Customer In-Store Behavior Analysis Using Beacon Data at a Home Improvement Retailer

Original Scientific Paper

Ayla Gülcü*
Bahçeşehir University,
Faculty of Engineering and Natural Sciences,
Department of Software Engineering
Istanbul, TÜRKIYE
ayla.gulcu@bau.edu.tr

İnanç ONUR
Patika Global Technology,
Istanbul, TÜRKIYE
inanc.onur@patikaglobal.com

Sümeyra Öztop
Patika Global Technology,
Istanbul, TÜRKIYE
sumeyra.oztop@patikaglobal.com

*Corresponding author

Enes Uğurlu
Patika Global Technology,
Istanbul, TÜRKIYE
enes.ugurlu@patikaglobal.com

Remzi Emre Sain
Patika Global Technology,
Istanbul, TÜRKIYE
remzi.sain@patikaglobal.com

Abstract *– In this study, we aimed to analyze the in-store behavior of customers at a home improvement retail company using data collected from Bluetooth Low Energy beacon devices installed on shelves and shopping carts within a selected store. The beacons were strategically placed on store shelves to ensure complete coverage, leaving no blind spots. To cover 18 departments spanning a total area of approximately 4,800 square meters, 99 beacons were deployed. The duration of stay in each department, the order of visits, and the absolute visit date and time were recorded in the database. To investigate the relationship between in-store behavior and purchase data, we combined customers' behavioral data with their purchase information. Correlation analysis revealed a positive relationship between visit duration and purchase amount, particularly in the Floor Deco, Paint, and Taps departments. Additionally, we visualized store-wide data using a network diagram, highlighting key shopping areas, customer flow patterns, and high-revenue departments. The problem was also formulated as a multi-class classification task, and LSTM and XGBoost algorithms were applied for comparative analysis. Experiments were conducted on both the original dataset and a cleaned version, utilizing two distinct data modeling approaches: one based solely on sequential department visits and another incorporating visit duration. The results showed that both models performed similarly on the noisy dataset, indicating that adding duration information did not improve learning. However, when trained on the cleaned dataset where short-duration visits were removed, LSTM models outperformed XGBoost, demonstrating a stronger ability to capture meaningful sequential patterns. These findings highlight the potential of BLE beacon technology in retail analytics, offering deeper insights into customer behavior and informing data-driven decision-making for store optimization and personalized marketing. Future work will focus on expanding the dataset and refining predictive models to further enhance the accuracy and applicability of in-store behavior analysis.*

Keywords: *BLE Beacon localization, data analysis, internet of things, machine learning*

Received: December 28, 2024; Received in revised form: February 21, 2025; Accepted: March 31, 2025

1. INTRODUCTION

Companies can gain valuable insights by analyzing customer behavior data. By leveraging data mining and machine learning techniques, businesses can identify high-value customers, capture their personalized needs, and implement effective marketing strategies.

Zhao [1] highlights that personalized marketing, based on extensive customer data analysis, can significantly enhance marketing campaigns, with conversion rates being 40% higher compared to non-personalized approaches. Traditionally, customers' purchase histories and Point of Sales (POS) data have been the primary tools for analyzing customer segments, especially when combined with demographic or geographic data. However, incorporating behavioral factors such as shopping duration, wandering routes, and visiting patterns can provide deeper insights into customer preferences and behaviors.

Behavioral data can help answer critical business questions, as discussed in [2], such as: What are the most frequent paths taken by customers? Which areas are skipped? Where do they spend the majority of their time? Do they primarily navigate the outer ring of the store, or do they focus on specific sections?

The Internet of Things (IoT) has become an essential tool, particularly in the retail sector, for gathering customer behavior data (see [3] for a review of the digital transformation across sectors). IoT sensors placed on store shelves or shopping carts can be utilized to analyze customers' shopping behaviors. Additionally, smart mobile applications can assist customers with in-store navigation. Various IoT technologies are employed for in-store customer tracking and localization, including: i) Bluetooth Low Energy (BLE) Beacon technology, ii) Ultra-Wideband (UWB) technology, iii) Wayfinding technologies, and iv) Camera and image processing technologies. BLE Beacon technology is particularly favored for its low energy consumption and cost-effectiveness.

Ke et al. [4] compare various Radio-Frequency (RF)-based indoor positioning technologies. They propose a BLE Beacon-based localization system designed to detect human locations in smart homes for power management purposes. Another application of indoor positioning is discussed in [5]. Spachos and Plataniotis [6] utilize BLE Beacons in an IoT-based smart museum to deliver interactive and personalized museum tours. Similarly, Shipkovenski et al. [7] present an indoor positioning system for hospital environments, enabling patient tracking and preventing unauthorized departures. Pangriya [8] provides a comprehensive literature review on Beacon technology, offering valuable insights into its future in the retail sector. The author also presents the results of a SWOT analysis (Strengths, Weaknesses, Opportunities, and Threats), conducted through two-phase interviews with experts from the technology and retail industries. Among the strengths of Beacon technology are its cost-effectiveness and ability to function indoors with high precision. However, its primary weakness is the requirement for customers to keep Bluetooth enabled and the store's mobile application running while in the store.

In this study, we analyze the in-store behavior of customers at a home improvement retailer using data collected from BLE beacon devices installed on shelves and shopping carts. We integrate this behavior data with purchase data to uncover key customer characteristics. We formulate the problem as a multi-class classification task and apply two distinct algorithms comparatively using two different models. Our approach differs from prior studies by introducing a novel data modeling technique that integrates both sequential department visit data and visit duration. Unlike studies that focus on a single department, our analysis spans 18 departments, providing a more comprehensive view of customer behavior. Unlike studies that focus on customer behavior within a single department, this research considers data spanning 18 departments, providing a more comprehensive understanding of customer behavior.

The remainder of the paper is organized as follows: Section 2 reviews related studies. Section 3 details beacon installation, data collection, and modeling approaches. Section 4 presents correlation analysis, visualizations, and algorithm comparisons. Finally, Section 5 concludes with key findings and remarks.

2. LITERATURE REVIEW

This section focuses on studies that utilize data collected from Beacon devices to infer user or customer behavior. While earlier studies primarily relied on Radio Frequency Identification (RFID) tags, recent advancements have led to the adoption of Bluetooth Low Energy (BLE) Beacons due to their cost-effectiveness, ease of deployment, and ability to function efficiently in real-world retail environments.

One of the earliest studies on customer in-store analysis using RFID was conducted by Larson et al. [2], who identified frequent shopping routes in a grocery store. They analyzed 27,000 shopper paths, with path lengths varying from short 2-minute routes to longer 2-hour shopping patterns. Their findings demonstrated that certain store layouts influence customer movement, providing insights into strategic shelf placements.

Recent advancements have led to the adoption of BLE Beacons, which offer a more scalable and flexible approach to in-store tracking. Unlike RFID, BLE Beacons can continuously monitor customer movement without requiring direct scanning or RFID-tagged objects. Instead, they rely on smartphone interactions or strategically placed beacons to track customer flow. Jain et al. [9] propose a comprehensive framework for evaluating the adoption and readiness of beacon technology. They highlight a significant gap in the field of proximity marketing and beacon technology and suggest that their study can help retail managers better understand the added value of implementing this technology.

Lemsieh and Abarar [10] conducted a study on customer perceptions of proximity marketing using BLE beacon technology in Moroccan malls. The results indicate that customers are eager to adopt this technology, as it helps them locate stores independently, provided their privacy concerns are adequately addressed. Due to these concerns, most studies focus on using BLE beacon technology for indoor localization of customers or personnel within a facility rather than for delivering personalized experiences. Garcia and Inoue [11] presents a study on the indoor localization of caregiving and nursing staff in a nursing Care facility using BLE Beacon technology. Spachos and Plataniotis [6] demonstrated the effectiveness of BLE Beacons in guiding museum visitors through location-aware tours. In another study, Shende et al. [12] utilized BLE Beacons to analyze customer movements in a shopping mall, enabling real-time personalized promotions based on movement patterns.

For personalized marketing, BLE beacon technology which can be seamlessly integrated with mobile phones is essential. In this study, we used this technology to collect customer in-store behavior data in a home improvement retailer. Analyzing customer behavior data helped to answer many business questions. We also combined the behavior data with the purchase data in order to investigate the relationship between them. We formulated the problem as a multi-class classification task, and employed LSTM and XGBoost algorithms. We introduced a novel data modeling approach that captures both the sequence of department visits and the duration of those visits.

Most studies in the literature that apply machine learning algorithms to analyze customer behavior data rely on data collected using RFID technology. More importantly, many of these studies focus on a single department within a store, whereas our study utilizes data from 18 departments. For instance, Zhao et al. [13] proposed a sequential classification-based model that integrates RFID data with point-of-sale (POS) data to classify and identify consumers' purchasing behavior, but their analysis was limited to a specific island area of a supermarket. Similarly, Zuo et al. [14] combined RFID-based customer visit data with POS transactions in a Japanese supermarket, focusing solely on the fish section. Extending this approach, Zuo et al. [15] placed RFID tags on shopping carts and installed receptors beneath store shelves to track customers' real-time interactions with different product aisles, though their study was confined to the bread section.

There are also studies that employ classification algorithms to analyze customer behavior in the e-commerce domain. For example, Li [16] utilized customers' past product browsing activity data to construct a classification model based on the XGBoost algorithm to predict the customer purchasing behavior in the e-commerce domain. Similarly, Park et al. [17] applied the same algorithm to predict the product a customer would purchase based on their browsing activity.

Our study differs from existing literature by incorporating sequential data into the classification task. Additionally, we introduce a novel modeling approach that integrates both sequential data and numerical duration information. Furthermore, rather than focusing on a single department, we utilize data from 18 departments, enabling a comprehensive analysis across the entire dataset.

3. METHODOLOGY

In this section, we first describe the BLE beacon installation process within the retail store. Next, we outline the localization and data collection procedures. Finally, we detail the data modeling approach employed and the machine learning algorithms applied in this study.

3.1. BEACON INSTALLATION AND DATA COLLECTION

Noise Filtering Process: Beacons transmit signals at regular intervals using Bluetooth Low Energy (BLE) technology. In this study, the system includes two types of beacons: shelf beacons and cart beacons. Shelf beacons act as transmitters, emitting signals within a defined area and sending them to nearby devices. Cart beacons function as receivers, capturing and recording signals emitted by the shelf beacons that correspond to the aisle where the customer's cart is located. This setup enables the system to accurately track the customer's location within the store.

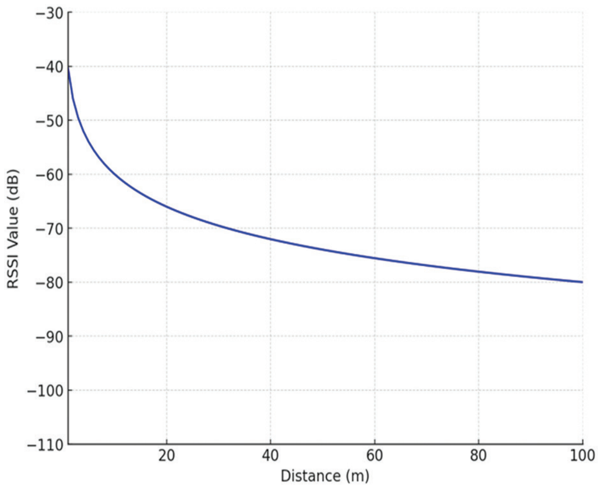

Fig. 1. The relationship between RSSI values and distance

The Received Signal Strength Indicator (RSSI) value, measured in decibels (dB), represents signal strength and is used to estimate the distance between a device and a beacon. As the distance between the transmitting beacon and the receiving device increases, the RSSI value decreases (becomes more negative) due to signal attenuation, a phenomenon where the signal's energy dissipates over distance. Fig. 1 illustrates how RSSI values vary with distance. RSSI data obtained from beacon devices provides a measure of signal strength; however, it is susceptible to environmental factors. This phenomenon, known as noise, can cause fluctuations in signal strength, reducing accuracy. Noise may originate from various sources, such as other electronic devices, structural obstacles (e.g., walls and shelves), human crowds, and even continuous variations in the distance between the beacon and the receiver.

The optimal placement of transmitter beacon devices on store shelves, which communicate with receiver beacon devices in customers' shopping carts, is crucial for ensuring effective communication. Extensive experiments were conducted in a controlled physical environment to determine the optimal signal range while minimizing interference from environmental noise. Table 1 summarizes the data collected by receiver beacons during the experimental visits.

Table 1. Experiments to determine the RSSI filter value

Id	RSSI Filter (dB)	Beacon Region	Interaction Duration	Diff Time
D1	-84	Lighting	1727	899
D1	-84	Promo	1727	899
D2	-82	Lighting	1160	922
D2	-82	Promo	1160	922
D3	-80	Lighting	402	437
D3	-80	Promo	402	437

Two departments, namely Lighting and Promo, were visited three times during the experiments. D1, D2, and D3 represent these visits, each conducted under different RSSI filter values. The Interaction Duration column indicates the total time the beacon received signals from transmitter beacons in the surrounding area, while the Diff Time column reflects the duration the beacon received signals irrespective of the number of beacons nearby. In the experimental visit D1 with a filter value of –84 dB, the Interaction Duration value was significantly higher than the Diff Time value. This discrepancy suggests that the beacon was receiving noisy signals from beacons installed in other departments. In contrast, experimental visit D3 with a filter value of –80 dB yielded similar values for Interaction Duration and Diff Time, indicating minimal noise interference. Thus, a filter value of –80 dB was determined to be optimal for this environment.

Beacon Localization: After determining a noise filter value of –80 dB, the distance between any two beacons needed to be defined. Since the areas of the departments varied, the number of beacons installed in each department also differed. For a given department, beacon placement was performed manually by measuring signal strength as follows: The tangent point was identified as the location where the RSSI value, measured while moving away from the first beacon, dropped to the predefined threshold of –80 dB. From this tangent point, a second beacon was gradually moved forward, with its signal strength at the tangent point continuously measured. Once the measured value dropped below the threshold, the beacon was fixed at its current position. Subsequent devices were positioned similarly, ensuring that their coverage areas were tangent to one another. Care was taken to ensure that blind spots - areas not covered by any beacon- did not exceed 2 meters. Depending on the department, beacons were placed approximately 8 to 14 meters apart. In total, 99 beacons were installed to provide coverage for an area of approximately 4,800 square meters.

Data Collection: As a shopping cart equipped with a beacon moves around the store, it records the MAC addresses of the beacons it interacts with, along with the interaction start and end times, connection durations, and RSSI values. This data is stored in the database, with the MAC addresses used to identify the departments visited by the cart, thereby mapping the shopping route. Checkout beacons and gateways were employed to match carts with specific checkout counters. Once these cart-checkout matches were established, cart-receipt pairings were conducted to associate the shopping route data with the corresponding purchase details for each cart.

This process allowed the data collected from a customer's cart beacon during their in-store journey to be matched with their shopping details and stored in the database for further processing. At the end of each day, the data was inspected for noise. To clean this noisy data, beacon connection durations and RSSI value information were analyzed. Finally, the departments visited during each trip were identified using the MAC addresses of the shelf beacons. This process is illustrated in Fig. 2.

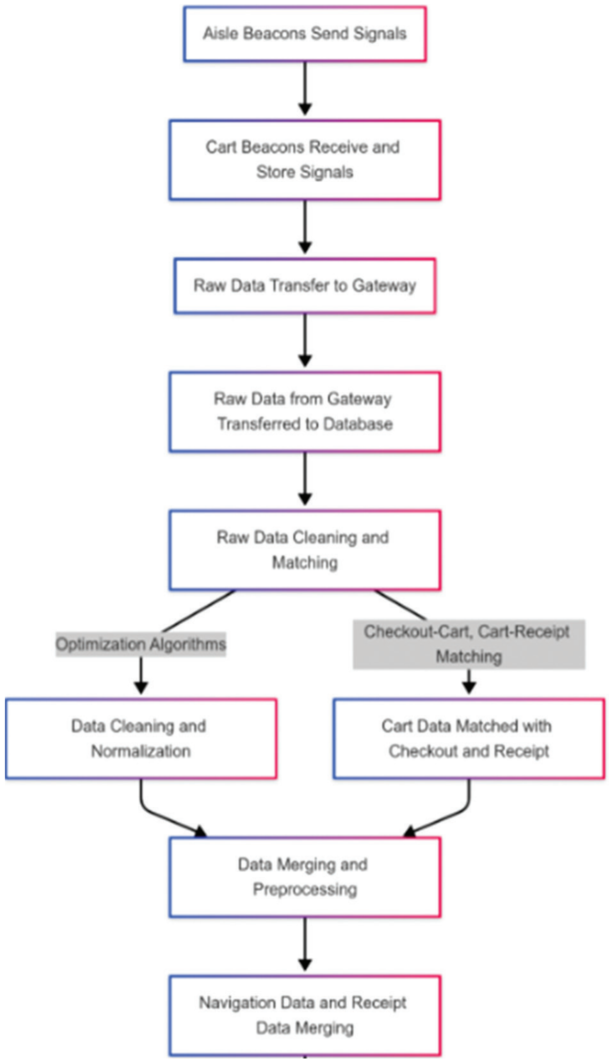

Fig. 2. Illustration of the steps in the data collection and analysis phase

3.2. DATASET

Dataset Format: This study primarily utilizes two types of tables to store customer data. The first table, illustrated in Table 2, records customer visit data. Each customer visit begins with a signal from the entrance beacon and concludes with a signal from the register beacon.

Table 2. Customer behavior Data

User id	Session id	Time	Dep. id	Duration (min.)
U1	S1	2023-07-31 15:00:10	Entrance	0
		2023-07-31 15:00:10	A	9
		2023-07-31 15:09:20	D	21
		2023-07-31 15:30:05	Z	5
		2023-07-31 15:35:05	Register	0
U2	S1	2023-07-31 15:01:10	Entrance	0
		2023-07-31 15:01:10	B	14
		2023-07-31 15:15:20	E	25
		2023-07-31 15:40:30	F	10
		2023-07-31 15:50:30	Register	0
...	...	...	...	...

The data collected from cart beacons undergo preprocessing steps to remove irrelevant information. For instance, if a customer merely passes through a department without spending time there, the beacon signals for that department are excluded. Table 2 provides an example with two visit sessions performed by two different customers. Customer U1 initially visits department D, spends 9 minutes there, and then proceeds to department Z.

The second type of data pertains to customer point-of-sale (POS) information, as shown in Table 3. Each row represents the POS data for a single customer visit, including the products purchased during that session and the total amount spent. To derive meaningful insights from customer behavior, the POS data table is combined with the customer visit data table. All subsequent analyses are conducted using this integrated dataset.

Dataset Details: In this subsection, we present details about the real-world data collected from a selected branch of the home improvement company. The dataset includes customer data recorded between January 9, 2024, and June 10, 2024. Due to a limited number of available shelf beacons, the study focused on 18 selected departments: Lighting, Window Deco, Floor Deco, Wall Deco, Bathroom Accessories, Tiles, Kitchen Accessories, Gardening, Hardware, Heating, Tools, Houseware Storage, Plumbing, Paint, Taps, Furniture, Electrical, and Hard Flooring. Although the entrance and register are not considered departments, they are included in each visit's data. After completing data cleaning and preprocessing phases, the combined dataset, illustrated in Fig. 3, was created. The column labeled checkout receipt id was used to merge the customer visit data with the POS data. For instance, the row with a checkout receipt id value of 316507 indicates that the customer visited six departments. The 'categories' column in the merged dataset table contains department IDs. The 'categories_x' column originates from the customer visit data table and represents the departments visited by a customer, while the 'categories_y' column comes from the POS data and indicates the departments where the customer made purchases.

Table 3. Customer POS Data

User id	Session id	Time	Dep. id	Duration (min.)
U1	S1	2023-07-31 15:35:05	P1, P1, P8, P9, P70	122
U1	S2	2023-07-31 15:50:30	P4, P16, P16	36
U2	S1	2023-08-02 14:15:50	P2	50
...	...	...	...	...

The sequence in which these departments were visited is preserved in the table, along with the duration of the visit (in seconds) for each department. Instead of listing specific product names, the table includes the product categories purchased during the visit. This category information was used to map purchases to corresponding department names. In the example row, the customer visited departments with category IDs 49, 7, 47, 200, 54, and 300 but made purchases only in the department with category ID 49. The purchase amounts associated with each category are also provided, enabling the total purchase amount to be easily computed. During the data collection process, no customer identity-related information was collected; therefore, the data may include multiple visits from the same customer. The final combined dataset comprises 1,447 rows, each representing a distinct customer visit session.

3.3. CLASSIFICATION ALGORITHMS

In this study, we have used two well-known classification approaches comparatively. The first method we employ is a Long Short-Term Memory (LSTM) network, which is a specialized type of Recurrent Neural Networks (RNNs). RNNs are a class of neural networks designed to process sequential data by maintaining a hidden state that captures temporal dependencies. LSTM networks, a specialized type of RNN, address the vanishing gradient problem by incorporating gating mechanisms, namely, input, output, and forget gates to selectively update and retain information over long sequences [18].

While a single-layer LSTM effectively captures sequential dependencies, its ability to extract complex hierarchical features is limited. To enhance performance, researchers often use stacked LSTMs (deep LSTMs), where multiple LSTM layers are stacked, with the hidden state of one layer serving as the input to the next. Another variant, the Bidirectional LSTM (BiLSTM), improves performance by processing sequences in both forward and backward directions. LSTMs are widely used in tasks such as time-series forecasting, natural language processing, and sequence classification due to their ability to model long-term dependencies effectively.

The second method that we employ is Extreme Gradient Boosting (XGBoost), which is a scalable and efficient implementation of gradient-boosted decision trees, designed for both classification and regression tasks [19]. It employs advanced regularization techniques like Lasso regression (L1 regularization) and Ridge regression (L2 regularization) to prevent overfitting and uses an optimized gradient descent approach to minimize loss iteratively. Traditional boosting algorithms, such as Gradient Boosting Machines (GBM), optimize the loss function using first-order gradient information. In contrast, XGBoost improves upon this by incorporating second-order gradient information, which includes both the gradient (first derivative) and the Hessian (second derivative). This allows for better approximation of the loss function. XGBoost is renowned for its speed and accuracy in handling structured data and has been widely adopted in machine learning competitions and real-world applications.

In this study, we formulate our problem as a multiclass classification task, aiming to predict customer sales volumes based on their store visit data. Input vectors representing customer visits were constructed using two distinct modeling approaches, each of which is detailed in this section.

Model 1: The dataset used in this study captures in-store customer behavior along with Point-of-Sale (POS) data. Each row represents a unique shopping session associated with a checkout receipt ID. For instance, the first row (Receipt ID: 321787) indicates that the customer visited 8 departments during their session. The sequence of department visits in this example is as follows: 200, 49, 300, 36, 300, 54, 60, and 54. Each department is represented using an 18-dimensional vector derived from one-hot encoding. Since there are 18 unique departments in total, each department is assigned a unique integer ID ranging from 0 to 17. Examples of department encoding are shown in Table 4.

Each customer session consists of a sequence of department visits. The encoded one-hot vectors for each visited department are concatenated to create the input representation. For example, a session with the sequence: "Kitchen Accessories → Gardening → Tools → Hardware → Tools" would be represented as a matrix of dimensions 5×18. All input vectors were padded to match the length of the longest session. The maximum observed session length is 56 departments. Shorter sessions were padded with the following zero vector:

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Thus, each session is represented as a 56×18 matrix. As the dataset contains a total of 1,447 shopping sessions, the complete dataset is represented as a 3-dimensional matrix with dimensions (1447, 56, 18) after encoding and padding.

Table 4. Model 1 Input Representation

Dep. Name	Id	One-Hot Encoding
Kitchen Accessories	0	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Heating	1	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Model 2: Model 1 preserves the department visit order; however, it does not consider the duration of the visit. In Model 2, the duration of stay in each department is incorporated into the encoding process. A repetition interval of 100 seconds was defined to model department visit durations. The total visit duration for each department was divided by this interval to determine the number of repetitions as shown in Table 5.

Table 5. Model 2 Input Representation

Dep. Id	Visit Dur. (sec.)	Repetition Interval (100s)	Transformed Sequence
54	222	3	[54, 54, 54]
37	36	1	[37]
18	110	2	[18, 18]

checkout_receipt_id	start_time	department_cnt	departments	categories_x	durations	categories_y	unit_price	quantity	sales_amount
321787	2024-06-07 11:26:15	8	Kitchen Acc-Gardening-Tools-Hardware-Tools-Lig...	200-49-300-36-300-54-60-54	8-836-176-100-270-2-18-16	49-300-60-60-60-60-300	949.9-71.99-79.99-37.99-304.9-79.99-31.99	1-1-1-1-1-1-1	949.9-71.99-79.99-37.99-304.9-79.99-31.99
321427	2024-06-10 14:56:07	7	Gardening-Kitchen Acc-Taps-Bathroom&Accessorie...	49-200-33-37-86-402-54	52-882-6-32-132-2-22	49-49	599.9-499.9	1-3	599.9-1499.7
316507	2024-06-09 12:12:53	6	Gardening-Electrical-Houseware&Storage-Kitchen...	49-7-47-200-54-300	932-84-14-48-222-98	49	112.9	1	112.9
308767	2024-01-12 12:27:14	12	Lighting-Wall Deco-Paint-Bathroom&Accessories-...	54-400-60-37-37-200-47-200-200-50-37-54	222-18-56-40-134-132-54-132-28-134-16-192	56-200-54-17,54-60	399.9-249.9-54.99-39.99-24.99	1-1-2-4-1	399.9-249.9-109.98-159.96-24.99
307807	2024-05-08 11:59:31	6	Lighting-Gardening-Houseware&Storage-Bathroom&...	54-49-47-37-54-7	252-78-10-242-82-140	54-47-7-54	149.9-32.99-239.9-149.9	2-1-1-2	299.8-32.99-239.9-299.8

Fig. 3. Combined data table where each row refers to a single customer session behavior and POS data

After transforming the sequences based on durations, the same one-hot encoding and padding steps as in Model 1 were applied. For example, a session with department IDs [54, 37, 18] and durations [222, 36, 110] would be transformed into: [54, 54, 54, 37, 18, 18, 50, 50]. The final input is a padded matrix with dimensions 56×18.

The duration-adjusted dataset contains 1,447 sessions with a maximum transformed session length of 119 departments. The complete input matrix dimensions are: (1447, 119, 18)

This enhanced encoding approach ensures that both visit order and duration are captured effectively, providing a richer representation for sequential modeling tasks.

4. RESULTS AND DISCUSSION

In this section, we first analyze the customer data, including both visit data and POS data, to derive meaningful insights. Subsequently, we present a comparative evaluation of the LSTM and XGBoost algorithms using the two models described in Section 3.

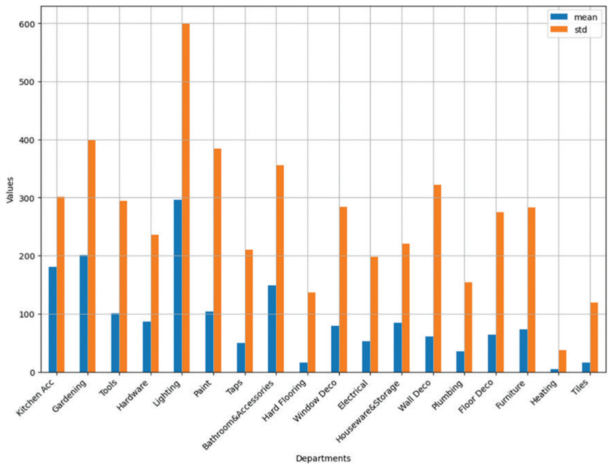

Fig. 4. Department visit duration average and standard deviation

4.1. CORRELATION ANALYSIS

We present the visit duration statistics for each department in Fig. 4. Notably, the 'Lighting' department has the highest average visit duration among all departments. However, it also exhibits the greatest standard deviation, indicating substantial variation in the time customers spend in this section. This suggests that while some customers may spend considerable time in the 'Lighting' department, others may pass through more quickly, resulting in a wide range of visit durations.

The purchase amount statistics for each department are detailed in Fig. 5. The 'Lighting' department has the highest average shopping amount, making it the top-performing section in this regard. Conversely, the 'Taps' department shows a notably high standard deviation, suggesting considerable variability in the prices of purchases made in this section.

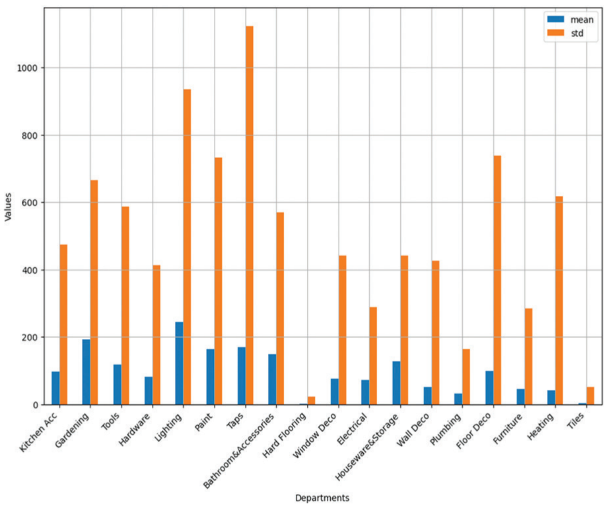

Fig. 5. Department sales amount average and standard deviations

To examine the relationship between the duration of stay in each department and the corresponding purchase amount, we conducted a correlation analysis using Pearson correlation coefficients. A high correlation coefficient indicates a positive relationship between the time spent in a department and the purchase amount. As shown in Fig. 6, the Floor Deco, Paint, and Taps departments exhibit a strong positive correlation, where increased time spent is associated with higher recorded shopping amounts.

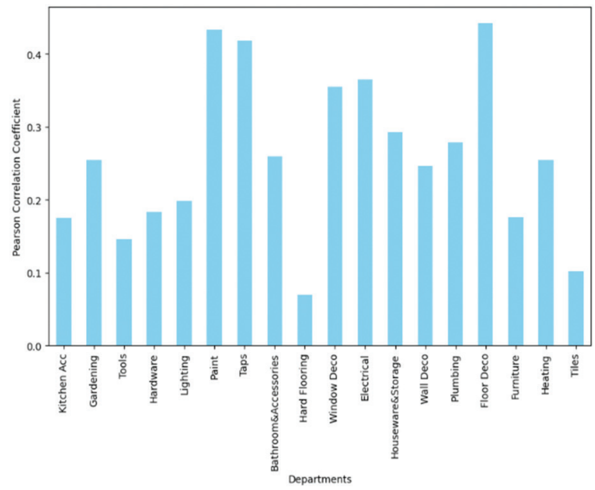

Fig. 6. Correlation between visit duration and purchase amount

4.2. VISUALIZATION

Heatmaps were utilized to visualize the most visited corridors better within the beacon installed areas of the store. The heatmap, based on beacon data collected on June 1, 2024, is shown in Fig. 7. Five different colors were employed on the map: light yellow represents the least dense corridors, while dark red indicates the most dense corridors. We have also created a network graph shown in Fig. 8 using the information from the entire dataset as follows:

- Each node represents a department.
- The size of each node corresponds to the total time spent in that department.
- An edge is drawn between departments with transitions, with edge weights determined by the number of transitions.
- The color of each node corresponds to the total shopping amount recorded in that department.

The network graph indicates that the most frequent transitions occur between the Lighting and Window Deco departments. The Lighting department, represented in the darkest color, records the highest total shopping amount and is also the department where the most time was spent overall. Similarly, the Gardening department shares comparable characteristics with the Lighting department in terms of total time spent and total shopping amount.

4.3. MULTI-CLASS CLASSIFICATION

We approached the problem as a multi-class classification task, where the total purchase amount at the end of each visit session was categorized and used as the response variable. Following the methodology in [13], we carefully analyzed statistical distributions and the histogram of total sales amounts (Fig. 9) to define appropriate category boundaries. Accordingly, three distinct categories for total sales were defined as follows:

- 0 - Low volume: (0, 1000)
- 1 - Medium volume: [1000, 10000)
- 2 - High volume: [10000, 100000)

We applied both algorithms to the input data generated using Model 1, which preserves the order of department visits but does not account for visit duration.

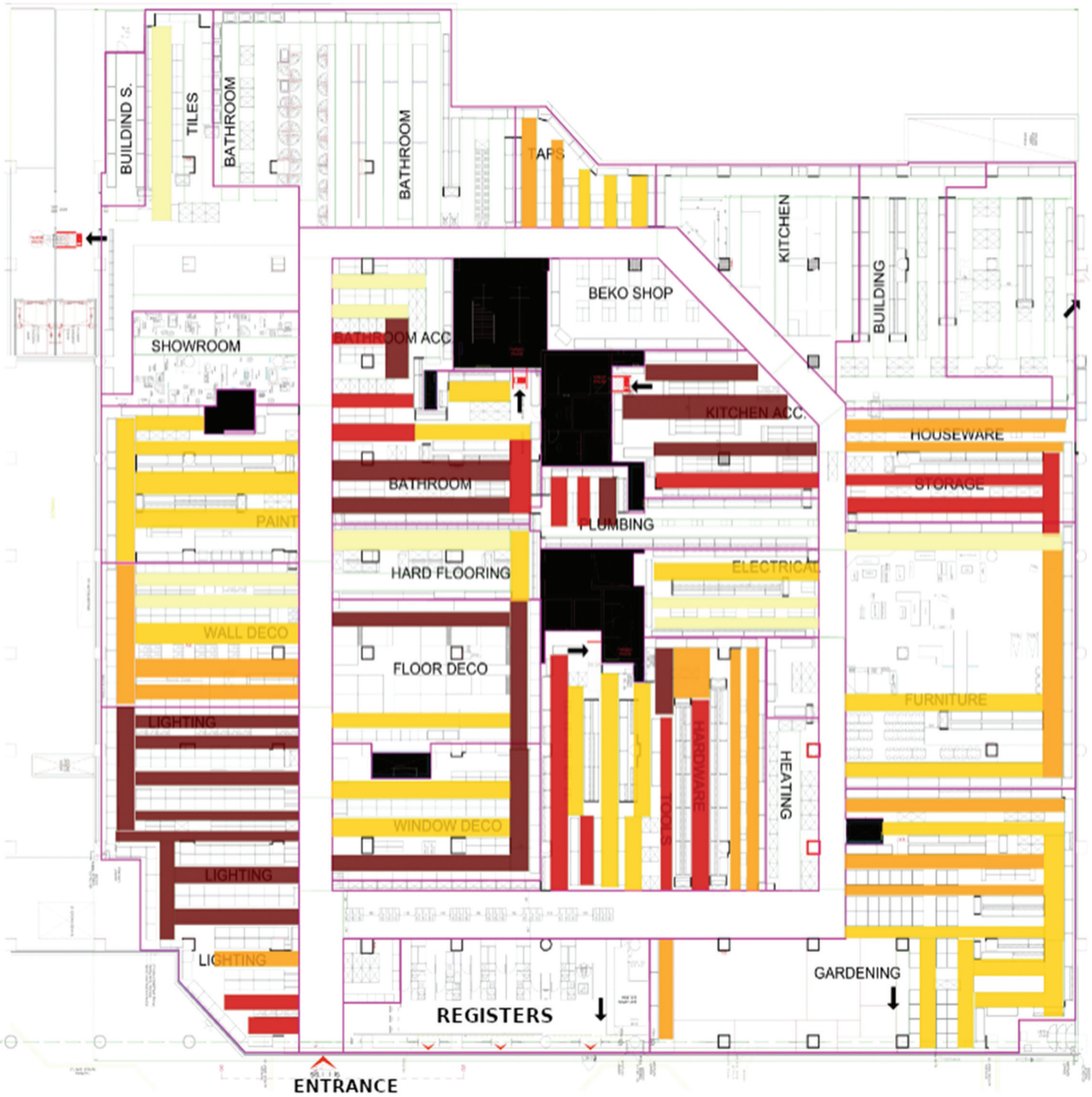

Fig. 7. Heatmap that shows the most visited corridors on 01.06.2024

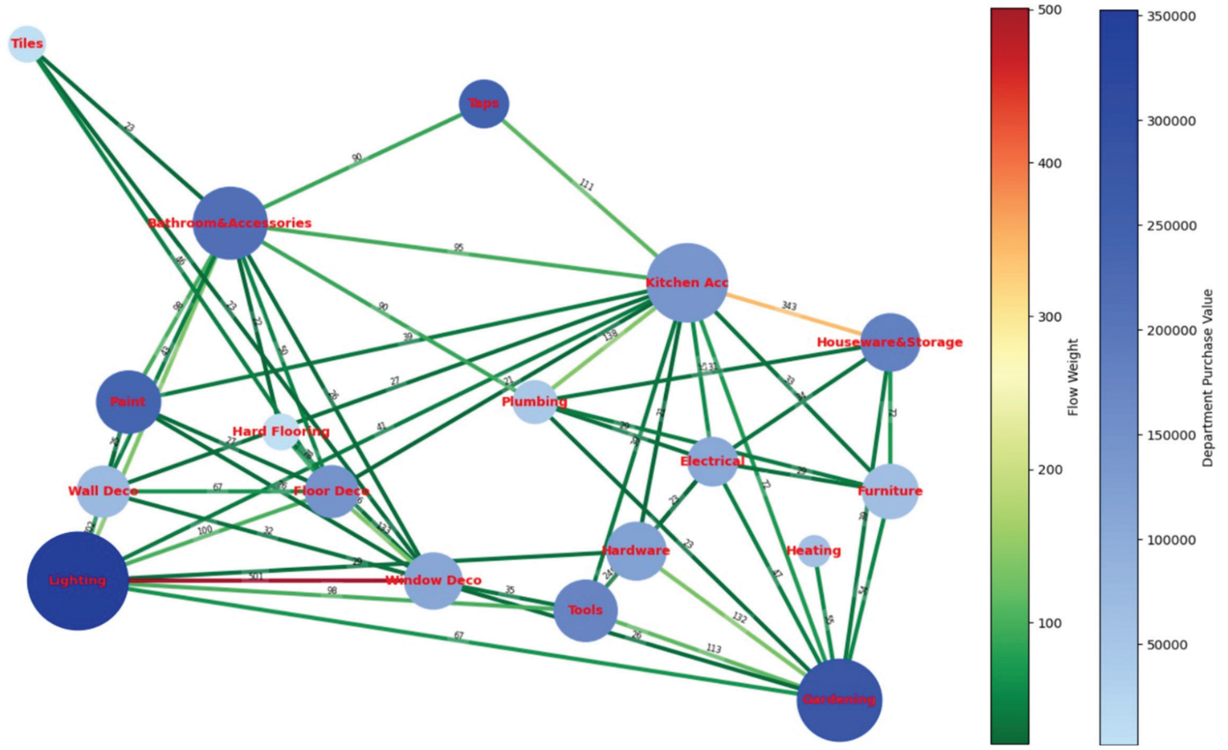

Fig. 8. In the network graph, each node refers to a department and the size of a node corresponds to the total time spent in that department. The darker color refers to a department with a larger total purchase. The edges represent the transitions from one department to another, while the edge color refers to the frequency of that transition

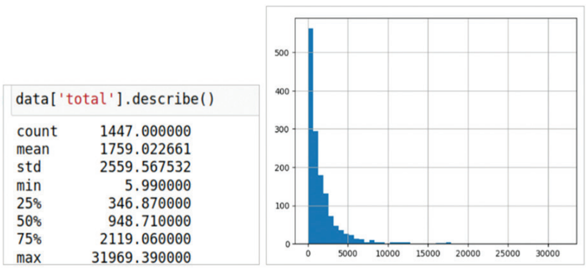

Fig. 9. Total sales data descriptive statistics and the histogram

The dataset was split into 80% for training and 20% for testing, and the classification performance of each method was evaluated in terms of the accuracy under various hyperparameter settings. The results for Model 1 and Model 2 are presented in Table 6 and Table 7, respectively.

Table 6. Model 1 performance comparison of LSTM and XGBoost under different hyper-parameter settings

Model	Configuration	Loss	Accuracy (%)
LSTM1	nl: 1; shl: 128; lr: 0.001; ni: 100; bs: 8	0.79	54.83
LSTM2	nl: 2; shl: 128; lr: 0.001; ni: 100; bs: 8	0.79	54.83
LSTM3	nl: 2; shl: 256; lr: 0.001; ni: 100; bs: 8	0.79	54.83
LSTM4	nl: 2; shl: 512; lr: 0.001; ni: 100; bs: 8	0.79	54.83
XGB1	ne: 2; maxd: 2; lr: 0.1	-	46.21
XGB2	ne: 3; maxd: 2; lr: 0.1	-	46.9
XGB3	ne: 2; maxd: 3; lr: 0.1	-	53.1
XGB4	ne: 3; maxd: 3; lr: 0.1	-	46.21

Table 7. Model 2 performance comparison of LSTM and XGBoost under different hyper-parameter settings

LSTM	Loss	Accuracy (%)	XGBoost	Accuracy (%)
LSTM1	0.79	54.83	XGB1	57.93
LSTM2	0.79	54.83	XGB2	57.24
LSTM3	0.79	54.83	XGB3	56.9
LSTM4	0.8	54.83	XGB4	57.59

The results state that the best LSTM model achieves an accuracy of 54.83% whereas XGB achieves 53.1% for model 1. When it comes to model 2, the results state that XGB achieves a better accuracy than LSTM.

As the initial results were inconclusive, we conducted additional experiments using a modified version of the dataset. Specifically, we removed departments with visit durations of less than 30 seconds in each visit sample. While this adjustment did not reduce the number of samples, it altered the composition of each visit by filtering out short-duration departments. Both algorithms were then applied to each data model.

Additionally, we repeated the experiments after further refining the dataset by excluding departments with visit durations of less than 60 seconds. The results, presented in Table 8, indicate that when using the refined dataset with a 60-second threshold, LSTM models achieved the highest accuracy. This cleaner dataset allowed LSTM models to better capture data characteristics, leading to improved learning. Notably, the loss trajectory of LSTM models on this refined dataset differed significantly from that observed on the original noisy dataset, as illustrated in Fig. 10.

When comparing Model 1 and Model 2 in terms of their impact on algorithm learning, the results suggest that this enhanced data representation did not positively affect classification accuracy. While a richer data representation appeared to improve XGBoost's performance on the noisy dataset, the same effect was not observed when using the cleaned dataset.

Table 8. Accuracy comparison of the algorithms on each of the models using different visit duration thresholds

Model	visit duration<30 removed		visit duration<60 removed	
	Model 1 Acc. (%)	Model 2 Acc. (%)	Model 1 Acc. (%)	Model 2 Acc. (%)
LSTM1	54.83	54.83	58.28*	54.83
LSTM2	54.83	55.86	54.83	54.83
LSTM3	54.83	54.83	56.55	55.17
LSTM4	54.83	54.83	54.83	55.17
XGB1	55.8	56.55	53.79	54.14
XGB2	55.5	56.2	55.17	53.4
XGB3	54.48	55.52	53.45	53.4
XGB4	52.7	54.14	52.76	54.48

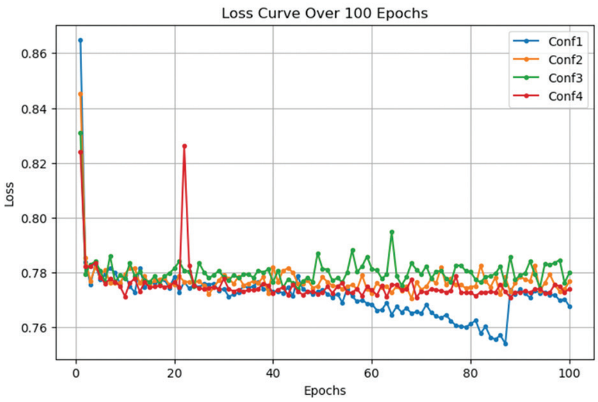

Fig. 10. LSTM loss curve for model 1 when departments with duration <60 secs are ignored

LSTM models appear to have a stronger learning capability than XGBoost for the sequential data used in this study. Throughout the data collection period, a total of 1,447 beacon sessions with valid in-store visit and POS data were gathered. We believe that with a larger dataset, LSTM models could achieve even better performance, further enhancing their ability to capture meaningful sequential patterns.

5. CONCLUSION

In this study, we aimed to observe the in-store behavior of customers at a home improvement retail company. Customer in-store visit data was collected using Bluetooth Low Energy (BLE) beacon devices installed on shelves and shopping carts within the selected store. Cart beacons functioned as receivers, capturing and recording signals from the shelf beacons. Beacons were strategically placed on store shelves to ensure complete coverage, leaving no blind spots. To cover 18 departments spanning approximately 4,800 square meters, 99 beacons were deployed. The duration of stay in each department, the sequence of visits, and the exact visit date and time were recorded in a database. Signals recorded by cart beacons were used to identify the departments where customers spent time, providing valuable insights into customer profiles that can inform marketing decisions.

The correlation analysis revealed a positive relationship between the duration of stay and the purchase amount, particularly in the Floor Deco, Paint, and Taps departments. This information can be leveraged to implement strategic actions, such as expanding these areas or assigning sales assistants to these departments, particularly on specific days of the week. Additionally, we visualized the entire store's data over a given time period using a network diagram. This diagram highlights the departments with the highest sales amounts and longest visit durations while also illustrating customer flow between departments. This visualization provides store managers with a comprehensive overview of in-store behavior, enabling data-driven decisions for optimizing department layouts. Furthermore, insights from customer flow patterns can be leveraged to design targeted marketing campaigns, enhancing the overall shopping experience and sales effectivene

To explore the relationship between in-store behavior and purchase data, we formulated the problem as a multi-class classification task and employed two machine learning models namely, Long Short-Term Memory (LSTM) networks and Extreme Gradient Boosting (XGBoost), for comparative analysis. Experiments were conducted on both the original noisy dataset and a cleaned version, using two distinct data modeling approaches. The first approach utilized only sequential department visit data, while the second model incorporated visit duration information into the sequence representation. The results indicate that both algorithms achieved similar accuracy on the noisy dataset across both data representations, suggesting that adding duration information did not enhance learning. However, when trained on the cleaned dataset, where short-duration department visits were removed, LSTM models demonstrated superior performance, highlighting their ability to better capture meaningful sequential patterns under refined data conditions

These findings underscore the potential of BLE beacon technology for gaining deeper insights into customer behavior, optimizing store layouts, and enhancing personalized marketing strategies. By analyzing in-store movement patterns, retailers can refine product placements, streamline customer journeys, and improve overall shopping experiences. In future work, we plan to collect additional customer visit session data and repeat the experiments with a larger dataset.

6. ACKNOWLEDGEMENTS

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under grant number 3225040

We sincerely thank Gürsu Gülcü for his invaluable insights and expertise, which played a crucial role in the data analysis process. We also extend our appreciation to İsmail Taha Samed Özkan for his assistance during the data analysis and visualization phase.

7. REFERENCES

[1] F. Zhao, "Analysis of consumer behavior and discussion of personalized marketing strategy in the era of big data", Finance & Economics, Vol. 1, No. 2, 2023.

[2] J. S. Larson, E. T. Bradlow, P. S. Fader, "An exploratory look at supermarket shopping paths", International Journal of Research in Marketing, Vol. 22, No. 4, 2005, pp. 395-414.

[3] E. Haktanır, C. Kahraman, S. C. Onar, B. Öztayşi, S. Cebi, "A state of the art literature review on digital transformation", Intelligent Systems in Digital Transformation: Theory and Applications, Springer, 2022, pp. 3-31.

[4] C. Ke, M. Wu, Y. Chan, K. Lu, "Developing a BLE beacon-based location system using location fingerprint positioning for smart home power management", Energies, Vol. 11, No. 12, 2018, p. 3464.

[5] Q. Liu, X. Yang, L. Deng, "An ibeacon-based location system for smart home control", Sensors, Vol. 18, No. 6, 2018, p. 1897.

[6] P. Spachos, K. N. Plataniotis, "BLE beacons for indoor positioning at an interactive IoT-based smart museum", IEEE Systems Journal, Vol. 14, No. 3, 2020, pp. 3483-3493.

[7] G. Shipkovenski, T. Kalushkov, E. Petkov, V. Angelov, "A beacon-based indoor positioning system for location tracking of patients in a hospital", Proceedings of the 2nd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Ankara, Turkey, 26-27 June 2020, pp. 1-6.

[8] R. Pangriya, "Beacon technology the future of retail: A review of the literature and swot analysis", A Journal of Management, Vol. 1, No. 11, 2023, pp. 1-11.

[9] D. Jain, M. S. B. Uppal, S. K. Singh, "Exploring Consumers' Readiness And Acceptance Of Beacon Technology", Educational Administration: Theory and Practice, Vol. 30, No. 5, 2024, pp. 3366-3374.

[10] H. Lemsieh, I. Abarar, "Considering the Beacon technology, a Green and eco-friendly revolution in the Moroccan Commercial Malls for a proximity marketing strategy", Proceedings of the E3S Web of Conferences, Istanbul, Turkey, 29-31 October 2024, p. 6.

[11] C. Garcia, S. Inoue, "Relabeling for Indoor Localization Using Stationary Beacons in Nursing Care Facilities", Sensors, Vol. 24, No. 2, 2024, p. 319.

[12] P. Shende, S. Mehendarge, S. Chougule, P. Kulkarni, U. Hatwar, "Innovative ideas to improve shopping mall experience over e-commerce websites using beacon technology and data mining algorithms", Proceedings of the International Conference on Circuit, Power and Computing Technologies, Kollam, India, 20-21 April 2017, pp. 1-5.

[13] L. Zhao, Y. Zuo, K. Yada, "Sequential sequence-to-sequence learning with gated-attention neural networks", Advances in Data Analysis and Classification, Vol. 17, No. 3, 2023, pp. 549-581.

[14] Y. Zuo, A. S. Ali, K. Yada, "Consumer purchasing behavior extraction using statistical learning theory", Procedia Computer Science, Vol. 35, 2014, pp. 1464-1473.

[15] Y. Zuo, K. Yada, E. Kita, "A Bayesian network approach for predicting purchase behavior via direct observation of in-store behavior", Proceedings of the Advanced Methodologies for Bayesian Networks: Second International Workshop, Yokohama, Japan, 16-18 November 2015, pp. 61-75.

[16] L. Li, "Analysis of e-commerce customers' shopping behavior based on data mining and machine learning", Soft Computing, 2023, pp. 1-10.

[17] S. J. Park, C. U. Kang, Y. C. Byun, "Extreme gradient boosting for recommendation system by transforming product classification into regression based on multi-dimensional word2vec", Symmetry, Vol. 13, No. 5, 2021, p. 758.

[18] S. Hochreiter, J. Schmidhuber, "Long short-term memory", Neural Computation, Vol. 9, No. 8, 1997, pp. 1735-1780.

[19] T. Chen, C. Guestrin, "Xgboost: A scalable tree boosting system", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13-17 August 2016, pp. 785-794.