it is difficult to train the forecasting model $f\theta$ because a sufficient historical sales dataset.

### 4.2. Uncertainty-aware forecasting

Here, we explain the variations of the demand forecasting model $f\theta$ with the TrainModel function and the inference with the model $f\theta$ in an online environment. To handle the uncertainty of demand for new products, we incorporate two types of deep learning models. In the first approach, we treat $f\theta$ as a probabilistic model using a BNN. The model parameter $\theta$ is expressed as a probability distribution instead of a deterministic value, and the neural network model $f\theta$ is trained by Bayesian inference. The second approach is based on meta-learning. This approach attempts to learn $f\theta$ from dataset using MAML

and immediately apply it to the demand forecasting task on a new dataset. Next, we describe inference in an online environment using the learned model $f\theta$.

#### 4.2.1. Planning with RS

In the online planning phase, the order quantity $At$ is determined by the Planning($St$,$D$) function using the current state $St$ and demand $D$ estimated by the upper confidence limit as inputs. In this process, planning is performed using RS, which is a model-based RL method. As shown in Fig. 5, this planning process generates $K$ random action sequences up to $h$ steps ahead based on a uniform distribution beginning from the current inventory level $q$ in state $St$ = [$q$,$ui$,$vj$,$wt$].

## 5. Experimental evaluation

To evaluate the effectiveness of the proposed method, i.e., model based DRL with offline learning and online planning, we performed an inventory management simulation for new products using real-world historical sales datasets for multiple products and multiple stores.

### 5.1. Simulation settings

In the following, we explain the datasets and compared methods that were considered in the numerical simulation.

#### 5.1.1. Experimental datasets

Beginning from $t$ = 0, i.e., the time at which new products are released, the agent starts to control and determine the daily order quantity until day $T$ = 200. Using the historical dataset of the daily sales of each new product in the target stores and the order quantity determined by the agent, the environment returns the next state $St$+1 = $s'$. The percentage of the actual total sales of the three new products (A, B and C) in all 10 stores during the relevant period. As can be seen, the demand for product C is the highest among these products, and the sales demand differs depending on the product.

#### 5.1.2. Baseline algorithms

In this inventory management simulation, the heuristic-based and model-free DRL approaches are used as baselines for comparison with the proposed method. Heuristic approach This is a simple rule-based policy that determines the order quantity by determining the recommended inventory level by $\beta \times \tilde{} d$ using the MA$\tilde{} d$ of the past 28 days' sales for product $i$ at store $j$ and the safety factor $\beta$. In this simulation, the safety factor is set to a large value, i.e., $\beta$ = 6, which represents a passive strategy that focuses on avoiding the occurrence of stock-outs. Model-free DRL We apply TRPO (Schulman et al., 2015), which is a policy-based approach, to the inventory management simulation of new products $i\in$ by learning the policy in the environment of dataset. Model-based DRL (proposed method) A demand forecasting model is trained using a BNN or MAML on dataset. For the inventory management of new products, the order quantity is determined by RS. Oracle Here, the sales demand $d$ for new products is known, and perfect inventory control is performed to maximize the number of sales while maintaining the minimum amount of inventory. Seen as a maximum performance in the simulation.

### 5.2. Evaluation results

To verify the effectiveness of the proposed algorithm for inventory management of new products in a multiproduct and multistore environment, the simulation results are discussed from the following perspectives: (1) the overall results are discussed to compare the performance of each algorithm according to the total reward, inventory turnover, and stock-out rate as evaluation metrics; (2) an analysis by product is discussed to evaluate how the inventory level is maintained for each product; (3) an analysis by product and store is described to see how inventory level is maintained for each product and store; and (4) model accuracy is investigated to compare the forecasting trends and accuracy of each demand forecasting model.

#### 5.2.1. Overall results

For each algorithm, the inventory management simulations for four new products and 10 stores were iterated 10 times for a total of 400 episodes. Here, the total reward is normalized via min–max normalization among all episodes. As can be seen, the proposed model-based DRL method using RS and BNN obtained the highest inventory turnover of 13.2, which represents a substantial improvement of > 5% from 12.72 obtained with the heuristic method. Furthermore, the proposed BNN algorithm reduced the stock-out rate to 0.5%, which is the same as the heuristic method (i.e., the passive strategy).

### 5.2.2. Analysis by product

Fig. 4 shows the time-series changes in the average inventory level at all stores at time $t$ for each of the four new products (A, B and C) for heuristic algorithm. As can be seen, for product C, for which demand is highest, the Bayesian and TRPO methods always maintained an excessive inventory level of ~10 units. With the TRPO method, when the inventory level was close to zero, the order-taking action was repeated with a large order quantity at once to ensure that the inventory level was sufficient. Fig. 5 and 6 plots the distribution of the average inventory levels for each of the four products (A, B and C) over the entire period $T$ for each algorithm. With the heuristic method, as mentioned previously, the inventory levels were very high for the high-demand product C.

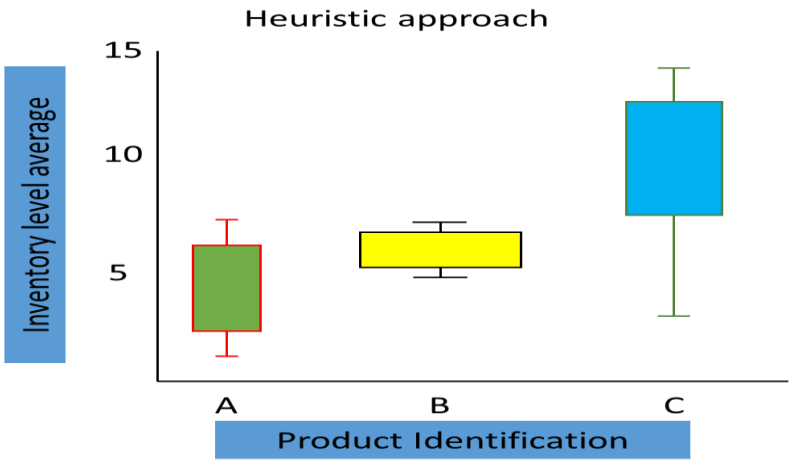


**Fig 4.** Distribution of inventory as per Heuristic approach

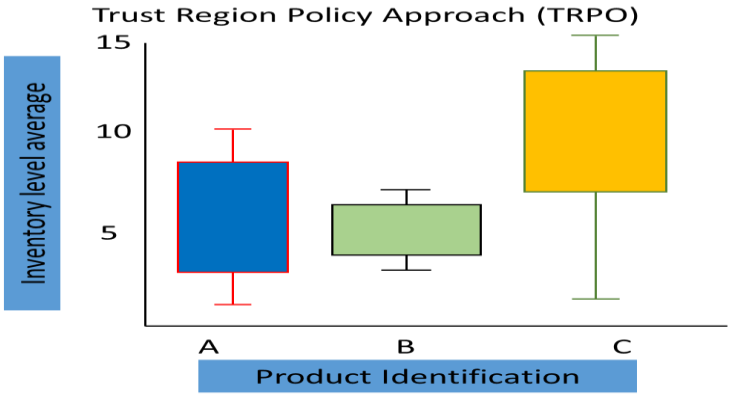


**Fig 5.** Distribution of inventory as per TRPO approach

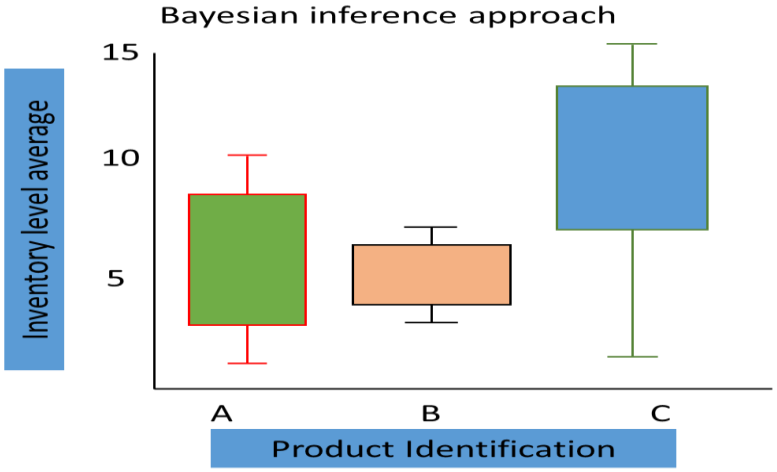


**Fig 6.** Distribution of inventory as per Bayesian approach

### 5.2.3. Analysis by product and store

The average inventory levels for each product and store combination over the entire period obtained by each algorithm are shown in Fig. 7. Here, the $x$-axis denotes the store ID, and the $y$-axis denotes the inventory level.

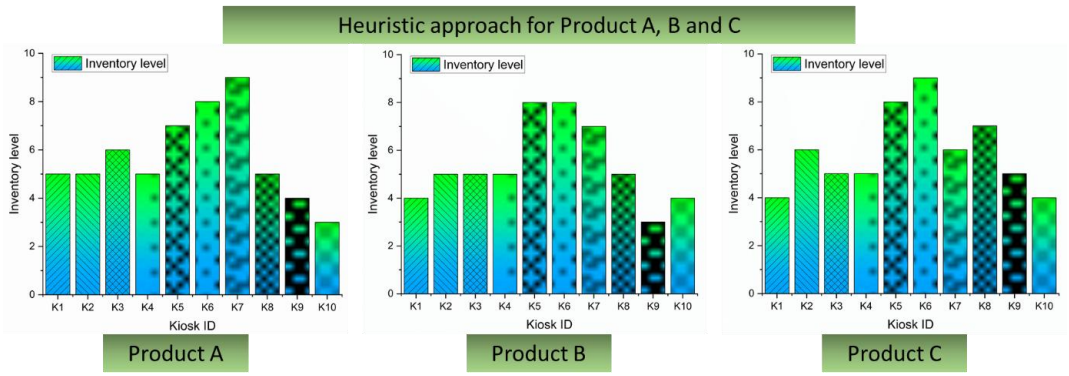


**Fig 7.** Average inventory levels as per Heuristic approach

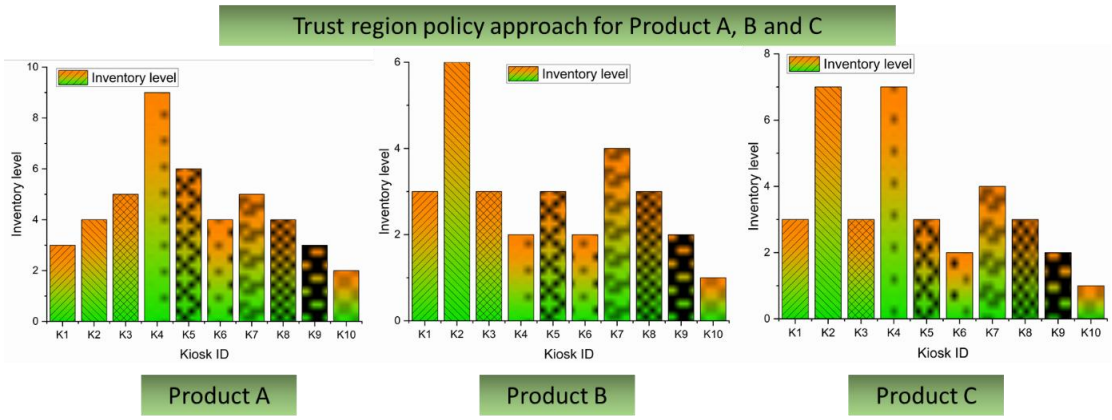


**Fig 8.** Average inventory levels as per TRPO approach

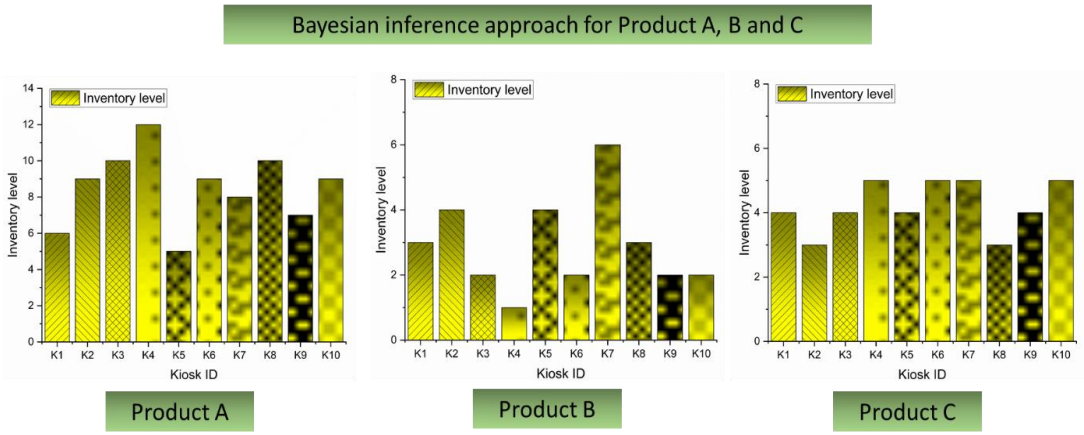


**Fig 9**. Average inventory levels as per Bayesian approach

The heuristic technique and model-free TRPO method resulted in significant variability in inventory levels across different products and stores. In the inventory management of product C using the heuristic technique, inventory levels fluctuate significantly among retailers. Conversely, the suggested Bayesian algorithms maintained virtually uniform inventory levels across both items and shops. The findings indicate that the suggested algorithms may effectively manage new product inventories in a multiproduct and multistore context.

### 5.2.4. Model accuracy

Here, we discuss the prediction performance results of the demand forecasting model, which is a component of the proposed method.

| Model | MAE | RMSE | $R^2$ |
|---|---|---|---|
| Heuristic | 0.59 | 0.34 | 0.82 |
| TRPO | 0.52 | 0.27 | 0.86 |
| Bayesian Neural Network | 0.49 | 0.22 | 0.95 |

Table 1 presents the prediction errors for the Bayesian Neural Network (BNN), TRPO and heusristic neural network model trained using Model-Agnostic Meta-Learning on the historical sales dataset. For each store of new products, we evaluate the mean absolute error (MAE), root mean square error (RMSE), and $R2$ score for the subsequent day's forecasted demand value$\hat{}$ $d$ and the actual demand value $d$. Furthermore, the moving average used in the heuristic approach, which forecasts the following day's sales demand based on the moving average of the preceding 28 days of sales, was taken into account for comparison. The BNN approach achieved the lowest MAE of 0.49, an improvement from the MA technique's 0.59. The heuristic technique achieved the optimal RMSE and $R2$ score, notably a $R2$ value of 0.82. Figure 9 illustrates the temporal variations in both the forecasted and actual demand for each product, with the x-axis representing the time step and the y-axis indicating the sales ratio of each product, normalised as a % of total sales across all shops, items, and periods. The demand trends for the four manufacturing tool goods (A, B, and C) are markedly distinct. Demand for product C, the most sought-after manufacturing tool, was substantial soon after its debut and shown little drop over time. Conversely, demand for the product was up shortly after its debut but then diminished swiftly.

## 6.DISCUSSION

The efficacy of the suggested strategy was validated by a simulation of inventory management over 200 days after the introduction of our new items across 10 outlets, utilising new manufacturing tool products as a case study. The suggested technique outperformed the comparative methods in total reward, inventory turnover, and stock-out rate by assessing daily demand for each product and shop using a trained model and finding the optimal supply amount via planning. The suggested strategy will be effective for inventory management of items with short lifecycles, as shown by these findings.The demand forecasting model used in the proposed technique was trained only on a historical sales record for previous goods. The model was not calibrated according to the sales of new items. This technique was used to enable the deployment of a functional inventory management system as a straightforward means for using the trained model in an online setting. Incorporating fine-tuning during the online phase is anticipated to enhance forecasting accuracy and, therefore, inventory management performance. A further constraint of the suggested technique pertains to the scope of appropriate product coverage. This research examined manufacturing tools as an exemplar of items with short life cycles and assessed the efficacy of the suggested strategy for the inventory management of four new products. Nevertheless, if the suggested strategy is to be used for tens to hundreds of thousands of goods in a typical retail establishment, it is essential to account for data sparsity. Furthermore, to manage forecasting for items with minimal sales quantities, product segmentation and hierarchical forecasting are essential.

## 7. CONCLUSION

This research presents a model-based deep reinforcement learning strategy that enhances inventory management for new goods across many products and stores. Model-free deep reinforcement learning requires substantial data and several trials, rendering it inappropriate for the inventory management of novel items. Consequently, this work concentrated on model-based Deep Reinforcement Learning (DRL), characterised by its great learning efficiency, and the suggested technique integrates model learning in an offline setting with planning using Reinforcement Sampling (RS) in an online context. A demand forecasting model derived from past sales records was used as a simulator to replicate a realistic supply chain environment. To mitigate the uncertainty around product demand, we used two categories of deep learning models: Bayesian Neural Networks (BNNs), which provide probabilistic predictions, and heuristic Artificial Neural Networks (ANNs), a conventional meta-learning approach. The efficacy of the suggested strategy was validated by numerical simulations utilising actual historical sales information to address the inventory management job for manufacturing tools, exemplifying a supply chain characterised by short product life cycles. The suggested strategy demonstrated superior performance across all measures including profitability, efficiency, and customer satisfaction. Furthermore, we verified

that the suggested strategy maintained approximately uniform inventory levels for each product across all locations. This suggests that the suggested strategy may efficiently account for the varying sales needs of items and retailers, hence ensuring adequate inventory levels. Consequently, we assert that the suggested inventory management technique may be promptly implemented on the day of product launch, which is especially crucial in supply chains managing items with short product life cycles. Additional performance enhancements include optimising the demand forecasting model in the live phase and mitigating data sparsity via product segmentation and hierarchical forecasting.

**REFERENCE**


[1] Arango, S. P., Heinrich, F., Madhusudhanan, K., & Schmidt-Thieme, L. (2021). Mul timodal meta-learning for time series regression. In International workshop on advanced analytics and learning on temporal data (pp. 123–138). Springer, http: //dx.doi.org/10.1007/978-3-030-91445-5

[2] Asoh, H., Akaho, M. S. S., Kamishima, T., Hasida, K., Aramaki, E., & Kohro, T. (2013). An application of inverse reinforcement learning to medical records of diabetes treatment. In ECMLPKDD2013 workshop on reinforcement learning with generalized feedback.

[3] Boute, R. N., Gijsbrechts, J., van Jaarsveld, W., &Vanvuchelen, N. (2021). Deep reinforcement learning for inventory control: A roadmap. European Journal of Operational Research.

[4] Camacho, E. F., & Alba, C. B. (2013). Model predictive control. Springer science & business media. Dillon, M., Oliveira, F., & Abbasi, B. (2017). A two-stage stochastic programming model for inventory management in the blood supply chain. International Journal of Production Economics, 187, 27–41.

[5] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning (pp. 1126–1135). PMLR, http://dx.doi.org/10.48550/arXiv.1703.03400.

[6] Malik, A., Kuleshov, V., Song, J., Nemer, D., Seymour, H., & Ermon, S. (2019). Calibrated model-based deep reinforcement learning. In International conference on machine learning (pp. 4314–4323). PMLR, http://dx.doi.org/10.48550/arXiv.1906. 08312.

[7] Meisheri, H., Baniwal, V., Sultana, N. N., Khadilkar, H., & Ravindran, B. (2020). Using reinforcement learning for a large variable-dimensional inventory management problem. In Adaptive learning agents' workshop at AAMAS. Mnih, V., Badia, A. P., Mirza, M., Graves,

[8] A., Lillicrap, T., Harley, T., Silver, D., &Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In International conference on machine learning (pp. 1928–1937). PMLR, http://dx. doi.org/10.48550/arXiv.1602.01783. Moerland, T. M., Broekens,

[9] J., & Jonker, C. M. (2020). Model-based reinforcement learning: A survey. http://dx.doi.org/10.48550/arXiv.2006.16712, arXiv preprint arXiv:2006.16712. Rao, A. V. (2009). A survey of numerical methods for optimal control. Advances in the Astronautical Sciences, 135(1), 497–528.

[10] Richards, A. G. (2005). Robust constrained model predictive control (Ph.D. thesis), Massachusetts Institute of Technology. Rojas, F. (2017). A methodology for stochastic inventory modelling with ARMA triangular distribution for new products. Cogent Business & Management, 4(1), Article 1270706.

[11] Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. In International conference on machine learning (pp. 1889–1897). PMLR, http://dx.doi.org/10.48550/arXiv.1502.05477.

[12] Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. Machine Learning, 22(1), 123–158.

[13] Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. ACM Sigart Bulletin, 2(4), 160–163. http://dx.doi.org/10.1145/122344. 122377. Sutton,

[14] R. S., & Barto, A. G. (2011). Reinforcement learning: An introduction. Cambridge, MA: MIT Press.

[15] Szepesvári, C. (2010). Algorithms for reinforcement learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 4(1), 1–103.

[16] Tavakoli, A., Pardo, F., & Kormushev, P. (2018). Action branching architectures for deep reinforcement learning. In Proceedings of the AAAI conference on artificial intelligence, Vol. 32, no. 1.