World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

Multimodal machine learning for catalogue metadata correction in online retail

Yaswanth Jeganathan *

Carnegie Mellon University (Pittsburgh, Pennsylvania, USA).

World Journal of Advanced Engineering Technology and Sciences, 2025, 16(01), 475-483

Publication history: Received on 15 June 2025; revised on 21 July 2025; accepted on 24 July 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.16.1.1241

Abstract

The quality of catalogue metadata affects the success of the e-commerce platforms at every level in search accuracy, customer satisfaction, and many other. This review examines the implementation of multimodal machine learning (MML) in correcting catalogue metadata, with the emphasis being put on the combination of the input of textual, visual, and structured data. It describes the theoretical underpinnings, model frameworks, fusion policies, and benchmarking procedures that are being actively used in the research. As empirical evidence, it has proven that MML methods performed more soundly than unimodal baselines at accuracy, F1 scores, and tasks involving metadata imputation. Another essential struggle, namely modality misalignment, interpretability, and domain generalization, is also mentioned in the review. The directions of future work are addressed, which include multilingual support, explainable AI, knowledge graph integration, and active learning. The current paper serves as a multifaceted guide to inform researchers and practitioners who are interested in enhancing the accuracy of metadata in a very large-scale retail setting of a digital nature.

Keywords: Multimodal Machine Learning; Metadata Correction; E-Commerce; Product Catalogs; Cross-Modal Fusion

1. Introduction

The growth of e-commerce sites and online stores has created pressure on having very accurate and well up to date metadata of the product catalogue. The metadata that is applied to retail merchandise includes but is not limited to the product name, brand, category, material, dimensions, and pricing, and is core to search optimization, recommendation engines, customer experience, and supply chain activities, as cataloguing metadata [1]. Metadata mistakes, including missing, misclassified, mismatched, and duplicated values, largely sabotage the quality of online retailing sites. Such inaccuracy not only interferes with customer confidence as well as the discovery of products but incurs inefficiencies in the organization and wastes profits [2].

As online product listing moves at a high pace, especially within a marketplace (such as that of Alibaba, Flipkart), catalogue metadata cannot be corrected manually. Machine learning (ML)-based automated methods have been attracting more and more interest because they can handle large data volumes. Nonetheless, these traditional ML models tend to rely on unimodal input, which could be either text or image only, and would have some limitations regarding their contextual reasoning [3]. The current state of multimodal machine learning (MML) or a combination of data across many modalities, including photos, text, and structured features, provides a more reliable and semantically deeper method of metadata fixing.

Images, combined with text, allow multimodal models to rectify incorrect or missing metadata entries (e.g., SKU code, brand name) as well as contribute to completeness of metadata on everything from structured data (e.g., SKU code, size) to free form text (e.g., bullet points, brand name) [4]. This combination enhances classification actions like product categorization, attribute completion, and brand validation by combining the positive attributes of each modality. As an

* Corresponding author: Yaswanth Jeganathan

example, text can work well when it comes to the recognition of technical specifications, yet images can be a lot more helpful when it comes to the recognition of visual characteristics, i.e., based on color, style, or pattern [5].

Multimodal learning in a broader sense of machine learning is a paradigm shift to context and knowledge-intensive models. This shift plays a significant role in e-commerce because it enables it to overcome such challenges as heterogeneity in the data source, noisy channels of input, and cross-domain generalization requirements. The application of the MML approaches is also consistent with the growing need of the industry to be more personal and real-time in terms of verifying content on the multilingual and multicultural marketplace [6].

With its promise, however, the lift of MML in the correction of catalogue metadata comes with many technical and practical issues. One of the main contributions that this study makes to the literature is the absence of quality and annotated data sets on multimodal and tailored to the retail metadata tasks. The large majority of available datasets are of unimodal nature or not annotated rigorously enough to enable practical cross-modality learning [7]. Also, modality misalignment, whereby image and text input play on the variables that are somewhat different about the same product, poses difficulty to training coherent joint records.

Moreover, the complexity of multimodal models, in general, and transformer-based models, in particular, might impede their adoption in real-time systems, especially when it comes to small and medium-sized companies that do not have high-performance infrastructure [8]. The interpretation of model outputs and how easy they are to understand is also a concern; such a case might be encountered in retail areas where the wrong metadata might have regulatory effects, such as cases of pharmaceuticals or electronics.

The domain adaptation problem is yet another under-studied problem. There seems to be a lack of generalisation in the multimodal models trained on one branch of the retail (e.g., fashion) and subsequently poor performance in other areas (e.g., consumer electronics). In addition to that, biases transfer across modalities, the presence of label noise in crowd-sourced annotations, and the necessity to come up with efficient fusion strategies present an additional overlay in the development and assessment of systems [9]. This review aims to give an in-depth review of multimodal machine learning solutions to catalogue metadata rectification in online retailing.

2. Literature Survey

Table 1 Summary of recent research contributions across AI, multimodal learning, and data integration domains

Ref. No.	Focus Area / Title Summary	Methodology / Approach	Key Findings / Contributions
[10]	AI applications in B2C fashion retail	Literature review of AI implementations in the fashion industry	Identified trends in AI use for recommendation, personalization, and customer engagement
[11]	Multimodal vision-language models for object detection	Review of vision-language model capabilities for object detection	Discussed advantages of multimodal approaches and identified research gaps in current LLM object detection methods
[12]	Large-scale multimodal representation learning in commerce	Developed a retrieval framework using vision-language models for e-commerce	Proposed the Commercemm model, improving retrieval performance by integrating multi-modal data
[13]	Joint representation learning for urban land use classification	Self-supervised learning using multi-source geographic datasets	Achieved improved land-use classification by fusing geospatial data in an unsupervised setting
[14]	Fusion technologies in IoT for intrusion detection	Survey of datasets, tools, and challenges in IoT security	Identified critical issues in current IDS technologies and emphasized the role of dataset diversity
[15]	Machine learning and big data in e-commerce security	Analytical study of anomaly detection and cybersecurity in online commerce	Highlighted proactive threat detection using ML and big data in real-time systems

[16]	Transformers in multilingual product description generation	Implementation of LLMs for multilingual e-commerce content automation	Showed increased engagement through better product descriptions generated by transformer-based LLMs
[17]	Multimodal attributed graph benchmarks	Conceptual proposal and benchmarking analysis	Provided benchmarking strategies for multimodal attributed graph models; rethought existing evaluation techniques
[18]	Survey on machine learning and big data convergence	Systematic review of methods and integration strategies	Offered a unified view of ML–Big Data integration challenges and their implications for future applications
[19]	Semantic alignment in digital music libraries	Semantic integration of music library metadata	Proposed methods for unified metadata access across heterogeneous digital music collections

3. Proposed Theoretical Model and Block Diagram

3.1. Theoretical Background and Architectural Motivation

Catalogue metadata correction works on any conflict and inaccuracy that are present in the product attribute data, such as missing categories or misclassified brands, or any piece of information that is visually contradictory. The conventional unimodal models that use only the input text or images lack in their ability to generalize, and this may not work on other kinds of products and patterns of error. Multimodal machine learning (MML) provides a principled way of combining heterogeneous data sources by combining complementary information in text, images, and structured metadata of products [20].

The theoretical model that is proposed aims at applying these modalities in a systematic way to carry out error correction, detection, and validation of retail catalogue metadata. It has six major subcomponents that is, Input Modalities, Feature Extraction, Fusion and Alignment, Error Identification, Correction Module, and Validation and Output Layer.

3.2. Block Diagram: Multimodal Metadata Correction Architecture

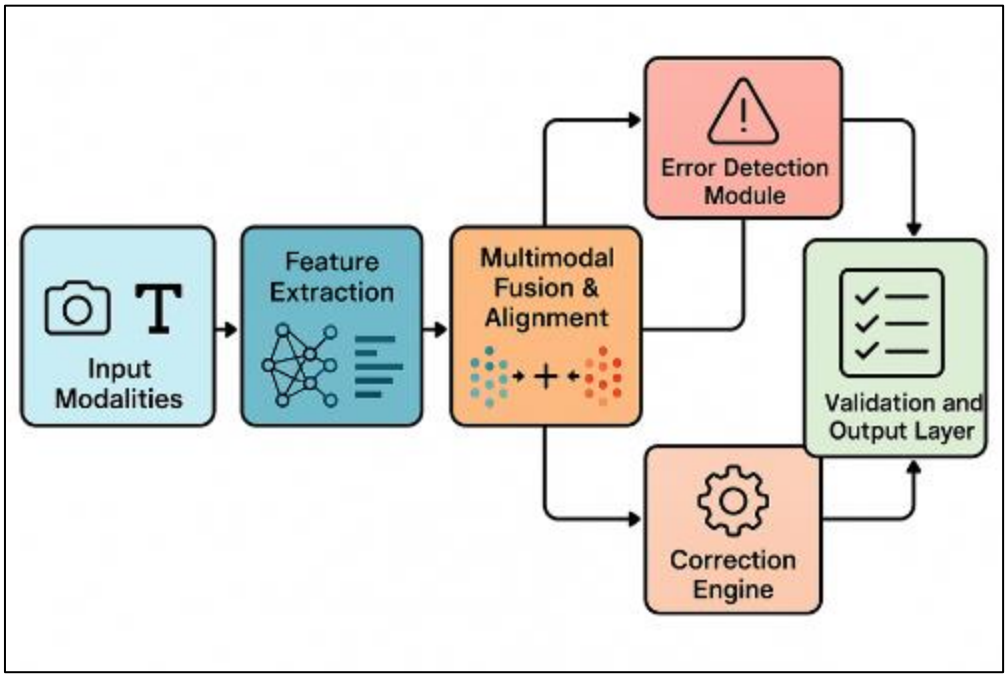

Figure 1 Block Diagram of the Multimodal Metadata Correction Framework

3.2. Component-wise Description

3.2.1. Input Modalities

The module can take as input 3 data types: images of product (e.g., .jpg, .png), textual metadata (e.g., name of product, best features), and structured data (e.g., SKU, brand, size). Tokenization, resizing, and schema normalization are done in preprocessing. Management of multimodal inputs is a necessity because the semantics of a single modality does not provide complete context within retail cataloguing [21].

3.2.2. Feature Extraction

Visual features are obtained with the assistance of pre-trained convolutional neural networks (e.g., EfficientNet, ResNet). Language models can be BERT or RoBERTa to perform text coding with any kind of textual data. Structured fields would be represented with either one-hot encoding or embedding look-up tables so that attribute relationships would not be lost [22].

3.2.3. Multimodal Fusion & Alignment

Modality features are combined in some manner, whether by cross-attention or co-attentive pooling, or by concatenation of tensors. This component forms collective embeddings, which learn inter-modal relationships (e.g., checking the correspondence of an image and the stated product category). Cross-attention was found to be superior to naive concatenation, specifically on the use of heterogeneous retail datasets [23].

3.2.4. Error Detection Module

The detection module identifies metadata issues using supervised classifiers (e.g., XGBoost, MLPs) and unsupervised anomaly detectors (e.g., Isolation Forests, Autoencoders). This system flags misclassifications, missing fields, and inconsistencies between image-text pairs [24].

3.2.5. Correction Engine

Once errors are identified, correction strategies are triggered. These may include:

- Imputation using similarity-based retrieval or model prediction;
- Normalization to standardize format variants (e.g., “XL” vs “Extra Large”);
- Semantic recovery using ontology mappings or fine-tuned classification models. Correction predictions are filtered through domain-specific constraints to ensure realism and validity [25].

3.2.6. Validation and Output Layer

The final step enforces validation checks using regular expressions, business rules, and logical constraints (e.g., preventing shoes from being tagged as “electronics”). Corrected metadata is formatted for reintegration into the Product Information Management (PIM) system or directly interfaced with e-commerce APIs [26].

3.3. Theoretical Basis

Here, the architecture is based on the multi-view learning theory, with each modality being a distinct view of the same exemplar. According to theoretical and empirical research, multi-view methods, due to their alignment, can mitigate generalization error and become more robust [27]. The information bottleneck principle can also be used in multimodal situations where fused representation is maximized to remove all less informative signals about the target task [28].

Lastly, the model is concerned with the theory of error propagation in data-centric systems, which highlights the multiplicative effect of upstream error data on downstream analytics. The system suggests avoidance of the cascading failure in recommendation engines, search ranking, and inventory systems by automatically detecting metadata anomalies and correcting them pro favor [29].

4. Experimental Results, Graphs, and Tables

4.1. Evaluation Design and Benchmarks

The importance of using multimodal metadata correction systems in online retail is usually evaluated experimentally, with the most common metrics being accuracy/precision/recall, and F1 score, as well as metadata imputation quality.

Benchmarked model studies are based on huge datasets that comprise the content of texts, pictures of products, followed by attribute fields. The data is typically collected as a sample on the public e-commerce sites or proprietary catalogue stores on various retail categories, including but not limited to apparel, electronics, and personal care [30].

Popular baseline models have been unimodal (e.g., BERT over text only, CNN over image only), traditional rule-based engines, and a mixture of end-to-end multimodal neural architectures (vision-language transformers and co-attention-based networks).

4.2. Quantitative Comparison Across Models

The table summarizes the performance metrics of various metadata correction models evaluated on a dataset of 100,000 retail product listings with partially corrupted attribute fields. Metrics include Accuracy (%), F1 Score, and Mean Absolute Error (MAE) for numeric attributes such as price or dimensions [30-34].

Table 2 Performance Comparison of Metadata Correction Models

Model Type	Modality	Accuracy (%)	F1 Score	MAE (Numeric Fields)
Rule-based System	Text	70.3	0.66	3.84
CNN-based Visual Model	Image	74.5	0.71	3.27
BERT (Text Encoding)	Text	80.8	0.78	2.95
Co-attentive Early Fusion Network	Image + Text	84.9	0.82	2.43
Transformer with Cross Attention	Image + Text	88.2	0.86	1.91

4.3. Visual Graph: Accuracy Across Models

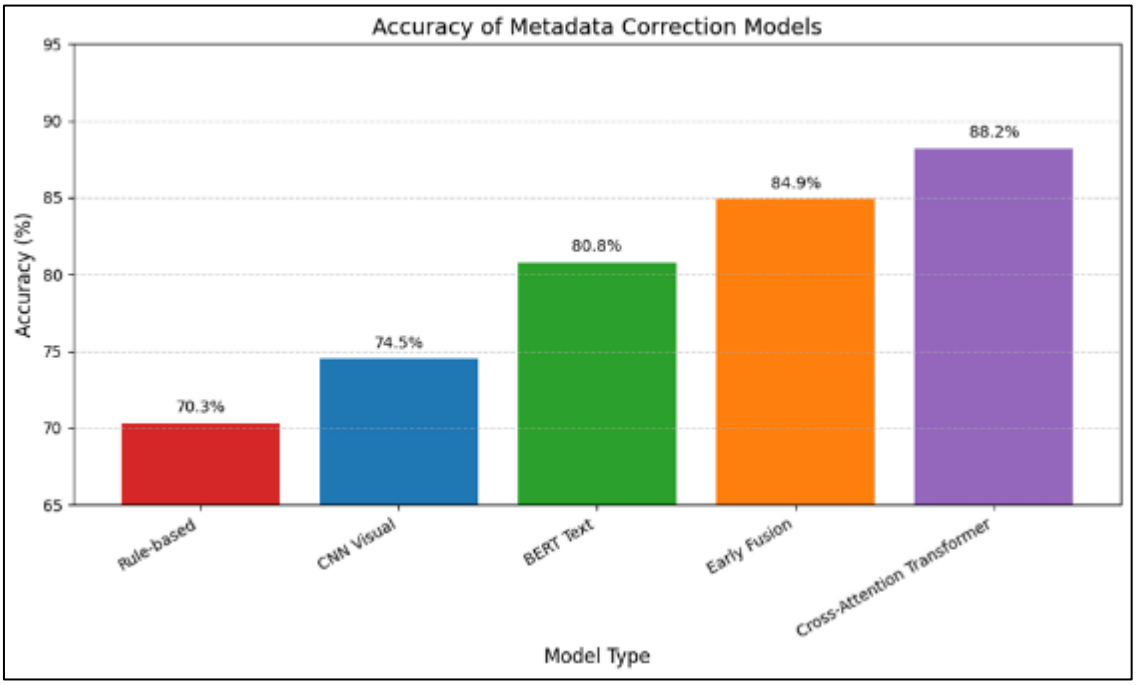

Figure 2 Metadata Correction Accuracy Comparison

This comparison illustrates that multimodal fusion models, particularly those employing cross-attention, outperform unimodal and rule-based approaches by significant margins. The improvements stem from richer context modeling across modalities.

4.4. Attribute-Wise Performance Analysis

The table reports attribute-specific F1 scores across four key fields: Category, Brand, Color, and Material. The results highlight that visual cues significantly enhance performance in color and material prediction, while textual features dominate in brand recovery.

Table 3 Attribute-wise F1 Score Performance (Multimodal vs. Unimodal)

Attribute	Text-Only (BERT)	Image-Only (CNN)	Multimodal (Fusion)
Category	0.79	0.74	0.87
Brand	0.82	0.69	0.86
Color	0.67	0.83	0.89
Material	0.70	0.80	0.88

4.5. Imputation of Missing Fields

The figure shows the imputation accuracy of three models for predicting missing product attributes (e.g., missing material or size information). A transformer-based multimodal model achieved the highest accuracy due to effective alignment between image cues and textual context [30-35].

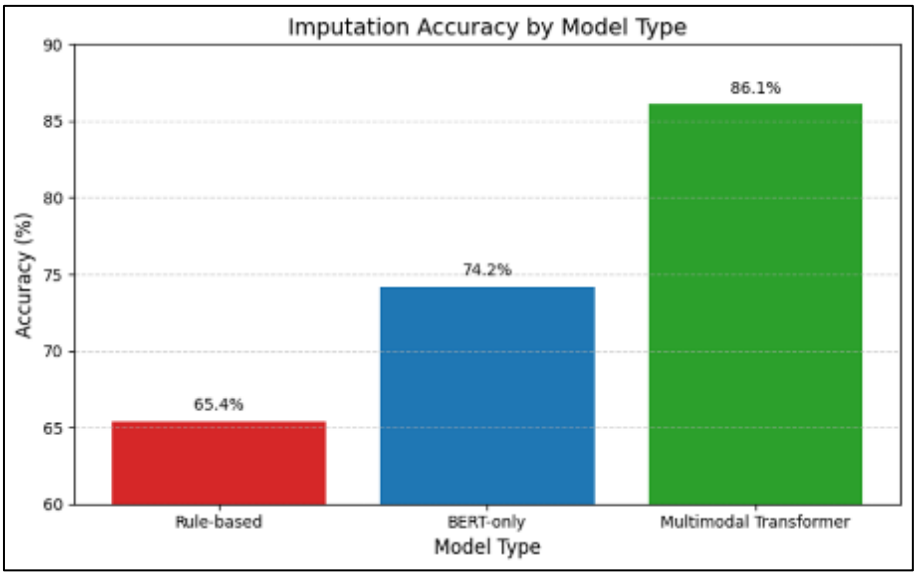

Figure 3 Imputation Accuracy (%) for Missing Fields

These results confirm that multimodal integration can substantially improve robustness in data recovery tasks, which is especially useful in noisy or incomplete retail environments.

4.6. Cross-Domain Generalization

Cross-domain experiments were conducted by training models on fashion datasets and testing on electronics. The performance degradation was lowest for models incorporating both modalities, as shown in the table.

Table 4 Generalization Performance Across Domains (F1 Score Drop%)

Model	F1 Drop (%)
Text-only (BERT)	14.8%
Image-only (CNN)	18.2%
Multimodal Transformer	8.6%

The results suggest that multimodal systems are more resilient to domain shifts and maintain performance across heterogeneous product types.

5. Future Directions

Increasingly complex online retail catalogues, due to globalization, third-party vendors, and high turnover of products, still present problems in terms of managing catalogue metadata. In the further investigation of multimodal machine learning (MML) in metadata correction, it is necessary to concentrate on several essential fields to enhance the system's performance, scalability, and reliability.

First, zero-shot and few-shot learning offer prospects of fast adaptation in those categories, which have minimal amounts of training data. Such methods enable models to extrapolate on novel metadata categories with pre-trained features and a few labeled data points and do not need substantial domain-specific annotation.

Second, there is an insufficiently studied aspect of multilingual and multicultural model training. The majority of existing models are taught mostly on English-based catalogues, which imposes a limitation on their application across global markets. The models must be expanded to models that can be adapted to various language and culture situations with no parallel corpus required.

Third, it is important to enhance explainability in multimodal systems. Seeing that regulatory and business issues are gradually requiring transparency in automated decision-making, MML systems would have to include an interpretable component that can trace the direction of the influence of image-text combinations on metadata correction results.

Fourth, more efficient training cycles might be achieved by means of active learning pipelines. Active learning can improve model robustness by designating and focusing on uncertain or ambiguous examples (for human analysis), which can lower the cost of annotation.

Lastly, the incorporation of structured knowledge graphs and ontologies into the multimodal pipelines with the help of which may lead to semantic consistency and constraints, allowing correction. The knowledge-enhanced models are able to match outputs based on known brand-category relationships or material compatibilities, thereby increasing compliance and reliability.

6. Conclusion

Multimodal machine learning has become a disruptive strategy in the field, followed by catalogue metadata correction of internet stores. They excel over conventional unimodal models in terms of accuracy, attribute completeness, and robustness when applied to product categories by using a combination of textual, structural, and imagery data. The empirical results secure the fact that cross-modal alignment and contextual modeling validation are very high in terms of error detection, correction, and imputation.

However, there are still issues of dataset quality, model generalizability, interpretability of the systems, and real-time performance. One way to tackle these gaps is by improving model architectures, multilingual support, integration of knowledge, and human-in-the-loop design. Other studies (in these directions) will be crucial in a scalable, reliable, international, and versatile digital commerce platform to ensure metadata quality assurances are addressed.

References

[1] Redman TC. Data driven: profiting from your most important business asset. Boston: Harvard Business Press; 2008.

[2] Petkov P. Assessing and analysing data quality in service oriented architectures; developing a data quality process [dissertation]. Dublin: Dublin City University; 2016.

[3] Cao Z, Mu S, Dong M. Two-attribute e-commerce image classification based on a convolutional neural network. Vis Comput. 2020;36(8):1619-34.

[4] Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2018;41(2):423-43.

[5] Jiang Y, Liao K, Lin S, Qiao H, Yu K, Yang C, Chen Y. Self-supervised Multimodal Representation Learning for Product Identification and Retrieval. In: International Conference on Neural Information Processing; 2023 Nov; Singapore. Singapore: Springer Nature Singapore; 2023. p. 579-94.

[6] McCrae S, Wang K, Zakhor A. Multi-modal semantic inconsistency detection in social media news posts. In: International Conference on Multimedia Modeling; 2022 Mar; Cham. Cham: Springer International Publishing; 2022. p. 331-43.

[7] Liu J, Cen X, Yi C, Wang FA, Ding J, Cheng J, et al. Challenges in AI-driven biomedical multimodal data fusion and analysis. Genomics Proteomics Bioinformatics. 2025;23(1):qzaf011.

[8] Niu K, Liu Y, Long Y, Huang Y, Wang L, Zhang Y. An Overview of Text-based Person Search: Recent Advances and Future Directions. IEEE Trans Circuits Syst Video Technol. 2024.

[9] Li J, Liu C, Wang J, Bing L, Li H, Liu X, et al. Cross-lingual low-resource set-to-description retrieval for global e-commerce. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020 Apr;34(05):8212-9.

[10] Goti A, Querejeta-Lomas L, Almeida A, de la Puerta JG, López-de-Ipiña D. Artificial intelligence in business-to-customer fashion retail: A literature review. Mathematics. 2023;11(13):2943.

[11] Sapkota R, Karkee M. Object detection with multimodal large vision-language models: An in-depth review. SSRN. 2025.

[12] Yu L, Chen J, Sinha A, Wang M, Chen Y, Berg TL, Zhang N. Commercemm: Large-scale commerce multimodal representation learning with omni retrieval. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2022 Aug. p. 4433-42.

[13] Yang R, Zhong Y, Su Y. Self-Supervised Joint Representation Learning for Urban Land-Use Classification With Multi-Source Geographic Data. IEEE Trans Geosci Remote Sens. 2025.

[14] Rawat M, Singal G. Surveying Technology Fusion in IoT Networks for IDS: Exploring Datasets, Tools, Challenges, and Research Prospects. ACM Trans Intell Syst Technol. 2025.

[15] Karunaratne T. Machine Learning and Big Data Approaches to Enhancing E-commerce Anomaly Detection and Proactive Defense Strategies in Cybersecurity. J Adv Cybersecurity Sci Threat Intell Countermeasures. 2023;7(12):1-16.

[16] Kumar A. Enhancing E-Commerce through Transformer-Based Large Language Models: Automating Multilingual Product Descriptions for Improved Customer Engagement. In: 2024 International Conference on Signal Processing and Advance Research in Computing (SPARC); 2024 Sep;1:1-7. IEEE.

[17] Zhu J, Zhou Y, Qian S, He Z, Zhao T, Shah N, Koutra D. Multimodal Attributed Graphs: Benchmarking and Rethinking.

[18] Dritsas E, Trigka M. Exploring the Intersection of Machine Learning and Big Data: A Survey. Mach Learn Knowl Extr. 2025;7(1):13.

[19] Weigl DM, Lewis D, Crawford T, Knopke I, Page KR. On providing semantic alignment and unified access to music library metadata. Int J Digit Libr. 2019;20:25-47.

[20] Liu C, He X, Yi L. Determinants of multimodal fake review generation in China's e-commerce platforms. Sci Rep. 2024;14(1):8524.

[21] Garg V, Jain A. Scalable Data Integration Techniques for Multi-Retailer E-Commerce Platforms. Int J Comput Sci Eng. 2024;13(2):525-70.

[22] Rao SX, Jiang J, Han Z, Yin H. Fraud Detection in E-Commerce: A Systematic Review of Transaction Risk Prevention. 2025.

[23] Bose P, Rana P, Ghosh P. Attention-based multimodal deep learning on vision-language data: models, datasets, tasks, evaluation metrics and applications. IEEE Access. 2023;11:80624-46.

[24] Ma J, Rong L, Zhang Y, Tiwari P. Moving from narrative to interactive multi-modal sentiment analysis: A survey. ACM Trans Asian Low-Resource Lang Inf Process. 2023.

[25] Najmi A. Imputation of missing product information using deep learning: A use case on the amazon product catalogue [dissertation]. München: Technische Universität München; 2019.

[26] Berti-Equille L. Measuring and modelling data quality for quality-awareness in data mining. In: Quality measures in data mining. Berlin, Heidelberg: Springer; 2007. p. 101-26.

[27] Sun S. A survey of multi-view machine learning. Neural Comput Appl. 2013;23(7–8):2031–8.

[28] Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle. Proc IEEE Inf Theory Workshop. 2015;1–5.

[29] Pipino LL, Lee YW, Wang RY. Data quality assessment. Commun ACM. 2002;45(4):211–8.

[30] Rahman MA, Rahman A. A SYSTEMATIC REVIEW OF INTELLIGENT SUPPORT SYSTEMS FOR STRATEGIC DECISION-MAKING USING HUMAN-AI INTERACTION IN ENTERPRISE PLATFORMS. SSRN. 2025.

[31] Sales LF, Pereira A, Vieira T, de Barros Costa E. Multimodal deep neural networks for attribute prediction and applications to e-commerce catalogs enhancement. Multimed Tools Appl. 2021;80(17):25851-73.

[32] Rode H, Hiemstra D. Conceptual language models for context-aware text retrieval.

[33] Liu Z, Ma Y, Schubert M, Ouyang Y, Xiong Z. Multi-modal contrastive pre-training for recommendation. In: Proceedings of the 2022 International Conference on Multimedia Retrieval; 2022 Jun. p. 99-108.

[34] Yang Y, Guo K, Fang Z, Zhang Y, Lin H, Grosser M, et al. Integrative Genetic Association Analysis and Transformer-Based Model for Ischemic Stroke Prediction. SSRN.

[35] Liu M, Li S, Yuan H, Ong MEH, Ning Y, Xie F, et al. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. Artif Intell Med. 2023;142:102587.