Predictive optimization in automotive supply chains: a BiLSTM-Attention and reinforcement learning approach

Asmae Amellal [a,1,*], Issam Amellal [b,2], Mohammed Rida Ech-Charrat [a,3], Hamid Seghiouer [a,4]

[a] ENSATÉ, Abdelmalek Essaadi University, Tétouan, 93000, Morocco
[b] ENSATB, Hassan 1st Univesity, Berrechid, 26202, Morocco
[1] asmae.amellal57@gmail.com; [2] amellal.issam@gmail.com; [3] charrat.mohammed@gmail.com; [4] folio_hamid@yahoo.fr
* corresponding author

ARTICLE INFO

Article history
Received September 28, 2023
Revised March 27, 2024
Accepted March 29, 2024
Available online August 31, 2024

Keywords
Supply chain management
BiLSTM-Attention model
Reinforcement learning
Game theory
Decision making

ABSTRACT

Effective supply chain management is pivotal for enhancing customer satisfaction and driving competitiveness and profitability in the automotive service and spare parts distribution sector. Our research introduces an innovative approach, integrating game theory, BiLSTM-Attention deep learning, and Reinforcement Learning (RL) to refine supply and pricing strategies within this domain. Focusing on Moroccan automobile companies, we utilized Enterprise Resource Planning (ERP) system data to forecast customer behavior using a BiLSTM model enhanced with an Attention mechanism. This predictive model achieved a Mean Squared Error (MSE) of 0.0525 and an R^2 value of 0.896, indicating high accuracy and an ability to explain substantial variance in customer behavior. To further our analysis, we incorporated reinforcement learning, evaluating three algorithms: Q-learning, Deep Q-Networks (DQN), and SARSA. Our findings demonstrate SARSA's superior performance in our context, attributed to its adeptness at navigating the dynamic environment of the automotive supply chain. By synergizing the predictive power of the BiLSTM-Attention model with the strategic optimization capabilities of reinforcement learning, particularly SARSA, our study offers a comprehensive framework for automotive companies to enhance their supply chain strategies, balancing profitability and customer satisfaction effectively in a rapidly evolving industry sector.

1. Introduction

In the modern era of globalization and swift market changes, the effectiveness of supply chain management (SCM) has emerged as a pivotal component in achieving and maintaining a competitive edge. This is even more pronounced in niche sectors like automotive service and spare parts distribution. Here, the margins for error are slim as success depends on precise pricing strategies coupled with intelligent inventory management methodologies [1], [2]. While robust SCM frameworks have the potential to streamline inventory flows, curtail unnecessary expenses, and amplify customer service quality, ill-conceived strategies can trigger a cascade of challenges. These include the peril of stockouts, dealing with unwarranted inventory surplus, contending with extended lead times, and grappling with surging operational costs. Such challenges pose severe threats, often undercutting the profitability matrix and competitive market stance [3].

Our study builds on the foundational role of SCM in the automotive parts and service industry, introducing a novel methodology that merges game theory, BiLSTM-Attention deep learning, and

reinforcement learning to tackle the sector's complex dynamics, aiming to enhance business interactions, profitability, and customer satisfaction. Various methods, including simpler ones like random forest, lasso, GRU, RNN, CNN, LSTM, and BiLSTM, and hybrid models like RNN-LSTM, CNN-RNN, and CNN-LSTM, have been used across disciplines for forecasting in recent years [4]–[7]. Attention mechanisms, in particular, have revolutionized deep learning by enabling models to focus dynamically on relevant data segments, improving data representation and addressing long-range dependencies [8]. While these technologies have advanced various domains, including natural language processing [9]–[11], their application in the automotive industry requires specific customization to meet the distinct challenges and opportunities of this evolving sector.

Transitioning from a broad examination of technological advancements to their particular relevance in our research, we delve into addressing the pivotal gaps highlighted by previous studies in the realm of behavior prediction and strategic decision-making within the automotive sector [12] delved into predicting customer purchasing behavior using various machine learning classifiers, attaining notable accuracy. Nevertheless, their reliance on static datasets might not capture the dynamic nature of the automotive service and spare parts sector, where specific industry variables are crucial. Their research also touches upon the unexplored potential of deep learning for larger datasets. [13] investigated multiscale adaptive object detection with contrastive feature learning in retail, a method adept at processing spatial and visual data for analyzing customer behavior. However, the method's high computational and memory demands and initial design tailored to retail settings may limit its applicability in the more variable and resource-strained automotive supply chain environments. [14] provided valuable insights within the banking sector, yet their study's reliance on conventional feature engineering and classification falls short of capturing the temporal dynamics that deep learning architectures, particularly BiLSTM-Attention, can offer architectures that are crucial for dissecting and forecasting the time-sensitive behaviors in dynamic sectors like automotive. [15] employed a GRU network to forecast occupancy in intelligent buildings, an approach that might only partially capture the complex behavioral patterns over time, which is essential in the automotive context. Similarly, [16] work in consumer behavior prediction, utilizing the SVM classification algorithm, showcases high accuracy and recall, underscoring the need for further enhancement in predictive depth and generalizability across various industries. The exploration of BiLSTM-attention mechanisms by [17] and [18] highlights their potential yet underscores the necessity for bespoke adjustments to effectively tackle the distinct challenges of the automotive spare parts and service industry.

Extending our analysis further, we target the automotive spare parts industry with game theory and reinforcement learning to enhance stock and pricing predictions. This focus requires adapting existing models from various domains to address the unique challenges of this sector, underscoring the need for specialized methodologies in this context. For instance, [19] integrates game theory and reinforcement learning for construction bidding to counteract the winner's curse. However, this model's reliance on specific algorithms and the low bid method may not suit the automotive sector's varied pricing and bidding strategies. Its primary focus on cost estimates, without considering factors like markup values or a fluctuating number of competitors, could limit its adaptability to the dynamic automotive supply chain environment. [20] explore energy management in microgrids using real-time pricing and reinforcement learning, optimizing electricity distribution. While compelling within microgrids, this model might struggle in the automotive supply chain due to its design for specific energy system dynamics, contrasting with the automotive sector's supply chain logistics, demand variability, and production planning. [21] develop a multi-agent deep reinforcement learning framework for community virtual power plants (cVPPs), focusing on energy management and bidding strategies.

Adapting this framework to the automotive supply chain might require broadening the reinforcement learning algorithms and rethinking the bidding strategy beyond the low bid method to account for the automotive industry's complexity. [22] propose a deep reinforcement learning framework for managing

photovoltaic-battery systems, emphasizing load forecasting. While promising for energy management, its suitability for the automotive supply chain may be limited by the substantial training time and computational resources needed, which could hinder rapid deployment in the industry's dynamic setting. [23] MARL framework, designed for energy system demand response, may not fully align with the automotive supply chain's diverse operational constraints and market dynamics. Lastly, [24] study stock price formation through a multi-agent reinforcement learning model, SYMBA, which might not capture the full complexity of the automotive market due to its narrow focus on single-stock trading and lack of features such as short-selling, potentially limiting its applicability to real-world market analysis. These studies highlight the need for tailored adaptations to leverage these advanced methodologies effectively in the automotive spare parts industry. To navigate these challenges, our approach is to customize the integration of game theory and reinforcement learning, tailoring it to the distinctive requirements of the automotive spare parts and service industry. Our strategy employs game theory to forecast outcomes from strategic interactions, factoring in the sector's dynamic and complex nature.

This is complemented by reinforcement learning, which will iteratively learn and refine policies over time, enhancing decision-making capabilities and facilitating precise stock and price strategy forecasting. Such an approach is vital for addressing the dynamic challenges of pricing and inventory management, underscored by the effectiveness of multi-agent RL in developing adaptive pricing strategies, as demonstrated in the research by [25]–[27]. This tailored integration is designed to provide a robust framework that not only overcomes sector-specific challenges but also drives strategic and operational excellence in the automotive spare parts and service industry.

Our fusion of a BiLSTM-Attention model with game theory and RL culminates in a framework tailored for the automotive industry, enhancing its forecasting and strategic agility. This initiative promises a supply chain responsive to the sector's evolving needs, ensuring stakeholders can adeptly navigate its dynamic landscape.

2. Method

First, In this section, we initially present the proposed framework as described in Fig.1.

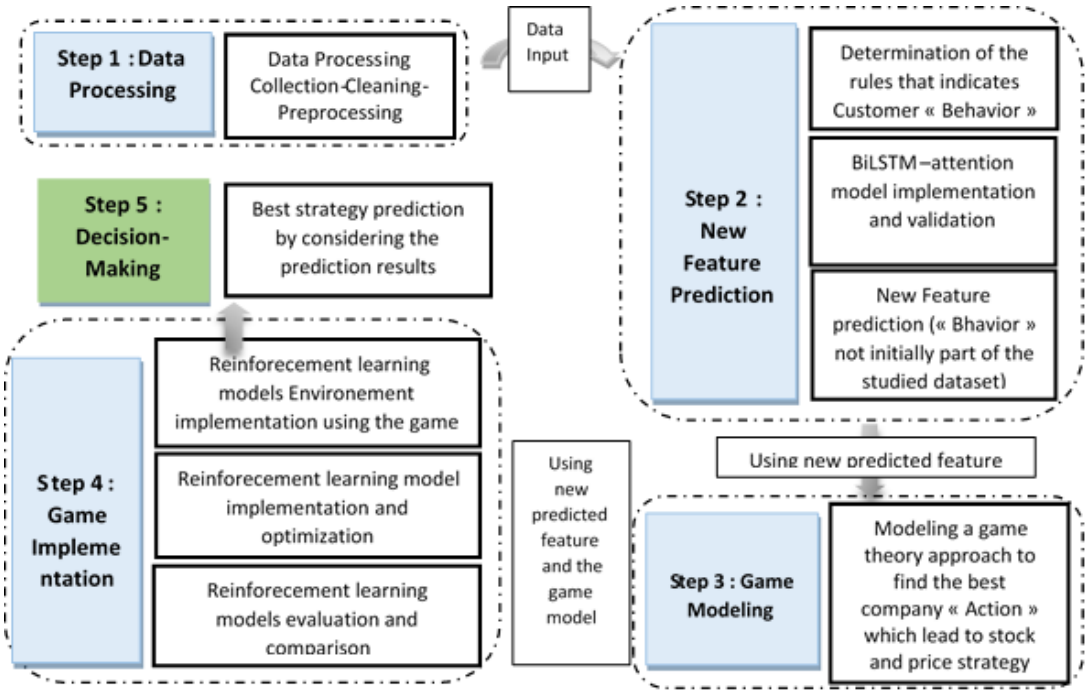

Fig. 1. Proposed Framework

2.1. Data Processing

In this study, we meticulously analyzed a dataset sourced from the ERP system of a Moroccan car distributor, concentrating on the after-sales service data from 2019 to 2022. Following an extensive data cleansing phase where we removed duplicates, instances of negative margins, blank cells, and outliers, our dataset was refined to encompass 91,769 records, each described by 12 detailed columns, as presented in Table 1.

Table 1. Features Description

Feature	Description
Customer	A unique ID represents each customer in the dataset.
Mileage	The vehicle's odometer reading in kilometers at the time of service.
Workshop	The name or location of the Workshop where the service was performed.
Date of Passage	The date when the vehicle was serviced.
Brand	The make of the vehicle.
Model	The model of the vehicle.
Date of Circulation	The initial registration date of the vehicle.
Type of Repair	The category or description of the repair.
Invoice Amount	The total cost charged for the service.
Margin	The profit made from the service.
Warranty Status	Indicates whether the vehicle was under warranty at the time of repair.
Repair Duration	The time taken to complete the repair.

During preprocessing, categorical data were encoded into numerical values and numerical data were normalized using the Min-Max method, scaling them to a 0-1 range with the formula.

$$x_{\min} = \frac{(x - x_{\min})}{(x_{\max} - x_{\min})} \quad (1)$$

Once normalized, the data were stored as either 32-bit or 64-bit integers. Subsequently, the dataset was divided into two subsets: one for training with 73,415 entries and another for testing with 18,354 entries, facilitating the development and validation of models. This methodology is essential for ensuring the dataset's suitability for in-depth analysis and modeling.

2.2. New Feature Prediction and Model Integration

After processing the data, we identified a critical feature, 'Customer Behavior,' which reflects the customers' maintenance choices, hinting at their loyalty or potential switch to competitors. We devised an algorithm to analyze and categorize this behavior based on two key metrics.

- Visit Intervals: The time between consecutive visits, indicating how regularly customers seek maintenance services.
- Mileage Increase: The change in vehicle mileage between visits, providing insights into vehicle usage and potential service needs

2.2.1. Selection of BiLSTM-Attention Model

We subsequently employed a BiLSTM-Attention model to forecast 'Customer Behavior .'The LSTM structure, created by [28] in 1997, addressed long-term dependency issues in traditional RNNs by incorporating a "memory" concept [29]. Fig. 2 encapsulates the LSTM and BiLSTM configurations utilized in our model.

- LSTM cell

  The input of an LSTM unit is composed of three distinct vectors. The first, named input vector, is external to the LSTM unit and is injected into it at instant t (The vector xt in the diagram above); the two of them are generated by the LSTM unit at the previous instant (t-1) (In the diagram, they are the vector ht-1 indicating the hidden state, and the vector ct-1 representing cell

state). The long-term memory cell (ct) at time t is updated using external input as well as short-term memory from the previous time step (t-1), denoted by xt-1 and ht-1, respectively [29].

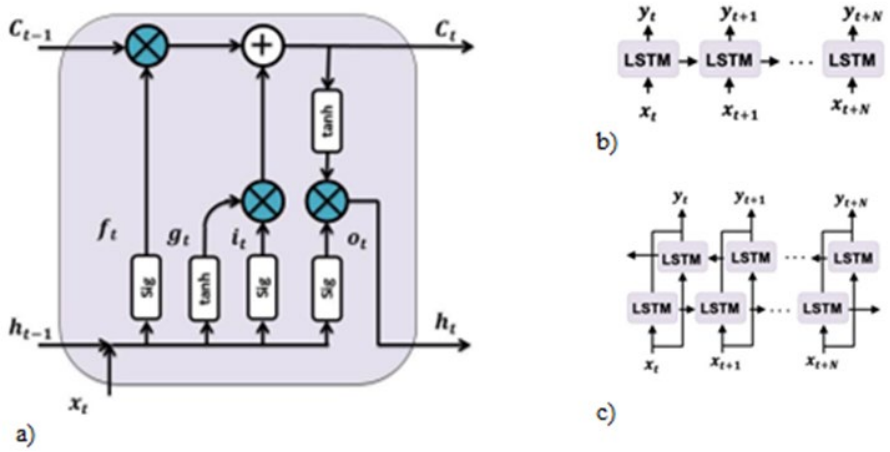

Fig. 2. a) An LSTM cell; b) Unidirectional LSTM; c) BiLSTM

- Unidirectional LSTM

  The LSTM gate handles information flow, resolving the vanishing gradient problem

- BiLSTM

  This type of LSTM combines two unidirectional LSTMs moving in opposite directions. It operates using both recent and older information.

The attention mechanism in deep learning allows models to focus more on specific features during training. It calculates attention weights, indicating each feature's importance and creating a weighted input representation. Mathematically [30], given an input sequence $X = [x_1, x_2, \ldots, x_n]$ and a target vector y, these inputs are transformed into key vectors K and value vectors V using trainable parameters W_k and W_v: $K = [W_k * x_1, \ldots, W_k * x_n]$; $V = [W_v * x_1, W_v * x_n]$. Target vector y is transformed into a query vector Q using parameters W_q: $Q = W_q * y$. The attention scores, a, are computed as the dot product between Q and K: a = [Q *k_1, Q * k_2, ..., Q * k_n], and normalized via the softmax function: a = softmax(a). The attention weights, w, are computed as a weighted sum of the value vectors V with weights given by a: $w = [a_1 * v_1, a_2 * v_2, \ldots, a_n * v_n]$. The final output z is the concatenation of w: $z = [w_1 + w_2 + \ldots + w_n]$, which is then passed through a fully connected layer for the final prediction.

2.2.2. The proposed BiLSTM-Attention model implementation

We implemented the network structure using the Python programming language and the Keras neural network framework. The combination of Python and Keras allowed us to create and train deep learning models in a user-friendly and flexible environment, which proved highly effective for our task. Table 2 shows the structure of a Bidirectional Long Short-Term Memory (BiLSTM)-Attention model that has been tuned by optimizing its hyperparameters.

In this instance, the "Mean Squared Error" (MSE) is employed as the loss function, which can be described as,

$$MSE = \frac{\Sigma(y_i - y_p)^2}{n} \quad (2)$$

In addition to the MSE, R2 is another parameter used to measure the performance of our model:

$$R^2 = 1 - \frac{\Sigma(y_i - y_p)^2)/n}{\Sigma(\hat{y}_i - y_p)^2)/n} \quad (3)$$

Where y_i is the predictive value, y_p is the actual value, $\hat{y}_i$ is the average value, and n is the number of observations or rows. Finally, « Grid search » is a hyperparameter optimization technique that involves searching over a pre-defined set of hyperparameter values to find the set of hyperparameters that result in the best performance on a validation set.

Table 2. BiLSTM-attention model structure

Component	Role/Calculation	Input value
Input	Inputs are tokenized using Keras' Tokenizer and padded to a fixed length.	Dataset
Input layer	Takes the padded sequences as input	padded sequences
Embedding layer	Depict tokens within a compact vector space, with each dimension in the vector signifying a distinct attribute or aspect	Tokens
Bidirectional LSTM	The output of the embedding layer is fed into a Bidirectional LSTM layer with dropout regularization.	num_units = 64 ; dropout_rate=0.2 ; recurrent dropout= dropout rate
Attention layer	Calculates attention weights for each timestep of the input sequence.	Activation function= Tanh
Dense output layer	Uses the transformed features from the preceding layers to make the final predictions	function= Softmax
Optimizer	Is the mathematical function used to minimize the error function	Adam
Loss Function	Determines the dissimilarity between the predicted and actual outputs	MSE (Mean Squared Error)
Batch_size	Is the number of samples that are processed in a single forward/backward pass of the neural network during training	32
Learning rate	It is the value used to adjust a model's parameters towards minimizing the error function.	0.01
Epochs	Refers to the number of times the entire training operation will be performed	20

The BiLSTM-Attention model is our research's foundation for game theory and reinforcement learning, focusing on the 'Customer Behavior' feature. This model creates probabilities of customers' adherence or non-adherence to regular maintenance, setting the initial conditions for our game. The model's choice is based on its ability to handle time sequences and long-term dependencies, which is crucial for using historical customer data. Its attention mechanisms help determine the importance of different time steps, enhancing prediction accuracy. These predictions are key for our reinforcement learning model, guiding it towards an optimal policy to maximize the cumulative reward for the branch workshop.

2.3. Integration of BiLSTM-Attention Mechanism and Reinforcement Learning

Integrating the BiLSTM-Attention mechanism with reinforcement learning is essential for accurately predicting customer behavior and influencing decision-making in supply chain management.

2.3.1. BiLSTM-Attention for Enhanced Feature Representation

The BiLSTM-Attention mechanism processes data bidirectionally to capture both past and future contexts. Its attention layer assigns weights to different time steps, highlighting critical information for the task. This process enhances feature representation, improving customer behavior prediction accuracy, a crucial element in the reinforcement learning model.

2.3.2. Role in Reinforcement Learning

The improved feature representations from BiLSTM-Attention serve as inputs for the reinforcement learning model, enabling more informed decisions in the dynamic supply chain environment. The detailed understanding of customer behavior enriches the model's state space, enhancing policy effectiveness.

2.3.3. Practical Application

In scenarios where various market factors influence customer behavior, the attention mechanism identifies and emphasizes these dependencies, aiding the reinforcement learning model in dynamically adjusting strategies to market demands. This capability is vital for optimizing inventory and pricing, thus enhancing the automotive supply chain's profitability and efficiency. Fig. 3 visually demonstrates how the attention mechanism supports strategic adaptation within the reinforcement learning framework.

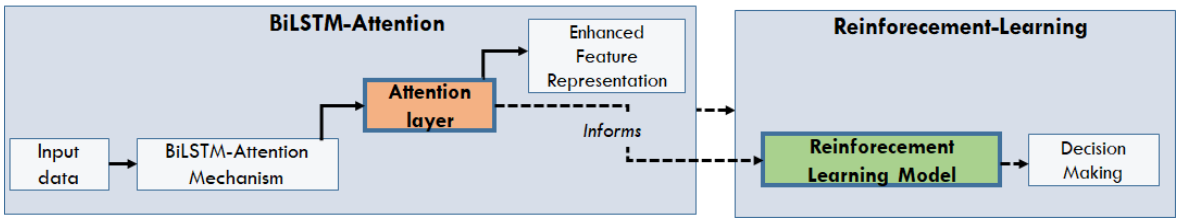

Fig. 3. BiLSTM-Attention Aided Reinforcement Learning for Market Demand Adaptation

2.4. Game Modeling

2.4.1. Game Theory Fundamentals

Introduced by John von Neumann in 1928, game theory models strategic player interactions, particularly in zero-sum games [31]. In the context of normal-form games, a game is characterized as a tuple $(n, A_1, \ldots, A_n, R_1, \ldots, R_n)$, where n is the number of players, Ak represents available actions to player k, and $R_k: A_1 \times \ldots \times A_n \rightarrow R$ is the player k's reward function, determining the payoff for a play a $\in\in A_1 \times \ldots \times A_n$ [32]. Game theory studies various game types and solution concepts like Nash equilibrium (no player gains by unilaterally changing strategy) and Pareto optimality (no player can improve without hurting others) to predict rational player behavior.

2.4.2. Game Model Setup

The game under consideration is a non-cooperative, repeated, and stochastic three-player engagement with imperfect information. Let us first introduce the symbols utilized for representing game models, as shown in Table 3.

Table 3. Notations and definitions

Variable	Definition
M	The Profit Margin for each Repair
M_c	Marketing cost
A	Average Number of Annual Visits (Customer visits to the Branch Workshop)
R	Remain Period of Warranty Coverage (Of customer car)
C_{BW}	The cost associated with adjusting the prices and managing the stock levels by Branch Workshop (Company)
S	This represents the cost savings for the customer when either the Branch Workshop or the competitor adjusts their pricing and inventory strategy. The savings are calculated as the percentage discount applied to the invoice amount, supplemented by a value δ, which stems from the successful implementation of an efficient sourcing strategy.
α_{BW}	It represents the adaptive response when the branch workshop modifies its inventory and pricing strategy. Specifically, a value greater than one ($\alpha_{BW} > 1$) indicates a significant adjustment in strategy.
α_c	Represents the strategic response when a competitor increases marketing and service enhancements investments. Specifically, an αc value greater than one ($\alpha_c > 1$) indicates a significant shift due to these competitive actions.
P	This denotes the probability of a customer choosing the competitor over the Branch Workshop. The agility variable, denoted as β, penalizes players in scenarios where any participant exerts no effort. Specifically, a β value greater than one ($\beta > 1$) indicates significant punitive measures in the event of inactivity.
D_l	Refers to the date of the last reparation in the Workshop
D_f	Refers to the date of the first use of the car
W	Refers to the total warranty period
Y_i	Represent the expected payoffs for the Branch Workshop under different conditions.

The three key players – "Customers," "Branch Workshops," and "Competitors" – act independently, repeatedly interacting over time. The stochastic nature of the game is attributed to the variable '*P*,' which represents the unpredictable decision-making of customers choosing between the Workshop and the competitor. While customers' interests may align with either the workshops or the competitors, the workshops and competitors typically conflict. Importantly, this game does not strictly follow a zero-sum dynamic, as the total gains or losses can fluctuate depending on the strategies and choices made by the players.

Based on the competitive intelligence analysis (benchmarking) conducted by the Branch Workshop, it is inferred that competitors usually adopt one of two strategic paths. These strategies involve either proactive adaptation of stock levels and pricing in response to market trends or maintaining their existing strategies in terms of stock and pricing. The specific strategies and objectives for each player within this game are detailed in Table 4.

Table 4. The game description

Players	Strategies	Objectives
Customer(Cu)	Choosing the Branch Workshop (CBW) or the Competitor (CC) refers to a customer's decision between these two service options.	Minimize the Cost
Competitors(K)	Adapts Price and Stock(APS) or Maintains Price and Stock(MPS)	Maximize the profit
Branch Workshop(BW)	Adapts Price and Stock(APS) or Maintains Price and Stock(MPS)	Maximize the profit

2.4.3. Payoff Function Definition

We detail the payoff functions for the Branch Workshop, considering various strategic interactions and outcomes. These functions incorporate factors like profit margins, marketing efforts, and customer loyalty indicators:

- Payoff Calculations: Payoffs are calculated using defined formulas, incorporating elements like the probability of customer choices, the impact of competitive actions, and operational costs.
- Objective: The primary objective is to maximize the Branch Workshop's payoff by optimizing its strategies in the competitive landscape.

The Branch Workshop payoffs are presented in Table 5.

Table 5. Branch Workshop' Payoffs

Customer	Competitors	Branch Workshop	Branch Workshop Payoffs
CBW	APS	APS	Y_1
	MPS		Y_2
	APS	MPS	Y_3
	MPS		Y_4
CC	APS	APS	Y_5
	MPS		Y_6
	APS	MPS	Y_7
	MPS		Y_8

In our study, our objective is to maximize the payoff for the Branch Workshop. For this purpose, we employed a single-agent Markov Decision Process (MDP) model, which necessitates determining only the payoffs for the Branch Workshop. We plan to formulate the other payoffs in future research, aiming for a more comprehensive analysis of the entire situation. Hence, the payoffs for the Branch Workshop are as follows:

$$Y_1 = (M - C_{BW})^* A^* R^* (1 - P)^* \beta^{2*} (1/\alpha_c)^* \alpha_{BW} \quad (4)$$

$$Y_2 = (M - C_{BW})^* A^* R^* (1 - P)^* \alpha_c{}^* \beta^{3^*} \alpha_{BW} \quad (5)$$

$$Y_3 = M^* A^* R^* (1 - P)^* (1/\alpha_c)^* (1/\beta)^3 (1/\alpha_{BW}) \quad (6)$$

$$Y_4 = M^* A^* R^* (1 - P)^* (1/\beta)^2 \alpha_c^* (1/\alpha_{BW} \quad (7)$$

$$Y_5 = (-M - C_{BW})^* A^* R^* P^* \sqrt{(1/\beta)^* \alpha_c{}^* (1/\alpha_{BW})} \quad (8)$$

$$Y_6 = (-M - C_{BW})^* A^* R^* P^* (1/\alpha_c)^* (1/\beta)^* (1/\alpha_{BW}) \quad (9)$$

$$Y_7 = (-M)^* R^* P^* \alpha_c{}^* \beta^3 \alpha_{BW} \quad (10)$$

$$Y_8 = (-M)^* R^* P^* (1/\alpha_c)^* \beta^3 \alpha_{BW} \quad (11)$$

'A' in the equation denotes the annual average of a customer's visits to the Branch Workshop, calculated by scaling their total visits since their car's first use to a yearly rate. This is achieved by multiplying the visit count by 365 and dividing by the days elapsed since the car's first use to the last repair, adjusting for non-calendar-year usage periods.

$$R = W - (\frac{D_l - D_f}{365}) \quad (12)$$

In the equation, 'R' represents the Remaining Warranty Coverage of the customer's car, computed by subtracting the elapsed time since the car's first use from the total warranty period 'W .'Time is converted to years for consistency. This calculation informs us about the remaining warranty, influencing potential repair costs and visit frequency.

$$M_c = (Annual\ Marketing\ Budget)/(Annual\ Number\ of\ Repairs) \quad (13)$$

'M_c' in the equation denotes the Marketing Costs per repair, calculated by dividing the total annual marketing budget by the annual number of repairs. This measurement gives a detailed view of marketing cost-efficiency, indicating the company's marketing spending per repair service, which is crucial for analyzing workshop profitability.

$$S = Discount\% * Invoice\ amount + \delta \quad (14)$$

'S' in the equation denotes the customer's Cost Savings when the Workshop adjusts its price and stock. It sums the discount savings (Discount % * Invoice amount) and ' δ,' representing extra savings from the Workshop's effective sourcing strategy. ' δ ' includes changes in costs of goods sold (COGS), logistics, and inventory carrying costs. Therefore, 'S' represents total customer savings via discounts and strategic sourcing.

$$C_{BW} = COGS + Operational\ Costs + Overhead\ Costs + Mc +$$

$$Logistics\ and\ Supply\ Chain\ Costs \quad (15)$$

Equations (4) to (11) represent the payoffs for the Branch Workshop under different scenarios, factoring in strategies of all players: customers, the Workshop, and competitors. The Workshop's payoff (Y_i) includes parameters α_{BW} and α_c, reflecting the Workshop's and competitors' actions, respectively. Competitor's adaptations act as a punishment factor, modifying payoffs negatively by 'α_c' or positively by '$1/\alpha_c$'. The agility parameter 'β' rewards the Workshop's good decisions and penalizes poor ones.

Payoffs (Y_5) to (Y_8) show potential losses when customers choose a competitor. The Workshop's decision quality and competitor's actions influence the loss magnitude. The agility parameter 'β' mitigates losses when good decisions are made and exacerbates it for poor decisions.

The payoffs are calculated considering the repair margin 'M,' the efforts 'C_{BW}' on stocks and prices, the annual visits 'A,' the remaining warranty period 'R,' and the probability 'P' of customers choosing the competitor.

2.4.4. Markov Decision Process (MDP) Framework

Our study uses a Markov Decision Process (MDP) to model interactions between the 'Branch Workshop,' its customers, and competitors. MDP optimizes the Workshop's decisions, factoring in customers and competitors through transition probabilities and reward functions.

This methodology reduces the complexity of the competitive environment, permitting strategic choices within the action space. The Workshop can ascertain and implement the optimal strategy for maximum long-term rewards despite external influences by harnessing the MDP in reinforcement learning. The MDP models an environment's dynamics and consists of four elements: state space, action space, transition probability function, and reward function.

Our proposed MDP components are:

- ***State space (S):*** Our model comprises all potential customer and competitor strategies combinations. Each game state is represented as a pair: (customer_strategy, competitor_strategy). We define the state space S as

$$S = \{(x, z) | \in X, z \in Z\} \quad (16)$$

  Where ' x ' denotes a specific customer strategy from set X, and ' z ' indicates a selected competitor strategy from set Z. Thus, S includes all possible pairs (x, z), providing a complete view of the strategic landscape. This state space represents potential outcomes of strategic interactions, laying the foundation for our MDP model by illustrating the 'Branch Workshop's decision-making environment

- ***Action space (A):*** Represents all possible strategies the 'Branch Workshop' can employ, including 'Adapts Price and Stock' or 'Maintains Price and Stock .'These actions represent the tactics the Workshop may choose based on the current game state. We define the action space A as:

$$A = \{a_1, a_2)\quad (17)$$

  Where A includes all possible actions or strategies for the 'Branch Workshop .' Action space A is crucial to the MDP as it outlines the Workshop's possible choices under different states. It influences the reinforcement learning agent's learning and decision-making as it seeks the optimal strategy to maximize rewards

- ***Payoff function (P):*** Quantifies the profits for the 'Branch Workshop' based on the game state (s) and the chosen action (a). It maps a state-action pair to an actual number, representing the payoff, shown as $P: S\ x\ A -> \mathbb{R}$. So, for each state-action pair (s, a), there is a real-valued payoff, expressed as:

$$P(s, a) = f(s, a) \quad (18)$$

  The payoff function f(s, a) calculates the 'Branch Workshop's profit (or regret) based on the current state (s) and selected action (a). It factors in market conditions, customer behavior, competitor actions, and operational costs, thus providing a comprehensive measure of potential profit for every possible decision

- ***Transition function (T):*** The transition function T defines how the game's state changes based on the 'Branch Workshop's actions. If the Workshop takes a certain action in state s, the game moves to a new state s'. This transition depends on factors like customer preference, competitor reaction, and market trends. So, we have

$$T(s,a) = s_{new} \quad (19)$$

However, the exact deterministic rule is data-dependent and influenced by factors such as customer preference, competitor reaction, and market dynamics. Here, (s) belongs to the state space S, and (a) pertains to the action space A. Despite R and P being mathematically alike, they have different roles. P indicates the Workshop's economic profit or loss, while R guides the learning agent, enabling it to refine its strategy via reinforcement learning.

Reinforcement learning, part of machine learning, educates agents to maximize rewards via environmental interactions. Agents optimize future actions by getting feedback and modifying their behavior. They either select the action with the top Q-value for a state or try new actions with an ε probability. Over time, the agent refines its strategy, identifying the best actions for the 'Branch Workshop.' The goal is to comprehend the best action-value function $Q(s,a)$, with 's' denoting state and 'a' action. The Q-value estimates the expected cumulative reward for an action in a state under the best policy, and the best action carries the highest Q-value. Employed techniques include:

- **Based on the Bellman equation, the Q-learning** algorithm updates the Q-table [33].

$$Q(s,a) = Q(s,a) + \alpha[R(s,a) + \gamma max_a\, Q(s',a') - Q(s,a)] \quad (20)$$

Here, α represents the learning rate, γ is the discount factor for future rewards, (s) and s' denote the current and subsequent states and a and a' signify the current and subsequent actions, respectively.

- **Based on the equation, the SARSA (State-Action-Reward-State-Action)** algorithm updates the Q-table [34].

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma\, Q(s',a') - Q(s,a)] \quad (21)$$

In this case, the next state's action value, $Q\ (s',a')$, is used instead of the maximum value among all possible following states as Q-Learning does.

- **The Deep Q-Network (DQN)** algorithm updates the Q-table through a neural network, and the Bellman equation is replaced by a loss function L that minimizes the difference between the predicted Q-values and the target Q-values [35].

$$L = E[\left(r + \gamma\, max_{a'}\, Q'(s',a';\theta_{target}\right) - Q(s,a;\theta)\Big) 2] \quad (23)$$

In the equation, r is the reward for action a in state s, γ is the future reward discount factor, and s' and a' are the next state and action. $Q(s,a;\ \theta)$ and $Q'(s,'\ a';\ \theta_{target})$ are the predicted and target Q-values, using weights θ and θ_{target}. E [ ] is the expected value over a sample of experiences.

2.4.5. Enhancing Reinforcement Learning with BiLSTM-Attention

Integrating the BiLSTM-Attention model with Q-learning enhances the precision of state-action value function, $Q(s,a)$, computations by providing a richer feature set for a more profound state understanding, leading to more accurate Q-value estimates. This detailed input helps prevent over-generalization and aids in the optimal policy convergence of the Q-learning algorithm, requiring adjustments to the learning rate, α, and discount factor, γ. Additionally, the BiLSTM-Attention model improves the DQN's ability to approximate the optimal action-value function by supplying attention-weighted inputs, increasing decision-making accuracy. This necessitates network architecture and loss function refinements to ensure stable training and prevent overfitting. Similarly, the attention mechanism benefits SARSA's on-policy approach by highlighting temporal sequences, influencing policy and value estimate updates, and allowing for a more nuanced policy space navigation, which prompts a review of policy exploration parameters to balance exploration and exploitation.

2.5. Reinforcement Learning Implementation

In alignment with the model framework detailed previously, our code implementation involves a sequence of defined steps to establish a dynamic reinforcement learning environment. This structured approach facilitates the training and assessing an agent's strategic capabilities using a suite of reinforcement learning algorithms, namely Q-Learning, Deep Q-Networks (DQN), and SARSA. To aid in interpreting the agent's strategic evolution and efficacy, the code includes mechanisms for visual representation. Here is a succinct elaboration of the code implementation process show as Fig. 4:

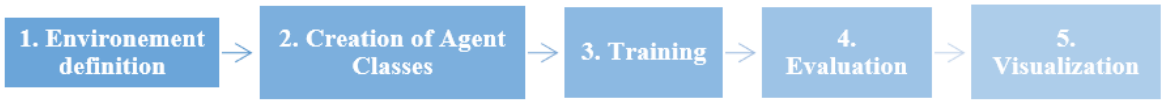

Fig. 4. Reinforcement learning models construction

Fig. 4 presents the reinforcement learning model construction, which is performed as follows:

- Environment Class Definition: This initial phase involves constructing the "WorkshopEnvironment" class, which is crucial for simulating the business context within our model. The environment is designed to integrate customer behavior predictions made by the Bi-LSTM-attention model. These predictions become part of the environment's state and influence the potential actions and resulting rewards. This step also includes the integration of the strategic decisions, modeled through the payoff functions (Y1 to Y8), which are informed by the predicted customer behavior.
- Creation of Agent Classes: Following the environment definition, the next step is the creation of agent classes, namely, "QLearningAgent," "DQNAgent," and "SARSAAgent." These classes are constructed to operate within the defined environment. The agents are crafted with the ability to interpret the enriched states, which now encapsulate both the traditional state variables and the new 'Behavior' feature predicted by the BiLSTM-attention model.
- Training the QLearningAgent: With the agents created, we move on to the training phase, where the "QLearningAgent" and other agents learn to navigate the environment. They make decisions based on the states, which include customer behavior predictions, and refine their strategies to optimize the expected payoffs according to the MDP framework. Training is an iterative process where agents update their policies in response to the rewards received from the environment.
- Performance Evaluation: Post-training, we evaluate the agents' performance within the environment, now with a fully informed perspective that includes the impact of customer behaviors on the payoffs. The agents' performance metrics are scrutinized to assess strategic depth, decision-making quality, and proficiency in maximizing the Branch Workshop's profits.
- Visualization of Results: We visualized the agent's learning progress and outcomes. We graphed the rewards (profits) obtained in each episode during the training and the average profits per action or strategy. The visualization gave us an understanding of the agent's decision-making progress over time and its ability to navigate the state-action space of our MDP, thereby assisting in further refinement of our models and business strategies

2.6. Decision Making

With the integration of our computational framework, powered by a BiLSTM-Attention deep learning model, into the business platform, businesses like "Branch Workshops" now have an advanced strategic decision-making tool at their disposal. This framework utilizes potential business actions as inputs, each representing a unique state in the dynamic business environment. These states cover a broad spectrum of scenarios, incorporating the following factors: market dynamics, customer behaviors, competitor actions, operational profits, and costs.

At the heart of this framework is a reinforcement learning algorithm that actively interacts with the business environment. It executes actions and subsequently receives rewards based on the derived profits. This feedback cycle guides the algorithm, promoting it to learn and select the most profitable actions.

The BiLSTM-Attention model plays a pivotal role in this integrated framework. It processes historical and real-time data to predict customer behaviors and market conditions, providing valuable input to the reinforcement learning algorithm. This predictive ability equips the algorithm to make more informed, efficient decisions, leading to optimized profitability.

The integrated framework, now part of the operational platform, outputs optimal business actions tailored to different situations. The learning algorithm's iterative process continually refines these actions, designed to maximize profitability amidst shifting market conditions.

This integration necessitates real-time access to the "Branch Workshops" and market data, enabling the framework to consistently update and fine-tune its advice. As the business interacts with the market, the model's learning capability facilitates the refinement of strategies, delivering progressively precise recommendations.

The model also offers visualization tools, including metrics like rewards per episode during the training phase and average rewards per action. Displayed on the company's platform, these insights assist management in understanding the implications of their decisions, serving as a resource for continuous strategy enhancement.

3. Results and Discussion

3.1. Results

3.1.1. Computational Environment

The computation was performed on a Jupyter Notebook platform equipped with an Intel(R) CORE(TM) i5 processor, 8 GB of RAM, Intel UHD Graphics GPU, and Windows 10 Pro 64-bit operating system.

3.1.2. In-depth Analysis of the BiLSTM-Attention Model

Fig. 5 demonstrates that our BiLSTM-Attention model effectively predicts customer behavior, significantly reducing errors in training and validation sets after just a few epochs. This model is particularly adept at capturing long-term dependencies and focusing on crucial sequence parts, which enhances its accuracy. The decline in error rates underscores the model's robust generalization capabilities and proficient handling of new data, effectively mitigating overfitting.

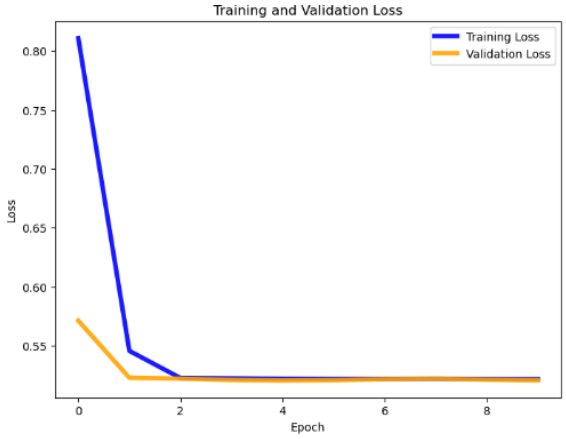

Fig. 5. Line plot of train and validation loss from the BiLSTM-Attention

Furthermore, the most striking result is the BiLSTM-Attention model, achieving the lowest MSE (0.0525) and the highest R2 value (0.896) when compared with traditional LSTM and BiLSTM models

(Table 6). This indicates a significant reduction in prediction errors and showcases the model's superior ability to account for nearly 90% of the variability in the target variable, which is a substantial improvement in predictive modeling accuracy.

Table 6. Errors results depend on the model

	LSTM	BiLSTM
MSE	0.0730	0.0723
R2	0.881	0.873

These findings align with the current body of literature, such as the studies by [36] and [37], which underscore the enhanced performance of BiLSTM over traditional LSTM models in sequence prediction tasks. This superiority is further supported by broader trends in deep learning research, where attention mechanisms have been recognized for their substantial contributions to improving model accuracy and overall performance [9], [38]. While our study does not directly compare our model to traditional forecasting models like AutoRegressive Integrated Moving Average (ARIMA), it is essential to note that our BiLSTM-Attention approach, with its inherent flexibility and ability to handle complex data uncertainties, represents a significant advancement over ARIMA's linear methodology, which often falls short in dealing with intricate data complexities [39].

This improvement is pivotal as it significantly enhances the model's ability to capture and predict complex patterns, outperforming existing models. It illustrates the effectiveness of integrating attention mechanisms with BiLSTM in capturing crucial long-term dependencies, a key advancement in predictive analytics.

3.1.3. Reinforcement learning models

This study compared three reinforcement learning algorithms: Q-Learning, Deep Q-Networks (DQN), and SARSA, using the metrics of total reward, average reward per episode, and cumulative reward. Each algorithm was tested in a simulated environment and evaluated on these metrics. Fig. 6 graphically depicts each algorithm's rewards and average rewards per episode. It shows that Q-Learning and SARSA performed better than DQN, yielding similar total and average rewards per episode results.

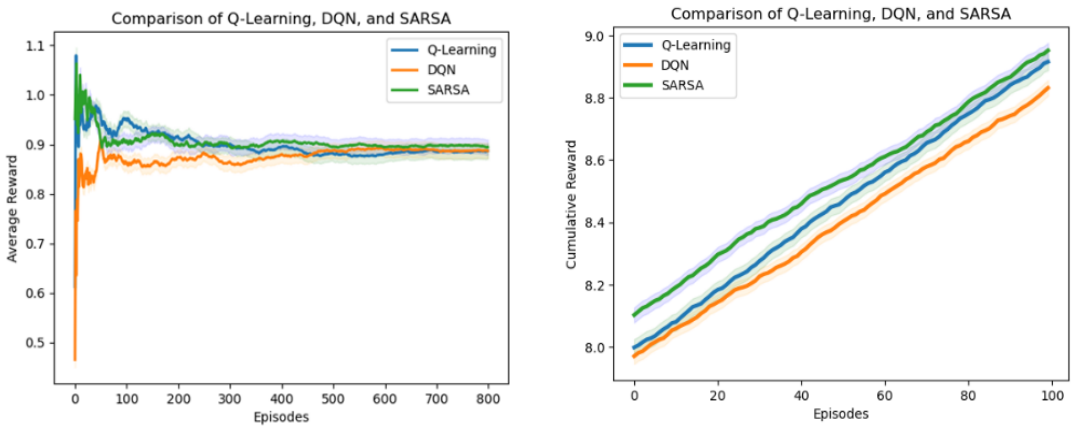

Fig. 6. Reward and the average reward per episode regarding the model

The study also used average and cumulative rewards to compare Q-Learning, DQN, and SARSA. The results in Fig. 7 revealed that SARSA performed best, mainly due to its stable on-policy learning approach, reduced overestimation of action values, and quick adaptation to changing environments.

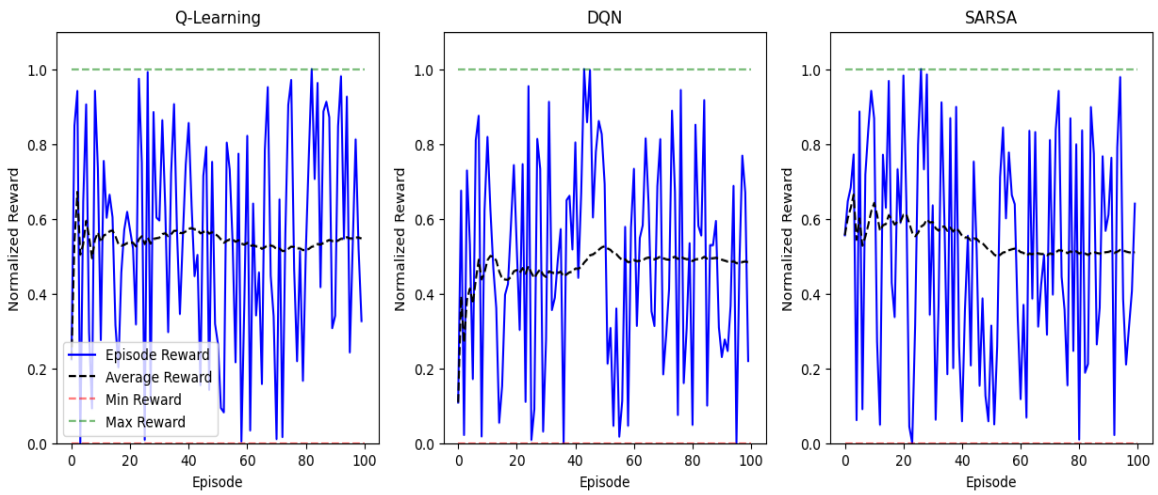

Fig. 7. Average and Cumulative Reward per episode

Fig.7 displays a clear trend of increasing cumulative rewards per episode for DQN, SARSA, and Q-Learning, showcasing their capability in learning and strategic decision-making. Notably, action 1 consistently outperforms action 0 in terms of cumulative rewards, with this preference being more pronounced in DQN and SARSA compared to Q-Learning. SARSA's on-policy learning framework stands out for its steady selection of action 1, underscoring its agility in adapting to environmental shifts and its proficiency in balancing the dual objectives of exploration and exploitation, thereby securing favorable outcomes.

These findings align with and expand upon the work of [40], [41], who have underscored the benefits of on-policy learning in dynamic environments. By comparing on-policy (SARSA) and off-policy (DQN and Q-Learning) algorithms, our research provides deeper insights into their strengths and applicability in real-world scenarios, highlighting the versatility and resilience of these methods. The adaptability and success of these reinforcement learning strategies in various industrial contexts illustrate their potential to offer dynamic and robust solutions across various applications. Among these, SARSA distinguishes itself, demonstrating superior performance in cumulative rewards, attributed to its strategic decision-making, and highlighting its potential as a preferred algorithm in environments where on-policy learning is crucial.

3.2. Discussion

3.2.1. Insights from the BiLSTM-Attention Model

Integrating the attention mechanism with the BiLSTM model significantly enhances the model's ability to interpret and predict complex behavioral patterns accurately, offering a considerable advancement over traditional LSTM models. This development is a testament to the model's increased technical sophistication and ability to handle more nuanced and detailed data. Furthermore, it signifies a broader shift within the industry toward more advanced and intricate predictive models. This evolution reflects a growing emphasis on leveraging deep insights and achieving higher precision in predictive analytics, aligning with broader industry trends toward sophisticated data-driven strategies, as discussed by various experts in the field [36], [37].

3.2.2. Strategic Implications of Reinforcement Learning Findings

Our study's exploration of reinforcement learning provides critical strategic insights, especially in selecting and applying algorithms to meet specific industry challenges. SARSA's exceptional performance in our analysis indicates a strategic preference for on-policy learning methods in dynamic and unpredictable settings. This preference aligns with strategic considerations in algorithm selection for real-world applications, as discussed by leading experts and researchers in the field, highlighting the importance of adaptable and robust learning strategies in complex environments [40], [41].

3.2.3. Implications of the proposed framework and future directions

The application of these models extends beyond theoretical implications, offering tangible benefits in strategic decision-making, resource allocation, and predictive analytics within the industry. These practical applications mirror the advancements in AI deployment in business settings. In addition, the findings contribute to a deeper understanding of how advanced modeling techniques can be tailored to specific industry needs, enhancing efficiency and strategic foresight. Adopting these models can lead to more informed decision-making processes, aligning with the industry trends toward data-driven strategies.

Our study's insights are framed within the boundaries of a specific computational setup and the chosen dataset. One area for improvement is the study's focus on a particular set of modeling techniques without exploring a more comprehensive array of potential methods, which may offer different insights or advantages. Additionally, our analysis should have extended to include multifaceted data elements like customer sentiment, which could provide a more comprehensive understanding of consumer behavior dynamics. Future research should broaden the scope of the methodologies examined, comparing the effectiveness of various predictive models in similar contexts. There is also a significant opportunity to delve into the impact of reinforcement learning models in more complex and dynamic environments. Integrating a broader spectrum of data factors, such as customer sentiment, in subsequent studies could enhance the depth and relevance of the models, offering a more nuanced perspective on the predictive patterns and behaviors observed.

4. Conclusion

specifically SARSA, in refining the decision-making processes within the automotive distribution industry. This optimization is supplemented by using BiLSTM-Attention models, contributing to accurate feature prediction. The strength of our study lies in its practical implications and potential to inform future research directions. One significant contribution is the opportunity to implement machine learning algorithms to discern patterns and correlations between variables such as customer loyalty, brand reputation, pricing strategies, and inventory decisions. By doing so, we can adjust our game modeling approach to incorporate the potential impact of these factors on the decision-making process and the inherent effects of market volatility and uncertainty. Another contribution is the potential use of deep learning models to anticipate market trends, competitor pricing strategies, demand forecasts, and inventory optimization under many scenarios. This allows us to account for the market's volatile nature and competitors' uncertain actions. This direction can lead to developing more robust models resilient to market uncertainties and capable of providing accurate predictions. Our proposed future research directions promise to surmount current limitations and further boost the performance of our models. This, in turn, leads to a more holistic understanding of optimal inventory and pricing actions in the automotive distribution industry, equipping it to navigate market volatility and uncertainty successfully. Ultimately, our work presents an integrated, data-driven strategy to revolutionize management in the automotive distribution industry, promoting efficiency, profitability, and competitiveness.

Declarations

Author contribution. Each listed Author has contributed to the work within their area of expertise
Funding statement. No funding from public, commercial, or non-profit agencies was allocated for this research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] S. Chopra and P. Meindl, “Supply Chain Management. Strategy, Planning & Operation,” in *Das Summa Summarum des Management*, Wiesbaden: Gabler, 2007, pp. 265–275, doi: 10.1007/978-3-8349-9320-5_22.

[2] C. Martin, *Logistics and Supply Chain Management*. p. 13, 2011. [Online]. Available at: https://industri.fatek.unpatti.ac.id/wp-content/uploads/2019/03/256-Logistics-Supply-Chain-Management-Martin-Christopher-Edisi-1.pdf.

[3] P. D. Larson, "Designing and Managing the Supply Chain: Concepts, Strategies, and Case Studies, David Simchi-Levi Philip Kaminsky Edith Simchi-Levi," *J. Bus. Logist.*, vol. 22, no. 1, pp. 259–261, Mar. 2001, doi: 10.1002/j.2158-1592.2001.tb00165.x.

[4] A. Amellal, I. Amellal, H. Seghiouer, and M. R. Ech-Charrat, "Improving Lead Time Forecasting and Anomaly Detection for Automotive Spare Parts with A Combined CNN-LSTM Approach," *Oper. Supply Chain Manag. An Int. J.*, vol. 16, no. 2, pp. 265–278, Jun. 2023, doi: 10.31387/oscm0530388.

[5] I. Amellal, A. Amellal, H. Seghiouer, and M. R. Ech-Charrat, "An integrated approach for modern supply chain management: Utilizing advanced machine learning models for sentiment analysis, demand forecasting, and probabilistic price prediction," *Decis. Sci. Lett.*, vol. 13, no. 1, pp. 237–248, 2024, doi: 10.5267/j.dsl.2023.9.003.

[6] F. Kurniawan, S. Sulaiman, S. Konate, and M. A. A. Abdalla, "Deep learning approaches for MIMO time-series analysis," *Int. J. Adv. Intell. Informatics*, vol. 9, no. 2, p. 286, Jul. 2023, doi: 10.26555/ijain.v9i2.1092.

[7] H. Haviluddin and R. Alfred, "Multi-step CNN forecasting for COVID-19 multivariate time-series," *Int. J. Adv. Intell. Informatics*, vol. 9, no. 2, p. 176, Jul. 2023, doi: 10.26555/ijain.v9i2.1080.

[8] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention Is All You Need In Speech Separation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, vol. 2021-June, pp. 21–25, doi: 10.1109/ICASSP39728.2021.9413901.

[9] Y. Oukdach, Z. Kerkaou, M. El Ansari, L. Koutti, A. Fouad El Ouafdi, and T. De Lange, "ViTCA-Net: a framework for disease detection in video capsule endoscopy images using a vision transformer and convolutional neural network with a specific attention mechanism," *Multimed. Tools Appl.*, pp. 1–20, Jan. 2024, doi: 10.1007/s11042-023-18039-1.

[10] X. Qiao *et al.*, "A Event Extraction Method of Document-Level Based on the Self-attention Mechanism," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13656 LNCS, Springer, Cham, 2023, pp. 609–619, doi: 10.1007/978-3-031-20099-1_50.

[11] N. Moritz, T. Hori, and J. Le Roux, "Triggered Attention for End-to-end Speech Recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, vol. 2019-May, pp. 5666–5670, doi: 10.1109/ICASSP.2019.8683510.

[12] G. Chaubey, P. R. Gavhane, D. Bisen, and S. K. Arjaria, "Customer purchasing behavior prediction using machine learning classification techniques," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 12, pp. 16133–16157, Dec. 2023, doi: 10.1007/s12652-022-03837-6.

[13] P. Kaushik, S. P. Singh Rathore, P. Kaur, H. Kumar, and N. Tyagi, "Leveraging Multiscale Adaptive Object Detection and Contrastive Feature Learning for Customer Behavior Analysis in Retail Settings," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 6s, pp. 326–343, Jun. 2023, doi: 10.17762/ijritcc.v11i6s.6938.

[14] M. Z. Abedin, P. Hajek, T. Sharif, M. S. Satu, and M. I. Khan, "Modelling bank customer behaviour using feature engineering and classification techniques," *Res. Int. Bus. Financ.*, vol. 65, p. 101913, Apr. 2023, doi: 10.1016/j.ribaf.2023.101913.

[15] N. Fatehi, A. Politis, L. Lin, M. Stobby, and M. H. Nazari, "Machine Learning based Occupant Behavior Prediction in Smart Building to Improve Energy Efficiency," in *2023 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, Jan. 2023, pp. 1–5, doi: 10.1109/ISGT51731.2023.10066411.

[16] Z. Zhang, "Consumer behavior prediction and marketing strategy optimization based on big data analysis," *Appl. Math. Nonlinear Sci.*, vol. 9, no. 1, pp. 1–14, Jan. 2024, doi: 10.2478/amns.2023.2.01630.

[17] J. Wen and Z. Wang, "Short-term load forecasting with bidirectional LSTM-attention based on the sparrow search optimisation algorithm," *Int. J. Comput. Sci. Eng.*, vol. 26, no. 1, p. 20, 2023, doi: 10.1504/IJCSE.2023.129154.

[18] W. Gomez, F.-K. Wang, and Z. E. Amogne, "Electricity Load and Price Forecasting Using a Hybrid Method Based Bidirectional Long Short-Term Memory with Attention Mechanism Model," *Int. J. Energy Res.*, vol. 2023, no. 1, pp. 1–18, Feb. 2023, doi: 10.1155/2023/3815063.

[19] M. O. Ahmed and I. H. El-adaway, "An integrated game-theoretic and reinforcement learning modeling for multi-stage construction and infrastructure bidding," *Constr. Manag. Econ.*, vol. 41, no. 3, pp. 183–207, Mar. 2023, doi: 10.1080/01446193.2022.2124528.

[20] G. Cui, Q.-S. Jia, and X. Guan, "Energy Management of Networked Microgrids With Real-Time Pricing by Reinforcement Learning," *IEEE Trans. Smart Grid*, vol. 15, no. 1, pp. 570–580, Jan. 2024, doi: 10.1109/TSG.2023.3281935.

[21] X. Li, F. Luo, and C. Li, "Multi-agent deep reinforcement learning-based autonomous decision-making framework for community virtual power plants," *Appl. Energy*, vol. 360, p. 122813, Apr. 2024, doi: 10.1016/j.apenergy.2024.122813.

[22] A. C. Real, G. P. Luz, J. M. C. Sousa, M. C. Brito, and S. M. Vieira, "Optimization of a photovoltaic-battery system using deep reinforcement learning and load forecasting," *Energy AI*, vol. 16, p. 100347, May 2024, doi: 10.1016/j.egyai.2024.100347.

[23] H. Markgraf and M. Althoff, "Safe Multi-Agent Reinforcement Learning for Price-Based Demand Response," in *2023 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*, Oct. 2023, pp. 1–6, doi: 10.1109/ISGTEUROPE56780.2023.10407281.

[24] J. Lussange, S. Vrizzi, S. Bourgeois-Gironde, S. Palminteri, and B. Gutkin, "Stock Price Formation: Precepts from a Multi-Agent Reinforcement Learning Model," *Comput. Econ.*, vol. 61, no. 4, pp. 1523–1544, Apr. 2023, doi: 10.1007/s10614-022-10249-3.

[25] R. May and P. Huang, "A multi-agent reinforcement learning approach for investigating and optimising peer-to-peer prosumer energy markets," *Appl. Energy*, vol. 334, p. 120705, Mar. 2023, doi: 10.1016/j.apenergy.2023.120705.

[26] Y. Han, X. Zhang, J. Zhang, Q. Cui, S. Wang, and Z. Han, "Multi-Agent Reinforcement Learning Enabling Dynamic Pricing Policy for Charging Station Operators," in *2019 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2019, pp. 1–6, doi: 10.1109/GLOBECOM38437.2019.9013999.

[27] L. Yu, C. Zhang, J. Jiang, H. Yang, and H. Shang, "Reinforcement learning approach for resource allocation in humanitarian logistics," *Expert Syst. Appl.*, vol. 173, p. 114663, Jul. 2021, doi: 10.1016/j.eswa.2021.114663.

[28] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[29] K. Smagulova and A. P. James, "A survey on LSTM memristive neural network architectures and applications," *Eur. Phys. J. Spec. Top.*, vol. 228, no. 10, pp. 2313–2324, Oct. 2019, doi: 10.1140/epjst/e2019-900046-x.

[30] G. Brauwers and F. Frasincar, "A General Survey on Attention Mechanisms in Deep Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3279–3298, Apr. 2023, doi: 10.1109/TKDE.2021.3126456.

[31] V. N. John and M. Oskar, *Theory of games and economic behavior*, 2nd ed. United States: Princeton University Pres, p. 625,1947. [Online]. Available at: https://psycnet.apa.org/record/1947-03159-000.

[32] A. Nowé, P. Vrancx, and Y.-M. De Hauwere, "Game Theory and Multi-agent Reinforcement Learning," in *Adaptation, Learning, and Optimization*, vol. 12, Springer, Berlin, Heidelberg, 2012, pp. 441–470, doi: 10.1007/978-3-642-27645-3_14.

[33] C. J. C. H. Watkins and P. Dayan, "Technical Note: Q-Learning," *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, 1992, doi: https://doi.org/10.1023/A:1022676722315, doi: 10.1023/A:1022676722315.

[34] R. S. Sutton and A. G. Barto, *Reinforcement learning : an introduction*, 2nd ed. Cambridge: The MIT Press, p. 526, 2015. [Online]. Available at: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf.

[35] V. Mnih *et al.*, "Playing Atari with Deep Reinforcement Learning," *arxiv Mach. Learn.*, pp. 1–9, Dec. 2013. [Online]. Available: https://arxiv.org/abs/1312.5602v1.

[36] A. El Zaar, N. Benaya, T. Bakir, A. Mansouri, and A. El Allati, "Prediction of US 30-years-treasury-bonds movement and trading entry point using the robust 1DCNN-BiLSTM-XGBoost algorithm," *Expert Syst.*, vol. 41, no. 1, p. e13459, Jan. 2024, doi: 10.1111/exsy.13459.

[37] Z. Jamshidzadeh, M. Ehteram, and H. Shabanian, "Bidirectional Long Short-Term Memory (BILSTM) - Support Vector Machine: A new machine learning model for predicting water quality parameters," *Ain Shams Eng. J.*, vol. 15, no. 3, p. 102510, Mar. 2024, doi: 10.1016/j.asej.2023.102510.

[38] S. Lv, K. Wang, H. Yang, and P. Wang, "An origin–destination passenger flow prediction system based on convolutional neural network and passenger source-based attention mechanism," *Expert Syst. Appl.*, vol. 238, p. 121989, Mar. 2024, doi: 10.1016/j.eswa.2023.121989.

[39] G. E. P. Box, G. C. Reinsel, G. M. Jenkins, and G. M. Ljung, *Time series analysis: forecasting and control*, 5th ed. Canada: John Wiley & Sons, Inc, p. 720, 2016. [Online]. Available at: http://link.springer.com/10.1007/978-3-319-59379-1%0Ahttp://dx.doi.org/10.1016/B978-0-12-420070-8.00002-7%0Ahttp://dx.doi.org/10.1016/j.ab.2015.03.024.

[40] J.-Y. Lee, A. Rahman, S. Huang, A. D. Smith, and S. Katipamula, "On-policy learning-based deep reinforcement learning assessment for building control efficiency and stability," *Sci. Technol. Built Environ.*, vol. 28, no. 9, pp. 1150–1165, Oct. 2022, doi: 10.1080/23744731.2022.2094729.

[41] T. Cui, N. Du, X. Yang, and S. Ding, "Multi-period portfolio optimization using a deep reinforcement learning hyper-heuristic approach," *Technol. Forecast. Soc. Change*, vol. 198, p. 122944, Jan. 2024, doi: 10.1016/j.techfore.2023.122944.