Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.068619

ARTICLE

Privacy-Preserving Gender-Based Customer Behavior Analytics in Retail Spaces Using Computer Vision

Ginanjar Suwasono Adi[1], Samsul Huda[2,*], Griffani Megiyanto Rahmatullah[3], Dodit Suprianto[1], Dinda Qurrota Aini Al-Sefy[3], Ivon Sandya Sari Putri[4] and Lalu Tri Wijaya Nata Kusuma[5]

[1]Artificial Intelligence of Things Research Group, Department of Electrical Engineering, Politeknik Negeri Malang, Malang, 65141, Indonesia
[2]Interdisciplinary Education and Research Field, Okayama University, Okayama, 700-8530, Japan
[3]Department of Electrical Engineering, Politeknik Negeri Bandung, Bandung, 40559, Indonesia
[4]Department of Business Administration, Politeknik Negeri Bandung, Bandung, 40559, Indonesia
[5]Department of Industrial Engineering, Faculty of Engineering, Universitas Brawijaya, Malang, 65145, Indonesia

*Corresponding Author: Samsul Huda. Email: shuda@okayama-u.ac.jp

Received: 02 June 2025; Accepted: 17 September 2025; Published: 10 November 2025

ABSTRACT: In the competitive retail industry of the digital era, data-driven insights into gender-specific customer behavior are essential. They support the optimization of store performance, layout design, product placement, and targeted marketing. However, existing computer vision solutions often rely on facial recognition to gather such insights, raising significant privacy and ethical concerns. To address these issues, this paper presents a privacy-preserving customer analytics system through two key strategies. First, we deploy a deep learning framework using YOLOv9s, trained on the RCA-TVGender dataset. Cameras are positioned perpendicular to observation areas to reduce facial visibility while maintaining accurate gender classification. Second, we apply AES-128 encryption to customer position data, ensuring secure access and regulatory compliance. Our system achieved overall performance, with 81.5% mAP@50, 77.7% precision, and 75.7% recall. Moreover, a 90-min observational study confirmed the system's ability to generate privacy-protected heatmaps revealing distinct behavioral patterns between male and female customers. For instance, women spent more time in certain areas and showed interest in different products. These results confirm the system's effectiveness in enabling personalized layout and marketing strategies without compromising privacy.

KEYWORDS: Business intelligence; customer behavior; privacy-preserving analytics; computer vision; deep learning; smart retail; gender recognition; heatmap; privacy; RCA-TVGender dataset

1 Introduction

Traditional retail strategies rely on conventional analysis techniques in brick and mortar stores, such as surveys and manual observations, are often limited by subjectivity, scalability, and an inability to grasp real-time behavioral data [1]. Although these methods provide fundamental insights, they fall short of the precision needed for deep customer analytics, particularly in dynamic retail environments where interactions are complex and fleeting. In response, the retail industry has increasingly incorporated computer vision technologies to automate behavior analysis, optimize store layouts, and refine marketing strategies [2]. Deep learning approaches, in particular, offer significant improvements in accuracy utilizing techniques such as facial recognition and gait analysis [3–5]. However, their comparative advantages, such as their robustness to

occlusions and lighting variations, must be systematically evaluated against conventional methods to justify widespread adoption.

In-store demographic analysis introduces unique obstacles compared to online retail analytics, primarily due to physical constraints such as occlusions, varying lighting, and non-cooperative subjects [6,7]. For example, while trajectory analysis and gaze tracking can map customer engagement [8–10], accurately correlating dwell time with purchasing behavior remains difficult. This limitation restricts the generation of actionable retail heatmaps, which are critical for strategic product placement and personalized marketing. Meanwhile, the other problems are the quality of demographic data and the degree of privacy preservation, which are heavily influenced by camera placement strategies [11–15]. The alternative solution is to implement end-to-end encrypted vision systems, although it introduces additional trade-offs that encryption can hinder real-time processing efficiency [16–20]. For instance, cryptographic and federated strategies could not only enhance security aspects hypothetically in similar scenarios but also increase computational cost. Addressing these challenges is essential for developing scalable, privacy-aware solutions that meet both analytical and ethical demands in modern retail environments [21–23].

To address these challenges, our study proposes an end-to-end privacy-preserving system for customer behavior analytics based on customer gender distribution in retail environments. Our proposed method achieves privacy preservation without relying on computationally intensive cryptographic primitives or federated communication protocols, thereby enabling efficient, real-time analytics while meeting ethical privacy standards. Essentially, it bridges the gap between the need for advanced analytics and fundamental ethical considerations in computer vision by demonstrating that real-time efficiency and strong privacy preservation are achievable without prohibitive computational overhead. Our key contributions include:

1. A novel retail analytics pipeline using a YOLOv9s deep learning framework for gender detection. By positioning cameras perpendicular to the observation area, it minimizes facial visibility, preserving privacy while maintaining detection accuracy.
2. The integration of AES-128 encryption to secure customer coordinate data, improving privacy and compliance with standard data protection while offering excellent compromise between security and latency.
3. Gender-based heatmap analytics to identify popular retail spots, enabling targeted marketing and store layout optimization.
4. RCA-TVGender dataset: a publicly available gender dataset captured via top-view cameras in a real-world retail environment that ensures privacy ethics.

2 Proposed Method

We present the overview of the proposed framework as a pipeline diagram in Fig. 1. In specific, the structure of our end-to-end privacy-aware system for gender mapping has three interconnected components: in-store system architecture, cloud-based service, and end-user application. To ensure secure communication, we utilize a Certificate Authority (CA) to manage the authentication and authorize all connections. Additionally, it distributes the predefined AES key to all nodes using Message Queuing Telemetry Transport (MQTT) over Transport Layer Security (TLS). The in-store system begins with strategically placed several cameras in a top-view setting that capture video footage of customers throughout the retail spaces. These cameras connect to a Network Video Recorder that functions as an Real-Time Streaming Protocol (RTSP) server, creating a continuous video stream of customer movements in the store. The video feeds are then processed by an edge device (AMD Ryzen 5 3.6 GHz 6-Core Processor, RAM 16 GB, with Graphic Card NVIDIA GeForce GTX 1650 4 GB) deployed on-premise, which runs several critical services including gender mapping and data encryption. Using YOLOv9 framework, the system performs real-time gender

recognition on detected customers while simultaneously mapping their positions within the store layout. To ensure data privacy protection, the edge computer processes the AES encryption of the data payload prior to transmission. The processed data payload (containing camera IDs, *xy* coordinates, gender, and timestamp parameters) is encrypted and sent to the cloud via a secure MQTT protocol with TLS.

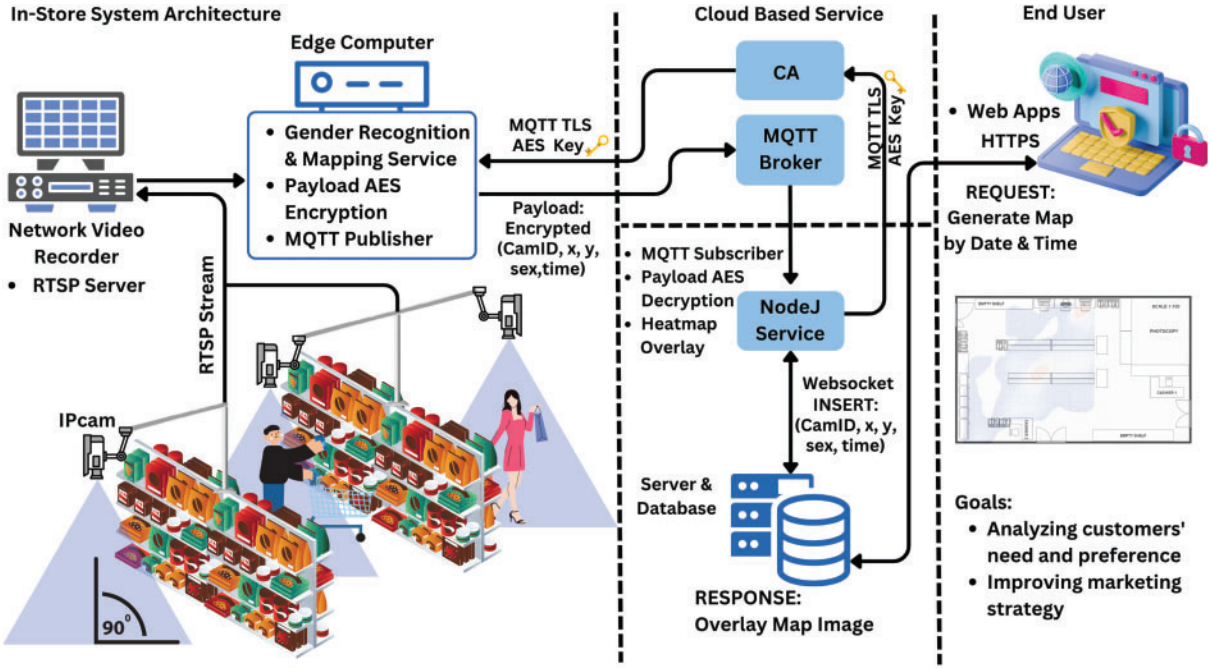

Figure 1: Pipeline design of proposed system architecture

Once the encrypted data reaches the cloud environment, multiple services collaborate to process it securely. The MQTT Broker serves as the central messaging hub, receiving the encrypted payloads from the store and routing them appropriately. A NodeJS service which acts as MQTT subscriber of same topic as in-store edge, processes these messages and forwards the decrypted data to the back-end database and processing systems. The Server and Database component performs several functions: it stores the incoming messages from NodeJS service and gives a response to an end-user web apps to give overlay image map based on the heatmap overlay generation from NodeJS feedback. This processed information is then made available to end users through secured HTTPS connections.

The end-user side provides retail analysts and managers with secure access to the processed customer heatmap data. Users interact with the system through web applications that request gender-based customer analytics based on specified date and time. Once responded by the Webservice, the application receives data to display visual map images showing customer traffic and gender distribution within the store. These visualizations overlay customer positions and demographic information in the store floor plan, enabling retailers to achieve key business goals: analyzing customer needs and preferences based on movement patterns and gender, and improving marketing strategies through better understanding of how different demographic groups interact with various store sections. The entire end-to-end process maintains privacy by ensuring that raw video never leaves the premises and that all transmitted data remains secure until it reaches authorized end users.

2.1 In-Store Setup Environment

The video footage for our system was captured in a retail store located within the Business Laboratory building of Politeknik Negeri Bandung, Indonesia. This Business Laboratory has an area of 12.75 and 8 m. We deployed four Hikvision 2 MP Fixed Bullet Network Cameras (model DS-2CD1023G0E-I) at strategic corridor locations where customers frequently visit to analyze their dwell time. To ensure a top-down perspective and minimize facial visibility, each camera was mounted perpendicular to the observation area, as illustrated in Fig. 2. The cameras were installed at a height of 2.8 m above the retail floor and positioned 1 m away from the center of the shelf to optimize coverage (see Figs. A1 and A2 in Appendix A).

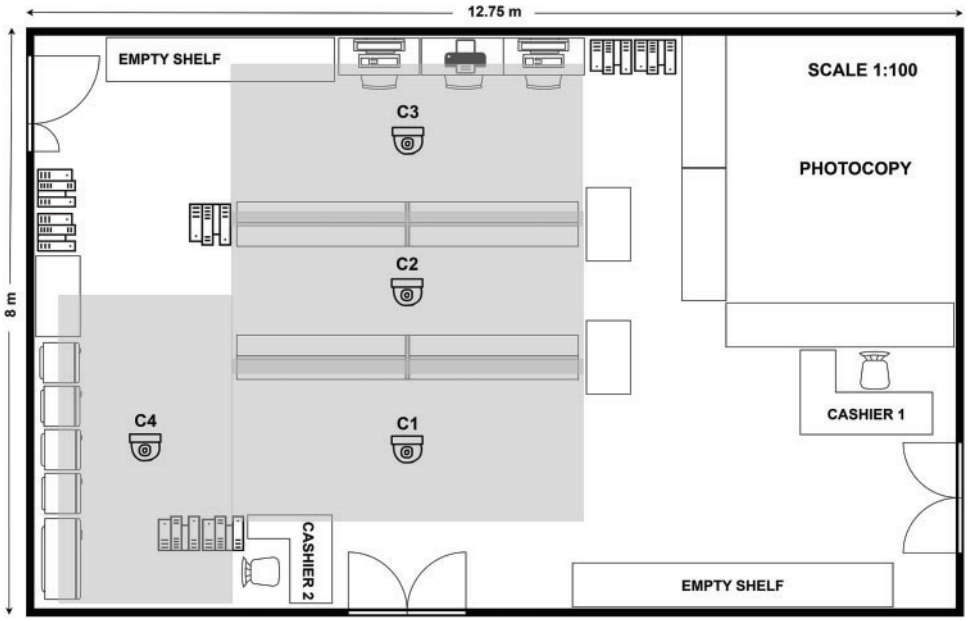

Figure 2: Floor retail plan and camera placement

To obtain a precision heatmap aligned with the coverage area of each camera, both extrinsic and intrinsic camera parameters must be considered. First, we calculate the physical coverage area captured by each camera using Eq. (1).

$$W_m = 2 \cdot D \cdot \tan\left(\frac{HFOV}{2}\right); \quad H_m = 2 \cdot D \cdot \tan\left(\frac{VFOV}{2}\right) \tag{1}$$

where D represents the height of camera placement to the ground, W_m and H_m specify the horizontal and vertical coverage areas (in cm) on the physical map, respectively. Each of $HFOV$ (Horizontal Field of View) and $V\,FOV$ (Vertical Field of View) define the angular coverage of the camera.

$$\alpha = \frac{W_p}{W_m} \quad \text{(width)}; \quad \beta = \frac{H_p}{H_m} \quad \text{(height)} \tag{2}$$

After obtaining a physical coverage representation of each camera, we then transform it onto scaled image map of the retail space. Given the map's 1:100 scale (1 pixel = 1 cm), we compute camera-specific scaling factors using Eq. (2). On Eq. (2), α and β denote scaling factors that relate pixel coordinates to physical space. As W_p and H_p correspond to the camera's width and height resolutions in pixels. It is worth to mention that one of our camera setup was rotate 90° (C4 in Fig. 2), so both W_p and H_p in Eq. (2) are swapped to account

for the altered axis orientation. By multiplying pixel coordinates with camera scaling factors we can obtain map coordinate in pixels as depicted in Eq. (3).

$$x = p_x \cdot \alpha; \quad y = p_y \cdot \beta \tag{3}$$

Eq. (4) shows the offset from camera center.

$$\Delta x = x - \frac{W_m}{2}; \quad \Delta y = y - \frac{H_m}{2} \tag{4}$$

Final area of each camera in map coordinate can be calculated using Eq. (5), where C_x and C_y represent centroid of camera in digital image map.

$$M_x = C_x \pm \Delta x; \quad M_y = C_y \pm \Delta y \tag{5}$$

The shaded regions in Fig. 2 illustrate the coverage areas of cameras C1–C4, calculated using the framework stated in Eqs. (1)–(5).

2.2 Deep Learning Framework for Gender Recognition

The comparative studies presented in Table 1 highlight current state-of-the-art approaches in vision-based customer behavior analysis within retail environments. One of the persistent challenges in this domain is the balance between behavioral insight and privacy awareness, especially when gender profiling is involved.

Table 1: State of the art comparison in analyzing customer behavior using vision-based techniques in retail environment

Authors	Algorithm	Camera view	Gender profiling	Customer analysis	Performance
Shili et al. [10]	YOLOv5 + DeepSORT	Tilted	No	Customer tracking & heatmap generation	Acc: 0.934
Nogueira et al. [24]	RetailNet: CNN + RGBP	Tilted	No	Heatmap generation	MAE: 1.232
Mendes et al. [25]	YOLOv5 + ByteTrack	Tilted	No	Customer tracking & Re-ID	Acc: 0.82
Ijjina et al. [3]	Wide ResNet 16-8 + mini Xception	Tilted	Yes	Customer age & gender	Acc: 0.708 (age), 0.829 (gender)
Nguyen et al. [26]	Multi-module: Deepsort + OpenPose + MobileNet	Front-View RGB-D	No	Tracks customer states (approach, pick, leave)	Precision: 0.71 (approach), 0.93 (leave), 0.71 (pick)
Del Carpio [27]	YOLOv5 + YOLOXn	Front-view	No	Customer detection	F1-score: 0.86 (YOLOv5), 0.74 (YOLOXn)

(Continued)

Table 1 (continued)

Authors	Algorithm	Camera view	Gender profiling	Customer analysis	Performance
Areni et al.* [5]	Viola-Jones + Gabor Filter 2-D + SVM	Front-view	Yes	Gender-based customer counting	Acc: 0.965
Frontoni et al. [13]	Water Filling Algorithm + Template Matching	Top-view RGB-D	No	People counting, shelf interaction analysis	Acc: 0.985 (counting), 0.972 (interactions)
Martini et al.[1*] [14]	Triplet Loss DCNN + ResNext-50	Top-view RGB-D	No	Person re-ID	mAP: 0.933
Paolanti et al.[1,2] [15]	VRAI-Net (Triple CNN)	Top-view RGB-D	No	Counting, interaction, re-ID	Acc: 0.995 (counting), 0.926 (interaction), 0.745 (re-ID)

Note: *Controlled environment; [1]TVPR2 dataset; [2]TVHeads and HaDa dataset.

As stated in Table 1, only a few studies incorporate gender recognition, which is critical for customer demographic analysis, but introduces a critical trade-off between demographic profiling and privacy preservation [3,5]. This privacy-exposure risk is further influenced by camera positioning, both tilted [3,10,24,25] and front-view configurations [5,26,27], where the risk of compromising customer privacy due to identifiable facial features captured by the system.

In contrast, top-view RGB-D systems [13–15] demonstrate a privacy-aware approach by eliminating facial data capture, achieving notable performance of their proposed systems. However, this architectural choice entirely sacrifices demographic profiling capabilities, creating a fundamental limitation for applications requiring customer personalization. This trade-off underlines the need for privacy-aware design in retail vision systems, where the solution must be tailored not only to accuracy and insight requirements, but also to legal and ethical considerations.

To bridge this gap, we propose a privacy-aware deep learning framework that enables gender detection without facial visibility. By leveraging perpendicular top-view camera placement and optimizing the YOLOv9 architecture, the system minimizes facial exposure while maintaining robust detection accuracy. This approach addresses the ethical dilemma of balancing demographic insights with the preservation of privacy. Furthermore, we contribute a novel publicly available dataset captured in real retail environments using top-view cameras, which facilitates research into privacy-conscious demographic analysis. This dataset not only aligns with emerging regulations on data privacy but also empowers the development of systems that respect customer anonymity while providing demographic insights.

2.2.1 RCA-TVGender Dataset

The RCA-TVGender dataset preparation began with a raw collection of 5930 images, comprising 3503 instances of men and 3510 instances of women, ensuring a balanced representation of the gender

categories [28]. This open access dataset was captured using four distinct camera positions in a top-view perspective during three daily observation sessions from 13th to 15th June 2023, at a retail store inside the Business Laboratory, Politeknik Negeri Bandung, Indonesia. The lighting conditions remained consistent throughout the data collection, as the retail store was located indoors, similar to most brick and mortar store environments. The characteristics of the RCA-TVGender dataset mainly focuses on top-head features. Based on customer demographics, for the man's class, the instances typically show no headwear, short haircuts, and black hair color. On the other hand, the woman's class shows that about 82.4% (2893 annotated objects) of instances wear a headscarf in various colors, while the remaining 17.6% (617 annotated objects) have no headdress, with hair colors ranging from brunette to black.

Each gender instance was independently labeled by five annotators. Only samples in which all annotators unanimously agreed on the assigned gender label were included in the final dataset. This strict unanimous agreement requirement serves as both a robust label validation mechanism and a stringent measure of inter-annotator agreement (effectively 100% agreement for included samples consists of a total 7013 instances). Disputed samples were excluded to ensure high label consistency and quality. This process resulted in the exclusion of approximately 0.1% of the initially annotated samples. The resulting RCA-TVGender dataset provides unique value for top-view gender classification in retail contexts, particularly in Muslim-majority regions where headscarves are culturally prevalent. However, this reflects our model's reliance on local demographics, which is a key consideration for its application scope.

To enhance diversity and robustness, a common augmentation strategy (geometric transformations included horizontal flips, 90° rotations (clockwise and counterclockwise), and random rotations between −45° and +45°, while photometric adjustments covered hue (±15°), saturation (±20%), brightness (±15%), exposure (±10%), and Gaussian blur (up to 2.5px)) were applied to our raw data, expanding the dataset to 10,118 images. These augmentations simulated real-world variations in lighting, orientation, and environmental conditions, improving the model's ability to generalize. Following the augmentation process, the dataset was split into training (70%, 7092 images), validation (20%, 1987 images), and test (10%, 1039 images) sets.

The training set exclusively incorporated with augmented data, whereas the validation and test sets retained only original, non-augmented images to preserve their integrity for unbiased evaluation. This approach prevented data leakage and ensured the model's performance was assessed on realistic, unaltered data. The balance of each class in the raw data was preserved post-augmentation through proportional sampling, maintaining fairness across categories. By performing augmentation processes before splitting, the pipeline maximized training set diversity while isolating the validation and test sets as reliable benchmarks. The final RCA-TVGender dataset of 10,118 images combined synthetically enriched training samples with authentic evaluation data, fostering robustness and generalization in downstream model applications.

2.2.2 YOLOv9

You Only Look Once (YOLO) marks a significant shift in computer vision by treating object detection as a single-stage regression problem. Instead of using separate steps for region proposal and classification, YOLO divides the input image into an $S \times S$ grid and each cell predicts B bounding boxes. Each bounding box includes a confidence score and class probabilities, which allow the model to perform object localization and classification in a unified process. Within YOLO design [29], real-time object detection with high spatial coherence and computational efficiency could be achieved. Over successive iterations, YOLO has evolved significantly: YOLOv2 [30] introduced anchor boxes and multi-scale training; YOLOv3 [31] added residual connections and multi-scale feature fusion; YOLOv4 [32] enhanced the architecture with a CSP-Darknet backbone, Spatial Pyramid Pooling (SPP), and Path Aggregation Network (PANet). YOLOv5 [33]

emphasized deployment efficiency and modularity with a PyTorch implementation, while YOLOv6 [34] and YOLOv7 [35] focused on optimization for real-time performance through reparameterization and compound scaling. Later, YOLOv8 [36] adopted an anchor-free architecture with a decoupled detection head, achieving improved accuracy and model simplicity.

Later, the YOLOv9 [37] introduces Programmable Gradient Information (PGI), a novel mechanism designed to preserve and enhance gradient propagation throughout the network depth. PGI establishes auxiliary gradient pathways that supplement the primary gradient flow. The design maintains effective information transmission in scenarios involving saturation or inefficiency of the main gradient route. In parallel, more recent YOLO versions, such as YOLOv10 to YOLOv12, have introduced notable innovations. YOLOv10 [38] eliminates the need for Non-Maximum Suppression (NMS) by adopting a consistent dual-assignment strategy during training, while YOLOv11 [39] enhances backbone efficiency through Cross Stage Partial with kernel size two (C3k2) blocks and Convolutional block with Parallel Spatial Attention (C2PSA) spatial attention modules. The latest development, YOLOv12 [40], represents the first attention-centric approach in the YOLO series by incorporating Area Attention Modules and Residual Efficient Layer Aggregation Networks (R-ELAN) to enhance detection capabilities while maintaining real-time inference.

Based on these foundational models, we adapt YOLOv9 for gender classification by systematically modifying the output layer and training strategy to suit a binary classification task. YOLOv9 is offered in multiple architectural variants designed to accommodate diverse computational and performance requirements. It starts with YOLOv9t (tiny) and YOLOv9s (small), which is optimized for resource-constrained environments such as mobile and edge devices and includes YOLOv9m (medium), which balances accuracy and efficiency for moderately demanding tasks. Larger variants such as YOLOv9l (large) and YOLOv9x/xl (extra-large) provide deeper architectures with higher parameter counts and are capable of delivering state-of-the-art detection performance. We adopt the YOLOv9 by configuring the number of classes to be reduced from the Common Objects in Context (COCO) standard of 80 to 2 (male and female), which requires adjustments to the output tensor dimensions to reflect binary predictions. Specifically, the number of outputs per anchor is computed as:

$$output\_channels = B \times (C + 5) \tag{6}$$

where B is the number of anchor boxes, $C = 2$ is the number of gender classes, and 5 represents 4 bounding box coordinates (x, y, w, h) including 1 confidence score. The adjustment will ensure compatibility with the binary classification objective. Anchor boxes are then optimized for gender detection by defining custom dimensions that correspond to the typical scales of top-view regions in the dataset. While the backbone, neck, and detection head components of YOLOv9 remain unchanged, the model benefits from their strong feature extraction and multi-scale aggregation capabilities. Through extensive ablation studies spanning YOLOv7 to YOLOv12 and further evaluations across architectural variants, YOLOv9s was identified as the optimal configuration to deliver the best trade-off between detection accuracy and computational efficiency for gender classification in our system.

2.3 Edge Deployment and Execution Framework

The edge computing pipeline in our system, shown in Algorithm 1, enables real-time customer tracking across four camera feeds while optimizing computational efficiency and data security. The process starts with initialization of a YOLOv9s model for gender detection and establishing secure connections to both RTSP camera streams and an MQTT broker. Detected individuals are processed through three key steps: first, bounding box centroids are converted to standardized map coordinates using camera-specific field-of-view parameters, including axis adjustments for rotated cameras; second, these coordinates are combined

with gender classifications and timestamps; and third, the data is aggregated into batched in JSON format as payloads with unique incremental IDs for end-to-end traceability.

Algorithm 1: Gender recognition & data encryption

Input: 4 RTSP streams, YOLOv9s model, AES-128 key, MQTT config
Output: Batch encrypted MQTT payload
Initialize:
• Connect to MQTT broker
• Load YOLOv9s model
• Initialize batch $\mathbf{B} \leftarrow \varnothing$, $id \leftarrow 1$
• Define $\forall C_i$: $[M_x^{min} M_y^{min}, M_x^{max}, M_y^{max}]$, rotation
• Initialize $frame\_count \leftarrow 1$ for each camera
• Define FRAME_SKIP=10
while *Running* **do**
  $\mathbf{B} \leftarrow \varnothing$ **foreach** *camera* $C_i \in \{1, 2, 3, 4\}$ **do**
    if $frame\_count[C_i] \mod FRAME\_SKIP = 0$ **then**
      $detections \leftarrow$ YOLOv9s(C_i.frame)
      foreach *detection d in YOLOv9(frame)* **do**
        $(cx, cy) \leftarrow$ GetCentroid($d.bbox$) **if** *rotated* **then**
          $x \leftarrow M_x^{min} + (cy/H) \cdot \Delta M_x$
          $y \leftarrow M_y^{min} + (cx/W) \cdot \Delta M_y$
        else
          $x \leftarrow M_x^{min} + (cx/W) \cdot \Delta M_x$
          $y \leftarrow M_y^{min} + (cy/H) \cdot \Delta M_y$
        $\mathbf{B} \leftarrow \mathbf{B} \cup \{id : id++, cam : C_i, x : x_{round}, y : y_{round}, sex : d.class, ts : time()\}$
      $frame\_count[C_i] \leftarrow 1$
    else
      $frame\_count[C_i] \leftarrow frame\_count[C_i] + 1$
  if $\mathbf{B} \neq \varnothing$ **then**
    $payload \leftarrow$ AES128(JSON($\mathbf{B}$))
    MQTT.publish("topic", *payload*)
    Wait until next processing cycle

The processed data from all cameras go through AES-128 encryption before published using MQTT protocol. To support real-time edge processing with limited resources, we applied lightweight techniques by encrypting only plaintext payload, omitting personally identifiable information, and reducing computation through frame-skipping process (FRAME_SKIP = 10), producing 2.5 Frames Per Second (FPS) per camera. This sampling rate adequately captures in-store macro-movement behavioral patterns without aliasing and exerts minimal impact on dwell time estimation or heatmap quality, as customer movements are generally slow, deliberate, and often persist for several seconds at specific locations. In general, the architecture balances computational efficiency, reduction in network bandwidth, and preservation of privacy by keeping gender recognition on-premise and transmitting only encrypted data payloads.

2.4 Cloud Deployment and Execution Framework

The cloud service manages two primary procedures: data decryption and web service handling. As outlined in Algorithm 2, the cloud process begins by initializing connections to both the MQTT broker and the MySQL database. Upon receiving encrypted data via MQTT, the system decodes the payload and decrypts it using AES-128 in CBC mode. The decrypted JSON is then validated against a predefined schema that includes fields such as id, cam, x, y, sex, and ts. Entries that pass validation are written to the gender map table using transactional control. Any failed insert triggers a rollback of all related changes to preserve data accuracy and consistency. Our method ensures a secure flow from encrypted transfer to structured database storage, which aims to reinforce both reliability and data integrity. The system will then run continuously to listen for incoming messages while isolating and logging errors. Specifically, the logging errors will be comprised of decryption failures, JSON parsing issues, or database exceptions, which are recorded without interrupting ongoing operations. Connections to MQTT and MySQL remain active during runtime and are properly closed during shutdown to prevent memory leaks. Our modular architecture promotes maintainability and scalability and reduces the risk of data loss or corruption under high-throughput conditions.

On the other hand, Algorithm 3 portrays heatmap generation and Web service mechanism. The heatmap service initializes by a listening phase of a Web server with WebSocket support. When a client requests data for a specific time range, the service queries the database for coordinates and gender metadata within that window. A Kernel Density Estimation (KDE) algorithm processes these data, generating spatial density patterns segmented by gender using Eq. (7). Given 2D data points of designated coordinates $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the kernel density estimate at a point (x, y) utilizing a Gaussian kernel is given as follows,

$$\hat{f}(x,y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} \frac{1}{2\pi} \exp\left(-\frac{1}{2} \left[\left(\frac{x - x_i}{h_x} \right)^2 + \left(\frac{y - y_i}{h_y} \right)^2 \right] \right) \tag{7}$$

where n is the number of data points, h_x, h_y are bandwidths in the x and y directions.

Algorithm 2: Data decryption & storage flow

Input: MQTT config, AES-128 key, Database credentials
Output: Processed data in MySQL
Initialize:
• Connect to MQTT broker, Subscribe to "topic"
• Establish MySQL connection
Function DecryptData*(encrypted_data)*
    try:
        raw ← base64Decode(encrypted_data)
        cipher ← AES.new(key, MODE_CBC, iv=raw [0:15])
        return unpad(cipher.decrypt(raw[16:]), AES.block_size).decode('utf-8')
    catch *Exception e*:
        Log("Decryption error: " + *e*.message)
Procedure InsertToDatabase(*item*)
    try:
        Execute INSERT INTO gendermap VALUES (:id, :cam, :x, :y, :sex, :ts)
        USING (*item*.id, *item*.cam, *item*.x, *item*.y, *item*.sex, *item*.ts)

(Continued)

Algorithm 2 (continued)

```
    catch DatabaseError e:
        Log("Insert failed: " + e.message)
while Running do
    if Message Received then
        decrypted ← DecryptData(message.payload)
        try:
            B ← jsonParse(decrypted)
            if B is List then
                try:
                    BeginTransaction()
                    foreach item ∈ B do
                        if hasKeys(item, [id, cam, x, y, sex, ts]) then
                            InsertToDatabase(item)
                    CommitTransaction()
                catch DatabaseError e:
                    RollbackTransaction() + Log("Batch failed: " + e.message)
            else
                Log("Parsed data is not a list")
        catch JSONError e:
            Log("JSON parsing failed: " + e.message)
Terminate: Close connections
```

Algorithm 3: Heatmap generation service

```
Input: Database credentials, Map image
Output: Image overlay
Initialize:
• Web server & WebSocket listening
• Load map image
Function GenerateHeatmap(coordinates, gender)
    try:
        Generate KDE plot(coordinates, gender)
        Render to image buffer
        return heatmap image
    catch Exception e:
        Log("Heatmap error: " + e.message)
Function HandleRequest(start, end)
    Query database:
    SELECT coordinates, gender WHERE time BETWEEN start AND end
    return coordinates, gender
while Running do
    if WebSocket message received then
        HandleRequest(start, end)
```

(Continued)

Algorithm 3 (continued)

heatmap ← GenerateHeatmap(coordinates, gender)
if *heatmap exists* **do**
    Imageoverlay ← Overlay(Map image, *heatmap*)
    Send via WebSocket
else
    Send clear signal
Terminate: Close connections

The resulting heatmap is rendered as an image overlay on the map image, which is instantly transmitted to the client via WebSocket. If there are no valid data, a clear signal is sent to reset the basic interface. Error handling is embedded throughout exceptions during heatmap rendering. Connections to the database and WebSocket are closed during shutdown. Designed for scalability, the service supports dynamic use cases for real-time processing in retail analytics for customer demographic heatmap visualization for Business Intelligence.

3 Results and Discussion

3.1 Performance of Gender Recognition

Our study explores the applicability of several YOLO object detection models for a gender classification system, specifically YOLOv7 to YOLOv12 series variants. YOLO models are recognized for their balance between detection speed and accuracy, which is suitable for real-time computer vision applications. Each variant (denoted with suffixes such as “t” (tiny), “n” (nano), “s” (small),“m” (medium), “c” (compact), and “e” (extended)) offers different trade-offs. In our experiment, we need to determine the most efficient and accurate model for gender identification tasks, considering both computational resource constraints and inference performance.

One of the important aspects of selecting a YOLO variant is the number of parameters. As represented in Table 2, of all the smallest variants, the YOLOv7t architecture has the highest parameter count among the tested models at 6.01 million, while YOLOv9t includes only 2 million. Reducing model size can significantly enhance deployment feasibility on edge devices, as it lowers memory consumption and enables faster inference. In addition to choosing the appropriate model variant, it is also crucial to consider performance evaluation metrics. Specifically, Precision (P) and Recall (R) are key indicators of classification reliability. In detail, Precision measures the proportion of correctly identified positive instances, while Recall indicates the ability of the model to capture all relevant instances. Later, the detection accuracy is assessed using F1-Score and mean Average Precision (mAP@50) at an Intersection over Union (IoU) threshold of 0.5. Additionally, another important factor is Giga Floating Point Operations per Second (GFLOPS), which reflects the computational complexity and resource demands during inference. All the YOLO model evaluations were conducted across multiple independent runs with different random initialization seeds to ensure the robustness and reliability of our experimental findings. In detail, each model variant was trained and evaluated five times using seeds 7, 24, 46, 79, and 93 while maintaining other configurations. This multiple-run approach allows us to assess model stability and provides confidence estimates for reported performance metrics. All results are presented as mean with standard deviation of ±0.1 to demonstrate the statistical reliability of our conclusions and enable proper comparison between model variants.

We conducted experiments on a gender detection model by performing an ablation study using the smallest parameter variants of YOLO. Based on the results as shown in Tables 2 and 3, we selected the optimal model and further evaluated it across all available model sizes. Our results show that YOLOv9t achieved the

optimal mAP@50 of 78% with an F1-Score of 73.1%. Therefore, we forward the experiments by varying the model size of YOLOv9, and we could obtain the optimum mAP@50 of 81.5% by utilizing YOLOv9s. High mAP models tend to generalize better to unseen data, which is particularly beneficial when deploying gender identification systems across diverse datasets or real-world environments. Later, we investigate each version of the YOLO to get more comprehensive information.

Table 2: Performance comparison of YOLO version variants on gender detection task. The bold values indicate the best performance for each parameter

Ver.	Params (M)	P (%)	R (%)	F1-S (%)	mAP@50 (%)	Inf. Time (ms)	GFLOPS
7t	6.01	78.4	65.0	71.1	77.1	1.6	13.2
8n	3.01	72.6	66.4	69.4	75.0	**1.4**	8.1
9t	**2.0**	**78.5**	68.4	73.1	**78.0**	2.2	7.6
10t	2.7	72.5	68.4	70.4	76.3	1.7	8.2
11n	2.59	74.8	**71.4**	73.1	77.5	1.6	6.3
12n	2.53	78.1	70.4	**74.1**	77.7	2.4	**5.8**

Table 3: Ablation of YOLO9 size variants on gender detection task. The bold values indicate the best performance for each parameter

Ver.	Params (M)	P (%)	R (%)	F1-S (%)	mAP@50 (%)	Inf. Time (ms)	GFLOPS
9t	**2.0**	78.5	68.4	73.1	78.0	**2.2**	**7.6**
9s	7.16	77.7	**75.7**	**76.7**	**81.5**	3.5	26.7
9m	20.01	79.7	69.1	74.0	76.7	6.2	76.5
9c	25.32	79.0	69.6	74.0	78.5	7.6	102.3
9e	57.37	**81.4**	72.3	76.6	80.6	16.5	189.1

Our experiments show that YOLOv9s attained the highest recall (75.7%), suggesting it is particularly effective at minimizing false negatives. High recall is desirable in gender detection systems where missing detections could lead to biased or incomplete data representation. Meanwhile, YOLOv9c and 9e also yielded high precision scores (79% and 81.4%, respectively), indicating a strong capacity to avoid misclassifications. YOLOv9s demonstrated the highest mAP@50 score at 81.5% with F1-Score at 76.7%, followed by YOLOv9e at 80.6% with F1-Score at 76.6%. Meanwhile, inference time directly impacts the responsiveness of real-time systems. In our comparison, YOLOv8n achieved the fastest inference time at 1.4 ms. Its number makes YOLOv8n well-suited for time-sensitive applications such as live video analysis and interactive systems. Meanwhile, YOLOv11n and YOLOv7t also demonstrated competitive inference time speeds by achieving 1.6 ms. In contrast, larger models, such as YOLOv9e and YOLOv9c exhibited significantly higher inference times by obtaining 16.5 ms and 7.6 ms, respectively. At the same time, we also observed the performance based on computational efficiency. The lightweight models, such as YOLOv11n (6.3 GFLOPS) and YOLOv12n (5.8 GFLOPS) could be optimized for deployment on low-power devices. Meanwhile, high-performance variants of YOLOv9e (189.1 GFLOPS) and YOLOv9c (102.3 GFLOPS) require substantial computational resources, which makes those models more suitable for GPU-accelerated platforms.

Our evaluation findings provide actionable insights for selecting appropriate YOLO variants in gender identification systems. For deployment in mobile or embedded contexts, models such as YOLOv8n or YOLOv11n are optimal due to their lightweight nature and rapid inference. For high-accuracy applications

with less stringent hardware constraints, such as in our system, YOLOv9s and YOLOv9e demonstrate superior detection capabilities. Although YOLOv11n achieves faster inference (1.6 ms), YOLOv9s provides an incremental 4.0% improvement in mAP@50 (81.5% vs. 77.5%) and significantly better F1-score (76.7% vs. 73.1%), which are critical to reflect the need to both accurately and consistently classify gender while minimizing misclassification in our retail analytics solution. With its 3.5 ms inference time, YOLOv9s still exceeds real-time requirements while maintaining optimal accuracy-efficiency balance for privacy-preserving edge deployment. Based on these results, the future direction of this model is by aiming architectural optimizations to further reduce inference latency while preserving high accuracy. Fig. A3 in Appendix B shows the batch prediction sample of YOLOv9s used in our system where prediction boxes represented the inference of gender recognition based on customer's head-center. To strengthen future models and enhance global applicability, expanding the dataset to include more diverse global demographic characteristics or ambiguous gender cases would provide critical balancing data. This would offer deeper insight into robust feature learning and significantly improve model performance across varied cultural contexts.

3.2 Edge Computing Performance Evaluation

The experimental evaluation demonstrates the computational impact of AES encryption on edge-based gender detection performance using YOLOv9s architecture across four camera deployment points. Processing 6314 frames with 2692 successful detections distributed through 147 MQTT message batches, the system exhibits a measurable but manageable degradation in performance as encryption complexity increases. Described in Table 4, the baseline configuration without encryption achieved optimal processing efficiency at 29.65 FPS with a total runtime of 212.93 s, establishing the performance ceiling for the gender detection pipeline. This baseline performance indicates that the YOLOv9s model effectively handles the computational demands of top-view gender classification in real-time edge environments, maintaining sufficient throughput for practical deployment scenarios.

Table 4: Comparison of edge performance metrics across different AES encryption levels

Metric	No encryption	AES-128	AES-192	AES-256
Total runtime (s)	212.93	214.85	217.74	221.42
Avg CPU usage (%)	33.48	33.80	34.70	39.00
RAM usage (MB)	552.91	561.52	568.74	574.63
Processing efficiency (FPS)	29.65	29.39	29.00	28.52

The introduction of AES encryption schemes reveals a systematic trade-off between data security and computational efficiency, with performance degradation scaling proportionally to encryption key length. AES-128 encryption introduces minimal overhead, reducing processing efficiency by 0.26 FPS while increasing total runtime by 1.92 s and CPU usage by 0.32%. However, the progression to AES-256 encryption demonstrates more substantial computational costs, with processing efficiency dropping to 28.52 FPS (3.8% reduction from baseline) and total runtime extending to 221.42 s. The increase in CPU usage to 39% under AES-256 represents a 16.5% elevation from the unencrypted baseline, while RAM consumption rose modestly from 552.91 to 574.63 MB across all encryption levels. This resource utilization pattern suggests that encryption overhead is mainly exhibited as CPU-bound processing rather than that in memory-intensive operations.

The practical implications of these performance metrics indicate that all tested encryption configurations maintain real-time processing capabilities suitable for edge deployment, even the most computationally

intensive AES-256 configuration sustaining nearly 30 FPS throughput. The relatively modest performance penalty associated with stronger encryption schemes makes AES-256 a viable option for security-critical applications where data protection requirements outweigh marginal efficiency losses. The consistent detection rate of 2692 identifications across 6314 processed frames (42.6% detection ratio) remains stable across all encryption levels, confirming that cryptographic operations do not compromise the accuracy or reliability of the underlying YOLOv9s gender detection algorithm. The efficient MQTT message batching strategy, consolidating detections into 147 transmission packets, demonstrates effective bandwidth utilization while maintaining low-latency communication suitable for real-time monitoring applications.

3.3 Privacy-Preserving Trade-Offs

As depicted in Table 5, the choice of AES-128 in the proposed solution was driven by its low latency and minimal computational overhead, which are essential for maintaining real-time performance in high-traffic retail settings and facilitating future scalability to multi-store ecosystems. However, previous studies identified that comprehensive privacy preservation requires layered strategies [16,19,22,41]. For instance, differential privacy (DP) can anonymize analytics outputs by adding calibrated noise to aggregated data, reducing inference-time exposure risks [16]. Federated learning (FL) frameworks, on the other hand, could enable distributed model training across retail stores without centralizing raw data, addressing storage vulnerabilities [19,23]. Nonetheless, FL introduces communication overhead and requires careful client selection to handle data heterogeneity [22]. For end-to-end protection, homomorphic encryption (HE) represents a promising frontier for securing data during both storage and computation phases. Unlike AES-128, HE enables direct computations on encrypted data without decryption, theoretically eliminating exposure risks during inference. However, current HE implementations remain computationally prohibitive for real-time video analytics, with performance overheads much slower than plaintext processing and significant ciphertext expansion [41].

Table 5: Comparative analysis of privacy-preserving techniques for retail analytics

Technique	Data protection scope	Computational overhead	Latency impact	Feasibility for real-time edge computing
Proposed solution (AES-128)	**Transmission**	**Low**	**Minimal**	**High**
Differential privacy (DP) [16]	Output data	Low-moderate	Low	Moderate
Federated learning (FL) [19,23]	Distributed training	Moderate	High	Limited
Homomorphic encryption (HE) [41]	End-to-end	Very high	Significant	Low

Although AES-128 encryption effectively secures data during the transmission phase, which is a critical feature for future scalable edge-based deployment across multiple retail stores, it does not fully mitigate potential risks associated with on-device storage or inference processes. To address privacy concerns, the proposed system is intentionally designed to exclude the collection of biometric facial data, thereby reducing the risk of sensitive data exposure. Furthermore, particular attention is given to data governance in practical deployment contexts. In alignment with prevailing regulatory frameworks, including GDPR,

PDPA, and CCPA, the system ensures that no personally identifiable information (PII) is stored. While more advanced methods, such as DP, FL, or HE, offer stronger privacy guarantees, their computational overhead and deployment complexity present significant challenges. Therefore, integrating such advanced privacy-preserving methods remains a key direction for future work.

3.4 Heatmap Implication and Business Analytics

3.4.1 Localization Detection Accuracy

As shown in Table 6, to validate the effectiveness of heatmap generation, we conduct quantitative metrics to support customer profiling in our visual analysis. We utilized the error distance performance metric for the predictive centroid and the actual location using sample data in the test set. Our system achieved an overall mean error distance of 0.17 ± 0.14 m across all cameras. The location prediction system demonstrates reasonable accuracy across all four camera coverage areas, achieving an overall mean error distance of 0.17 m based on 400 distinct position measurements with an error distribution ranging from 0.03 to 0.46 m. Performance varies considerably across different areas, with C4 showing the best accuracy at 0.14 m mean error, while C1 and C2 both recorded higher mean errors of 0.19 m, and C3 performed moderately well with a mean error of 0.16 m.

Table 6: Prediction error analysis for localization detection accuracy

Camera	Mean (m)	Std Dev (m)	Min error (m)	Max error (m)
C1	0.19	0.11	0.09	0.35
C2	0.19	0.17	0.05	0.46
C3	0.16	0.14	0.03	0.40
C4	0.14	0.14	0.05	0.38
Overall	**0.17**	**0.14**	**0.03**	**0.46**

The systematic nature of the observed errors can be largely attributed to the fundamental difference between the prediction methodology and ground truth measurement, where the system generates predictions based on head-centered centroids while actual location markers are positioned at foot level, creating a consistent anthropometric offset that explains much of the observed error magnitude. This head-to-foot displacement of approximately 0.17 m aligns well with expected human body proportions and suggests that simple calibration adjustments could significantly improve overall system performance for future work. The location prediction system demonstrates practical utility for indoor positioning applications, with its 0.17-m mean accuracy falling within acceptable ranges for most use cases, and the systematic nature of errors provides clear pathways for improvement through calibration adjustments and environmental optimizations that account for customer demographic characteristics.

In addition, we evaluated the impact of our frame-skipping mechanism (2.5 FPS) against a 25 FPS baseline as depicted in Fig. 3. Our quantitative analysis observed that the heatmap consistency differences were significantly small, ranging from 0.000 to 0.010 across cameras and genders, indicating that the spatial heatmap structure and major hotspots are preserved. However, the frame-skipping process systematically underestimates absolute dwell times: the absolute deviations per camera are 61.1 s (C1), 229.1 s (C2), 176.0 s (C3), and 0.0 s (C4), with a mean absolute deviation of about 116.5 s between cameras. In relative terms, the 2.5 FPS dwell estimates are on average 56% underestimation of the 25 FPS values. Although the relative ranking of camera activity and overall macro-movement trends remain unchanged, these results indicate that 2.5 FPS is adequate for preserving macro-movement patterns and heatmap integrity for comparative analysis such

as hotspot detection and ranking by activity. Given the substantial gains in computation, network efficiency, and privacy preservation point of view, this trade-off is analytically acceptable for in-store behavior analysis focused on relative patterns rather than exact timestamps.

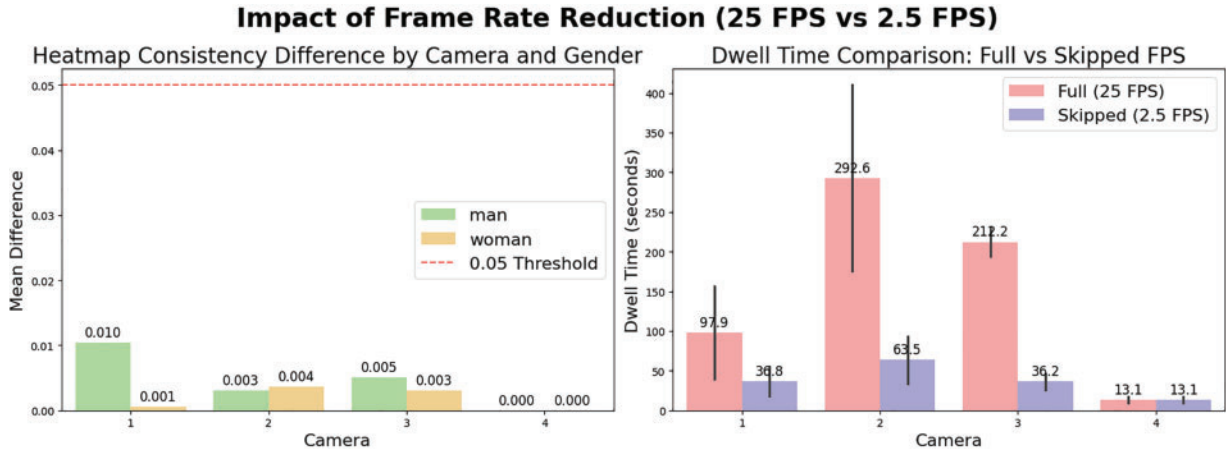

Figure 3: Quantitative analysis on the impact of frame rate reduction (25 FPS vs 2.5 FPS)

3.4.2 Heatmap and Business Analysis

The web interface of our end-user application is depicted in Fig. 4a; in this figure, our app successfully generates a heatmap during a one-minute time frame as a proof-of-concept sample of our end-to-end system. With the aid of this app, business owners or analysts can monitor customer demographic distribution, categorized by gender (male and female), within the store for any desired observation period. For enhanced analytical rigor, we performed heatmap analysis of a 90-min period (9:00 to 10:30 a.m.) captured during the holiday snapshot on 16th June 2023, using recordings from all four overhead cameras, where (C1–C4, Fig. 2) showing the store behave as a two-tier ecosystem displayed in Fig. 4b–d. Tier 1 is a "service anchor", camera C3 covers the self-service printer counter where every shopper, male or female, queues for around 20 s. In our setting, the enforced dwell at the counter highlights a zone of potentially high commercial relevance, which could be further examined in controlled studies. Tier 2 splits by gender, in C2's cone, the women-only heatmap blooms bright red, echoing work showing that female shoppers examine more items and build larger baskets than males in comparable zones. Men, by contrast, sweep along the left-hand perimeter caught by C1 and C4, a pattern retailers increasingly observe as "perimeter shopping". Overlaying the male and female maps reveals a stark "white corridor" on the men's plot at left part of C4 area, exactly where sanitary pads and panty-liners sit at eye level. The corridor not only hides the feminine-care shelf from men, it also shadows adjacent neutral items such as cotton wipes whose sales potential is gender-agnostic.

Because the video footage was built on faceless customers, the general data protection regulation (GDPR)-compliant analytics that streams only encrypted *xy*-coordinates and gender, merchandising tweaks can be tested without enlarging the privacy footprint.

1. Printer queue (C3)
   Attach a 30 cm rail with different brands of mineral water, and 32 GB flash drives while running a micro-LED crawl displaying the message "While you wait > Study Pack −5%". The offer monetizes compulsory dwell time and aligns with the high attachment propensity of short-mission shoppers reported in convenience-store studies.

2. Female hotspot (C2)
   Replanogram the gondola face with ready to drink (RTD) teas, pastel gummy packs, and an "Exam-Survival Kit" bundle. Longer female dwell times correlate with larger and more hedonic baskets; supplying colorful, low-price SKUs exploits this link.
3. White corridor and male sweep (C1 and C4)
   Slide sanitary products one facing deeper, relabel the bay as "Personal Care", and place a waist-high "Study-Fuel" dump bin (nuts, energy bars) in front. Softening the gender cue may reduce embarrassment costs and could potentially improve sales of neutral personal-care items, though this hypothesis requires validation in controlled field experiments.

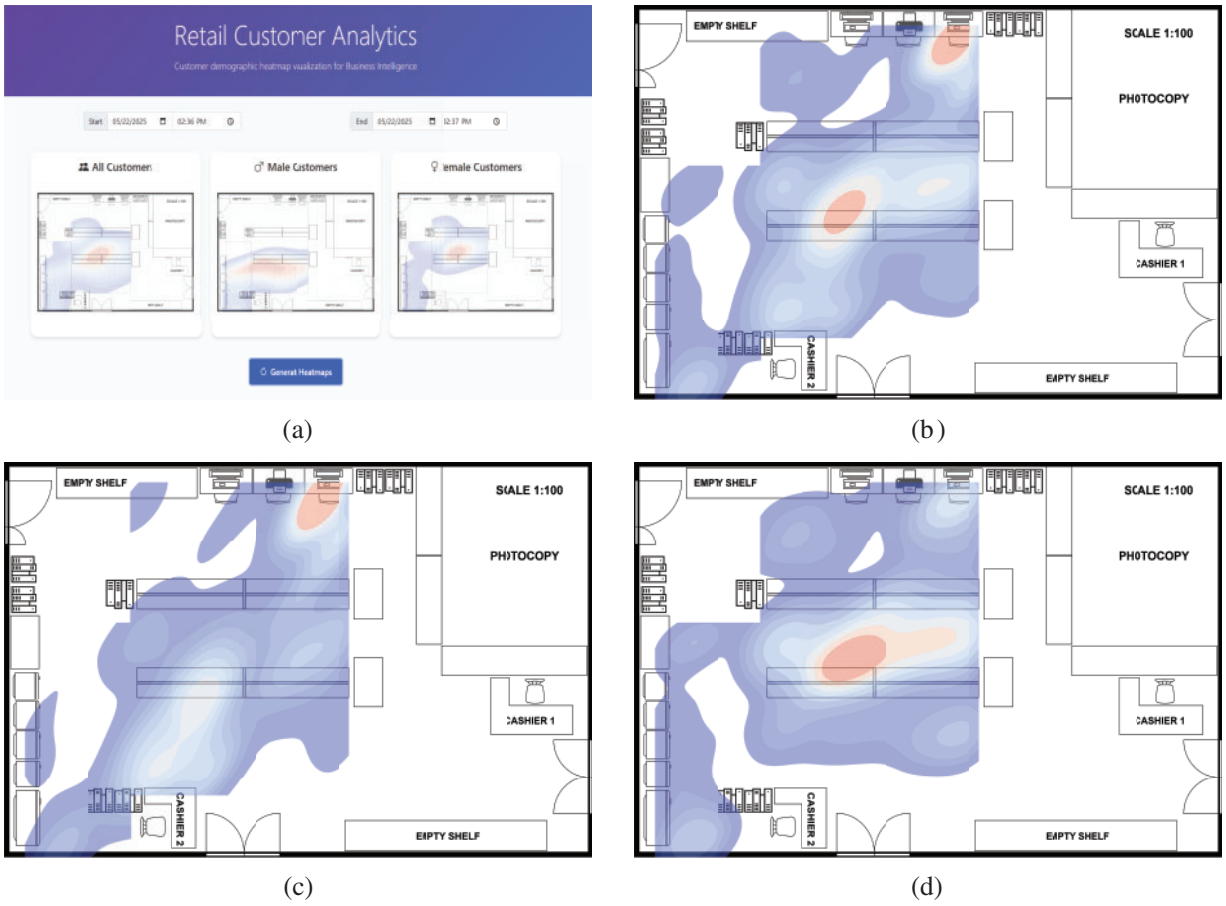

Figure 4: User interface of retail customer analytics of one minute time frame sample (**a**); Customer demographic heatmaps utilizing cool-warm KDE plot during 90-min time frame: (**b**) All customers distribution; (**c**) Male customers distribution; (**d**) Female customers distribution

If similar adjustments were tested, even small incremental changes per basket could translate into noticeable revenue gains. However, quantifying such outcomes requires integration with POS data and structured experimental trials. Performance could be further evaluated within the same privacy-aware loop. For example, one-week heatmap deltas may reveal whether male avoidance patterns change after interventions, but this requires systematic validation. In this way, the heatmap evolves from a purely diagnostic image into a closed-loop optimization tool that is both commercially potent and regulation-safe [2].

The present analysis is constrained to a single holiday morning; subsequent data collection across peak lecture breaks and evening periods is needed to model time-of-day effects. While current heatmap visualizations provide valuable business intelligence, their use cases remain observational rather than experimental. In addition, integrating the encrypted gender heatmap feed with SKU-level margin data will enable real-time dynamic planograms, a capability highlighted but not yet operationalized in related privacy-aware retail studies. Quantitative validation through field trials such as A/B testing of sales lift post-planogram adjustments would strengthen business impact claims but require store-level partnerships not presently available. These findings demonstrate that the proposed novel end-to-end framework not only protects patron privacy, but also provides monetizable insights, bridging the gap between ethical computer vision analytics and tactical retail decision making.

4 Conclusion

This paper presented a privacy-preserving customer analytics system for extracting gender-specific behavioral insights in retail environments without relying on facial recognition. The system integrates a YOLOv9s-based gender classification model, trained on the RCA-TVGender dataset, with AES-128 encryption of customer positional data. The RCA-TVGender dataset plays a critical role by providing side-profile and non-frontal imagery. This makes it well suited for developing gender classification models that minimize facial visibility, which is essential for privacy-preserving analytics. This combination effectively addresses key privacy and ethical concerns commonly associated with computer vision-based analytics. Experimental results show that the YOLOv9s model provides the best performance in terms of balance between accuracy and computational efficiency, with an mAP@50 of 81.5%, a precision of 77.7%, and a recall of 75.7%. The system is still able to process data in real-time even utilizing encryption, with a slight decrease in FPS that is still within reasonable limits, even with AES-256 encryption configuration.

Heatmap analysis based on the distribution of male and female customers during 90 min of observation also revealed different patterns of shopping behavior based on gender. For example, women tend to spend more time in certain areas and show interest in different products than men. These insights are used to recommend more personalized and effective product rearrangement strategies (planograms) and marketing without compromising customer privacy. In general, the developed system demonstrates that the tracking of customer behavior and data-based business decision making can be performed while maintaining ethics and privacy, thus providing a safe, efficient, and applicable solution for optimizing the management of retail space. Future work will focus on extending the system to action recognition and other demographic attributes along with exploring real-time implementation in diverse retail environments.

Acknowledgement: We extend our sincere gratitude to the Business Laboratory of Politeknik Negeri Bandung, Indonesia, for providing access to the experimental facilities and data collection that is essential for this study. We also thank Vipkas for his insightful feedback on the technical parts of our proposed system, which significantly strengthened its conceptual framework. Finally, we deeply appreciate the contributions of Dede, Manda, Ica, Khurnia, Navita, and Syafinas for their assistance in in-store camera installation and dataset annotation, which were instrumental in this research.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Ginanjar Suwasono Adi and Samsul Huda; methodology, Ginanjar Suwasono Adi, Griffani Megiyanto Rahmatullah, and Dodit Suprianto; software, Ginanjar Suwasono Adi, Griffani Megiyanto Rahmatullah, Dodit Suprianto, and Dinda Qurrota Aini Al-Sefy; validation, Ginanjar Suwasono Adi, Samsul Huda, Griffani Megiyanto Rahmatullah, and Dodit Suprianto; formal analysis, Ginanjar Suwasono Adi, Samsul Huda, Griffani Megiyanto Rahmatullah, Ivon Sandya Sari Putri,

and Lalu Tri Wijaya Nata Kusuma; resources, Ginanjar Suwasono Adi, Griffani Megiyanto Rahmatullah, and Dodit Suprianto; data curation, Ginanjar Suwasono Adi, Samsul Huda, Griffani Megiyanto Rahmatullah, Ivon Sandya Sari Putri; writing—original draft preparation, Ginanjar Suwasono Adi, Griffani Megiyanto Rahmatullah, Dodit Suprianto, and Ivon Sandya Sari Putri; writing—review and editing, Samsul Huda and Lalu Tri Wijaya Nata Kusuma; visualization, Ginanjar Suwasono Adi, Griffani Megiyanto Rahmatullah, and Dinda Qurrota Aini Al-Sefy; supervision, Samsul Huda and Lalu Tri Wijaya Nata Kusuma. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset that support the findings of this study are openly available in Roboflow Universe Repository at https://universe.roboflow.com/aiot-research-group/rca-tvgender (accessed on 16 September 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Declaration of Generative AI and AI-Assisted Technologies: During the preparation of this work, the authors used Grammarly to ensure the grammar of information delivery. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix A

Figure A1: Placement of camera installation in business laboratory

Figure A2: Camera field of view: (**a**) FoV C1; (**b**) FoV C2; (**c**) FoV C3; (**d**) FoV C4

Appendix B

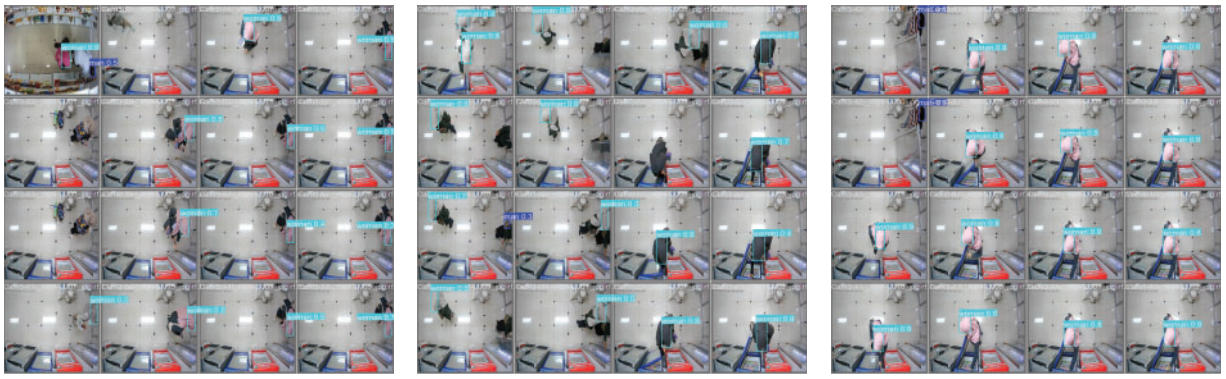

Figure A3: Sample of batch prediction of YOLOv9s

References

1. Larsen NM, Sigurdsson V, Breivik J. The use of observational technology to study in-store behavior: consumer choice, video surveillance, and retail analytics. Behav Anal. 2017;40(2):343–71. doi:10.1007/s40614-017-0121-x.
2. Becker J, Müller K, Cordes AK, Hartmann MP, von Lojewski L. Development of a conceptual framework for machine learning applications in brick-and-mortar stores. In: Proceedings of the 15th International Conference on Wirtschaftsinformatik. Potsdam, Germany; 2020. p. 225–41.
3. Ijjina EP, Kanahasabai G, Joshi AS. Deep learning based approach to detect customer age, gender and expression in surveillance video. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 2020 Jul 1–3; Kharagpur, India: IEEE. p. 1–6.

4. Kikuchi H, Miyoshi S, Mori T, Hernandez-Matamoros A. A vulnerability in video anonymization-privacy disclosure from face-obfuscated video. In: 2022 19th Annual International Conference on Privacy, Security & Trust (PST); 2022 Aug 22–24; Fredericton, NB, Canada. IEEE. p. 1–10.
5. Areni IS, Safitri TN, Indrabayu I, Bustamin A. Gender-based customer counting system using computer vision for retail stores. J Comput Sci. 2020 Apr;16(4):439–51.
6. Baek NR, Cho SW, Koo JH, Truong NQ, Park KR. Multimodal camera-based gender recognition using human-body image with two-step reconstruction network. IEEE Access. 2019;7:104025–44. doi:10.1109/access.2019.2932146.
7. Nguyen DT, Kim KW, Hong HG, Koo JH, Kim MC, Park KR. Gender recognition from human-body images using visible-light and thermal camera videos based on a convolutional neural network for image feature extraction. Sensors. 2017;17(3):637. doi:10.3390/s17030637.
8. Ganokratanaa T, Ketcham M, Pramkeaw P, Chumuang N, Yimyam W. Vision-based consumer behavior analysis system for retail optimization. In: 2024 IEEE/ACM 17th International Conference on Utility and Cloud Computing (UCC); 2024 Dec 16–19; Sharjah, United Arab Emirates. IEEE. p. 472–6.
9. Liu X, Krahnstoever N, Yu T, Tu P. What are customers looking at?. In: 2007 IEEE Conference on Advanced Video and Signal Based Surveillance; 2007 Sep 5–7; London, UK. p. 405–10.
10. Shili M, Jayasingh S, Hammedi S. Advanced customer behavior tracking and heatmap analysis with YOLOv5 and DeepSORT in retail environment. Electronics. 2024;13(23):4730. doi:10.3390/electronics13234730.
11. Liciotti D, Zingaretti P, Placidi V. An automatic analysis of shoppers behaviour using a distributed rgb-d cameras system. In: 2014 IEEE/ASME 10th International Conference on Mechatronic and Embedded Systems and Applications (MESA); 2014 Sep 10–12; Senigallia, Italy. IEEE. p. 1–6.
12. Senior A. Privacy enablement in a surveillance system. In: 2008 15th IEEE International Conference on Image Processing; 2008 Oct 12–15. San Diego, CA, USA. IEEE; p. 1680–3.
13. Frontoni E, Raspa P, Mancini A, Zingaretti P, Placidi V. Customers' activity recognition in intelligent retail environments. In: New Trends in Image Analysis and Processing–ICIAP 2013; 2013 Sep 9–13; Naples, Italy. p. 509–16.
14. Martini M, Paolanti M, Frontoni E. Open-world person re-identification with rgbd camera in top-view configuration for retail applications. IEEE Access. 2020;8:67756–65. doi:10.1109/access.2020.2985985.
15. Paolanti M, Pietrini R, Mancini A, Frontoni E, Zingaretti P. Deep understanding of shopper behaviours and interactions using RGB-D vision. Mach Vis Appl. 2020;31(7–8):1–21. doi:10.1007/s00138-020-01118-w.
16. Zhang X, Wang T, Ji J, Zhang Y, Lan R. Privacy-preserving face attribute classification via differential privacy. Neurocomputing. 2025;626(11):129556. doi:10.1016/j.neucom.2025.129556.
17. Ahmad J, Larijani H, Emmanuel R, Mannion M, Javed A, Ahmadinia A. An intelligent real-time occupancy monitoring system with enhanced encryption and privacy. In: 2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC); 2018 Jul 16–18; Berkeley, CA, USA. IEEE. p. 524–9.
18. Geng Y, Zhang Z. Privacy-preserving face recognition using a cryptographic end-to-end optoelectronic hybrid neural network. Appl Phys B. 2025;131(4):1–16. doi:10.1007/s00340-025-08442-x.
19. Zhang J, Zhou J, Guo J, Sun X. Visual object detection for privacy-preserving federated learning. IEEE Access. 2023;11:33324–35. doi:10.1109/access.2023.3263533.
20. Goel A, Agarwal A, Vatsa M, Singh R, Ratha N. DeepRing: protecting deep neural network with blockchain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2019 Jun 16–17; Long Beach, CA, USA. IEEE.
21. Li H, Lo JTY. A review on the use of top-view surveillance videos for pedestrian detection, tracking and behavior recognition across public spaces. Acc Anal Prevent. 2025;215(3):107986. doi:10.1016/j.aap.2025.107986.
22. Constantinou G, You S, Shahabi C. Towards scalable and efficient client selection for federated object detection. In: 2022 26th International Conference on Pattern Recognition (ICPR); 2022 Aug 21–25; Montreal, QC, Canada. IEEE; 2022. p. 5140–6.
23. Li Q, He B, Song D. Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. IEEE. p. 10713–22.

24. Nogueira V, Oliveira H, Silva JA, Vieira T, Oliveira K. RetailNet: a deep learning approach for people counting and hot spots detection in retail stores. In: 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI); 2019 Oct 28–30; Rio de Janeiro, Brazil. IEEE. p. 155–62.
25. Mendes D, Correia S, Jorge P, Brandão T, Arriaga P, Nunes L. Multi-camera person re-identification based on trajectory data. Appl Sci. 2023;13(20):11578. doi:10.3390/app132011578.
26. Nguyen TD, Hihara K, Hoang TC, Utada Y, Torii A, Izumi N, et al. Retail store customer behavior analysis system: design and implementation. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. Cham, Switzerland: Springer Nature; 2024. p. 305–18.
27. Del Carpio AF. Analyzing computer vision models for detecting customers: a practical experience in a mexican retail. Int J Adv Intell Inform. 2024;10(1):131–47. doi:10.26555/ijain.v10i1.1112.
28. AIoT Research Group. RCA-TVGender dataset [Open Source Dataset]. Roboflow; 2025. [cited 2025 May 29]. Available from: https://universe.roboflow.com/aiot-research-group/rca-tvgender.
29. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. IEEE. p. 779–88.
30. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. IEEE. p. 7263–71.
31. Redmon J, Farhadi A. Yolov3: an incremental improvement. arXiv:1804.02767. 2018.
32. Bochkovskiy A, Wang CY, Liao HYM. Yolov4: optimal speed and accuracy of object detection. arXiv:2004.10934. 2020.
33. Jocher G, Chaurasia A, Stoken A, Borovec J, Kwon Y, Michael K, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. Zenodo; 2022. doi:10.5281/zenodo.7347926.
34. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, et al. YOLOv6: a single-stage object detection framework for industrial applications. arXiv:2209.02976. 2022.
35. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. IEEE. p. 7464–75.
36. Sohan M, Sai Ram T, Rami Reddy CV. A review on yolov8 and its advancements. In: International Conference on Data Intelligence and Cognitive Informatics. Singapore: Springer; 2024. p. 529–45.
37. Wang CY, Yeh IH, Mark Liao HY. Yolov9: learning what you want to learn using programmable gradient information. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2024. p. 1–21.
38. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. Yolov10: real-time end-to-end object detection. Adv Neural Inform Process Syst. 2024;37:107984–8011.
39. Khanam R, Hussain M. Yolov11: an overview of the key architectural enhancements. arXiv:2410.17725. 2024.
40. Tian Y, Ye Q, Doermann D. Yolov12: attention-centric real-time object detectors. arXiv:2502.12524. 2025.
41. Chen Y, Duan F, Zhao Y, Han T, Hu J, Liu X, et al. Efficient face information encryption and verification scheme based on full homomorphic encryption. Sci Rep. 2025;15(1):11388. doi:10.1038/s41598-025-95383-2.