JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Page: 845-854
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.517

Unsupervised Clustering of Handwritten Essay Answer Images Using Vision Transformer

Mohamad Asyqari Anugrah[1], Yaya Wihardi[2,*], Rani Megasari[3]

[1,2,3] Faculty of Mathematics and Science Education, Computer Science, Universitas Pendidikan Indonesia, Bandung, Indonesia
Email: [1]asyqari@upi.edu, [2]yaya@upi.edu, [3]megasari@upi.edu
(*Email Corresponding Author: yaya@upi.edu)
Received: July 30, 2025 | Revision: July 31, 2025 | Accepted: July 31, 2025

Abstract

This study explores the use of deep clustering methods to automatically group handwritten essay answer sheets based on their visual patterns. Feature extraction was performed using three backbone models: ResNet-50, Vision Transformer (ViT-base), and Tr-OCR. These features were then clustered using two unsupervised algorithms—K-means (with k=5) and HDBSCAN (with minimum cluster size = 10). To enhance clustering performance, a deep clustering approach was implemented by applying K-means iteratively to refine feature representations. Evaluation was conducted both quantitatively, using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score, and qualitatively, through t-SNE visualizations and cluster content inspection. The ViT and Tr-OCR backbones outperformed CNN-based ResNet-50, achieving higher cluster cohesion and separation. Notably, the final clustering result using ViT with HDBSCAN reached a Silhouette Score of 0.772, Davies-Bouldin Index of 0.369, and Calinski-Harabasz Score of 408.006. The findings indicate that vision transformer-based models are more effective for unsupervised grouping of handwritten visual data. This approach can assist educators in accelerating and objectifying the grading process and may serve as a foundation for future automated essay evaluation systems integrating OCR and NLP techniques.

Keywords : Image Clustering, DeepCluster, Vision Transformer, Convolutional Neural Network, Handwritten-Essay

1. INTRODUCTION

In the field of education, teachers play a vital role in nurturing student's development across cognitive, emotional, and practical domains [1]. To asses the extent of this development, it is essential to conduct learning evaluations. Learning evaluation refers to a series of activities aimed at collecting, analyzing, and interpreting data regarding both the process and outcomes of student learning [2]. A robust evaluation system provides insights into instructional quality and guides educators in designing more effective teaching strategies [3]. Among the various methods of student assestment, essay tests are considered one of the most effective. According to study [4], essay-based assestments encourage students to respond using their own reasoning, enabling deeper expression of knowledge. While essays are relatively easy for teachers to prepare, they present significant challenges during grading. Evaluating essay answers manually is time-consuming and labor-intensive, requiring teachers to interpret and assess each response based on the content and quality of arguments [5]. This makes essay evaluation less efficient compared to multiple-choice test, which are easier to score. Therfore, a more efficient solution is needed to support teachers in assessing essay responses. One promising evenue is the use of Artificial Intelligence (AI) based technologies.

Recent advancements in AI have opened new possibilities in processing both image and text data. AI refers to systems that exhibit intelligent behaviour by analyzing their environment and taking actions-often autonomously to achieve specific goals [6]. Deep Learning (DL) has emerged as a powerful technique capable of learning complex data representastions using layered neural architectures [7]. Deep Learning, particularly in the field of Computer Vision (CV), allows machine to interpret and analyze visual information in a way that simulates human visual perception [8]. One important task in computer vision beside classification task is image clustering, which groups similar data points based on feature similarity.

Image Clustering has been explored in several studies, highlighting the inherent difficulty of clustering high-dimensional visual data. A common approach involves extracting features using convolutional neural network (CNN) as the backbone model, followed by clustering using algorithms such as K-Means [9]. However, these approaches often yield suboptimal clustering perfomance due to the challenges posed by high-dimensional feature spaces, where distance-based similarity metrics become less effective [10]. Traditional algorithm like K-means also require the predefinition of the number of clusters (K), which often impractical in real-world, unlabeled datasets and may lead to poorly separated clusters. To address these limitations, density-based clustering methods such as DBSCAN [11] have been proposed. DBSCAN automatically determines the number of clusters based on the density of data points, thus eliminating the need to pre-specify K. However, its perfomance is highly sensitive to parameter selection. HDBSCAN [12], an extension of DBSCAN, was introduced to simplify this process by reducing the number of required hyperparameters primarly focusing on the *minimum_cluster_size* while providing more robust clustering results in high-dimensional settings. In previous research, CNN-based feature extraction combined with standard clustering techniques showed limited effectiveness [13] while ViT-based approaches yielded more promising results [13].

Nevertheless, improvements were reported after incorporating a self-supervised learning framework such as DeepCluster. Caron et al [14], could improve clustering perfomance by iteratively assigning pseudo-labels through clustering and using them to train the CNN bakbone. While this approach improved clustering quality, its effectiveness

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Page: 845-854
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.517

remains contrained by the representaional capacity of CNNs [14]. Building on this line of work, a more recent method Vision Transformer for Contrastive Learning (VTCC) [15] was proposed by You et al., integrating the representational strength of ViT with contrastive learning for self-supervised image clustering. Experimental results demonstrated that VTCC outperforms state-of-the-art methods in terms of clustering accuracy and consistency.

Inspired by these findings, the present study aims to extend the DeepCluster framework by replacing the CNN backbone with a ViT. Additionally, we propose the use of HDBSCAN as the clustering algoritm, leveraging its ability to discover based on density cluster without requiring the number of clusters to be predefined. This experimental setup is expected to reveal the perfomance advantages of ViT-based feature extraction and density-aware clustering for the task of grouping handwritten essay answer images. By grouping these images, it is hoped that the essay assessment process will be faster, more objective, and more consistent than manual methods.

2. RESEARCH METHODOLOGY

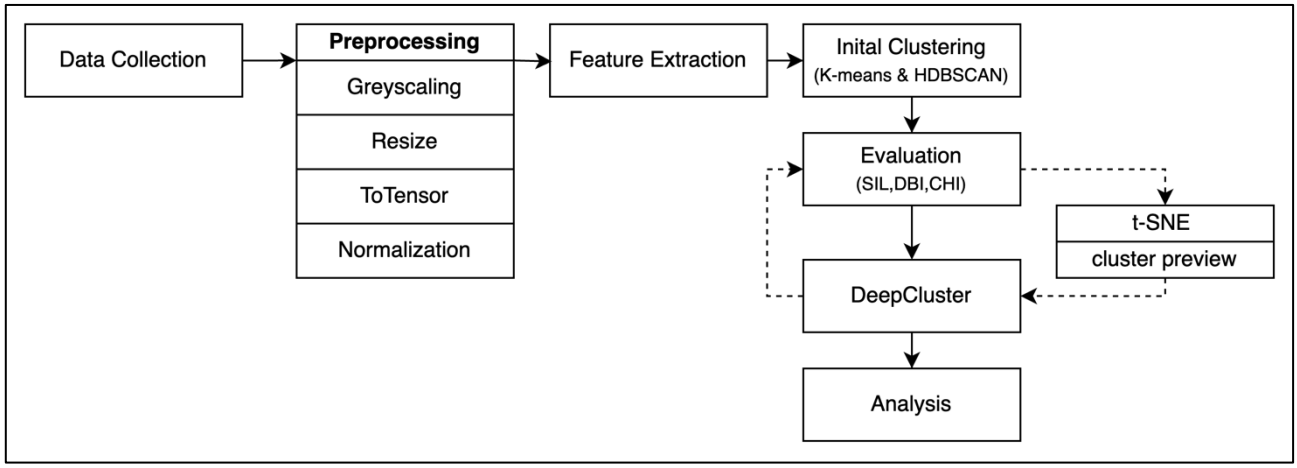

Figure 1. Research Process Flow

2.1 Type and Approach of Research

This study adopts a quantitative experimental approach, focusing on the systematic investigation of feature extraction and clustering performance on handwritten essay answer images. The research aims to explore and evaluate the effectiveness of Vision Transformer-based feature representations when combined with clustering algorithms. Through controlled experiments, clustering performance is assessed using internal evaluation metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. In addition to the quantitative evaluation, a qualitative visual analysis is also conducted. This includes visualizing the distribution of clusters in a two-dimensional space using t-SNE, as well as examining the actual content of images grouped within the same cluster. These qualitative observations serve to complement the metric-based assessment by providing insight into the visual consistency and interpretability of the resulting clusters. The combination of quantitative metrics and visual analysis justifies the use of an experimental research design, making it well-suited for understanding the behavior and performance of unsupervised learning techniques in the context of visual essay answer data.

2.2 Object and Scope of Research

The object of this research is a collection of handwritten essay answer photos submitted by university students as part of academic assessments. These images originate from two different courses—Algorithm Design and Analysis, and Database Systems—and were obtained through two primary sources: a learning management system (SPADA) and direct photo documentation. The research focuses on clustering these essay answer images based on visual similarity without relying on predefined labels or semantic content. The scope of this study is limited to the domain of educational technology, particularly in the context of automating and enhancing the assessment process in higher education. The study investigates the effectiveness of using Vision Transformer (ViT) [16] as a backbone feature extractor within a clustering pipeline, employing unsupervised learning techniques. The clustering is performed using deep clustering methods (e.g., DeepCluster) [9], [12] and evaluated with both classical and density-based clustering algorithms (e.g., K-Means, HDBSCAN). The primary aim is to analyze whether visual similarities among student responses can be meaningfully grouped, thereby opening possibilities for semi-automated feedback and answer grouping.

2.3 Data Collection Techniques

The dataset used in this study was collected from two main sources. The first source consists of handwritten essay answers submitted by students of the “Algorithm Analysis and Design” course via the SPADA (*Sistem Pembelajaran Daring UPI*) online learning platform. These images were downloaded and categorized based on the question numbers assigned by the lecturer. The second source consists of essay answer sheets from the “Database Systems” course, which were captured manually by photographing each answer sheet one by one, and then grouped

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Page: 845-854
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.517

according to their respective question numbers. Images that contain noise such as unnecessary objects will be cropped and only the answer will be left and then each iamges will be greyscaled. In both cases, the lecturers created open-ended essay questions that required students to respond with handwritten answers. After completing their answers, students were instructed to take a photo of their responses and upload them individually to the SPADA platform (for the first data source). The resulting dataset contains images of handwritten responses corresponding to different questions. These images vary in handwriting style, layout, and formatting across students. Notably, the dataset does not include predefined class labels, as the objective of this research is to group similar visual patterns of the answers through unsupervised image clustering rather than evaluating their semantic content.

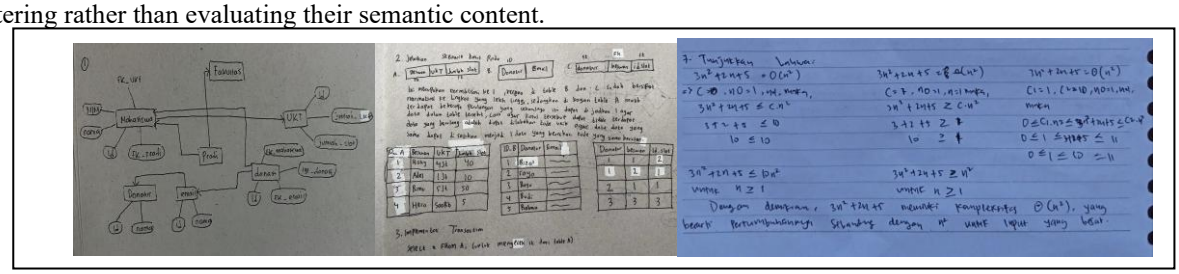

Figure. 2. Dataset Examples

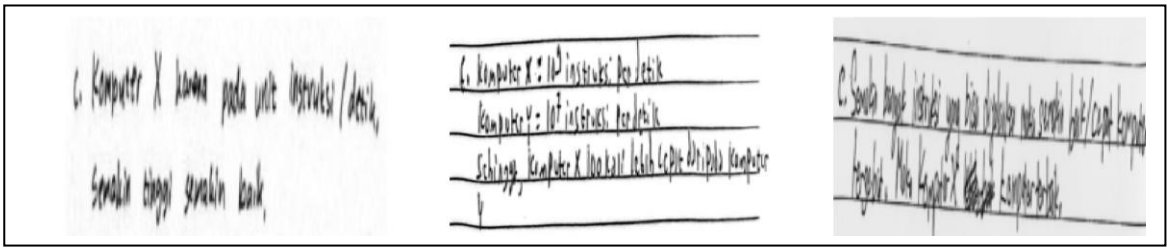

Figure 3. Dataset After Preprocess Examples

2.4 Tools and Materials Used

This study employed both hardware and software tools to support feature extraction and clustering of handwritten essay answers. The experiments were primarily conduMatplotlib a laptop with an Apple M1 processor, 8 GB RAM, and 512 GB SSD. Python 3.10 was the main programming language, with most development and training performed on Google Colaboratory, which provided GPU T4 acceleration. Visual Studio Code was used for local development, while documentation and diagrams were created using Microsoft Office, draw.io, and Google Workspace. Key Python libraries included PyTorch for deep learning, HuggingFace Transformers (AutoImageProcessor and ResNetForImageClassification) for using pre-trained models, and TorchVision for image handling. Clustering and evaluation utilized scikit-learn (K-means, silhouette score, Davies-Bouldin, Calinski-Harabasz) and HDBSCAN. Additional tools such as NumPy, Matplotlib, t-SNE, PIL, and OS supported data processing and visualization.

2.5 Research Procedures or Stages

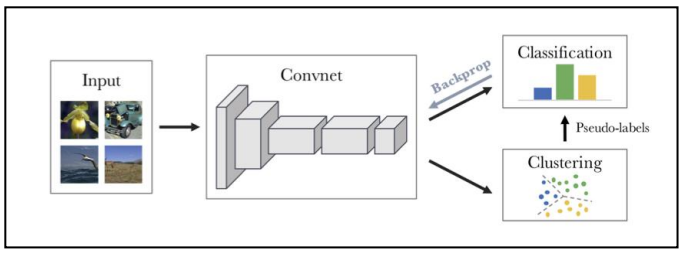

Figure 4. DeepCluster Pipeline (Caron et al) *[14]*

1) Data Collection

The dataset used in this study was sourced from two main origins. The first consisted of handwritten essay answer images submitted by university students via the SPADA e-learning platform, where each student uploaded a photo of their answer for every question. The second source involved manually captured photos of handwritten essay sheets, categorized by question number. The collected data contains diverse handwriting styles, layouts, and writing formats, with no predefined class labels, as the study focuses on clustering based on visual similarity rather than semantic content. The result of this process is as shown in **Figure 2.**

2) Preprocessing

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Page: 845-854
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.517

The preprocessing stage was conducted to standardize the format of the input images and facilitate effective feature extraction in subsequent steps. Several procedures were applied during this stage: **(a) Resizing** – each image was resized to 224×224 pixels for ResNet-50 [17] and ViT-base [18], and to 384×384 pixels for Tr-OCR [19], in accordance with the input requirements of each model; **(b) Greyscaling** – all images were converted to grayscale to reduce color-related noise and focus on structural patterns; **(c) Normalization** – image pixel values were normalized using the common mean and standard deviation from ImageNet, specifically Normalize(mean = [0.5, 0.5, 0.5], std = [0.5, 0.5, 0.5]); and **(d) ToTensor** – each image was converted into a PyTorch tensor format for compatibility with deep learning pipelines. The result of this process is as shown in **Figure 2.**

3) Feature Extraction

To extract visual representations from essay answer images, two types of backbone models were employed: a Vision Transformer (ViT) encoder and a convolutional neural network (CNN) based on ResNet-50. Both models processed preprocessed image batches and produced fixed-length feature embeddings suitable for unsupervised clustering. For the ViT-based approach, images were passed through a pretrained ViT encoder. During inference, the model was set to evaluation mode and transferred to the GPU. Each image batch was forwarded through the encoder to obtain the last_hidden_state tensor, from which the class token was excluded. Mean pooling was then applied across the remaining patch tokens, resulting in one embedding vector per image. These embeddings were collected into a matrix of shape (n_samples, embedding_dim). In parallel, a CNN-based approach using Resnet-50 was applied. The model was likewise moved to the GPU and inference was conducted with gradient computation disabled. Each image batch was passed through the CNN, and the output tensor was squeezed to remove singleton spatial dimensions, yielding vectors of shape (batch_size, feature_dim). The outputs were transferred to the CPU, converted to NumPy arrays, and concatenated into a full feature matrix. Finally, L2 normalization was applied to ensure each feature vector had unit length. These extracted feature representations from both ViT and CNN backbones served as the input for subsequent clustering and evaluation stages.

4) Clustering

Clustering was performed using two algorithms: K-Means and HDBSCAN. The K-Means algorithm was applied with the number of clusters set to **k = 5**, while HDBSCAN used a **minimum_cluster_size = 10**. To ensure fair and comparable results, the same clustering parameters were applied across all question categories.

5) Evaluation

The evaluation metrics used in this study include Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score, which aim to assess the quality of the clustering results by considering aspects such as distance between points, cluster density, and separation between clusters. Additionally, t-SNE visualization was employed as a complementary evaluation method to observe the distribution of data points across clusters, providing further insights into the clustering performance.

6) DeepCluster

The DeepCluster method was used in this study as an unsupervised learning approach to enhance the visual feature representations of handwritten essay answer sheets. This method combines two main iterative stages: clustering and model weight updating for the feature extractor. The process begins by extracting initial features from the images using backbone models such as Resnet-50, ViT-base, and Tr-OCR. These features are then clustered using the K-Means algorithm to generate pseudo-labels. These labels do not represent actual answer categories but serve as temporary targets to retrain the model, allowing it to produce more representative features. After pseudo-labeling, the backbone model is retrained using these labels so that its weights adapt to the natural structure of the data formed during clustering. This process is repeated several times—two iterations for ViT and Tr-OCR, and ten iterations for Resnet-50—until the model produces stable and high-quality features. Through this iterative process, the feature distribution becomes more compact and structured, which ultimately helps clustering algorithms like K-Means or HDBSCAN form more meaningful and well-separated clusters. The results show a significant improvement in clustering evaluation metrics after applying DeepCluster compared to the initial clustering stage.

7) Analysis

The evaluation was conducted quantitatively using clustering evaluation metrics including Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score. The performance was assessed by comparing the results of the initial clustering and the final clustering after the DeepCluster process. In addition to the quantitative evaluation, a qualitative assessment was also performed through t-SNE visualizations and by examining the contents of each resulting cluster to better understand the visual characteristics grouped by the model.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Page: 845-854
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.517

2.6 Data Analysis Techniques

The clustering results in this study were evaluated using several quantitative metrics and visualization techniques. First, the Silhouette Score [20] was used to measure both intra-cluster cohesion and inter-cluster separation. A high Silhouette value indicates that samples within a cluster are close to each other while being well separated from samples in other clusters, reflecting good clustering performance. Next, the Davies-Bouldin Index (DBI) [21] was employed to assess the average similarity between each cluster and its most similar one, where lower DBI values suggest more compact and well-separated clusters. In addition, the Calinski-Harabasz Index (CHI) [22] was applied to evaluate clustering quality based on the ratio of between-cluster dispersion to within-cluster dispersion. A higher CHI score indicates that the cluster centroids are more widely spaced and the data within each cluster is tightly grouped. Lastly, t-distributed Stochastic Neighbor Embedding (t-SNE) [23] was used for two-dimensional visualization of the feature space, allowing for a qualitative inspection of whether visually similar images were grouped effectively into the same cluster.

3. RESULTS AND DISCUSSION

The experiment utilizes each backbone to extract features from images of students' handwritten essay answer sheets. These extracted features are then used in two main stages: (1) initial clustering, which is performed on the features obtained directly from the backbone without any further training, and (2) final clustering, which is conducted after the features have undergone representation refinement through the DeepCluster process.

3.1 Presentation of Research Results

3.1.1 Initial Feature Extraction

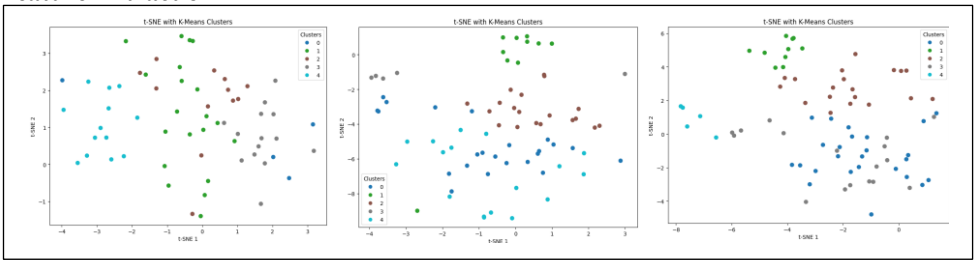

Figure 5. t-SNE visualization for ViT initial features distributions

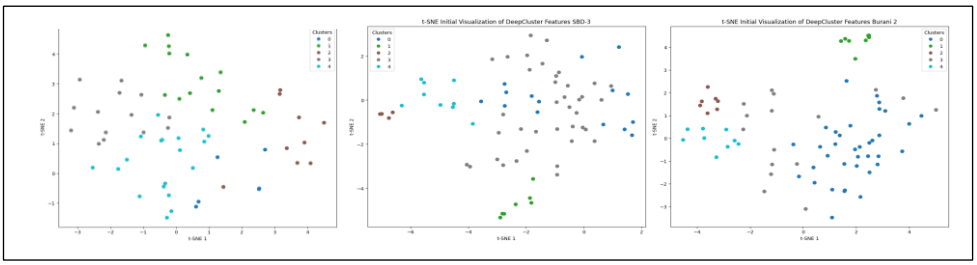

Figure 6. t-SNE visualization for Resnet-50 initial features distributions

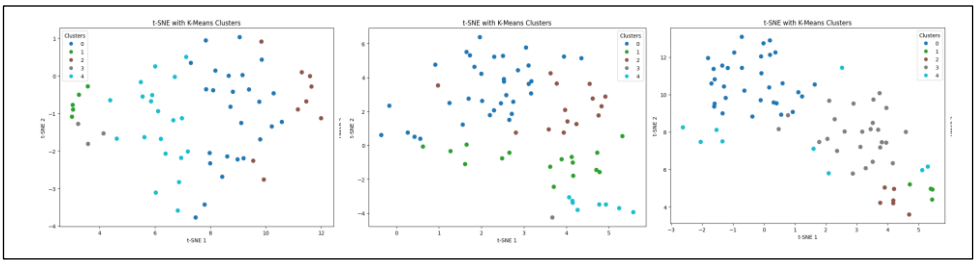

Figure 7. t-SNE visualization for Tr-OCR initial features distributions

As illustrated in **Figures 5, 6,** and **7**, the features extracted by the three backbone models appear to be insufficiently representative, resulting in scattered visualizations where data points remain far apart from each other. This indicates that the features have not yet effectively captured meaningful relationships within the data. The dispersed nature of the cluster distributions aligns with the low evaluation scores obtained during the initial clustering stage, as shown in **Tables 1** and **2**. These results collectively suggest that the initial feature representations lacked the compactness and separability required for effective clustering.

3.1.2 Final Feature Extraction

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Page: 845-854
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.517

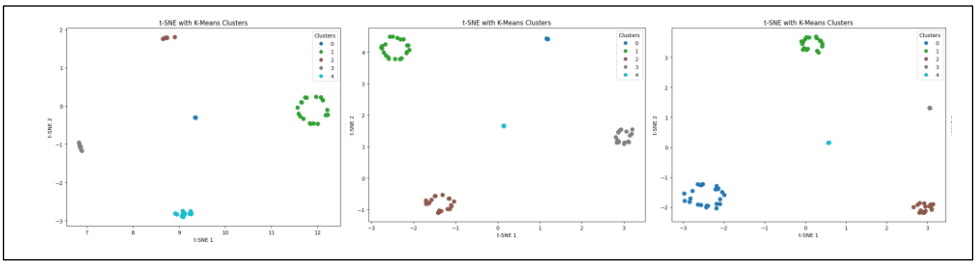

Figure 8. t-SNE visualization for ViT final features distributions

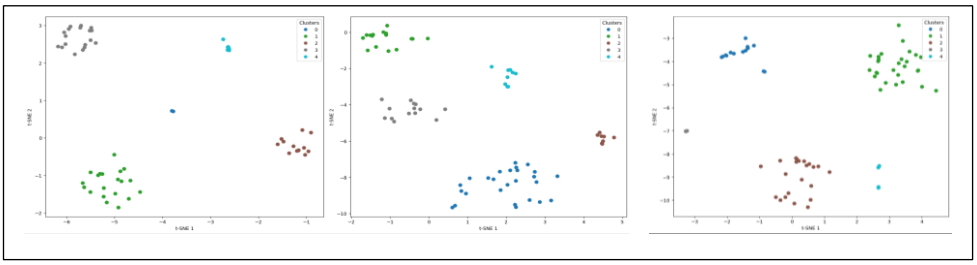

Figure 9. t-SNE visualization for Resnet-50 final features distributions

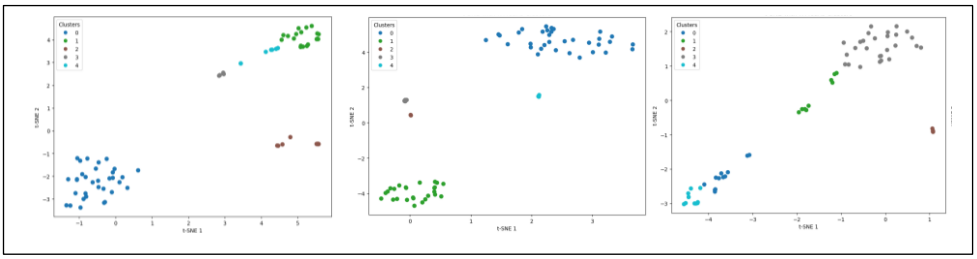

Figure 10. t-SNE visualization for Tr-OCR final features distributions

Based on the visualizations presented in **Figures 8, 9,** and **10**, it can be observed that the data points begin to form more compact and distinct groups within their respective regions. This indicates that the features have become more representative after undergoing the DeepCluster process. The improved visual distribution aligns with the results of the evaluation metrics on **Table 1** and **2**, which also show a significant improvement, further supporting the enhanced quality of the extracted features.

3.1.3 Clustering Result

Table 1. K-means Clustering Result

Backbone	Clustering K-means						Iteration
	Initial Clustering			*Final Clustering*			
	SIL	***DBI***	***CHI***	***SIL***	***DBI***	***CHI***	
Resnet-50	**0.16**	2.128	8.767	0.339	1.314	21.225	10
ViT-base	0.083	2.3	6.587	**0.783**	**0.424**	**381.47**	2
Tr-OCR	0.14	**1.77**	**24.16**	0.567	0.697	153.94	2

Table 1 presents the average values of clustering evaluation metrics for each backbone using the K-means algorithm with **k = 5**. The results demonstrate a significant improvement from the initial clustering stage to the final clustering stage, indicating the effectiveness of the DeepCluster process. Among the models, the Tr-OCR backbone achieved the best overall performance, followed by ViT-base and Resnet-50. Notably, despite Resnet-50 undergoing 10 iterations—compared to only 2 iterations for ViT-base and Tr-OCR—it still produced inferior clustering results. This highlights the superior capability of ViT based models in learning more meaningful feature representations for this particular task.

Table 2. HDBSCAN Clustering Result

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Page: 845-854
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.517

Backbone	Clustering HDBSCAN							
	SIL	*DBI*	*CHI*	*n-cluster*				Iteration
				n=0	*n=2*	*n=3*	*n=4*	
Resnet-50	0,333	1,248	26,565	**0**	10	13	3	10
ViT-base	**0,772**	**0,369**	**408,006**	**0**	**4**	13	8	2
Tr-OCR	0,618	0,575	144,751	1	10	12	3	2

Table 2 presents the average values of clustering evaluation metrics for each backbone using the HDBSCAN algorithm with a ***min_cluster_size*** of **10**. The results shown represent only the final clustering stage, as the algorithm struggled to form any clusters during the initial stage due to the sparsity of the features. After the DeepCluster process, the features became more representative **Figures 8, 9, and 10**, enabling the algorithm to successfully form meaningful clusters.

Figure 11. Images inside clusters (*each row means a different cluster*)

As shown in **Figure. 11**, the images within Cluster 1, which contains only four members, still exhibit significant visual variation—for instance, one image includes a table, while the others do not. Cluster 2 is characterized by text layouts that appear more spaced out, with noticeable gaps between lines or words. Cluster 3 shares similarities with Cluster 2, but the text appears denser, although gaps are still present. Cluster 4 consists primarily of answers that include tables, making it the most distinct and clearly defined cluster due to the prominent presence of tabular elements. Cluster 5 resembles Cluster 3, with relatively dense text and some spacing, but overall lacks strong distinguishing visual features.

3.2 Analysis of Findings

3.2.1 Clustering with K-means

The experimental results from each backbone demonstrate a significant improvement across all evaluation metrics between the initial clustering and the final clustering stage. This indicates that the features extracted after undergoing

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Page: 845-854
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.517

the DeepCluster process become more refined and representative, enabling better clustering performance. The improvement is further supported by the visual distribution of the clusters. Initially, the data points appeared scattered without discernible structure; however, after the DeepCluster refinement, the embeddings began forming more compact and well-defined groups. This alignment between the evaluation metrics and the visual t-SNE cluster distribution confirms that the feature representations learned through DeepCluster positively impact clustering outcomes (Caron et al., n.d.). **Table 1** presents the average values of clustering metrics for each backbone using the K-means algorithm with k = 5.

In the initial clustering results, all three backbones exhibited relatively low clustering quality. This is evidenced by the low Silhouette Score (SIL) and high Davies-Bouldin Index (DBI), suggesting that the clusters were poorly separated and internally dispersed. A key contributing factor to this suboptimal result is the domain gap between the pretraining data and the target dataset. The Resnet-50 and ViT-base backbones were pretrained on the ImageNet dataset [24], which contains natural images of animals, objects, and landscapes—vastly different in appearance from the handwritten essay answer sheets used in this study. When such pretrained models are applied without fine-tuning, the resulting features often fail to capture meaningful semantics in the new domain, leading to underwhelming clustering performance. This finding is consistent with the observations reported by Lowe et al. [13], who found that pretrained CNN and ViT models tend to yield degraded clustering quality when applied to out-of-domain (OOD) datasets.

Interestingly, Tr-OCR outperformed the other backbones during the initial clustering stage in terms of DBI and the Calinski-Harabasz Index (CHI). This can be attributed to Tr-OCR's two-phase pretraining process, which incorporates both synthetic and real handwritten data [19]. As a result, Tr-OCR learns visual representations that are better aligned with the characteristics of handwritten content. Given that the dataset in this study comprises handwritten essay answers, the domain similarity provides a clear advantage to Tr-OCR in capturing more relevant feature representations.

In contrast, during the final clustering stage, ViT-base surpassed the other backbones, achieving the best performance across all three evaluation metrics. This outcome suggests that Vision Transformer-based backbones are better equipped than convolutional neural networks such as Resnet-50 for producing discriminative features in an unsupervised clustering scenario, even though Resnet-50 underwent more DeepCluster iterations [25].

3.2.2 Clustering with HDBSCAN

Based on the results presented on Table 2, the ViT-base model once again outperformed the other backbones by achieving the highest scores across all three evaluation metrics. This suggests that the feature representations generated by the Vision Transformer architecture are generally more effective than those produced by the Resnet-50 model. Furthermore, the analysis of the number of clusters formed by each model reveals that Tr-OCR failed to produce any clusters using the HDBSCAN algorithm. This observation is supported by the visualization results, which show that the feature distributions generated by Tr-OCR remained relatively dispersed and lacked compactness. As HDBSCAN is a density-based clustering algorithm, it requires sufficiently dense regions to identify clusters. In this case, the absence of high-density areas in the Tr-OCR feature space made it difficult for HDBSCAN to detect and form valid clusters.

3.2.3 Evaluation Metrics and t-SNE Visualization Results

The experimental results reveal that some models produce dense and well-separated visual clusters, yet receive relatively low scores in quantitative evaluation metrics. This discrepancy arises because visualizations generated using algorithms such as t-SNE do not always directly reflect the clustering quality measured by metrics such as Silhouette Score (SIL), Davies-Bouldin Index (DBI), or Calinski-Harabasz Score (CHI). This is due to the fundamental difference between the approaches used by visualization techniques and clustering evaluation methods.

t-SNE is a dimensionality reduction technique that does not preserve the absolute distances from the original high-dimensional space. Instead, it maps the probabilistic relationships among data points into a two- or three-dimensional space. Specifically, t-SNE transforms pairwise distances into probability distributions that represent similarities, making the distances in the visualized space indicative of proximity rather than true metric distances [26]. As clustering metrics are calculated in the original high-dimensional space, this explains why data points that appear well-grouped in a t-SNE plot may still yield low evaluation scores [27]. Therefore, two-dimensional visualization using t-SNE is more appropriate for providing a qualitative impression of how well the extracted features represent the data and whether the data tends to form separable clusters. Well-structured, compact, and clearly defined visual clusters may serve as an early indication of the representational quality of the features, even if this is not immediately reflected in quantitative clustering scores.

3.3 Implications of the Results

This study provides both practical and theoretical contributions. Practically, the proposed clustering system can assist educators and academic institutions in automatically grouping handwritten essay answer sheets, helping speed up and simplify the review and grading process objectively. Theoretically, the findings show that the Vision Transformer architecture outperforms CNN in representing unstructured visual data, especially when combined with deep clustering. These results lay the groundwork for developing automated essay evaluation systems and offer insights for future research in computer vision and education. In the future, the system could be expanded by integrating OCR and NLP to understand answer content and improve generalization across institutions.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Page: 845-854
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.517

3.4 Limitations of the Study

The clustering process was solely based on the visual similarity of the answer sheet images, without considering the actual content or semantic meaning of the students' written responses. The dataset used in this experiment consisted of 1,966 essay answer sheet images from two different courses: (1) Algorithm Analysis and Design, which included 23 different types of questions, and (2) Database Systems, which consisted of 3 types of questions. The number of samples per question type was relatively limited, averaging around 65 to 70 images per category. The models used in this study were ResNet-50 and ViT-base, both pretrained on the ImageNet dataset. These models were not specifically trained on handwritten answer sheet images. In contrast, the Tr-OCR backbone used was pretrained on handwritten datasets, giving it a potential advantage in representing handwriting-based visual features. Clustering was conducted directly on the extracted features in their original high-dimensional space, without applying any dimensionality reduction techniques. This approach was chosen to preserve the complete information contained in the feature representations, although it may increase the complexity of the clustering process.

4. CONCLUSION

This study demonstrates the effectiveness of deep clustering methods combined with different backbone architectures for the task of unsupervised grouping of handwritten essay answer sheets. By extracting features using CNN, ViT-base, and Tr-OCR models, and applying DeepCluster with K-means and HDBSCAN algorithms, it was found that ViT-based backbones, particularly Tr-OCR, outperform CNN in generating more representative and discriminative feature embeddings. This was evident through improved clustering evaluation metrics and clearer cluster structures in t-SNE visualizations. For example, the final clustering results using ViT and HDBSCAN achieved a Silhouette Score of 0.772, Davies-Bouldin Index of 0.369, and Calinski-Harabasz Index of 408.006, indicating strong intra-cluster cohesion and well-separated clusters. Moreover, the study highlights that clustering handwritten answer sheets based solely on visual features can already provide meaningful groupings, such as grouping answers with tables or similar text layouts. This shows potential for practical applications in academic settings, where educators could benefit from automated essay clustering to support quicker, more objective grading processes. Future work may explore the integration of text-based content understanding using OCR and NLP, as well as testing the system on more diverse datasets to enhance its generalizability and robustness.

REFERENCES

[1] A. Sopian, D. Fungsi Guru, and A. Sopian Sekolah Tinggi Ilmu Tarbiyah Raudhatul Ulum, “Tugas, Peran dan Fungsi Guru dalam Pendidikan,” 2016.

[2] N. Nurhasanah *et al.*, “Evaluasi Pembelajaran Dikelas Universitas Islam Negeri Sumatera Utara,” vol. 1, no. 2, p. 6, 2023, doi: 10.59581/jmpb-widyakarya.v1i2.485.

[3] I. Magdalena, H. N. Fauzi, and R. Putri, “PENTINGNYA EVALUASI DALAM PEMBELAJARAN DAN AKIBAT MEMANIPULASINYA,” 2020. [Online]. Available: https://ejournal.stitpn.ac.id/index.php/bintang

[4] L. R. Wachidah, Y. Laila, A. Irmawati, S. Amin, T. Bahasa Indonesia, and I. Madura, “Implementasi Penggunaan Tes Essay dalam Evaluasi Pembelajaran Daring pada Siswa Kelas VII SMP Negeri 1 Tlanakan KONFERENSI NASIONAL LALONGÉT II.”

[5] Siswanto, “PENGGUNAAN TES ESSAY DALAM EVALUASI PEMBELAJARAN,” *JURNAL PENDIDIKAN AKUNTANSI INDONESIA Vol. V No. 1 – Tahun 2006 Hal. 55 - 61*.

[6] H. Sheikh, C. Prins, and E. Schrijvers, “Artificial Intelligence: Definition and Background,” 2023, pp. 15–41. doi: 10.1007/978-3-031-21448-6_2.

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*.

[8] Y. Chen, S. Wang, L. Lin, Z. Cui, and Y. Zong, “Computer Vision and Deep Learning Transforming Image Recognition and Beyond,” *International Journal of Computer Science and Information Technology*, vol. 2, no. 1, pp. 45–51, Mar. 2024, doi: 10.62051/ijcsit.v2n1.06.

[9] J. A. Hartigan and M. A. Wong, “A K-Means Clustering Algorithm,” *J R Stat Soc Ser C Appl Stat*, vol. 28, no. 1, pp. 100–108, 1979, doi: https://doi.org/10.2307/2346830.

[10] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, “On the Surprising Behavior of Distance Metrics in High Dimensional Space.”

[11] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, “DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN,” *ACM Transactions on Database Systems*, vol. 42, no. 3, Jul. 2017, doi: 10.1145/3068335.

[12] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.

[13] S. C. Lowe, J. B. Haurum, S. Oore, T. B. Moeslund, and G. W. Taylor, “An Empirical Study into Clustering of Unseen Datasets with Self-Supervised Encoders,” Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.02465

[14] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep Clustering for Unsupervised Learning of Visual Features,” Jul. 2018, [Online]. Available: http://arxiv.org/abs/1807.05520

[15] H.-B. Ling, B. Zhu, D. Huang, D.-H. Chen, C.-D. Wang, and J.-H. Lai, “Vision Transformer for Contrastive Clustering,” Jul. 2022, [Online]. Available: http://arxiv.org/abs/2206.12925

[16] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Page: 845-854
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.517

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[18] B. Wu *et al.*, "Visual Transformers: Token-based Image Representation and Processing for Computer Vision," 2020.
[19] M. Li *et al.*, "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models," 2021.
[20] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," 1987.
[21] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans Pattern Anal Mach Intell*, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.
[22] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, Jan. 1974, doi: 10.1080/03610927408827101.
[23] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," 2008.
[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
[25] X. Zhou and N. L. Zhang, "Deep Clustering with Features from Self-Supervised Pretraining," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.13364
[26] T. T. Cai and R. Ma, "Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2105.07536
[27] Z. Yang, Y. Chen, and J. Corander, "T-SNE Is Not Optimized to Reveal Clusters in Data," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.02573