

Article

Hierarchical Product Recognition Under Partial Supervision for Retail Environments

Mohammad Hori ^{1,*}, Bernd Noche ¹  and Abbas Horri ²

¹ Department of Transport Systems and Logistics, Faculty of Engineering, University of Duisburg-Essen, Keetmanstraße. 3-9, 47058 Duisburg, Germany; bernd.noeche@uni-due.de

² Department of Computer Engineering, Faculty of Technology and Engineering, Shahrekord University, Shahrekord 64165478, Iran; horri@sku.ac.ir

* Correspondence: mohammad.hori-najafabadi@stud.uni-due.de

Abstract

Background: Retail product recognition is difficult in practice because products often look similar, class distributions are unbalanced, and annotations usually contain only one observed product label, even though products belong to broader category structures. This study addresses this setting as classification under incomplete supervision rather than as a fully annotated multi-label problem. **Methods:** We evaluate a lightweight hierarchy-aware recognition framework on a laboratory-based retail dataset containing real supermarket products. Visual features are extracted with a pretrained ResNet-50, reduced by truncated singular value decomposition (TSVD), and classified using class-conditional kernel density estimation (KDE). A coarse-to-fine refinement step first assigns each sample to one of seven product groups and then applies a group-specific classifier. **Results:** The hierarchical TSVD-KDE model achieved 83.7% Top-1 accuracy, 95.8% Top-5 accuracy, and 71.8% macro-averaged F1-score, improving over the flat TSVD-KDE variant and the convolutional neural network (CNN) linear baseline. For rare classes, the model reached 62.4% Top-1 accuracy while maintaining an average inference time of 6.1 ms per sample. **Conclusions:** The results suggest that combining compact visual representations with probabilistic classification and a simple product hierarchy can improve balanced recognition performance without relying on a large end-to-end architecture.

Keywords: retail product recognition; hierarchical classification; incomplete supervision; TSVD; kernel density estimation; probabilistic classification; class imbalance; long-tail distribution; visual recognition; retail logistics



Academic Editor: Rameshwar Dubey

Received: 13 May 2026

Revised: 29 June 2026

Accepted: 30 June 2026

Published: 2 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In the last ten years, visual recognition has made significant advancements, mainly due to deep convolutional architectures and large-scale annotated datasets like ImageNet [1–3]. Yet, the assumptions that these benchmarks are based on do not align with what is typically seen in real-world retail situations. The challenges of retail product recognition stem from large and often changing product catalogs, fine-grained visual similarities between products, unbalanced class distributions, and variations in product appearance due to factors like packaging, orientation, shelf placement, and lighting conditions.

One major challenge with retail datasets is that product images are usually annotated at just one observable level, like a stock keeping unit (SKU) or a dominant product class. Concurrently, products can be characterized by wider semantic attributes such as category,

brand, packaging type, or product variant. Usually, these extra semantic relations are not annotated exhaustively. We adopt a practical definition of partial supervision in this study: the training data consists of one observed class label for each instance, and the broader semantic structure of the product space is only partially represented by a manually defined retail hierarchy.

It is important to differentiate this scenario from full multi-label learning or latent label reconstruction. The aim of this work is not to deduce all hidden labels of a product or to discover an unknown taxonomy from the data. Rather, the research examines if a lightweight probabilistic classifier can take advantage of compact visual representations and a predefined coarse-to-fine product hierarchy when annotations are lacking and class distributions are imbalanced.

The relevance of hierarchical structure in retail recognition is underscored by the fact that product taxonomies naturally arrange items that are visually and semantically related. By employing a coarse-to-fine classification strategy, we can minimize the effective search space at the fine-grained level, which may aid in reducing confusion between unrelated product groups. Nevertheless, a lot of hierarchy-aware learning approaches depend on complicated end-to-end architectures, or they assume that label structures are fully specified during training [4–6]. It is tough to meet these assumptions in operational retail environments, where product catalogs change often, and we also need to consider computational efficiency.

According to recent research on weakly supervised and noisy-label learning, it appears that recognition performance is affected not just by model capacity, but also by the structure and quality of the supervision [7–9]. In retail applications, simply raising architectural complexity may not be enough if the available annotations are incomplete, imbalanced, or inconsistently structured. This leads to the assessment of recognition pipelines that are simpler and more adaptable, which decouple visual feature extraction from probabilistic classification.

Driven by these factors, this work puts forward and assesses a hierarchy-aware TSVD–KDE framework for retail product recognition in the context of incomplete supervision. A pretrained convolutional backbone extracts visual features, which are then projected into a compact latent space through truncated singular value decomposition (TSVD). Kernel density estimation (KDE) is used to estimate class-conditional distributions. Coarse-to-fine refinement is achieved using a predefined product hierarchy by initially assigning an image to a broad product group and then utilizing a group-specific classifier.

This work has three key contributions. To begin with, the research offers a precise definition of retail product recognition under incomplete semantic supervision, making it clear how this differs from full multi-label learning and hidden-label reconstruction. Secondly, it presents a lightweight integration of TSVD-based latent representation, KDE-based probabilistic classification, and predefined hierarchical refinement for fine-grained retail recognition. Third, it assesses the proposed framework using a laboratory-based retail dataset containing actual products, detailing overall accuracy, macro-F1 performance, rare-class behavior, and inference time to evaluate both recognition quality and computational efficiency.

2. Related Work

Retail product recognition is located at the intersection of visual recognition, weakly supervised learning, hierarchical classification, and retail logistics. Existing studies have shown that computer vision can support operational retail processes such as shelf monitoring, product identification, inventory auditing, warehouse sorting, and checkout assistance. However, retail recognition remains difficult because product catalogs are large, visually

similar items are common, class distributions are often highly imbalanced, and annotation protocols usually provide only limited semantic information. This section reviews the main research streams relevant to the present study and clarifies how the proposed work differs from full multi-label learning, latent-label reconstruction, and automatic hierarchy learning.

2.1. Multi-Label Classification in Visual Recognition

Multi-label classification allows one instance to be associated with several semantic labels at the same time [10]. This formulation is useful in structured visual domains because an object may simultaneously belong to different categories, attributes, or taxonomic levels. Early approaches often decomposed the problem into independent binary classification tasks [11], while later methods explicitly modeled label dependencies and co-occurrence patterns to improve recognition performance [12].

Deep learning has further improved multi-label recognition by learning shared visual representations together with label relationships [12,13]. However, most multi-label approaches assume that training annotations are sufficiently complete. This assumption is difficult to satisfy in retail settings, where products may be described by category, brand, packaging type, flavor, size, and SKU, but only one of these labels is usually available in the dataset.

The present study does not attempt full multi-label prediction. Instead, it treats the dataset as an incomplete semantic supervision setting: each product instance has one observed class label, while broader product relations are only partially represented through a predefined hierarchy. This distinction is important because the proposed method is not designed to recover all hidden labels, but to evaluate whether hierarchy-aware probabilistic refinement can improve fine-grained recognition under limited annotation.

2.2. Visual Feature Representation and Latent Feature Compression

Retail product images contain several visual cues, including color, texture, packaging shape, typography, logo placement, and layout patterns. These cues are often subtle, especially when products differ only by flavor, size, or variant. Modern convolutional networks can extract rich visual representations from such images, but high-dimensional feature vectors may contain redundancy and noise, particularly when the available dataset is imbalanced or limited.

Dimensionality reduction methods provide a practical way to obtain compact feature representations before classification. Truncated singular value decomposition (TSVD) is commonly used to project high-dimensional data into a lower-dimensional latent space while preserving dominant variance components [14]. In this study, TSVD is used as a feature compression step rather than as a multimodal fusion method. This terminology is important: the present work uses only visual image data and does not combine images with text, metadata, or transactional records.

Compact latent representations are relevant for retail logistics because recognition systems may need to operate with limited computational resources and frequent product catalog updates. Prior studies in logistics and supply chain analytics emphasize that practical AI systems must be evaluated not only by predictive performance, but also by computational efficiency, adaptability, and robustness in operational environments [15,16].

2.3. Learning Under Incomplete Supervision

Incomplete annotation is a recurring challenge in visual recognition. In partial-label or single-positive learning, only part of the true label information is observed, while other potentially valid labels remain unknown [17]. Treating all unobserved labels as negative can introduce bias and may reduce generalization performance [17,18].

Several approaches have been proposed to address incomplete supervision, including loss reweighting, pseudo-label correction, regularization methods, and label enhancement strategies [8,19–21]. These methods are especially relevant when the objective is to infer missing labels or reconstruct latent multi-label structures.

The present work is more limited in scope. It does not estimate hidden labels or reconstruct a complete label matrix. Instead, it evaluates a classification pipeline that uses one observed product label per instance and a manually defined coarse product hierarchy. The term incomplete supervision is therefore used to describe the limited semantic annotation available in the dataset, not to claim full partial-label recovery.

2.4. Hierarchical Multi-Label Classification

Hierarchical classification uses structured relationships among labels to organize prediction tasks [4]. In many domains, labels follow a taxonomy in which broad categories contain more specific classes. Such structures can support coarse-to-fine reasoning and may reduce confusion among unrelated classes [5,22].

In retail recognition, hierarchical organization is natural because products are commonly grouped by category, department, or product family. A coarse-to-fine strategy can first identify a broad product group and then perform fine-grained recognition within that group. This reduces the effective class space at the second stage and may be useful when many classes are visually similar.

However, hierarchy-aware methods differ in their assumptions. Some approaches require complete hierarchical annotations or enforce strict consistency across multiple semantic levels [4–6]. The present study uses a simpler setting: seven manually defined product groups are used as a predefined hierarchy for refinement. The framework therefore does not solve automatic taxonomy learning or unknown hierarchy discovery.

2.5. Learning with Unknown or Latent Hierarchies

Recent studies have investigated methods for learning label hierarchies directly from data. These approaches use latent embeddings, structured label representations, or probabilistic modeling to infer relationships among classes [23,24]. Such methods are theoretically attractive for retail environments because product taxonomies may change due to new products, seasonal assortments, supplier changes, and inconsistent catalog structures.

Nevertheless, learning unknown hierarchies is a different problem from the one addressed in this paper. The present work assumes that a coarse product grouping is available before training and uses it only for hierarchical refinement. Unknown hierarchy learning is therefore discussed as related literature and as an important direction for future work, but it is not claimed as a contribution of this study.

2.6. Intelligent Retail Logistics and Operational Visual Recognition

Beyond computer vision, product recognition is connected to broader developments in intelligent retail logistics and operations management. Visual recognition can support inventory visibility, shelf auditing, automated sorting, product search, and check-out assistance. In these applications, model performance must be considered together with scalability, inference speed, ease of updating, and integration into operational workflows [15,16,25–28].

Operations and logistics research increasingly emphasizes the role of data-driven automation, uncertainty handling, and resilient information systems in modern supply chains [29,30]. These issues are relevant to retail recognition because product data are often incomplete, product catalogs change frequently, and operational systems require timely decisions under imperfect information.

The present study therefore positions retail product recognition as a practical visual recognition problem with operational relevance for retail logistics. However, it does not claim to develop a complete decision-support system. Rather, it contributes a lightweight recognition component that may be integrated into broader retail automation and logistics workflows.

Recent retail operations research further emphasizes that technology-driven retail systems should be evaluated not only by technical performance, but also by their contribution to operational visibility, service quality, responsiveness, and decision-making in retail processes. Fisher and Raman [31] summarize three decades of retail operations research and highlight the importance of data-driven innovation for improving retail execution. Similarly, Rooderkerk et al. [32] show that retail analytics has evolved from descriptive analysis toward predictive and prescriptive approaches that support assortment, inventory, pricing, and operational decisions. Related studies also demonstrate the growing value of external and digital data sources in retail forecasting [33], the operational effects of logistics service recovery in stores [34], and the broader role of digital transformation in innovation platforms [35]. Furthermore, recent research on online retailing shows that operational design choices, such as delivery speed, can influence downstream customer behavior and returns [36], reinforcing the need to view recognition technologies within broader retail operations systems.

2.7. Data Quality and Label Noise in Visual Recognition

Data quality strongly affects the reliability and reproducibility of machine learning systems. Large-scale visual datasets may contain annotation errors, duplicate samples, missing labels, inconsistent metadata, and class imbalance [37]. These issues can distort both model training and performance evaluation.

Retail datasets are particularly sensitive to such problems because products may differ only in small visual details. Annotation inconsistencies can therefore lead to unstable comparisons between models and may hide weaknesses in rare or underrepresented classes. For this reason, dataset description, preprocessing, annotation protocol, and evaluation design are important components of reproducible retail recognition research.

2.8. Positioning of the Present Work

The present work builds on these research streams while adopting a deliberately narrower and more operationally focused scope. Unlike full multi-label learning methods, it does not attempt to predict all semantic attributes of a product. Unlike latent-label reconstruction methods, it does not infer missing labels. Unlike unknown hierarchy learning approaches, it does not learn a taxonomy from data.

Instead, the study evaluates whether a lightweight combination of pretrained visual features, TSVD-based latent representation, KDE-based probabilistic classification, and predefined hierarchical refinement can improve retail product recognition under incomplete semantic supervision and class imbalance. The contribution lies in the integration and empirical evaluation of these components for a realistic retail recognition setting, with attention to accuracy, macro-F1 performance, rare-class behavior, and inference efficiency.

3. Methodology

3.1. Problem Formulation

Retail product recognition differs from standard supervised classification because product images are often annotated only with one observed product label, while broader semantic information such as product group, brand, packaging type, or variant is not exhaustively available. In this study, incomplete supervision refers to this limited annotation setting. The method does not attempt to reconstruct hidden labels or learn an unknown taxonomy; instead, it uses the observed class labels together with a predefined product hierarchy.

Let $\mathcal{X} \subset \mathbb{R}^D$ denote the visual feature space and let $\mathcal{Y} = \{1, \dots, C\}$ denote the set of fine-grained product classes. Each sample ($x_i \in \mathcal{X}$) is associated with one observed class label ($y_i \in \mathcal{Y}$). A manually defined mapping

$$(\pi : \mathcal{Y} \rightarrow \mathcal{G})$$

assigns each fine-grained class to one coarse product group ($g \in \mathcal{G}$).

The objective is to learn a classification function

$$[f : \mathcal{X} \rightarrow \mathcal{Y}]$$

that remains effective under incomplete semantic annotation, visual similarity between classes, and imbalanced class distributions.

3.2. Visual Feature Representation via TSVD

Each product image is first processed using a pretrained convolutional neural network. The final classification layer is removed, and the resulting feature vector is used as a visual representation of the image. No textual metadata, product descriptions, or transactional data are used in this study.

Let ($X \in \mathbb{R}^{N \times D}$) denote the matrix of extracted feature vectors, where (N) is the number of samples and (D) is the original feature dimensionality. To reduce redundancy and obtain a compact latent representation, truncated singular value decomposition (TSVD) is applied [14]:

$$[X \approx U_k \Sigma_k V_k^T]$$

where ($k \ll D$) denotes the number of retained components. The latent representation is obtained as

$$[Z = X V_k = U_k \Sigma_k]$$

and each sample is represented by a compact vector ($z_i \in \mathbb{R}^k$). This step reduces dimensionality, suppresses low-energy components, and provides a stable representation for subsequent probabilistic modeling.

Figure 1 illustrates the complete pipeline of the proposed retail product recognition framework. The pipeline consists of visual feature extraction, TSVD-based latent feature compression, class-conditional density estimation using KDE [38], and predefined hierarchy-guided coarse-to-fine refinement.

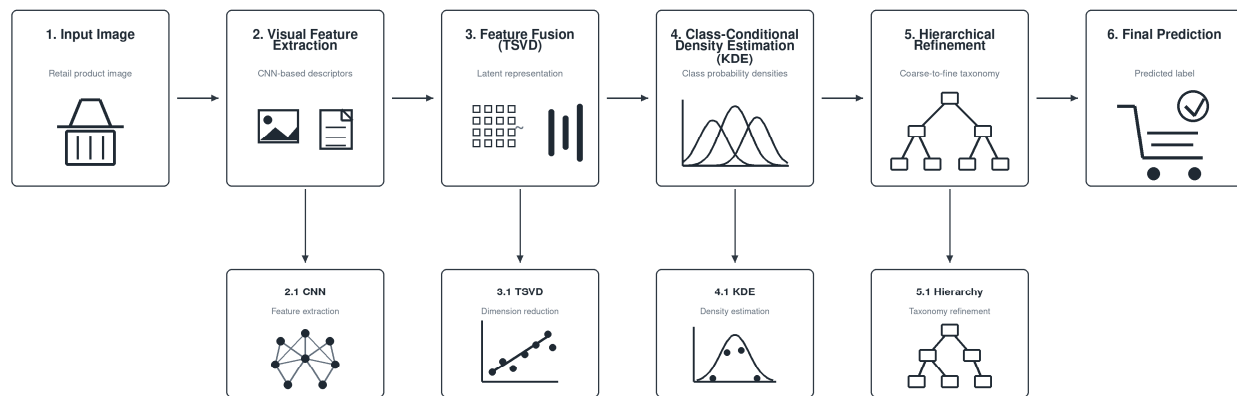


Figure 1. Complete pipeline of the proposed retail product recognition framework. The framework uses visual feature extraction, truncated singular value decomposition (TSVD)-based latent feature compression, kernel density estimation (KDE)-based class-conditional density estimation, and predefined hierarchy-guided coarse-to-fine refinement to predict the final product class.

3.3. Class-Conditional Density Modeling

After projection into the latent space, classification is formulated as a class-conditional density estimation problem. Instead of learning a direct mapping from features to labels, we model the distribution of feature vectors for each class.

For each class $c \in \mathcal{Y}$, we estimate the probability density $p(z | c)$ using kernel density estimation (KDE):

$$p(z | c) = \frac{1}{N_c h^k} \sum_{i=1}^{N_c} K\left(\frac{z - z_i^{(c)}}{h}\right)$$

In this formulation, N_c denotes the number of training samples associated with class c , h is a bandwidth parameter controlling the smoothness of the estimate, and $K(\cdot)$ is a kernel function, typically chosen as a Gaussian.

The use of KDE provides a flexible, non-parametric model of class distributions that does not impose restrictive assumptions such as Gaussianity. This is particularly important in retail settings, where visual features may exhibit complex and multi-modal distributions due to variations in packaging, lighting, and viewpoint.

3.4. Decision Rule

Given a test sample z , classification is performed by selecting the class that maximizes the posterior probability. This leads to the following decision rule:

$$\hat{y} = \operatorname{argmax}_{c \in \mathcal{Y}} p(z | c) \cdot p(c)$$

where $p(c)$ denotes the prior probability of class c .

In practice, class priors can be estimated from the empirical class distribution or set to uniform values to avoid bias toward frequent classes. The probabilistic formulation provides several advantages. It naturally provides confidence estimates for predictions, supports handling of class imbalance through priors, and avoids the implicit negative-label assumption present in standard discriminative losses.

3.5. Hierarchical Classification and Progressive Refinement

Retail product categories are inherently organized in hierarchical taxonomies, ranging from coarse product groups to fine-grained variants. Exploiting this structure can significantly reduce classification ambiguity.

Instead of performing flat classification over all classes, we adopt a coarse-to-fine hierarchical strategy. At the first stage, a global classifier predicts a coarse product group. This step captures dominant visual characteristics and reduces the search space.

In subsequent stages, classification is restricted to a subset of classes belonging to the predicted group. Dedicated TSVD–KDE models are trained for each subset, allowing more fine-grained discrimination among visually similar products.

This progressive refinement strategy has two key advantages. First, it reduces confusion between unrelated classes by limiting comparisons to semantically coherent subsets. Second, it enables the model to focus on increasingly subtle visual cues at deeper levels of the hierarchy.

Figure 2 illustrates the hierarchical decision process used in the refinement stage. A coarse product group is predicted first, which defines the candidate subset of fine-grained product classes. KDE-based classification is then applied only within this subset to obtain the final prediction.

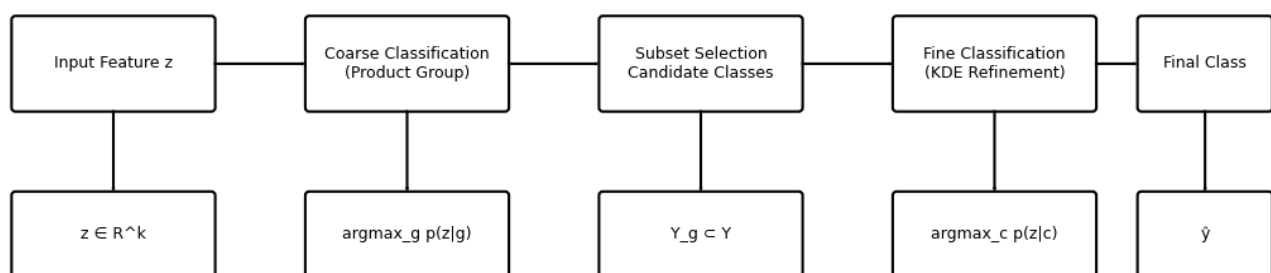


Figure 2. Hierarchical coarse-to-fine classification process. A coarse classifier first predicts a predefined product group, which defines the candidate subset of fine-grained classes. A refined kernel density estimation (KDE)-based probabilistic classifier is then applied within this subset to obtain the final product prediction.

3.6. Interpretation and Design Rationale

The proposed framework combines three components: TSVD-based feature compression, KDE-based probabilistic classification, and predefined hierarchical refinement. TSVD reduces the dimensionality of pretrained visual features. KDE models class-conditional feature distributions in the latent space. The hierarchy restricts fine-grained classification to semantically related candidate classes.

The contribution of the framework lies in this lightweight integration rather than in the invention of a new standalone algorithm. Compared with end-to-end deep architectures, the proposed design separates feature extraction from probabilistic inference, which can simplify model updates when product catalogs change. At the same time, the method remains limited by the quality of the pretrained features, the predefined hierarchy, and the selected KDE bandwidth.

4. Experimental Analysis

4.1. Dataset Description

The dataset was collected in a controlled laboratory environment designed to represent a small supermarket shelf setting. Real commercially available retail products were arranged on shelves and photographed under realistic retail-like conditions. The setup included different product categories, packaging formats, product sizes, shelf positions, and varying degrees of visual similarity between items.

Figure 3 shows the laboratory shelf setup used during data collection. The arrangement was designed to reproduce typical visual challenges encountered in retail product

recognition, including cluttered shelf layouts, partial occlusions, different object scales, similar packaging designs, and variations in product orientation.



Figure 3. Laboratory-based retail shelf setup used for image acquisition.

All images were captured by the authors as part of the experimental setup. No customer data, personal information, or human subjects were involved. The dataset contains only product images and object-level annotations.

The final dataset consists of 13,969 images covering 59 product classes. Images were annotated manually using the YOLO annotation format. Each visible product instance was labeled with a bounding box and assigned to one observed product class. Additional semantic attributes such as brand, packaging type, flavor, or product variant were not exhaustively annotated. This annotation structure motivates the incomplete supervision setting considered in this study.

The dataset contains multiple product instances per image and reflects several practical retail recognition challenges, including fine-grained visual similarity, class imbalance, occlusion, and heterogeneous packaging appearances.

4.2. Data Quality Assessment

A comprehensive validation of the dataset revealed substantial annotation inconsistencies. In total, 19,587 warnings and errors were identified, including

- Malformed annotation entries;
- Invalid bounding box coordinates outside the normalized range [0, 1];
- Duplicated bounding boxes;
- Missing or empty annotation files.

Such inconsistencies introduce uncontrolled noise and violate fundamental assumptions of object detection and classification models. Consequently, direct training on the raw dataset would lead to unreliable and non-reproducible results.

Additionally, a subset of images without corresponding annotation files was detected. These cases require careful interpretation, as they may correspond to valid negative samples or incomplete labeling. Treating them naively as background-only images may introduce bias into the learning process.

4.3. Data Cleaning and Preprocessing

To address these issues, an automated preprocessing pipeline was developed to enforce strict consistency with the YOLO annotation format. The pipeline performs the following operations:

- Validation and removal of malformed annotation entries;
- Correction of bounding box coordinates to the valid range;
- Elimination of duplicate annotations;
- Correction of zero-area bounding boxes;
- Verification of image–annotation consistency.

The cleaning procedure is conservative and only modifies annotations that violate formal constraints. No semantic relabeling is performed, ensuring that the original data distribution is preserved.

After preprocessing, all annotations conform to a consistent and machine-readable format. This step establishes a reliable foundation for subsequent training and evaluation and ensures reproducibility.

4.4. Class Imbalance Analysis

Following preprocessing, the dataset retains a pronounced long-tail distribution. High-frequency classes contain between 500 and 1300 instances, while several classes contain fewer than 10 samples.

This imbalance has significant implications for classification performance. Models trained without explicit mechanisms for handling rare classes tend to overfit dominant categories and exhibit poor generalization on underrepresented classes.

The hierarchical classification strategy proposed in this work mitigates this issue by restricting decision spaces to semantically coherent subsets. This reduces competition between unrelated classes and improves robustness in fine-grained recognition tasks.

4.5. Dataset Splitting and Leakage Prevention

The dataset is divided into training and validation subsets using an 80/20 split. To ensure a scientifically valid evaluation, particular attention is paid to preventing data leakage.

Duplicate and near-duplicate images are identified using MD5 hashing. Images with identical content are grouped and treated as indivisible units during splitting. This ensures that no duplicate image appears in both training and validation sets.

This step is critical in retail datasets, where identical product images may appear multiple times under different conditions. Without such precautions, performance metrics may be artificially inflated.

The final dataset structure follows the standard YOLO directory format, enabling seamless integration with existing training pipelines.

4.6. Evaluation Protocol and Quantitative Results

All experiments were conducted under the same feature-extraction and data-splitting setup to ensure comparability across methods. Visual features were extracted using a pretrained ResNet-50 backbone with ImageNet weights. The final classification layer was removed, and feature vectors were obtained from the global average pooling layer, resulting in a feature dimensionality of $D = 2048$.

For the proposed approach, TSVD was used to project the extracted feature vectors into a latent space of dimension $k = 128$. The KDE classifier used a Gaussian kernel. The bandwidth parameter was selected through validation-based grid search over $h \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$. The value $h = 0.5$ achieved the best validation macro-F1 score and was used in all experiments. Class priors were estimated from the empirical class distribution.

The hierarchy used in the experiment was manually predefined based on seven product groups: beverages, snacks, non-alcoholic drinks, baked goods, alcoholic beverages, cakes and pastries, and dairy products. At inference time, the model first predicts the coarse product group and then applies the corresponding group-specific TSVD-KDE classifier.

The evaluated baselines include a CNN with linear classifier, k-nearest neighbors with $k = 5$, an RBF-kernel SVM, and an ensemble-based classifier. All baseline methods used the same extracted ResNet-50 feature representations. This design isolates the effect of the classification strategy from the effect of the visual backbone.

Table 1 reports the classification performance of all evaluated methods.

Table 1. Classification performance comparison. CNN: convolutional neural network; kNN: k-nearest neighbors; SVM: support vector machine; RBF: radial basis function; TSVD: truncated singular value decomposition; KDE: kernel density estimation.

Method	Top-1 Accuracy (%)	Top-5 Accuracy (%)	Macro-F1 (%)	Inference Time (ms)
CNN + Linear Classifier	78.4	92.1	64.7	4.8
kNN ($k = 5$)	75.2	90.3	61.5	12.6
SVM (RBF)	80.1	93.4	66.8	9.3
Ensemble Model	82.3	94.6	67.8	21.7
TSVD-KDE (proposed)	81.5	94.1	69.5	5.2
TSVD-KDE + Hierarchy	83.7	95.8	71.8	6.1

The results indicate that the proposed TSVD-KDE model improves balanced recognition performance compared with the CNN + Linear Classifier baseline. Top-1 accuracy increases from 78.4% to 81.5%, while macro-F1 improves from 64.7% to 69.5%. The improvement in macro-F1 is particularly relevant because the dataset contains a pronounced long-tail class distribution.

The hierarchical extension provides further gains. Compared with the flat TSVD-KDE model, TSVD-KDE + Hierarchy improves Top-1 accuracy from 81.5% to 83.7%, Top-5 accuracy from 94.1% to 95.8%, and macro-F1 from 69.5% to 71.8%. These results suggest that restricting fine-grained classification to semantically related product groups can reduce ambiguity among visually similar products.

The ensemble model achieves competitive accuracy but requires substantially higher inference time. Its inference time is 21.7 ms per sample, compared with 6.1 ms for TSVD-KDE + Hierarchy. Therefore, under the present experimental setup, the hierarchical TSVD-KDE model offers a favorable accuracy-efficiency trade-off.

To further examine the effect of class imbalance, performance was analyzed across class-frequency groups. The hierarchical TSVD-KDE model achieved 88.9% Top-1 accuracy for frequent classes with more than 100 samples, 79.6% for medium-frequency classes with 10–100 samples, and 62.4% for rare classes with fewer than 10 samples. These results show that rare classes remain challenging, even for the hierarchy-aware model. A more detailed comparison of rare-class behavior across additional baselines is left for future work. To isolate the contribution of individual components, an ablation study was conducted. The results are shown in Table 2.

The ablation results show a gradual performance increase after adding each component. TSVD improves the CNN baseline from 78.4% to 79.6% Top-1 accuracy and from 64.7% to 66.2% macro-F1. Adding KDE further improves performance to 81.5% Top-1 accuracy and 69.5% macro-F1. The hierarchical refinement step provides the strongest final configuration, reaching 83.7% Top-1 accuracy and 71.8% macro-F1.

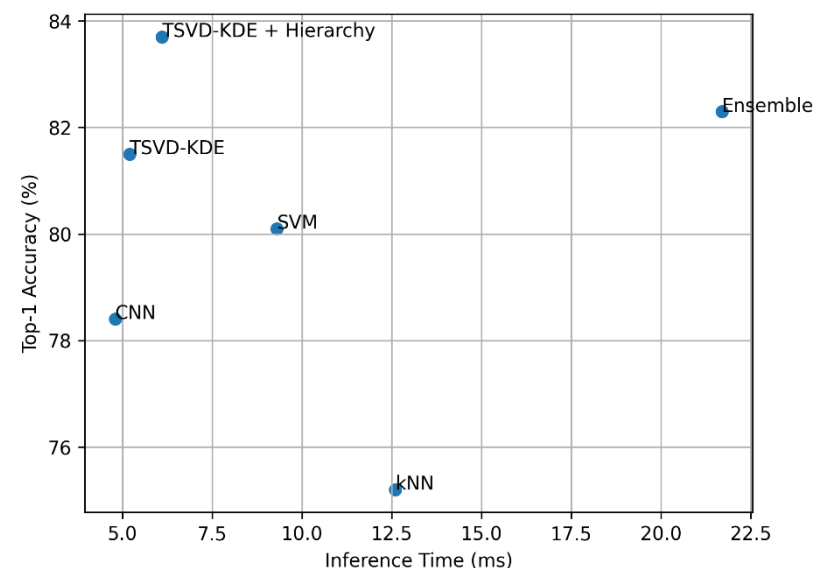
Table 2. Ablation study evaluating the contribution of truncated singular value decomposition (TSVD), kernel density estimation (KDE), and hierarchical refinement.

Configuration	Top-1 Accuracy (%)	Macro-F1 (%)
CNN + Linear Classifier	78.4	64.7
TSVD only	79.6	66.2
TSVD + KDE	81.5	69.5
TSVD + KDE + Hierarchy	83.7	71.8

These findings suggest that the performance improvement is not caused by a single component alone, but by the combined effect of compact latent representation, probabilistic class modeling, and predefined coarse-to-fine refinement.

4.7. Accuracy–Efficiency Trade-Off

Retail recognition systems require models that balance predictive performance with computational efficiency. Figure 4 illustrates the relationship between Top-1 accuracy and average inference time for the evaluated methods.

**Figure 4.** Trade-off between Top-1 accuracy and average inference time across the evaluated methods. Inference time was measured on an NVIDIA RTX 3090 graphics processing unit (GPU) using batch size 1.

Inference time was measured on an NVIDIA RTX 3090 GPU using batch size 1. All methods were evaluated under the same hardware and software conditions to ensure a consistent comparison.

The CNN + Linear Classifier provides the fastest inference time, but it achieves lower Top-1 accuracy and macro-F1 than the proposed TSVD–KDE-based models. The ensemble model achieves competitive Top-1 accuracy, but its inference time is substantially higher than that of the other evaluated methods.

The hierarchical TSVD–KDE model achieves the highest Top-1 accuracy and macro-F1 in this comparison while maintaining a low average inference time of 6.1 ms per sample. These results suggest a favorable accuracy–efficiency trade-off under the present experimental setup.

However, the computational analysis is limited to inference time. Training time, memory consumption, and scalability with respect to larger product catalogs were not systematically evaluated and are therefore discussed as limitations and future research directions.

5. Discussion

This research sheds light on how to create lightweight retail product recognition systems when faced with incomplete semantic supervision and class imbalance. The findings indicate that in the current dataset, recognition performance is affected by the choice of classifier as well as by the feature space structure and the label space organization. The macro-F1 performance is enhanced by the combination of TSVD-based latent representation, KDE-based probabilistic classification, and predefined hierarchical refinement when compared to the flat baselines that were assessed.

It seems that the hierarchical refinement step is particularly beneficial in this experimental context. The model limits the number of candidate labels in the second stage by initially assigning a product image to a coarse product group and then restricting fine-grained classification to the corresponding subset of classes. This might lessen the confusion between unrelated product groups, which could account for the observed improvement in the macro-F1 score. The hierarchy employed in this study, however, is a manually predefined one and does not reflect an automatically learned taxonomy. As a result, the findings should be seen as evidence for the potential value of predefined coarse-to-fine refinement, rather than as a universal answer to hierarchy learning.

The TSVD-KDE formulation offers a straightforward probabilistic approach from a methodological standpoint, serving as an alternative to a purely discriminative linear classifier using pretrained CNN features. While TSVD diminishes the dimensionality of the feature vectors that have been extracted, KDE is responsible for modeling class-conditional densities within the latent space. When product catalogs need updates without the necessity of retraining a complete end-to-end model, this division between feature extraction and probabilistic classification can be quite beneficial. Still, the quality of the approach relies on the pretrained feature extractor, the chosen latent dimensionality, and the bandwidth of the KDE.

As shown by the computational results, the proposed framework has a favorable inference-time profile given the current experimental setup. While the hierarchical TSVD-KDE model achieves higher Top-1 accuracy and macro-F1 scores than the baselines we evaluated, it also keeps an average inference time of 6.1 ms per sample on the tested hardware. Nonetheless, the analysis only considers inference time. Aspects such as training time, memory usage, and scalability concerning significantly larger product catalogs were not assessed systematically. Due to this, assertions regarding operational deployment should be approached with caution and need further validation.

5.1. Implications for Theory

By exploring the combination of a predefined product hierarchy with probabilistic classification under incomplete semantic supervision, this research adds to the body of work on retail product recognition. Instead of introducing a new deep architecture, this work assesses a lightweight combination of pretrained visual features, a TSVD-based latent representation, KDE-based class-conditional modeling, and coarse-to-fine refinement.

The results indicate that organizing the label space hierarchically can enhance balanced recognition performance when product classes are visually alike and not evenly distributed. This is consistent with prior research on hierarchy-aware classification and suggests that relatively simple probabilistic models can perform well when paired with compact visual representations.

It also makes clear the difference between incomplete semantic supervision and full partial-label learning. The proposed method does not reconstruct hidden labels or learn about unknown taxonomies. Rather, it utilizes the observed class labels that are available along with a predefined hierarchy. This differentiation is crucial for future studies because retail datasets frequently include incomplete semantic information, but not every situation requires or can accommodate full latent-label reconstruction.

5.2. Implications for Practice

In retail recognition settings, the proposed framework could be practically useful where computational simplicity and update flexibility are important. Since feature extraction is decoupled from probabilistic classification, new product classes may be incorporated by simply updating the classifier components instead of having to retrain the entire end-to-end model. This is pertinent for retail settings that have changing assortments and regular product updates.

These practical implications are consistent with the recent retail operations literature, which emphasizes the increasing role of data-driven technologies in improving retail visibility, responsiveness, and operational execution [31,32]. In this context, visual recognition can support retail analytics by providing more timely and structured information about products, shelves, and inventory states. However, as shown in prior work on retail forecasting, logistics service recovery, and digital transformation, the operational value of such systems depends not only on predictive accuracy, but also on integration with existing retail processes and decision routines [33–35].

In fact, the predefined hierarchy mirrors the organization of many retail product catalogs in practice. By categorizing products into coarse groups before fine-grained classification, this may enhance the transparency of the recognition process and potentially reduce confusion between unrelated product categories. This is important for applications like shelf monitoring, inventory auditing, product search, and assisting with checkout.

Nonetheless, the practical implications ought to be viewed with caution. Using a single laboratory-based retail dataset, the experiments were carried out, and the assessment centered on classification accuracy, macro-F1, rare-class behavior, and inference time. Before concluding on deployment in operational retail or logistics systems, further validation is necessary.

5.3. Limitations of the Study and Future Research Directions

Several limitations should be acknowledged. Initially, the assessment relies on a single laboratory-based retail dataset with 59 product classes. Even though the dataset comprises actual commercially available products and simulates various retail-like visual challenges, it fails to encompass the full range of variability found in operational super-market environments, such as customer movement, fluctuating light conditions, seasonal shelf arrangements, camera variations, and store-specific merchandising strategies. Future research should assess the framework using more public retail datasets and larger product catalogs.

Secondly, the design of the experiment relies on a fixed train–validation split. Even though all the methods evaluated share the same split, the lack of repeated experiments, cross-validation, confidence intervals, and statistical significance tests weakens the reliability of the performance claims. Future studies ought to incorporate repeated assessments across various random seeds and should include standard deviations or confidence intervals.

Third, the baseline comparison only includes conventional classifiers that utilize the same pretrained CNN feature representation. While this design allows for a controlled

comparison of classifier behavior, it does not prove superiority over recent state-of-the-art methods like vision transformers, contrastive learning models, metric learning approaches, few-shot learning methods, or sophisticated hierarchical classifiers. Future research should incorporate these methods as extra baselines.

Fourth, the computational analysis considers only inference time. There was no systematic evaluation of training time, memory consumption, and scalability concerning the increase in the number of classes. Since KDE-based methods might need to store class-specific feature representations, future research should look into the memory requirements and computational scaling when dealing with larger retail catalogs.

Fifth, the hierarchy this study relies on is predefined and set manually. The framework does not learn unknown taxonomies or reconstruct latent hierarchical structures from the data. Future studies ought to explore adaptive hierarchy learning and automatic taxonomy construction techniques for retail environments that have incomplete or evolving product structures.

In conclusion, this research relies solely on visual image information. There are no product descriptions, pricing details, or transactional data, nor is there any textual metadata. Extensions in the future could explore multimodal recognition settings, but this would necessitate additional data sources and a different experimental design.

6. Conclusions

A lightweight probabilistic framework for retail product recognition with incomplete semantic supervision was introduced in this paper. This method brings together several elements: extracting visual features using pretrained models, compressing latent features with TSVD, estimating class-conditional densities through KDE, and employing a predefined hierarchical coarse-to-fine refinement strategy. It is built to tackle challenges that are commonly seen in retail product recognition, such as fine-grained visual similarity, class imbalance, and the scarcity of semantic annotation.

As shown by the experimental results on a laboratory-based retail dataset, the hierarchical TSVD-KDE model enhances Top-1 accuracy and macro-F1 compared to the flat baselines that were assessed. Notably, the enhancement in macro-F1 indicates that predefined hierarchical refinement can aid in achieving a more balanced classification performance when dealing with long-tail class distributions. In the current experimental setup, the method keeps a low inference time, which indicates a positive accuracy-efficiency balance.

The results need to be understood in the context of the study. The assessment relies on a single dataset, a predetermined hierarchy, and a small selection of baseline methods. In addition, the analysis emphasizes classification accuracy, macro-F1, rare-class behavior, and inference time; however, training time, memory requirements, statistical robustness, and external validation on other retail datasets are still unresolved issues.

Future research should assess the framework on larger and more varied public retail datasets, incorporate repeated experiments and statistical significance testing, and compare it with recent deep learning, transformer-based, contrastive learning, and advanced hierarchical classification methods. Future expansions could explore adaptive hierarchy learning and multimodal recognition by incorporating extra data sources like product descriptions, metadata, or transactional information.

Author Contributions: Conceptualization, M.H.; Investigation, M.H.; Methodology, M.H.; Resources, M.H.; Software, M.H.; Supervision, B.N. and A.H.; Validation, M.H.; Writing—original draft, M.H.; Writing—review and editing, M.H., B.N. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Universität Duisburg–Essen. We acknowledge support from the Open Access Publication Fund of the University of Duisburg–Essen.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105.
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
3. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
4. Vens, C.; Struyf, J.; Schietgat, L.; Džeroski, S.; Blockeel, H. Decision Trees for Hierarchical Multi-Label Classification. *Mach. Learn.* **2008**, *73*, 185–214. [[CrossRef](#)]
5. Wehrmann, J.; Cerri, R.; Barros, R.C. Hierarchical Multi-Label Classification Networks. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 5225–5234.
6. Giunchiglia, E.; Lukasiewicz, T. Coherent Hierarchical Multi-Label Classification Networks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9662–9673.
7. Arpit, D.; Jastrzębski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M.S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. A Closer Look at Memorization in Deep Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 233–242.
8. Zhou, D.; Chen, P.; Wang, Q.; Chen, G.; Heng, P.-A. Acknowledging the Unknown for Multi-Label Learning with Single Positive Labels. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 423–440.
9. Kim, Y.; Kim, J.M.; Akata, Z.; Lee, J. Large Loss Matters in Weakly Supervised Multi-Label Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 14156–14165.
10. Tsoumakas, G.; Katakis, I. Multi-Label Classification: An Overview. *Int. J. Data Warehous. Min.* **2007**, *3*, 1–13.
11. Guo, Y.; Gu, S. Multi-Label Classification Using Conditional Dependency Networks. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Spain, 16–22 July 2011; pp. 1300–1305.
12. Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; Xu, W. CNN-RNN: A Unified Framework for Multi-Label Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2285–2294.
13. Jing, L.; Yang, L.; Yu, J.; Ng, M.K. Semi-Supervised Low-Rank Mapping Learning for Multi-Label Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1483–1491.
14. Halko, N.; Martinsson, P.-G.; Tropp, J.A. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Rev.* **2011**, *53*, 217–288. [[CrossRef](#)]
15. Min, S.; Zacharia, Z.; Smith, C.D. Defining Supply Chain Management: In the Past, Present, and Future. *J. Bus. Logist.* **2019**, *40*, 44–55. [[CrossRef](#)]
16. Gunasekaran, A.; Papadopoulos, T.; Dubey, R.; Wamba, S.F.; Childe, S.J.; Hazen, B.; Akter, S. Big Data and Predictive Analytics for Supply Chain and Organizational Performance. *J. Bus. Res.* **2017**, *70*, 308–317. [[CrossRef](#)]
17. Cole, E.; Mac Aodha, O.; Lorieul, T.; Perona, P.; Morris, D.; Jojic, N. Multi-Label Learning from Single Positive Labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 21–24 June 2021; pp. 933–942.
18. Durand, T.; Mehrasa, N.; Mori, G. Learning a Deep ConvNet for Multi-Label Classification with Partial Labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 647–657.
19. Kundu, A.; Bas, S.; Lam, M.; Chen, Z.; Modolo, D.; Saenko, K.; Bulò, S.R.; Tighe, J. Exploiting Weakly Supervised Visual Patterns to Learn from Partial Annotations. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33.

20. Huynh, D.; Elhamifar, E. Interactive Multi-Label CNN Learning with Partial Labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 9423–9432.
21. Xu, N.; Qiao, C.; Lv, J.; Geng, X.; Zhang, M.-L. One Positive Label Is Sufficient: Single-Positive Multi-Label Learning with Label Enhancement. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 4513–4521.
22. Cerri, R.; Barros, R.C.; de Carvalho, A.C.P.L.F.; Jin, Y. Reduction Strategies for Hierarchical Multi-Label Classification in Protein Function Prediction. *BMC Bioinform.* **2016**, *17*, 373. [[CrossRef](#)]
23. Patel, D.; Dangati, P.; Lee, J.-Y.; Boratko, M.; McCallum, A. Modeling Label Space Interactions in Multi-Label Classification Using Box Embeddings. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Conference, 25–29 April 2022.
24. Jo, S.; Shin, D.; Na, B.; Jang, J.; Moon, I.-C. Hierarchical Multi-Label Classification with Partial Labels and Unknown Hierarchy. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM), Birmingham, UK, 21–25 October 2023; pp. 1025–1034.
25. Qiu, X.; Luo, H.; Xu, G.; Zhong, R.; Huang, G.Q. Physical Assets and Service Sharing for IoT-Enabled Supply Hub in Industrial Park (SHIP). *Int. J. Prod. Econ.* **2015**, *159*, 4–15. [[CrossRef](#)]
26. Fosso Wamba, S.; Gunasekaran, A.; Akter, S.; Ren, S.J.-f.; Dubey, R.; Childe, S.J. Big Data Analytics and Firm Performance: Effects of Dynamic Capabilities. *J. Bus. Res.* **2017**, *70*, 356–365. [[CrossRef](#)]
27. Ivanov, M.; Dolgui, A. A Digital Supply Chain Twin for Managing the Disruption Risks and Resilience in the Era of Industry 4.0. *Prod. Plan. Control.* **2021**, *32*, 775–788. [[CrossRef](#)]
28. Dubey, R.; Gunasekaran, A.; Childe, S.J.; Papadopoulos, T.; Luo, Z.; Wamba, S.F.; Roubaud, D. Can Big Data and Predictive Analytics Improve Social and Environmental Sustainability? *Technol. Forecast. Soc. Change* **2019**, *144*, 534–545. [[CrossRef](#)]
29. Christopher, M.; Peck, H. Building the Resilient Supply Chain. *Int. J. Logist. Manag.* **2004**, *15*, 1–14. [[CrossRef](#)]
30. Lee, H.L. The Triple-A Supply Chain. *Harv. Bus. Rev.* **2004**, *82*, 102–112. [[PubMed](#)]
31. Fisher, M.; Raman, A. Innovations in retail operations: Thirty years of lessons from Production and Operations Management. *Prod. Oper. Manag.* **2022**, *31*, 4452–4461. [[CrossRef](#)]
32. Rooderkerk, R.P.; DeHoratius, N.; Musalem, A. The past, present, and future of retail analytics: Insights from a survey of academic research and interviews with practitioners. *Prod. Oper. Manag.* **2022**, *31*, 3727–3748. [[CrossRef](#)]
33. Fu, Y.; Fisher, M. The value of social media data in fashion forecasting. *Manuf. Serv. Oper. Manag.* **2023**, *25*, 1136–1154. [[CrossRef](#)]
34. Peinkofer, S.T.; Esper, T.L.; Smith, R.J.; Williams, B.D. Retail ‘save the sale’ tactics: Consumer perceptions of in-store logistics service recovery. *J. Bus. Logist.* **2022**, *43*, 238–264.
35. Fukawa, N.; Fisher, G.J. Managing digital transformation with blockchain: Distributed ledger innovation platforms. *J. Product. Innov. Manag.* **2025**, *42*, 766–788. [[CrossRef](#)]
36. Masuch, S.; Landwehr, J.R.; Flath, C.M.; Thiesse, F. The faster, the better? The impact of short delivery times on product returns in online retailing. *J. Retail.* **2024**, *100*, 475–485. [[CrossRef](#)]
37. Northcutt, C.G.; Athalye, A.; Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, Virtual Conference, 6–14 December 2021.
38. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman and Hall: London, UK, 1986.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.