



Engineering visual–textual fusion for automatic item categorization in Indonesian e-commerce platforms

Yuliana Melita Pranoto^{*1, 2)}, Anik Nur Handayani²⁾, Heru Wahyu Herwanto²⁾ and Yosi Kristian¹⁾

¹⁾Institut Sains dan Teknologi Terpadu Surabaya, Surabaya 60284, Indonesia

²⁾Universitas Negeri Malang, Malang 65145, Indonesia

Received 7 June 2025
Revised 5 December 2025
Accepted 11 March 2026

Abstract

In the modern digital era, e-commerce remains a fundamental pillar of the global economy. The rapid growth of online ventures has fundamentally reshaped how products are bought and sold. However, as the number of products on e-commerce platforms in Indonesia increases, the main challenge is accurately matching products to consumer preferences. While previous studies have primarily used unimodal approaches to product matching, relying on a single type of information, little research has examined how to recognize new products on e-commerce platforms. Utilizing a multimodal approach that integrates information from diverse data types, such as text and images, is increasingly appealing for improving product-matching quality. This study's main contribution is to create vector representations of image and text features using deep learning techniques, including Convolutional Neural Networks and Doc2Vec. It also involves merging features through a cross-modality approach using FeedForward Neural Networks and determining the appropriate parameters for clustering new products into existing clusters using Hierarchical Clustering. The research demonstrates that employing a multimodal approach that leverages text and image information can enhance product matching quality on e-commerce platforms in Indonesia, achieving a Normalized Mutual Information of 0.96.

Keywords: E-commerce, Multimodal, Deep learning, Cross-modality, Hierarchical clustering

1. Introduction

In today's digital era, e-commerce has emerged as a vital pillar in the global economy. The rapid growth of online businesses has significantly transformed how consumers purchase and sell products [1, 2]. A crucial aspect of the online shopping experience is finding products that meet the consumer's needs and preferences [3, 4]. However, as the number of products on e-commerce platforms increases, the primary challenge is achieving accurate, relevant product matching.

Previous research has predominantly employed unimodal approaches to product matching, relying solely on a single type of information, such as text [5] or images [6]. Some drawbacks include lower accuracy due to insufficient information from a single data type and the inability to handle incomplete data, as products may lack necessary information in one data type [7].

An intriguing approach to improving product matching quality is to leverage multimodal information, such as text and images [8, 9]. This multimodal approach enables e-commerce platforms to leverage diverse data types for more accurate, efficient product matching [10]. By considering textual information from product descriptions and product image features, e-commerce platforms can better understand consumer preferences and needs [11].

This research also aims to identify suitable parameter settings that can serve as a reference for grouping new products into existing clusters or, if needed, creating new ones. The main contributions of this study are:

1. Creating image feature vectors using Deep Learning with a pre-trained Convolutional Neural Network (CNN).
2. Creating text feature vectors from product titles using Doc2Vec.
3. Combining image and text features using a FeedForward Neural Network (FNN) through a cross-modality approach.
4. Finding the best parameter settings from a small sample of product data to help classify new products accurately.
5. Applying Hierarchical Clustering to organize new products into appropriate clusters based on the selected parameters.

2. Previous research

Product matching is an essential topic in e-commerce that focuses on finding and linking similar items across different sources [12]. In earlier studies, researchers often used predefined rules and traditional statistical approaches. Even so, machine learning methods have become more widely adopted because they offer better accuracy and performance [13, 14].

*Corresponding author.

Email address: ymp@stts.edu

<https://doi.org/10.64960/easr.2026.262497>

2.1 Unimodal approaches

The issue of product matching on e-commerce platforms has attracted the attention of many researchers. Some previous studies utilized unsupervised algorithms to match products based on their titles. Recommendation systems for product images, similar to those shown to consumers on e-commerce websites, have been developed using unsupervised machine learning [15]. However, this research is limited to datasets of clothing and accessory products, so the results may not apply to other product types.

Moreover, new consumer product recommendation systems have mainly been built on text similarity with past purchases [16]. These systems use machine learning techniques, such as Word2Vec, to create vector representations of product descriptions and reviews. Another method called Doc2Vec, an extension of Word2Vec, is also used to analyze numerical vectors from text documents [17].

A study by [18] uses a neural network approach to solve product-matching problems involving product titles on e-commerce platforms. The outcomes are evaluated against classical models such as SVMs and decision tree-based methods, and the analysis demonstrates that neural networks perform more effectively in both accuracy and training speed. In line with this, product matching is enhanced by fine-tuning pre-trained neural network models. As explained in our previous research [19], using pre-trained networks to analyze image features can improve matching accuracy compared to older methods that do not adapt well to changes in product data. These networks are also more effective at addressing issues such as different product names, missing information, and inconsistent formats across e-commerce sites.

Recently, many studies have focused on using machine learning, deep learning, to match products to unstructured data such as free-text descriptions [20]. For example, [21] applied deep learning to identify similar product images on store shelves. In general, deep learning has been successfully applied across various areas to address complex product-matching problems [22].

2.2 Clustering approaches

Additionally, clustering methods have gained attention for grouping similar text data. One commonly used technique is hierarchical clustering, which provides a hierarchical data structure [23]. These clustering techniques have been applied to automatically match product titles without supervision by assessing title similarity within each cluster. The goal is to enhance the efficiency of product searches on e-commerce platforms. However, many studies assume that the desired number of clusters is known in advance, whereas in reality, the actual number of similar products may not be predetermined.

Several studies related to e-commerce product matching have primarily relied on product text. Our previous research attempted to combine Doc2Vec and Hierarchical Clustering to cluster e-commerce products based on product titles [24]. Unfortunately, using short product titles without considering visual details may result in some similar products going undetected. To address the limitations of text-only approaches, [25] proposed using Deep Hierarchical Clustering based solely on image features for e-commerce product grouping.

2.3 Multimodal approaches

In recent years, several studies have employed deep learning and transfer learning techniques to extract multimodal features for product matching [26]. However, most existing multimodal approaches perform categorization at a coarse level (e.g., clothing, electronics, books), where a single category may contain thousands to millions of products spanning diverse types and brands [27]. Other work [28] has explored image–text matching using artificial neural networks based on a Multi-Modality Cross Attention Network (MMCA), which measures semantic–visual similarity between an image and its associated text description. In this setting, image and text representations are learned jointly within a cross-modal attention framework.

Building on these studies, the present work addresses the limitations of prior approaches by focusing on finer-grained product matching and by adopting architectures that are both effective and computationally efficient. For the visual modality, we employ the EfficientNetV2S architecture to process product images; this CNN model is compact and offers fast training while maintaining strong performance [29]. For the textual modality, we use Doc2Vec to obtain dense embeddings that capture the semantic content of product titles, thereby supporting more accurate product matching. The two unimodal embeddings are then integrated through a Feedforward Neural Network (FNN) to implement the cross-modality fusion. FNN is a simple yet powerful neural network architecture widely used in machine learning [30]. In our setting, it provides a lightweight alternative to attention-based multimodal fusion while still capturing interactions between image and text features.

Modern vision–language models (VLMs) such as CLIP, BLIP-2, CoCa, and ALIGN have recently become the dominant paradigm in multimodal research. These models are typically trained as large-scale image–text foundation models on web data and contain hundreds of millions or more parameters, requiring high-end GPU infrastructure for pretraining or effective fine-tuning [31]. They are primarily optimized for general-domain tasks such as zero-shot recognition and cross-modal retrieval on English image–text pairs.

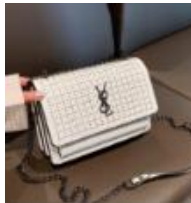
In contrast, the present study focuses on domain-specific product matching and hierarchical clustering on a relatively small, Indonesian e-commerce dataset (34,000 products) with short, noisy, and often code-mixed product titles. In this setting, fine-tuning very large VLMs would be computationally expensive and risk overfitting, while the language and domain mismatch further limit their expected benefit. For these reasons, we instead adopt more compact, modality-specific encoders (EfficientNetV2S for images and Doc2Vec for text), which can be trained directly on local marketplace data and then fused via a lightweight FNN-based cross-modality architecture.

3. Materials and methods

3.1 Data

This study used product data from the Kaggle Dataset [32], supplemented by data from several other e-commerce platforms, including Shopee, Tokopedia, and Blibli. These three platforms are widely used e-commerce platforms in Indonesia [33], with Tokopedia and Shopee prominent in Southeast Asia [34].

Table 1 Sample Collection of Products in the Dataset

Product		Group
	Baca Buku Ini Saat Engkau Patah Hati : Sesungguhnya Berpura-pura Tegar Itu Menyakitkan	A
	Baca Buku ini Saat Engkau Patah Hati - Anak Hebat	
	Baca Buku ini Saat Engkau Patah Hati	B
	BACA BUKU INI SAAT ENKKAU PATAH HATI: SESUNGGUHNYA BERPURA-PURA TEGAR ITU MENYAKITKAN	
	[Bayar Ditempat] - Jam Tangan MI 2 Band Watch/ Jam Tangan LED / Jam Tangan Unisex /Jam Sport Digital	B
	Jam Tangan Led Sport Digital Watch Wanita Pria Karet Anak Sekolah Murah Import Fashion Grosir/COD	
	Jam Tangan Led mi 2	C
	BAYAR DI TEMPAT - Jam Tangan LED Mi 2 Sport Digital	
	TAS IMPORT FASHION SELEMPANG 7075 93280 3254 1572 N2594 B1669 I07448 364 2594	C
	LT7215 READYCODE LTASRANSELIMPORT BATAM DIJAMIN BAGUS CR7012 B07448 OB1669 1572 LT1572 BQ2594 DJ3280	
	TAS IMPORT CANTIK E364/T07448/L1572/ C7012/B2594 TAS JAKARTA TAS CANTIK TAS MURAH TAS GROSIRCODE	C
	TAS SELEMPANG IMPORT MURAH JAKARTA TI7448/E364/L1572/ C7012/ B2594 MURAH CANTIK LAKU IMPORT KEKINIAN	

The Kaggle Dataset has become a valuable and diverse data source for researchers and practitioners across various fields. As the most significant global data competition and collaboration platform, Kaggle provides access to thousands of datasets spanning topics from social sciences to computer science, health, and finance. The diversity and quality of datasets offered by Kaggle enable users to explore real-world issues more effectively, test machine learning models, and develop innovative solutions [35].

The dataset used in this study comprises 34,450 products categorized into 8,758 groups, including household appliances, electronic devices, clothing, shoes, food, beauty products, baby supplies, cooking equipment, health equipment, and others. Examples of products are shown in Table 1. A subset of 11,700 products was used to train a deep learning model leveraging multimodal features, including product images and product titles. In the subsequent stage, two experiments were conducted with different data volumes for product matching: 3,450 products and 19,300 products. In these experiments, hierarchical clustering was used to group products by finding the best parameter settings from a small sample of product data.

3.2 Method

Figure 1 provides an overview of the proposed framework, consisting of (a) the model-building process and (b) the product-matching process. The model-building stage produces three trained models: image, text, and multimodal, which are subsequently used in the product-matching stage. The green blocks indicate components presented in our previous work [24], where a text-trained Doc2Vec model was combined with Hierarchical Clustering to group e-commerce products based solely on product titles. In addition, the yellow blocks correspond to components introduced in our earlier study [25], in which an EfficientNetV2-based image model trained on product images was integrated with Hierarchical Clustering to cluster e-commerce products according to their image collections.

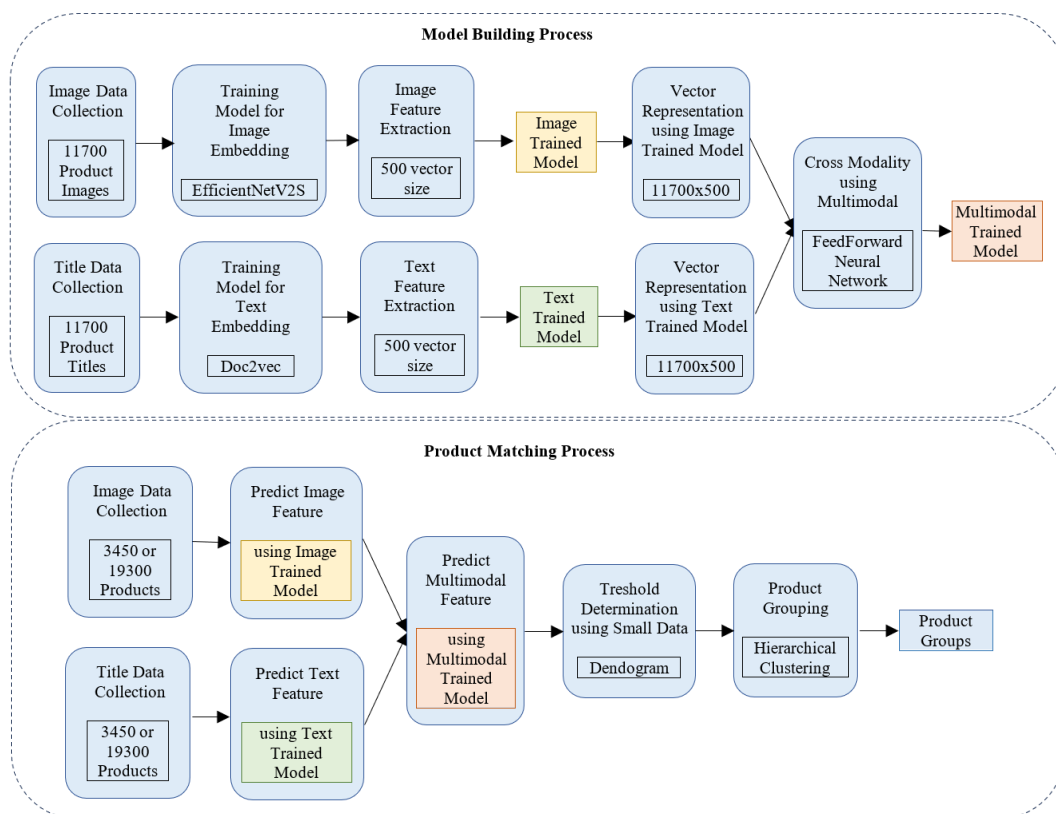


Figure 1 (a) Model Building Process; (b) Product Matching Process

In this study, the image model is constructed by training on 11,700 product images using transfer learning from a pre-trained EfficientNetV2S CNN, which is then further fine-tuned for the target domain. The text model is trained on 11,700 product titles using Doc2Vec. Both models produce 500-dimensional feature vectors. The resulting image and text representations are then combined through a cross-modality Feedforward Neural Network (FNN), which learns a shared latent space and yields the multimodal model.

The product-matching process begins by predicting product image features using the image model and text features using the text model. The predicted image and text feature matrices are then fed into the multimodal model to obtain multimodal feature vectors. Subsequently, a small subset of the evaluated products is selected to determine suitable clustering parameters (distance metric, linkage, and threshold) based on the dendrogram structure. These parameters are then used to cluster the complete set of evaluated products into semantically coherent groups using Hierarchical Clustering.

3.2.1 Modified Convolutional Neural Network

The Convolutional Neural Network (CNN) is a popular neural network architecture in deep learning, designed to handle image data [36] effectively. EfficientNetV2s, developed by Google, is a CNN model that builds on the previous EfficientNet model. This architecture is known for its efficiency, smaller model size, and faster training while delivering high accuracy. By combining the strengths of previous architectures with recent innovations in deep learning, EfficientNetV2s promises better performance in various image processing tasks [29]. We performed image embedding to compress visual information from images into lower-dimensional vectors (256x256) to expedite training while preserving essential information, ensuring the images remain well-recognizable. Here, we used a batch size of 32 and a categorical class mode.

EfficientNetV2s is a pre-trained network trained on ImageNet. Transfer learning from pre-trained CNN models to recognize various visual representations has enhanced system performance [37, 38]. In this study, we utilized transfer learning from the pre-trained network. We modified the layers at the end of the model (adding Dense, BatchNormalization, and Dropout layers), then fine-tuned the model. Subsequently, the entire modified network was retrained. The activation function used is softmax.

A summary of the modified network model is presented in Table 2. In the subsequent process, when representing vectors, the network used will be trimmed to the layer that exhibits image features (with a vector size of 500). Deep learning approaches in CNN architectures can be employed to learn visual features from images [39-41].

Table 2 Modifications to the EfficientNetV2s Model

Description	Modified EfficientV2s
Total Parameters	22,817,254
Added Layers	10
Vector size	500
Overfitting handler	Dropout, Batch Normalization, GlobalAveragePooling2D
Number of Epoch	105
Validation Accuracy	85%

3.2.2 Doc2Vec

Doc2Vec is a machine learning algorithm that transforms text into numerical vectors, making it useful for tasks such as categorization, grouping, and recommendation. Unlike Word2Vec, which processes individual words, Doc2Vec represents an entire document as a single vector. According to [17], Doc2Vec can produce better features than Word2Vec for automatic product classification.

One advantage of Doc2Vec is its ability to handle limited data and language variation in product descriptions. Research by [42] shows that Doc2Vec creates efficient and meaningful vector representations that help understand the content and context of documents. Studies [43, 44] suggest that its performance can be improved by using models trained on large datasets and pre-trained word embeddings. Study [45] found that Doc2Vec can generate product vectors that closely match existing clusters in recommendation systems. Equation 1 illustrates the mathematical approach used in Doc2Vec, which increases the likelihood of a target word (w) based on the context words (C) surrounding it in a document (D). In this formula, L stands for the likelihood that the model seeks to maximize.

$$L = \sum_{w \in D} \sum_{c \in C(w)} \log p(c|w, D) \quad (1)$$

Doc2Vec extracts features by learning from the words in a document, where each word contributes to the training process. In this study, we designed text features using 500-dimensional vectors to align with the image feature model. Doc2Vec has two primary model types: Distributed Memory (DM) and Distributed Bag-of-Words (DBOW). The DM model uses both document embeddings and word vectors to generate the next word, enabling it to understand word order and meaning. In contrast, the DBOW model relies solely on the document's numerical representation to suggest possible words, without considering word order [46, 47].

Building on this framework, our previous study [24] compared the two architectures to determine which variant is more suitable for the present task. The results showed that DBOW consistently outperformed DM. Consequently, this study adopts the DBOW variant, which uses a document vector to predict document words. The detailed configuration and adjustments applied during text model training are summarized in Table 3.

Table 3 Modifications to the Doc2Vec Model

Description	Modified Doc2Vec
Vector size	500
Alpha / Learning rate	0.025
Number of Epoch	100
Architecture	DBOW

In addition to this empirical evidence, the use of Doc2Vec, along with several methodological and practical considerations, supports the use of the textual embedding technique. The textual data in this study consists of Indonesian e-commerce product titles, which are generally short, keyword-centric, and often non-grammatical. For this type of text, large contextual models such as BERT or RoBERTa may not provide substantial additional benefits over simpler models [48], whereas Doc2Vec is effective for short, domain-specific documents [17,42-45]. Moreover, the proposed framework requires fixed-dimensional embeddings that can be fused directly with visual features via an FNN and subsequently used in Agglomerative Hierarchical Clustering. Doc2Vec naturally produces a single vector per product title, simplifying integration into the multimodal pipeline without additional pooling or complex fine-tuning.

From a computational standpoint, fine-tuning Transformer-based models such as BERT or RoBERTa would demand substantially greater resources [48]. At the same time, Doc2Vec can be trained unsupervised on 11,700 product titles at a considerably lower cost. Therefore, Doc2Vec is considered more appropriate for this study, as it yields compact, efficient semantic representations without requiring large-scale training or heavy computation, while remaining well-aligned with the overall multimodal fusion framework.

3.2.3 Cross-modality

Cross-modality refers to the interaction between two or more different types of information. In machine learning, this concept is applied when data from different modalities, such as text and images, are integrated into a single framework. Such integration enables the simultaneous use of information from multiple sources to obtain a more comprehensive understanding of a product. By leveraging cross-modality, deep learning models can learn complex relationships between visual features in images and semantic concepts in text [49].

In this study, we employ a Feedforward Neural Network (FNN) architecture to process data from different modalities. FNN is a relatively simple neural network architecture that is widely used in machine learning. According to [30], an FNN can extract and learn important features from two modalities and perform similarity matching with a target. The modifications introduced during the construction of the cross-modality model with the FNN architecture are summarized in Table 4. A dedicated FNN first processes each data modality, and the resulting feature vectors are then combined to obtain a joint representation of both modalities. This joint representation serves as the multimodal input to our model.

The selection of FNN as the architecture for cross-modality fusion is based on its suitability for the data characteristics and the objectives of this study. Both feature types, image features generated by EfficientNetV2S and text features obtained from Doc2Vec, have been normalized into fixed-dimensional vectors (500 dimensions), making FNN a stable and efficient architecture for learning the nonlinear relationships between the two modalities. As shown in Table 4, the modified FNN model achieves a validation accuracy of 88%, indicating its effectiveness in forming an informative joint representation. In line with [30, 49], FNN provides a competitive multimodal fusion baseline when each modality has been first reduced by its respective encoder.

Furthermore, compared with more complex multimodal models such as Transformer-based or cross-attention-based architectures [48], FNN is computationally lighter, more robust to the relatively high level of visual noise present in Indonesian marketplace data, and better suited for producing a practical model that can be deployed at an industrial scale. These considerations further justify adopting FNN as the core architecture for cross-modality fusion in this work.

Table 4 Modifications to the FNN Model

Description	Modified FNN
Total Parameters	5,911,394
Total Layers	13
Overfitting handler	Dropout, Batch Normalization
Number of Epoch	10
Validation Accuracy	88%

3.2.4 Hierarchical clustering

Hierarchical clustering is an algorithmic learning method used to group data samples with common characteristics. Such a technique produces a tree structure known as a dendrogram, in which each branch represents a grouping of data points. One key advantage of hierarchical clustering is its ability to visually show relationships among data points while allowing flexibility in determining the number of clusters [23]. However, its performance depends on the choice of distance metrics and linkage methods [50, 51].

Hierarchical clustering has two primary categories: agglomerative and divisive. In our research, we apply the agglomerative method, in which each data point starts as a separate cluster and is gradually combined into larger clusters [52, 53]. The results of this method can vary depending on the selected metric and linkage type. Several combinations have been tested, including Euclidean, Manhattan, and Cosine Distances, along with linkage methods such as Ward, Single, Complete, and Average Linkage [54].

Euclidean Distance calculates how far two clusters are from each other, while Cosine Similarity measures how similar two vectors are based on the angle between them. A smaller Euclidean distance or a higher Cosine value indicates greater similarity. Ward Linkage groups clusters by reducing variance within them, whereas Complete Linkage focuses on the maximum distance between data points in different clusters [54].

In this study, a small sample of 50 products was randomly selected from the 3,450-product test dataset to determine suitable clustering parameters. This subset was used to manually inspect the clustering patterns in the dendrogram without processing the entire dataset, thereby making parameter exploration substantially more efficient. Image and text features were extracted from these products using the models described in Sections 3.2.1 and 3.2.2, and then combined in the multimodal model (Section 3.2.3). The resulting multimodal vectors were used to generate a dendrogram using agglomerative clustering (see Figure 2). All combinations of distance metrics and linkage criteria were evaluated on this dendrogram, and the clustering threshold was selected at the cut point that produced the most stable partitions and best matched the reference label structure. The optimal threshold values for each parameter combination are reported in Table 5.

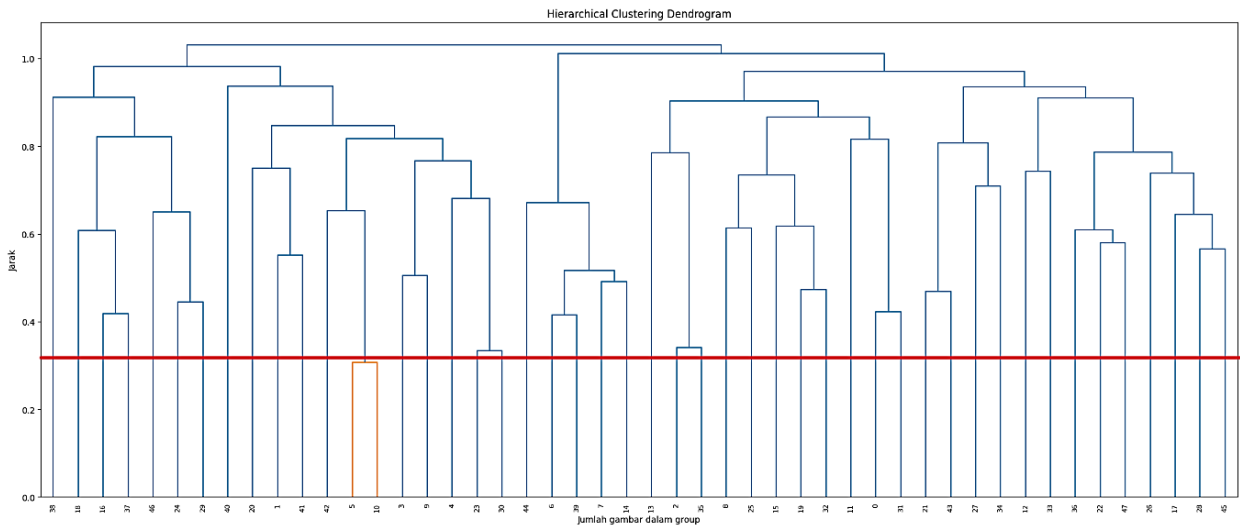


Figure 2 Dendrogram using DBOW Architecture with Cosine Similarity, Complete Linkage, and Threshold of 0.32

Table 5 Parameter Criteria in Agglomerative Clustering

Metric	Linkage	Best Threshold
Cosine Similarity	Single Linkage	0.32
	Complete Linkage	
	Average Linkage	
	Ward Linkage	
Euclidean Distance	Single Linkage	10
	Complete Linkage	
	Average Linkage	
	Ward Linkage	
Manhattan Distance	Single Linkage	280
	Complete Linkage	
	Average Linkage	

After determining the preferred threshold and linkage from the dendrogram analysis, this configuration was re-evaluated on the full dataset of 3,450 products and assessed using NMI. Validation with NMI confirmed that the selected parameters yielded the best performance when applied to the larger dataset. A sensitivity analysis was then conducted by observing how the clustering results changed as the threshold varied in the dendrogram and by comparing NMI values across different threshold settings (summarized in Table 6). The differences in performance across thresholds indicate that the model remains relatively stable in the neighbourhood of the optimal threshold region. To further verify the robustness of these findings, additional experiments were performed on an extended dataset comprising 19,300 products, as discussed in Section 4.

4. Results and discussion

Clustering performance is measured using the Normalized Mutual Information (NMI) metric. This metric helps determine how well the clusters reflect the data's actual structure. NMI is a widely used metric in clustering analysis to measure the similarity between two clustering results [55].

In this study, NMI values are calculated on a scale from 0 to 1 to evaluate the clustering performance under different parameter criteria. A greater NMI score reflects more accurate clustering, meaning the algorithm grouped the data more effectively. Therefore, NMI is used to measure how well the algorithm identifies patterns in the data, with higher values showing more accurate clusters [56, 57]. Compared to other methods, NMI performs better for comparing clustering results and labeling clusters.

$$NMI = \frac{2xMI(C,Y)}{H(C)+H(Y)} \quad (2)$$

Equation 2 computes the NMI value by determining the Mutual Information and Entropy values of the clustering C and Y. Here, C represents the predicted clusters. At the same time, Y denotes the ground truth clusters. When the Mutual Information is maximized, and the Entropy is minimized, the maximum value of 1 is achieved. Conversely, the minimum value of 0 is attained when the two clusterings are entirely dissimilar, resulting in the minimum Mutual Information and Maximum Entropy. Mutual information measures the amount of information shared between two variables, while Entropy quantifies the uncertainty in cluster labels [55].

Table 6 NMI Score for Each Parameter Criteria in Agglomerative Clustering

Metric	Linkage	NMI Score
Cosine Similarity	Single Linkage	0.808
	Complete Linkage	0.928
	Average Linkage	0.924
Euclidean Distance	Ward Linkage	0.909
	Single Linkage	0.902
	Complete Linkage	0.908
	Average Linkage	0.907
Manhattan Distance	Single Linkage	0.899
	Complete Linkage	0.909
	Average Linkage	0.907

We computed NMI scores for each combination of metric and linkage criteria on the 3450-product test dataset using the threshold from Table 5. The results are displayed in Table 6, where the combination yielding the highest NMI value is Cosine Similarity with Complete Linkage. Subsequent experiments were conducted using this optimal combination, and the Euclidean Distance with Ward Linkage combination was used as another comparison for the best metric. Euclidean and Ward were used as a comparison, referring to the best results from our previous research [24, 25].

Table 7 Test Result With Various Modalities

Tested Data	Modality	Metric	Linkage	NMI Score
3,450 product	Image	Euclidean Distance	Ward Linkage	0.90
		Cosine Similarity	Complete Linkage	0.88
	Text	Euclidean Distance	Ward Linkage	0.92
		Cosine Similarity	Complete Linkage	0.94
	Multimodal	Euclidean Distance	Ward Linkage	0.91
		Cosine Similarity	Complete Linkage	0.93
19,300 product	Image	Euclidean Distance	Ward Linkage	0.95
		Cosine Similarity	Complete Linkage	0.95
	Text	Euclidean Distance	Ward Linkage	0.96
		Cosine Similarity	Complete Linkage	0.96
	Multimodal	Euclidean Distance	Ward Linkage	0.96
		Cosine Similarity	Complete Linkage	0.96

To assess the robustness of the results and obtain a simple confidence interval approximation for NMI, the model was evaluated on two data subsets: 50 products from the 3,450-product test dataset and 200 products from the 19,300-product test dataset. Table 7 reports the experiments conducted with different modalities: image, text, and their combination (multimodal). The fact that the average NMI values exceed 0.90 (Table 7) indicates not only that the model performance is stable across variations in the sampled data, but also that the cluster assignments show strong agreement with the ground-truth product categories. This strong agreement provides empirical evidence that products grouped within the same cluster generally belong to semantically coherent categories. For each modality, results are reported using the best-performing combination of distance metric and linkage criterion, drawn from Cosine

similarity with Complete linkage and Euclidean distance with Ward linkage. On the 3,450-product test dataset, the text modality achieved the highest NMI of 0.94, while on the 19,300-product test dataset, both the text and multimodal configurations reached an NMI of 0.96.

To refine the statistical interpretation, the variation of NMI was examined across the three modalities (image, text, multimodal) on both test datasets (3,450 and 19,300 products). The NMI ranges for each modality, as reported in Table 7, are 0.88–0.95 for image, 0.92–0.96 for text, and 0.91–0.96 for multimodal. These ranges suggest that the clustering performance is consistent and does not depend on a single threshold setting. This analysis supports the claim that the model is robust to threshold variations and maintains stable performance across different evaluation scenarios.

To further quantify the model's stability and reliability, the mean and standard deviation of NMI were computed across the two test datasets (Table 8). The results show that the 19,300-product dataset is much more stable, with a minimal standard deviation of 0.0052, indicating a highly consistent cluster structure. By contrast, the 3,450-product dataset exhibits larger performance variability (standard deviation = 0.0216), suggesting that the resulting clusters are more sensitive to the choice of distance metric and linkage criterion.

Table 8 Mean and Standard Deviation of NMI

Tested Data	NMI Mean	Standard Deviation
3,450 product	0.9133	0.0216
19,300 product	0.9567	0.0052

Moreover, to evaluate the time efficiency of the clustering process, an additional experiment was conducted using the best parameter configuration (Cosine similarity and Complete linkage) on the 3,450-product dataset. The computation times were 4.15 seconds for the text-based representation, 4.12 seconds for the image-based representation, and 8.03 seconds for the multimodal representation. These computation times indicate that Agglomerative Clustering remains practically feasible for medium-sized datasets. Although the execution time for multimodal features is almost twice that of unimodal features, this is expected because the multimodal representation has twice the dimensionality of the unimodal vectors (1,000 vs. 500). Consequently, distance computations and hierarchical structure construction require additional processing time. Nevertheless, an execution time of 10 seconds or less for 3,450 products remains efficient in the context of offline batch processing commonly used in e-commerce backend systems.

To illustrate the product-matching results using multimodal criteria, we utilize Tables 1, 9, and 10. Table 1 presents a subset of the data used for testing, with the outcomes illustrated in Tables 9 and 10. Table 9 illustrates the products successfully grouped into the appropriate clusters, whereas Table 10 shows the products that failed to be grouped into the appropriate clusters and formed new clusters.

Table 9 Example of Successful Product Grouping

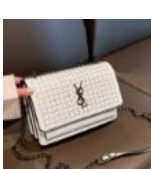
Product Images and Title		Group
	Baca Buku Ini Saat Engkau Patah Hati : Sesungguhnya Berpura-pura Tegar Itu Menyakitkan	A
	BACA BUKU INI SAAT ENGKAU PATAH HATI: SESUNGGUHNYA BERPURA-PURA TEGAR ITU MENYAKITKAN	
	[Bayar Ditempat] - Jam Tangan MI 2 Band Watch/ Jam Tangan LED / Jam Tangan Unisex /Jam Sport Digital	B
	BAYAR DI TEMPAT - Jam Tangan LED MI 2 Sport Digital	
	Jam Tangan Led mi 2	C
	TAS IMPORT FASHION SELEMPANG 7075 93280 3254 1572 N2594 B1669 I07448 364 2594	
	TAS SELEMPANG IMPORT MURAH JAKARTA TI7448/E364/ L1572/C7012/B2594 MURAH CANTIK LAKU IMPORT KEKINIAN	
	TAS IMPORT CANTIK E364/T07448/L1572/ C7012/B2594 TAS JAKARTA TAS CANTIK TAS MURAH TAS GROSIRCODE	

Table 10 Example of Products Failed to Cluster into a Single Group

Product Images	Product Title	Expected Group	Predicted Group
	Baca Buku Ini Saat Engkau Patah Hati	A	D
	Baca Buku ini Saat Engkau Patah Hati - Anak Hebat	A	E
	Jam Tangan Led Sport Digital Watch Wanita Pria Karet Anak Sekolah Murah Import Fashion Grosir/COD	B	F
	LT7215 READYCODE LT TASRANSELIMPORTBATAMDIJAMINBAGUS CR7012 B07448 OB1669 1572 LT1572 BQ2594 DJ3280	C	G

The results in Tables 9 and 10 underscore the importance of utilizing both modalities (text and images) in product matching. When the two modalities exhibit both similarities and differences, they are considered to represent distinct product groups. This observation supports the assertion by [49] that cross-modality enables deep learning models to comprehend the relationship between visual features in images and semantic concepts in text.

Moreover, Table 11 shows that the model achieves high accuracy in product matching, even when there are no perceptible differences in images or text. However, the model identifies these products as distinct. In the presence of subtle differences, the model creates a new cluster. Overall, these results indicate that the model can correctly assign new products to appropriate clusters or create new groups when necessary.

Table 11 Example of Product Included in the New Cluster

Product Images and Title		Expected Group	Predicted Group
	SOFTLENS KOKO JET BLACK MINUS - 0.50 s/d -5.00 / KOKO BLACK	M	M1
	SOFTLENS KOKO BLACK NORMAL / KOKO BLACK		M2
	PAKET CREAM RD JUMBO 30.GR KUALITAS SUPER ORI	N	N1
	CREAM RACIKAN RD 30GR 4IN1 KUALITAS ORI		N2

5. Conclusions

After conducting several experiments and analyses, this study arrives at several key conclusions. First, employing cross-modality with a simple FNN architecture improves the performance of the multimodal experiments, increasing NMI values by up to strong and thereby indicating its effectiveness in forming an informative joint representation. This architecture enables the model to capture the interaction between visual attributes in images and semantic meanings in text. In addition, compared with more complex multimodal architectures such as Transformers or cross-attention-based models [31], FNN is computationally lighter, more robust to the relatively high levels of visual noise present in Indonesian marketplace data, and better suited to developing practical models that can be deployed at an industrial scale.

Second, the experimental results on 3,450 products show that the image modality processed with the EfficientNetV2S architecture produces reasonably practical visual features but remains vulnerable to background clutter and variations in image quality. In contrast, the text modality encoded with Doc2Vec achieves higher NMI scores, indicating that product titles in this dataset are generally more stable and semantically informative. When both modalities are combined via cross-modality fusion with an FNN architecture, clustering performance improves consistently across both the 3,450- and 19,300-product datasets. A simple ablation analysis with three configurations: image only, text only, and image-text fusion, confirms that removing any single modality leads to a substantial decrease in NMI compared with the multimodal configuration; in particular, omitting the text modality reduces performance by up to 5%. As shown in Figure 3, multimodal fusion yields the most stable performance, especially on the larger dataset, by leveraging the complementary strengths of both modalities and mitigating their individual limitations.

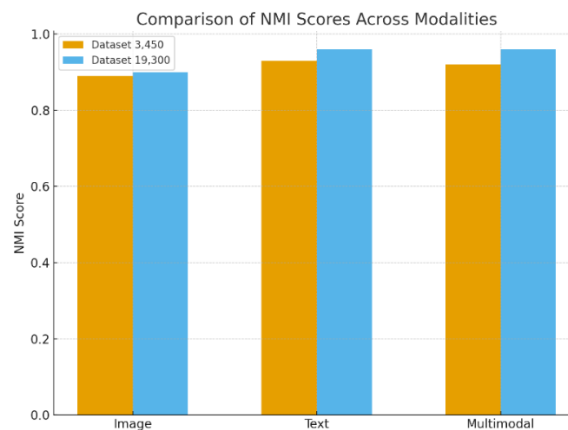


Figure 3 Comparison of NMI Scores across Modalities

This study also highlights the practical benefits of using a small subset of randomly selected test data to guide parameter selection for clustering larger datasets. Approximately 1% of the test data was randomly sampled to examine the resulting dendrogram. The experimental results indicate that the system can automatically assign new products to predefined groups or create separate clusters when they do not align well with existing groups, thus confirming the robustness and adaptability of the proposed clustering framework.

This research has significant implications for Indonesia's e-commerce sector by improving the accuracy of product matching on online platforms. It addresses the weaknesses of unimodal methods by combining text and image information within a multimodal framework and provides a comprehensive pipeline for more accurate and efficient product identification. The use of deep learning techniques—such as Convolutional Neural Networks and Doc2Vec—coupled with a cross-modality approach has proven effective in enhancing the quality of product matches. The impact of this research extends to the overall user experience on e-commerce platforms: accurate product matching enables consumers to find products that meet their preferences more quickly, leading to higher customer satisfaction and increased sales. The enhanced clustering and identification methods can also help e-commerce platforms better manage their product catalogues and personalize recommendations, fostering a more competitive and dynamic online marketplace.

Despite these contributions, this study is still subject to several limitations. Many aspects require improvement in future work, including increasing the quantity and diversity of product datasets and enhancing product image quality, particularly by reducing distracting background elements, to enable more informative feature extraction. It is also promising to explore Transformer-based architectures in subsequent studies, where the lightweight model proposed here can serve as an efficient baseline that may be extended with Transformers to handle more complex product-matching scenarios or to better capture contextual information from long textual descriptions and user reviews.

The background noise issue discussed in this paper is based on the authors' qualitative analysis rather than a dedicated controlled experiment. This observation arises from the fact that many product images contain diverse backgrounds, shop watermarks, and non-product objects, which, according to the literature, can affect the quality of visual representations. Previous studies have reported that non-uniform backgrounds can interfere with feature extraction in CNNs and reduce visual-matching accuracy, especially for products with similar shapes [21, 28]. Our experimental findings [19] are consistent with this, as lighter architectures, such as MobileNetV2, were more sensitive to background clutter and yielded lower accuracy and NMI scores than deeper models, such as EfficientNetV2. These considerations suggest that future work should investigate background cleaning, automatic object segmentation, and saliency-guided feature extraction to further improve visual representations and mitigate bias caused by noisy product images.

6. References

- [1] Fakieh B, Happonen A. Exploring the social trend indications of utilizing e-commerce during and after COVID-19's hit. *Behav Sci (Basel)*. 2022;13(1):5.

- [2] Afonso AP, Carneiro J, Azevedo AI. The impact of COVID-19 on e-commerce: a systematic review of the literature on the purchasing behavior of online retail consumers. *J Mark Res Case Stud.* 2024;2024:1-9.
- [3] Cui Z. Clustering-Based analysis of e-commerce customers' consumption behavior in the post-epidemic period. In: Gaikar V, Kandel BK, Mallick H, editors. 2023 4th International Conference E-Commerce Internet Technol (ECIT 2023); 2023 Mar 17-19; Nanchang, China. Zhengzhou: Atlantis Press; 2023. p. 30-5.
- [4] Chen H. Clustering analysis of online shopping behavior on the tiktok platform: revealing different user characteristics. *Adv Econ Manag Polit Sci.* 2024;91:46-56.
- [5] Cherednichenko O, Ivashchenko O, Cibák L, Lincenyi M. Item matching model in e-commerce: how users benefit. *Econ Cult.* 2023;20(1):77-90.
- [6] Alamdari PM, Navimipour NJ, Hosseinzadeh M, Safaei AA, Darwesh A. An image-based product recommendation for e-commerce applications using convolutional neural networks. *Acta Inform Pragensia.* 2022;11(1):15-35.
- [7] Feng C, Chen W, Chen C, Xu T, Chen E. Multimodal representation learning-based product matching. In: Zhang N, Wang M, Wu T, Hu W, Deng S, editors. CCKS 2022 - Evaluation Track. CCKS 2022. Communications in Computer and Information Science. Singapore: Springer; 2022. p. 180-90.
- [8] Ko E. Product matching through multimodal image and text combined similarity matching [thesis]. Stockholm: School of Electrical Engineering and Computer Science, 2021.
- [9] Gupte K, Pang L, Vuyyuri H, Pasumarty S. Multimodal product matching and category mapping: text+ image based deep neural network. 2021 IEEE International Conference on Big Data (Big Data); 2021 Dec 15-18; Orlando, USA. USA: IEEE; 2021. p. 4500-5.
- [10] Peeters R, Bizer C. Supervised contrastive learning for product matching. WWW '22: Companion Proceedings of the Web Conference 2022; 2022 Apr 25-29; Lyon, France. New York: ACM; 2022. p. 248-51.
- [11] Wilke M, Rahm E. Towards multi-modal entity resolution for product matching. 32nd GI-Workshop on Foundations of Databases; 2021 Sep 1-3; Munich, Germany. p. 1-5.
- [12] Sortur M, Rajpoot P, Manjunath, Subhanandh, Rao HC. E-commerce product matching at internet scale. ICEME '22: Proceedings of the 2022 13th International Conference on E-business, Management and Economics; 2022 Jul 16-18; Beijing, China. New York: ACM; 2022. p. 45-51.
- [13] Kinasih ANS, Handayani AN, Ardiansah JT, Damanhuri NS. Comparative analysis of decision tree and random forest classifiers for structured data classification in machine learning. *Sci Inf Technol Lett.* 2025;5(2):13-24.
- [14] Sutjiadi R, Sendari S, Wahyu HH, Kristian Y. Deep learning for segmentation and classification in mammograms for breast cancer detection: a systematic literature review. *Adv Ultrasound Diagn Ther.* 2024;8(3):94-105.
- [15] Addagarla SK, Amalanathan A. Probabilistic unsupervised machine learning approach for a similar image recommender system for e-commerce. *Symmetry.* 2020;12(11):1783.
- [16] Sharma AK, Bajpai B, Adhvaryu R, Pankajkumar SD, Gordhanbhai PP, Kumar A. An efficient approach of product recommendation system using NLP technique. *Mater Today: Proc.* 2023;80:3730-43.
- [17] Abubakar HD, Umar M, Bakale MA. Sentiment classification: review of text vectorization methods: bag of words, Tf-Idf, Word2vec and Doc2vec. *SLU J Sci Technol.* 2022;4:27-33.
- [18] Bai X, Duan L, Tang R, Batra G, Agrawal R. Improving text-based similar product recommendation for dynamic product advertising at yahoo. CIKM '22: Proceedings of the 31st ACM International Conference on Information & Knowledge Management; 2022 Oct 17-21; Atlanta, USA. New York: ACM; 2022. p. 2883-92.
- [19] Pranoto YM, Handayani AN, Kristian Y. Marketplace product image grouping using transfer learning of deep convolutional neural network in COVID-19 post-pandemic situation. In: Wibawa AP, editor. *The Spirit of Recovery.* Boca Raton: CRC Press; 2023. p. 55-63.
- [20] Alabdullatif A, Aloud M. AraProdMatch: A machine learning approach for product matching in e-commerce. *Int J Comput Sci Netw Secur.* 2021;21(4):214-22.
- [21] Tonioni A, Serra E, Di Stefano L. A deep learning pipeline for product recognition on store shelves. 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS); 2018 Dec 12-14; Sophia Antipolis, France. USA: IEEE; 2018. p. 25-31.
- [22] Li Y, Li J, Suhara Y, Wang J, Hirota W, Tan WC. Deep entity matching: challenges and opportunities. *J Data Inf Qual.* 2021;13(1):1-17.
- [23] Habib A, Akram M, Kahraman C. Minimum spanning tree hierarchical clustering algorithm: a new pythagorean fuzzy similarity measure for the analysis of functional brain networks. *Expert Syst Appl.* 2022;201:117016.
- [24] Pranoto YM, Handayani AN, Herwanto HW, Kristian Y. Optimizing product matching in e-commerce with DOC2VEC: leveraging hierarchical clustering parameters based on product titles. *ECTI Trans Comput Inf Technol.* 2024;18(3):396-405.
- [25] Pranoto YM, Handayani AN, Herwanto HW, Kristian Y. Optimized image-based grouping of e-commerce products using deep hierarchical clustering. *Int J Adv Intell Inform.* 2025;11(3):336-54.
- [26] Zuo Z, Wang L, Momma M, Wang W, Ni Y, Lin J, et al. A flexible large-scale similar product identification system in e-commerce. KDD'20 IRS Workshop; 2020 Aug 23-27; San Diego, USA. New York: ACM; 2020. p. 1-9.
- [27] Das N, Joshi A, Yenigalla P, Agrwal G. MAPS: multimodal attention for product similarity. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022 Jan 3-8; Waikoloa, USA. USA: IEEE; 2022. p. 2988-96.
- [28] Wei X, Zhang T, Li Y, Zhang Y, Wu F. Multi-modality cross attention network for image and sentence matching. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13-19; Seattle, USA. USA: IEEE; 2020. p. 10938-47.
- [29] Grd P, Tomicic I, Barcic E. Transfer learning with EfficientNetV2S for automatic face shape classification. *J Univers Comput Sci.* 2024;30(2):153-78.
- [30] Huang N, Liu J, Miao Y, Zhang Q, Han J. Deep learning for visible-infrared cross-modality person re-identification: a comprehensive review. *Inf Fusion.* 2023;91:396-411.
- [31] Li J, Li D, Savarese S, Hoi S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. ICLR'23: Proceedings of the 40th International Conference on Machine Learning; 2023 Jul 23-29; Hawaii, USA. New York: ACM; 2023. p. 19730-42.

- [32] Suresh. Shopee train images with labels dataset [Internet]. 2021 [cited 2022 Jun 4]. Available from: <https://www.kaggle.com/datasets/dharmiksv/shopee-train-images-with-labels>.
- [33] Farliana N, Rahmaningtyas W, Widhiastuti R. Development of e-commerce management and policy in Indonesia. *Am J Humanit Soc Sci Res*. 2022;6(1):155-60.
- [34] Bamansoor S, Pande B, Al Moaiad Y, Pathmanathan PR, El-Ebiary YAB, Latiff NAA, et al. Efficient online shopping platforms in Southeast Asia. 2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE); 2021 Jun 15-17; Cameron Highlands, Malaysia. USA: IEEE; 2021. p. 164-8.
- [35] Banachewicz K, Massaron L. *The Kaggle book : data analysis and machine learning for competitive data science*. Birmingham: Packt Publishing Ltd; 2022.
- [36] Wibawa AP, Handayani AN, Rukantala MRM, Ferdyan M, Budi LAP, Utama ABP, et al. Decoding and preserving Indonesia's iconic Keris via A CNN-based classification. *Telemat Inform Rep*. 2024;13:100120.
- [37] Kumar JS, Anuar S, Hassan NH. Transfer learning based performance comparison of the pre-trained deep neural networks. *Int J Adv Comput Sci Appl*. 2022;13(1):797-805.
- [38] Wibawa AP, Yudha Pratama WA, Handayani AN, Ghosh A. Convolutional Neural Network (CNN) to determine the character of wayang kulit. *Int J Vis Perform Arts*. 2021;3(1):1-8.
- [39] Karim MR, Beyan O, Zappa A, Costa IG, Rebholz-Schuhmann D, Cochez M, et al. Deep learning-based clustering approaches for bioinformatics. *Brief Bioinform*. 2021;22(1):393-415.
- [40] Ren Y, Pu J, Yang Z, Xu J, Li G, Pu X, et al. Deep clustering: a comprehensive survey. *IEEE Trans Neural Netw Learn Syst*. 2025;36(4):5858-78.
- [41] Dwiyanto FA, Dreżewski R, Wibawa AP. Dealing with COVID-19 using deep learning for computer vision. In: Wibawa AP, editor. *The Spirit of Recovery*. Boca Raton: CRC Press; 2024. p. 18-39.
- [42] Yoon B, Kim S, Kim S, Seol H. Doc2vec-based link prediction approach using SAO structures: application to patent network. *Scientometrics*. 2022;127:5385-414.
- [43] Dogru HB, Tilki S, Jamil A, Hameed AA. Deep learning-based classification of news texts using doc2vec model. 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA); 2021 Apr 6-7; Riyadh, Saudi Arabia. USA: IEEE; 2021. p. 91-6.
- [44] Supriyono, Wibawa AP, Suyono, Kurniawan F. Multimodal deep learning for youtube stand-up comedy transcription in Indonesian language. 2025 17th International Conference on Knowledge and Smart Technology (KST); 2025 Feb 26-Mar 1; Bangkok, Thailand. USA: IEEE; 2025. p. 358-63.
- [45] Kapoor P. Classification & clustering of text based on Doc2Vec & K-means clustering based similarity measurements. In: Mishra PK, Yadav SP, editors. *Demystifying Emerging Trends in Machine Learning*. Singapore: Bentham Science Publishers; 2025. p. 249-60.
- [46] Fujita Y, Ueda K. A method for selecting training data using Doc2Vec for automatic test cases generation. 2024 IEEE International Conference on Consumer Electronics (ICCE); 2024 Jan 6-8; Las Vegas, USA. USA: IEEE; 2024. p. 1-6.
- [47] Cheng J. Long text topic mining and clustering analysis base on Doc2Vec-LDA and K-Means. 2024 International Conference on Machine Learning and Cybernetics (ICMLC); 2024 Sep 20-23; Miyazaki, Japan. USA: IEEE; 2024. p. 9-13.
- [48] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach [Internet]. arXiv [preprint]. 2019 [cited 2025 May 9]. Available from: <https://arxiv.org/abs/1907.11692>.
- [49] Liu Y, Guo Y, Fang J, FAN J, Hao Y, Liu J. Survey of research on deep learning image-text cross-modal retrieval. *J Front Comput Sci Technol*. 2022;16(3):489-511. (In Chinese)
- [50] Li T, Rezaeipannah A, Tag El Din ESM. An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. *J King Saud Univ - Comput Inf Sci*. 2022;34(6):3828-42.
- [51] Ran X, Xi Y, Lu Y, Wang X, Lu Z. Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artif Intell Rev*. 2023;56:8219-64.
- [52] Shetty P, Singh S. Hierarchical clustering: a survey. *Int J Appl Res*. 2021;7(4):178-81.
- [53] Oti EU, Olusola MO. Overview of agglomerative hierarchical clustering methods. *Br J Comput Netw Inf Technol*. 2024;7(2):14-23.
- [54] Vijaya, Sharma S, Batra N. Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon); 2019 Feb 14-16; Faridabad, India. USA: IEEE; 2019. p. 568-73.
- [55] Zhang D, Nan F, Wei X, Li SW, Zhu H, McKeown K, et al. Supporting clustering with contrastive learning. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2021 Jun 6-11; Mexico. Stroudsburg: ACL; 2021. p. 5419-30.
- [56] Abbasi SO, Nejatian S, Parvin H, Rezaie V, Bagherifard K. Clustering ensemble selection considering quality and diversity. *Artif Intell Rev*. 2019;52:1311-40.
- [57] Yang X, Yan J, Cheng Y, Zhang Y. Learning deep generative clustering via mutual information maximization. *IEEE Trans Neural Netw Learn Syst*. 2023;34(9):6263-75.