EXPLORER
Journal of Computer Science and Information Technology
E-ISSN 2774-4647
Website: https://journal.fkpt.org/index.php/Explorer DOI: https://doi.org/10.47065/explorer.v6i2.2915

Multimodal Medical Clustering of Chest X-Ray Images and Clinical Reports Using Deep Visual and Semantic Text Features

Humasak Simanjuntak[1*], Lenni Hutapea[2], Jonathan Simorangkir[3], Yolanda Saragih[4], Jaden Panggabean[5]
[1*,2,3,4,5]Program Studi Sistem Informasi, Fakultas Informatika dan Teknik Elektro, Institut Teknologi Del, Laguboti, Indonesia
[1*]humasak@del.ac.id, [2]iss23023@students.del.ac.id, [3]iss23030@students.del.ac.id, [4]iss23050@students.del.ac.id, [5]jaden.panggabean@del.ac.id
[*]humasak@del.ac.id

Abstract-The limited availability of labelled medical data poses a significant challenge to developing supervised learning-based radiology analysis systems. This study proposes an unsupervised multimodal clustering framework that jointly exploits two complementary clinical information sources: chest X-ray images and radiology reports. Visual features are extracted from 1,000 chest X-ray images using a pretrained ResNet152, while semantic features are generated using a Dual TF-IDF Vectorizer that independently processes the findings and conclusion sections with weighting factors of 0.4 and 0.6, respectively, and is further enriched with ten latent topics derived from Latent Dirichlet Allocation (LDA). The visual and textual representations are integrated through weighted late fusion after L2 normalization. UMAP with 50 latent components and cosine distance is subsequently applied for dimensionality reduction before K-Means clustering. Experimental results demonstrate that the proposed multimodal framework achieves its best performance with 8 clusters, yielding Silhouette Scores of 0.7324, Davies–Bouldin Indexes of 0.3416, and Calinski–Harabasz Indexes of 3,789.10, indicating compact, well-separated cluster structures. The resulting clusters successfully capture clinically meaningful diagnostic patterns, including normal examinations, cardiomegaly, pneumonia, pleural abnormalities, and other thoracic conditions, demonstrating that integrating visual and semantic clinical information enables effective exploration of unlabeled radiology data. These findings suggest that the proposed framework provides a practical foundation for medical data organization and exploratory clinical decision support without requiring manual annotations.

Keywords: Multimodal Clustering, Chest X-Ray, Clinical Report, ResNet152, Dual TF-IDF, K-Means

1. INTRODUCTION

Artificial Intelligence (AI) has emerged as a transformative force in modern healthcare, reshaping how medical diagnostic systems operate, particularly in the interpretation of radiological imaging. Among the most widely used diagnostic tools, chest X-rays (CXRs) play a central role in detecting pulmonary and cardiovascular conditions, including pneumonia, tuberculosis, pulmonary oedema, pleural effusion, and cardiomegaly. The application of deep learning to CXR analysis, especially through Convolutional Neural Networks (CNNs) built on architectures such as ResNet, DenseNet, and EfficientNet, has yielded impressive results across disease classification and detection benchmarks [1], [2], [3], [4], [5].

Despite this progress, the predominant paradigm remains heavily dependent on supervised learning, which demands large volumes of precisely annotated data. In medical imaging, however, data annotation is far from straightforward. It requires the expertise of specialist radiologists, substantial financial investment, extended timelines, and is frequently subject to inter-expert variability [6] [7]. This challenge is compounded in resource-limited settings, particularly in developing nations such as Indonesia, where the scarcity of qualified radiologists makes the construction of high-quality labelled datasets an especially formidable undertaking. Layered on top of this are persistent concerns surrounding data privacy, regulatory compliance, and restricted access to clinical records, all of which hinder the large-scale collection and distribution of medical data [8].

As a consequence, the majority of medical AI research continues to depend on internationally curated public datasets such as NIH ChestX-ray14 [9], CheXpert [10], MIMIC-CXR [11], and PadChest [12]. These collections, predominantly sourced from hospitals in the United States and Europe, carry inherent demographic and clinical biases that limit their representativeness for other populations. Patient demographics, disease prevalence patterns, imaging protocols, report language, and clinical case distributions in these datasets differ substantially from the Indonesian healthcare context. Indonesia presents a distinct epidemiological landscape shaped by endemic diseases, including tuberculosis (TB), tropical respiratory infections, regionally variant community-acquired pneumonia, and pulmonary conditions tied to specific environmental exposures [13]. Relying uncritically on foreign datasets, therefore, introduces the risk of systematic model bias and diminished generalizability when these systems are deployed to the local Indonesian population.

Beyond the limitations of the dataset, AI-driven chest X-ray research in Indonesia remains nascent compared to more developed research ecosystems. Existing work has largely been confined to single-modality

EXPLORER
Journal of Computer Science and Information Technology
E-ISSN 2774-4647
Website: https://journal.fkpt.org/index.php/Explorer DOI: https://doi.org/10.47065/explorer.v6i2.2915

image classification, extracting insights solely from visual information without incorporating the broader clinical context that practicing radiologists routinely draw upon, including physician reports, medical history, and detailed descriptions of radiological findings [14]. However, in real clinical workflows, diagnosis is rarely a purely visual act. Radiologists synthesize visual observations with textual interpretation, patient-specific context, and clinical reasoning before arriving at a final assessment.

This gap motivates a shift toward multimodal approaches. The integration of chest X-ray images with medics' clinical reports holds considerable promise for constructing richer, more clinically faithful data representations. While visual features encode anatomical structures and pathological patterns, clinical text conveys semantic information about medical interpretation, diagnostic certainty, and disease severity [15], [16], [17]. Together, these modalities more closely mirror the actual cognitive process of radiological diagnosis than either alone. However, the prevailing literature on multimodal medical data remains concentrated on supervised classification tasks and image captioning, leaving unsupervised multimodal clustering, particularly in the Indonesian clinical setting, largely unexplored [18], [19], [20], [21].

Unsupervised clustering offers a compelling alternative, especially in situations with limited annotation. By grouping cases based on shared visual and semantic characteristics, clustering systems can facilitate the discovery of latent disease patterns, expose hidden clinical subgroups, support dataset organisation, and lay the groundwork for subsequent semi-supervised learning pipelines. In radiology specifically, multimodal clustering enables the identification of cases with distinctive abnormal signatures without requiring explicit diagnostic labels from clinicians.

Motivated by these challenges and opportunities, this study proposes a Multimodal Medical Clustering framework for Chest X-Ray Images and Clinical Reports Using Deep Visual and Semantic Text Features. The proposed approach combines visual representations extracted from local CXR images with semantic text features derived from a Dual TF-IDF Vectorizer applied to physicians' clinical reports. To further enrich the textual representation, latent topic structures are captured using Latent Dirichlet Allocation (LDA). The two modalities are subsequently integrated using a weighted late-fusion strategy, yielding a unified, comprehensive multimodal representation suitable for unsupervised medical clustering in the Indonesian clinical context.

2. METHODOLOGY

This research follows the CRISP-DM methodology, which comprises four systematic stages: Data Understanding, Data Preparation, Building Model, and Deployment. An overview of the methodology flow is explained in the subsection.

2.1 Dataset and Data Understanding

The dataset comprised 1,000 patient records, each consisting of a single chest X-ray image in PNG format and a corresponding clinical report containing the physician's findings and interpretation (conclusion), stored in CSV format. All personally identifiable information was removed prior to analysis.

Temuan:
Jantung tidak membesar. Aorta dan mediastinum superior tidak melebar. Trachea di tengah. Kedua hilus tidak menebal. Corakan bronchovaskuler kedua paru baik. Tidak tampak infiltrate di kedua lapangan paru. Kedua sinus costophrenicus lancip, kedua hemidiafragma licin. Tulang-tulang dinding dada intak, tidak tampak scoliosis.
Kesimpulan:
Tidak tampak kelainan radiologis pada jantung dan paru-paru.

Figure 1. Explorer Sample chest x-ray image and the clinical report

EXPLORER
Journal of Computer Science and Information Technology
E-ISSN 2774-4647
Website: https://journal.fkpt.org/index.php/Explorer DOI: https://doi.org/10.47065/explorer.v6i2.2915

The de-identified dataset was subsequently encrypted using the Fernet symmetric encryption scheme and stored on a secure institutional server, ensuring that the stored data remains inaccessible without the corresponding decryption key. All PNG images were uniform in size (1,466 × 1,706 pixels) in RGB colour mode, with a mean file size of 0.89 MB (range: 0.54–1.17 MB). The three sample chest X-ray images are shown in Figure 1.

Preliminary inspection confirms that there are no missing values or duplicate records in the dataset. Regarding textual content, the findings field has a mean length of 315.47 characters (SD=36.64). In comparison, the conclusion field averaged 65.74 characters (SD=27.89), reflecting the inherently more concise nature of diagnostic summaries relative to observational descriptions.

2.2 Data Preparation

The data preparation stage is performed to improve the dataset's quality before feature extraction and clustering. At this stage, text preprocessing is carried out on the findings and conclusions attributes through several stages: converting all text to lowercase, removing non-alphabetic characters and numbers using regular expressions ([^a-z\s]); tokenization, selective removal of Indonesian stopwords, where words that have no clinical meaning such as "tidak", "tanpa", and "bukan" are deliberately retained because they fundamentally change the meaning of the diagnosis (for example: "tidak terlihat besar" as opposed to "tampak besar"); and finally, removing short tokens (less than three characters). The results of the preprocessing are stored in a new attribute.

Furthermore, image preprocessing is applied to the chest X-ray images to ensure consistent and reliable feature extraction. Initial visualization of the PNG images revealed considerable variations in brightness, contrast, orientation, and aspect ratio across the dataset. To address these variations, all images are converted to RGB format and square-padded to standardize their aspect ratios while preserving the original anatomical structures. The images are subsequently resized to 224 × 224 pixels to meet the input requirements of DenseNet121. Finally, pixel values are normalized using the standard ImageNet mean and standard deviation to ensure compatibility with the pre-trained model and enhance the stability of the feature extraction process.

2.3 Building Model

The proposed technical pipeline comprises four tightly integrated steps. In the first step, feature extraction is performed independently for each modality: visual features are derived from chest X-ray images using a pretrained ResNet152 convolutional neural network, while textual features are constructed from clinical reports through a combination of Dual TF-IDF vectorization and Latent Dirichlet Allocation (LDA). The second step applies weighted late fusion, in which each modality's feature vector is first L2-normalised to ensure commensurate scaling, and the representations are then combined into a unified multimodal vector. The third step performs dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP), which projects the high-dimensional fused representation into a lower-dimensional space while preserving both local neighbourhood structure and global topological relationships inherent in the original data. The last step applies K-Means clustering to the reduced representation, grouping records according to their multimodal feature similarity to reveal latent clinical patterns across the dataset. The resulting cluster assignments are analyzed to characterize the dominant clinical within each group. The complete architecture workflow is illustrated in Figure 2.

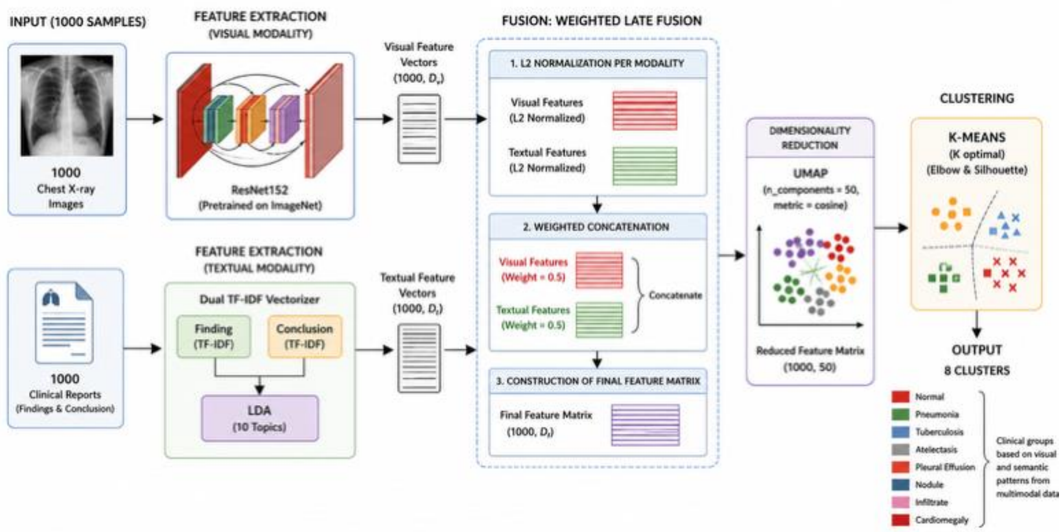

Figure 2. Multimodal Clinical Pattern Discovery from Chest X-Ray Images and Radiology Reports Using Deep Visual and Semantic Text Representations

EXPLORER
Journal of Computer Science and Information Technology
E-ISSN 2774-4647
Website: https://journal.fkpt.org/index.php/Explorer DOI: https://doi.org/10.47065/explorer.v6i2.2915

2.3.1 Visual Feature Extraction with ResNet152

Visual feature extraction is performed using ResNet152, a 152-layer Deep Convolutional Neural Network (DCNN) pretrained on the ImageNet dataset. ResNet152 is selected for its well-established capacity to produce stable, discriminative high-level visual representations across both medical and general image recognition tasks. Its residual learning mechanism, implemented through skip connections, facilitates effective gradient flow during backpropagation, mitigating the vanishing gradient problem that commonly afflicts very deep conventional CNN architectures. By learning identity mappings across layers, residual connections enable the network to retain salient information throughout its depth without degradation.

Transfer learning is employed to leverage the rich visual knowledge encoded in ImageNet-pretrained weights, compensating for the relatively limited size of the chest X-ray dataset. In this hierarchical feature-learning process, shallow layers capture low-level visual primitives such as edges, texture, and contrast, intermediate layers encode progressively more complex patterns, and the deepest layers produce high-level semantic representations corresponding to anatomical structures and radiological findings. This hierarchical architecture makes ResNet152 particularly well-suited for extracting clinically relevant features from chest radiographs.

The fully connected classification head is removed, leaving the network to function exclusively as a feature extractor, yielding a 2,048-dimensional feature vector per image. Prior to extraction, each chest X-ray image (originally 1,466 × 1,706 pixels) is resized to 224 × 224 pixels to conform to the ResNet152 input specification. Pixel intensities are then normalized using the ImageNet channel-wise statistics (mean = [0.485, 0.456, 0.406]; standard deviation = [0.229, 0.224, 0.225]) to align the input distribution with that observed during pretraining, ensuring compatibility with the learned feature representations.

Inference is conducted in batch mode using a DataLoader with a batch size of 32 to optimize GPU memory utilization and maintain consistent sample ordering. The extraction process produces a visual feature matrix of dimensions 1,000 × 2,048, denoted as where each row encodes the visual characteristics of a single chest X-ray image as a dense numeric vector. This matrix is subsequently passed to the multimodal fusion stage, along with the extracted clinical text features.

2.3.2 Report Feature Extraction with TF-IDF and LDA

Text feature extraction is designed to construct a semantically rich representation of radiology reports, each comprising two clinically distinct components: a findings section and a conclusion section. Rather than concatenating these sections prior to vectorization, the two components are processed independently using separate TF-IDF vectorizers, an approach referred to here as Dual TF-IDF, to preserve the distinct linguistic characteristics and differential clinical value of each section. The findings section typically contains detailed observational descriptions of radiological appearances, whereas the conclusion section, though more concise, carries greater diagnostic weight as it reflects the radiologist's final interpretive judgment [22].

Feature extraction for each text is performed using the Term Frequency–Inverse Document Frequency (TF-IDF) method, which weights terms according to their within-document frequency relative to their corpus-wide prevalence, thereby emphasising clinically discriminative terminology while suppressing common, uninformative terms [23]. The sublinear_tf=True parameter compresses term frequencies via a logarithmic transformation, mitigating the disproportionate influence of high-frequency terms characteristic of radiology reports. To capture multi-word clinical expressions, such as pleural effusion, mild cardiomegaly, and pulmonary infiltrate, the vectorizers are configured with ngram_range=(1, 2), enabling both unigram and bigram representations. Extraction yields 492 features from the findings section and 355 features from the conclusion section.
The two TF-IDF representations are subsequently merged through a weighted linear combination:

$$F_{text} = 0.4 * T_{Finding} + 0.6 * T_{Conclusion} \tag{1}$$

A higher weight is assigned to the conclusion, on the basis that it encapsulates the radiologist's final diagnosis and therefore carries greater clinical and discriminative value than the observational narrative of the findings. This produces a combined TF-IDF matrix of dimensions 1,000 × 847.

To further enrich the semantic representation beyond surface-level term co-occurrence, Latent Dirichlet Allocation (LDA) is applied to capture latent thematic structure within the report corpus. LDA is a generative probabilistic model that represents each document as a mixture of latent topics, where each topic is characterized by a distribution over vocabulary terms [24]. In the context of radiology reports, LDA enables the discovery of recurring clinical patterns that may not be explicitly recoverable from TF-IDF alone, such as co-occurring symptom clusters or disease-specific terminological distributions. The model is configured with n_components=10, max_iter=20, and random_state=42 under batch learning. This yields a topic-document probability matrix of dimension 1,000 × 10, where each entry represents the posterior probability that a given report belongs to each of the ten latent topics. Inspection of the highest-probability terms within each topic permitted clinical labeling of the ten discover topics as follows: (1) Normal general, (2) Normal variant, (3)

EXPLORER
Journal of Computer Science and Information Technology
E-ISSN 2774-4647
Website: https://journal.fkpt.org/index.php/Explorer DOI: https://doi.org/10.47065/explorer.v6i2.2915

Dextroscoliosis, (4) Pleural effusion, (5) Chronic emphysema, (6) Bronchopneumonia, (7) Cardiomegaly, (8) Lung mass/nodule, (9) Atherosclerosis, and (10) Cavitation/fracture. These labels are interpretive and provided for descriptive purposes only; the LDA model operates unsupervised, without access to diagnostic ground truth. The final text feature matrix is obtained by horizontally concatenating the combined TF-IDF matrix and the LDA topic matrix:

$$F_{text} = [T_{combined} | T_{LDA}] \in \mathbb{R}^{1000\times857} \tag{2}$$

comprising 847 TF-IDF features and 10 LDA topic features, for a total of 857 dimensions.

2.3.3 Weighted Late Fusion

After independent extraction of visual and textual features, the two modality-specific representations are integrated using a Weighted Late Fusion strategy. Late fusion, which performs at the feature representation level after each modality has been fully and independently processed, is preferred over early fusion approaches because it preserves the intrinsic characteristics of each modality while producing a complementary joint representation [25]. In this study, the two primary information sources are visual features derived from chest X-ray images, which encode anatomical structure and visual abnormality patterns, and semantic features derived from clinical reports, which provide explicit diagnostic context and interpretive judgment. The fusion process proceeds through two sequential steps: L2 normalization and weighted concatenation.

a. L2 Normalization

Prior to fusion, each modality's feature matrix is independently normalized using L2 normalization, such that every feature vector x is rescaled to unit Euclidean norm:

$$\hat{x} = \frac{x}{\|x\|_2} \tag{3}$$

This step is essential in multimodal learning because visual and textual feature vectors occupy fundamentally different numerical ranges. Without normalization, modalities with larger magnitude distributions would disproportionately influence downstream distance computations during dimensionality reduction and clustering, effectively obscuring the contribution of the lower-magnitude modality.

b. Weighted Concatenation

Following normalization, the two modality vectors are combined through weighted horizontal concatenation. Assuming that textual and visual information contribute equally to clinical decision-making, each modality is assigned the same scalar weight in constructing the final clinical representation:

$$F_{fusion} = [0.5 \times F_{visual} \mid 0.5 \times F_{text}] \tag{4}$$

where $F_{visual} \in \mathbb{R}^{N\times2048}$ denotes the ResNet152-derived visual feature matrix, $F_{text} \in \mathbb{R}^{N\times857}$ denotes the Dual TF-IDF and LDA text feature matrix, and F_{fusion} is the resulting multimodal representation. Equal weights of 0.5 are assigned to each modality to ensure a balanced contribution from both visual and semantic information sources, reflecting the complementary and equally informative nature of radiological images and clinical reports in diagnostic decision-making. This explicit weighting scheme allows the relative contributions of each modality to be systematically controlled, a key advantage of weighted late fusion over unweighted concatenation.

2.3.4 Dimension Reduction with UMAP and Clustering

Following weighted late fusion, the resulting high-dimensional multimodal feature matrix $F_{fusion} \in \mathbb{R}^{N\times2905}$ is subjected to dimensionality reduction prior to clustering. High-dimensional representations are known to degrade the performance of distance-based clustering algorithms due to the curse of dimensionality, in which inter-point distances become increasingly uniform as dimensionality increases, rendering cluster boundaries less distinguishable.

Dimensionality reduction is performed using Uniform Manifold Approximation and Projection (UMAP) [26], a nonlinear manifold learning method that projects high-dimensional data into a lower-dimensional embedding while preserving local neighbourhood structure and global topological relationships. Compared to linear methods such as Principal Component Analysis (PCA), UMAP is better suited to multimodal feature spaces with heterogeneous distributions, as it captures complex nonlinear dependencies that linear projections cannot recover. UMAP has demonstrated strong performance in both visualization and pre-clustering dimensionality reduction across a range of biomedical applications [27].

The key hyperparameter governing the output embedding is n_components, which determines the dimensionality of the reduced representation. Since this parameter directly influences the quality of the embedding

EXPLORER
Journal of Computer Science and Information Technology
E-ISSN 2774-4647
Website: https://journal.fkpt.org/index.php/Explorer DOI: https://doi.org/10.47065/explorer.v6i2.2915

and, consequently, clustering performance, it is selected through systematic empirical evaluation. UMAP is applied across five candidate values $d \in \{20, 30, 50, 75, 100\}$, and each resulting embedding is evaluated using K-Means clustering with the Silhouette Score as the quality criterion.

The evaluation results are summarized in Figure 3. The configuration n_components=50 yields the highest Silhouette Score of 0.7524, indicating the most well-separated and internally coherent cluster structure among all configurations tested. This finding is consistent with the well-established trade-off in manifold learning between information retention and noise suppression: insufficient dimensionality discards clinically relevant structural information, while excessive dimensionality preserves redundant features and noise, thereby impairing cluster separability. The value $d=50$ optimally balances these competing considerations for the fused multimodal representation employed in this study.

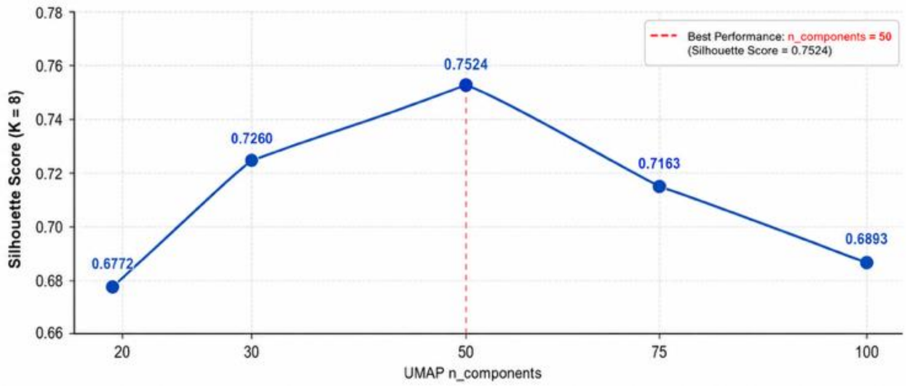

Figure 3. Silhouette-Based Evaluation of UMAP Dimensionality Reduction for Multimodal Medical Data

Accordingly, n_components=50 is adopted as the final UMAP configuration, along with the following parameter settings: n_neighbors=15, min_dist=0.1, and metric=cosine. The cosine distance metric is selected in preference to Euclidean distance because it measures angular similarity between feature vectors, a more appropriate dissimilarity measure for normalized, high-dimensional TF-IDF and deep visual feature representations, in which vector direction is more informative than vector magnitude. The resulting low-dimensional embedding $\mathbf{E} \in \mathbb{R}^{N\times 50}$ is subsequently used as input to the K-Means clustering stage.

3. RESULTS AND DISCUSSIONS

This section presents the implementation results and performance evaluation of the proposed multimodal medical clustering framework. The analysis focuses on integrating visual features extracted from chest X-ray images with semantic features derived from clinical radiology reports. Before clustering, the multimodal features were extracted, weighted late-fused, and reduced in dimensionality using UMAP to obtain a more compact and informative representation of the data. The discussion begins with word cloud visualizations to explore the distribution of dominant medical terms and semantic patterns captured from the clinical reports. Next, the optimal number of clusters for the K-Means algorithm is determined using the Elbow Method and the Silhouette Score analysis. Finally, the clustering results are evaluated to examine the quality of cluster separation, the stability of the generated clusters, and the clinical patterns identified from the multimodal representations. Through these analyses, this study demonstrates that combining visual and textual information can provide a richer, more meaningful representation of medical data for unsupervised clinical pattern exploration, without relying on manually annotated labels.

3.1 Semantic Exploration of Clinical Reports Using Word Cloud Visualization

Word cloud visualization is applied to examine the semantic content of the clinical report corpus prior to clustering. This analysis provides an interpretable overview of the dominant medical terminology captured by the Dual TF-IDF and LDA. Figure 4 presents word cloud visualizations for the ten latent topics identified by the LDA model from the chest X-ray radiology report corpus. Term size reflects contribution weight within each topic, providing an interpretable overview of the semantic structure captured by the Dual TF-IDF and LDA pipeline.

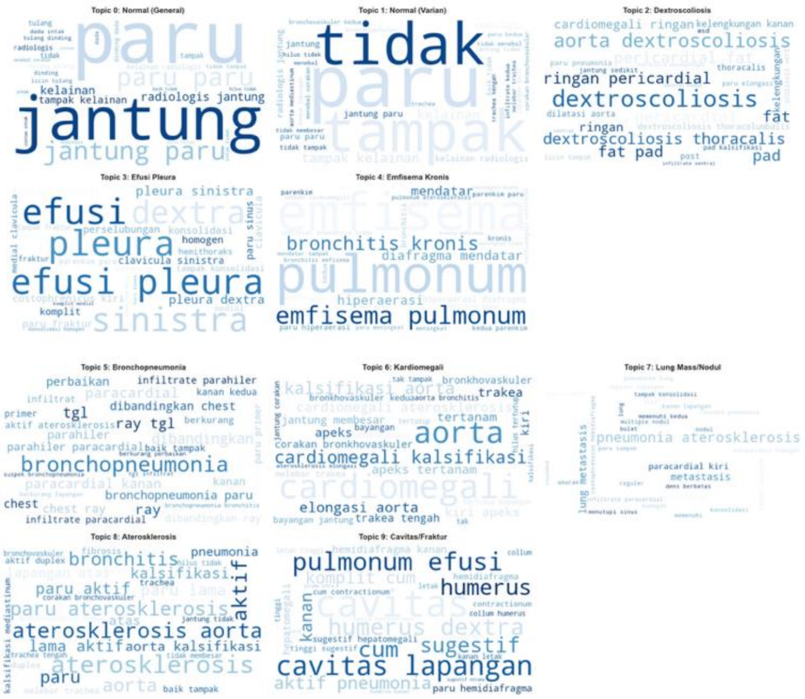

Figure 4. Word Cloud Visualization for each LDA Topic

Topics 0 and 1 both represent normal radiological findings, distinguished primarily by linguistic variation in reporting style, Topic 0 centers on organ references (“paru”, “jantung”) paired with negation terms, while Topic 1 more closely mirrors the formal phrase “tidak tampak kelainan radiologis” (no radiological abnormality detected). Topic 2 captures dextroscoliosis, with co-occurring pericardial fat pad terms indicating frequent incidental mediastinal findings. Topic 3 represents pleural effusion, with bilateral laterality terms and the presence of consolidation suggesting a subset of parapneumonic cases. Topic 4 reflects chronic emphysema, with hyperaeration and diaphragmatic flattening terms consistent with classic Chronic Obstructive Pulmonary Disease (COPD) radiological findings. Topic 5 captures bronchopneumonia through perihilar infiltrate terminology with right-sided predominance. Topic 6 identifies cardiomegaly alongside aortic elongation and calcification, suggesting an older cardiovascular patient subgroup. Topic 7 encompasses pulmonary masses and nodules, with metastasis terms indicating inclusion of both primary and secondary malignant lesions. Topic 8 characterizes atherosclerosis through aortic calcification terms, with concurrent active lung terminology suggesting overlap with chronic pulmonary infection. Topic 9 represents a mixed cluster of cavitary lung disease and thoracic trauma, with cavitation terms clinically significant in the context of the high TB burden in the Indonesian population.

Overall, the ten topics span the full clinical spectrum present in the corpus, normal variants, infectious and inflammatory conditions, cardiovascular and degenerative disease, malignancy, and structural abnormalities, confirming that the extracted text features carry sufficient semantic specificity to support the subsequent multimodal clustering stage.

3.2 Evaluation of Multimodal Clustering Performance

Determining the optimal number of clusters k is a critical step in unsupervised learning, as it directly governs the granularity of data partitioning and the interpretability of the resulting clinical groupings. To ensure a robust and unbiased selection, k is evaluated across the range $k \in \{2,3, \dots ,15\}$ using four complementary internal validation metrics: the Elbow Method (inertia), Silhouette Score, Davies-Bouldin Index (DBI), and Calinski Harabasz Indeks (CHI). Employing multiple criteria simultaneously reduces the risk of overfitting the cluster count to any single metric's bias.

The evaluation results are presented in Figure 5. The Elbow Method curve shows a pronounced reduction in inertia for $k < 8$, beyond which the rate of decrease diminishes markedly, forming a characteristic inflection point at $k = 8$. This indicates that additional clusters beyond this point contribute negligible improvement to within-cluster compactness.

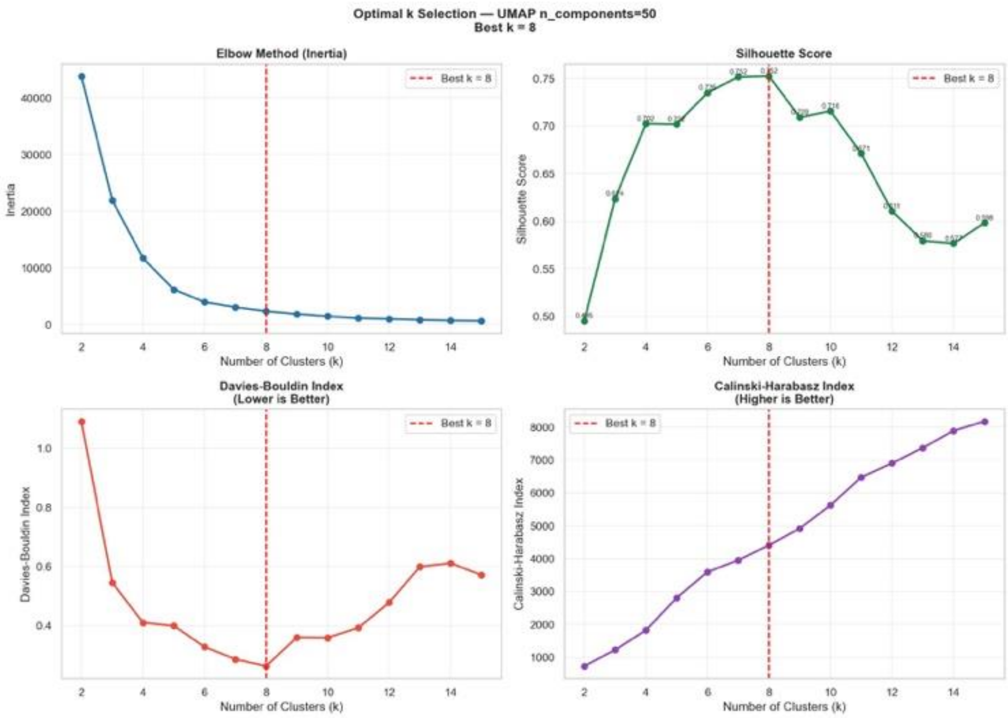

Figure 5. Optimal Cluster Selection Using Multi-Metric Internal Validation Analysis.

The Silhouette Score reaches its global maximum at *k=8*, with a value of 0.752, indicating well-separated, internally cohesive clusters. This score reflects that samples are substantially more similar to members of their own cluster than to those of the nearest neighbouring cluster. The DBI attained its minimum at *k=8* (approximately 0.30), confirming minimal inter-cluster overlap and maximum cluster distinctiveness relative to all other configurations tested. While the CHI exhibits a monotonically increasing trend with *k*, a known characteristic of this metric, the rate of increase begins to stabilize near *k=8*, and the value at this point reflects a favourable balance between inter-cluster dispersion and intra-cluster compactness.

The convergence of all four metrics at *k = 8* provides strong and consistent evidence that this configuration optimally captures the underlying structure of the fused multimodal representation. Accordingly, *k = 8*, is applied as the final cluster count for the K-Means algorithm in the subsequent analysis stage.

3.3 Visualization of Multimodal Clustering Result

K-Means clustering is applied to the 50-dimensional UMAP embedding derived from the 2,908-dimensional multimodal feature matrix, using the optimal cluster count of *k=8* established in the preceding evaluation. Figure 6 presents the 2D UMAP visualisation of the K-Means clustering results (*k=8, n=1,000*) applied to the fused multimodal representation. Each point represents a single patient record, with colours denoting cluster membership. The overall clustering structure achieves a Silhouette Score of 0.7524, a DBI of 0.2630, and a CHI of 4,409.60, indicating well-separated, internally cohesive partitions.

Seven of the eight clusters exhibit visually compact and spatially well-separated groupings, reflecting strong intra-cluster similarity and clear inter-cluster boundaries. Cluster 1 (n=280) and Cluster 3 (n=262) are the largest groups and occupy distinct regions of the embedding space, suggesting they capture the two most prevalent clinical patterns in the dataset, likely corresponding to the normal finding subgroups identified in the LDA topic analysis. Cluster 2 (n=201) forms a tight, isolated region in the lower-right quadrant, indicative of a highly homogeneous patient subgroup. Clusters 4 (n=50), 5 (n=30), 7 (n=18), and 6 (n=9) are notably smaller, with Clusters 5, 6, and 7 appearing as compact, well-defined islands, consistent with rare or clinically specific conditions such as lung mass, cavitation, or dextroscoliosis. Cluster 0 (n=150) occupies a spatially isolated region in the upper-left quadrant, further confirming its clinical distinctiveness from the remaining groups.

EXPLORER
Journal of Computer Science and Information Technology
E-ISSN 2774-4647
Website: https://journal.fkpt.org/index.php/Explorer DOI: https://doi.org/10.47065/explorer.v6i2.2915

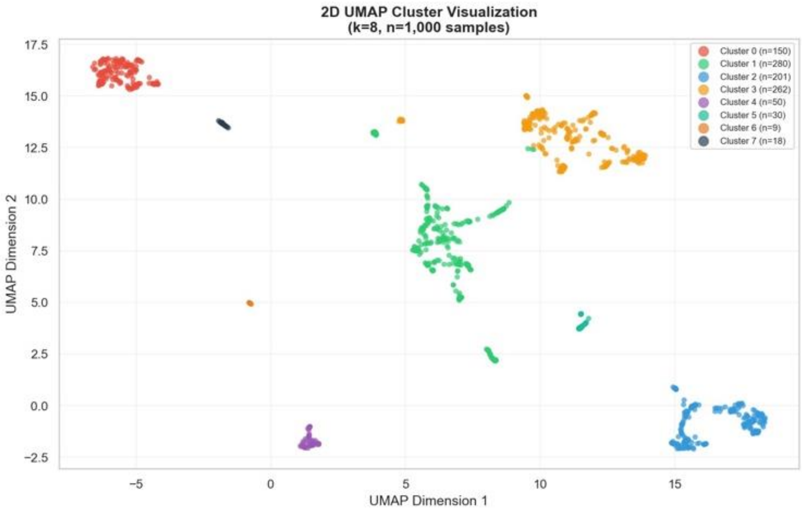

Figure 6. UMAP-Based Visualization of the Final Multimodal Medical Clustering Structure (K = 8)

The spatial separation observed across all eight clusters in the 2D projection is consistent with the high quantitative evaluation scores, collectively validating the effectiveness of the multimodal late fusion approach in producing a clinically meaningful, well-structured data partition.

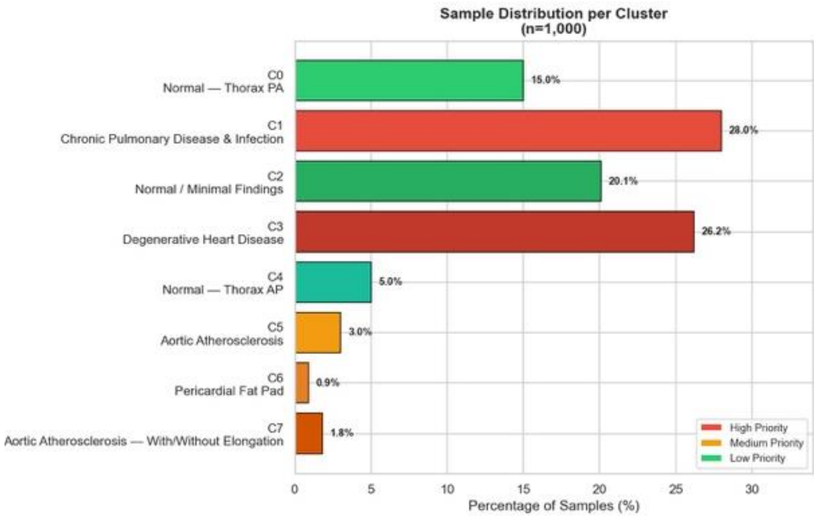

Figure 7. The distribution across eight K-Means clusters derived from multimodal chest X-Ray clustering ($k = 8$, $n = 1{,}000$), with clinical cluster label.

Figure 7 illustrates the distribution and clinical characterization of the eight clusters produced by the multimodal K-Means clustering framework. The two largest clusters: C1: Chronic Pulmonary Disease & Infection (28.0%) and C3: Degenerative Heart Disease (26.2%), collectively account for over half of the dataset and are classified as high priority, reflecting their clinical urgency and prevalence within the patient population. Normal findings are distributed across three separate clusters: C0: Normal - Thorax PA (15.0%), C2: Normal / Minimal Findings (20.1%), and C4: Normal - Thorax AP (5.0%), together representing 40.1% of all records and suggesting that the model successfully differentiated normal cases by radiological projection type and report linguistic pattern. The remaining three clusters, C5: Aortic Atherosclerosis (3.0%), C7: Aortic Atherosclerosis with/without Elongation (1.8%), and C6: Pericardial Fat Pad (0.9%), are smaller but clinically distinct, with C5 and C7 classified as medium priority and C6 as low priority. The clear separation of pathological from normal clusters, combined with the clinically coherent labelling of each group, demonstrates that the proposed multimodal fusion approach produces patient groupings that are not only statistically well-defined but also meaningfully interpretable in a clinical context.

4. CONCLUSION

This research successfully developed a multimodal medical clustering system for chest X-ray data by integrating visual features, clinical text features, and patient demographic information into a single multimodal representation. Visual features were extracted using ResNet152 with 2,048 dimensions, while text features were obtained using a Dual TF-IDF and LDA approach with a total of 857 dimensions. All features were then combined using a weighted late fusion strategy before being dimensionally reduced using UMAP with 50 components and the cosine metric. The reduced representations were then clustered using the K-Means algorithm to identify clinical patterns in the chest X-ray data without relying on label annotations. Based on evaluation results ranging from K=2 to K=15, the optimal configuration was obtained at K=8. The final evaluation demonstrated good clustering performance with a Silhouette Score of 0.7324, a Davies-Bouldin Index of 0.3416, and a Calinski-Harabasz Index of 3,789.10. The eight clusters formed successfully represented various clinical patterns that could be interpreted medically, including normal conditions, chronic pulmonary disease and infections, cardiovascular abnormalities, and special conditions such as pericardial fat pads. The results of this study demonstrate that multimodal integration can produce richer, more informative, and more clinically meaningful representations of medical data than unimodal approaches, thus potentially serving as the foundation for medical data exploration and clinical analysis systems on unlabeled radiology datasets. As a further development, future research could explore the use of more domain-specific pretrained medical models, such as CheXNet, BioViL-T, or MedCLIP, to improve the quality of visual feature representations. On the text side, contextual embedding-based approaches such as IndoBERT or the Indonesian-language BioBERT can capture the semantics of clinical reports more deeply than TF-IDF. Furthermore, clinical validation by a radiologist is crucial to ensure the resulting cluster interpretations align with actual medical conditions. Future research could also explore more adaptive clustering methods, such as HDBSCAN or spectral clustering, and expand the integration of multimodal data with laboratory information, electronic medical records, or genomic data to build a more comprehensive clinical analysis system.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the local partner hospital for providing institutional approval for the use of anonymized retrospective clinical data in this research. All chest X-ray images and clinical reports were de-identified prior to transfer to the research team and were securely stored using encryption mechanisms before analysis, in accordance with applicable data protection and patient confidentiality regulations. Furthermore, the authors also gratefully acknowledge the Institute for Research and Community Service (LPPM), Institut Teknologi Del, for providing financial support for the publication of this research.

REFERENSI

[1] P. G. Anderson, H. Tarder-Stoll, M. Alpaslan, N. Keathley, D. L. Levin, S. Venkatesh, E. Bartel, S. Sicular, S. Howell , R. V. Lindsey dan R. M. Jones, “Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays,” Scientific Reports , vol. 14, no. 1, p. 25151, 2024.

[2] R. Siddiqi dan S. Javaid, “Deep Learning for Pneumonia Detection in Chest X-ray Images: A Comprehensive Survey,” Journal of Imaging, vol. 10, no. 8, p. 176, 23 July 2024.

[3] M. Y. Ouis dan M. A. Akhloufi , “Deep learning for report generation on chest X-ray images,” Computerized Medical Imaging and Graphics, vol. 111, p. 102320, January 2024.

[4] B. D. Katzman, M. Alabousi, N. Islam, N. Zha dan M. N. Patlas, “Deep Learning for Pneumothorax Detection on Chest Radiograph: A Diagnostic Test Accuracy Systematic Review and Meta Analysis,” Canadian Association of Radiologists Journal, vol. 75, no. 3 , p. 525–533, 2024.

[5] B. Oltu, S. Güney, S. E. Yuksel dan B. Dengiz, “Automated classification of chest X-rays: a deep learning approach with attention mechanisms,” BMC Medical Imaging, vol. 25, no. 1, p. 71, 4 March 2025.

[6] T. Dhar, N. Dey, S. Borra dan R. S. Sherratt, “Challenges of deep learning in medical image analysis – improving explainability and trust,” IEEE Transactions on Technology and Society, vol. 4, no. 1, pp. 68-75, 4 January 2023.

[7] F. Gröger, L. Amruthalingam, S. Lionetti, A. A. Navarini, F. Ille dan M. Pouly, “A review and systematic guide to counteracting medical data scarcity for AI applications,” Computer Methods and Programs in Biomedicine Update, vol. 8, p. 100220, September 2025.

[8] Y. Liu, U. R. Acharya dan J. H. Tan, “Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation,” Computer Methods and Programs in Biomedicine, vol. 260, p. 108571, March 2025.

[9] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri dan R. M. Summers, “ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” dalam IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[10] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones dan Larson, “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 1, pp. 590-597, 17 July 2019.

[11] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark dan S. Horng , “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” Scientific Data, vol. 6, no. 1, p. 317, 12 December 2019.

[12] A. Bustos , A. Pertusa, J.-M. Salinas dan M. D. l. Iglesia-Vayá , “PadChest: A large chest x-ray image dataset with multi-label annotated reports,” Medical Image Analysis, vol. 66, p. 101797, December 2020.

[13] I. Renika, E. H. Pangaribowo dan A. J. Pitoyo, “Exposing Indonesia's Spatial Epidemiology of Pulmonary Tuberculosis: Geographic Pattern Analysis and Cluster Identification, 2000-2023,” Journal of Geosciences and Environmental Studies, vol. 3, no. 1, p. 22, 31 March 2026.

[14] A. Shelke, M. Inamdar, V. Shah, A. Tiwari, A. Hussain, T. Chafekar dan N. Mehendale, “Chest X-ray Classification Using Deep Learning for Automated COVID-19 Screenin,” SN Computer Science, vol. 2, no. 300, p. 2021, May 2021.

[15] N. Hayat, K. J. Geras dan F. E. Shamout, “MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images,” dalam In Machine Learning for Healthcare Conference, 2022.

[16] N. Aksoy, S. Sharoff, S. Baser, N. Ravikumar dan A. F. Frangi, “Beyond images: an integrative multi-modal approach to chest x-ray report generation,” Frontiers in Radiology, vol. 4, 15 February 2024.

[17] N. Kaur dan A. Mittal , “RadioBERT: A deep learning-based system for medical report generation from chest X-ray images using contextual embeddings,” Journal of Biomedical Informatics, vol. 135, p. 104220, November 2022.

[18] E. Khan, M. Z. U. Rehman, F. Ahmed, F. A. Alfouzan, N. M. Alzahrani dan J. Ahmad, “Chest X-ray Classification for the Detection of COVID-19 Using Deep Learning Techniques,” Sensors, vol. 22, no. 3, February 2022.

[19] F. D. Adhinata, N. G. Ramadhan, A. Amrulloh dan A. R. Bahtiar, “Comparison of Supervised Learning Methods for COVID-19 Classification on Chest X-Ray Image,” CommIT Journal, vol. 16, no. 2, p. 195–201, 2022.

[20] G. M. M. Alshmrani , Q. Ni , R. Jiang, H. Pervaiz dan N. M. Elshennawy, “A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images,” Alexandria Engineering Journal, vol. 64, pp. 923-935, 1 February 2023.

[21] M. Innat, M. F. Hossain, K. Mader dan A. Z. Kouzani, “A convolutional attention mapping deep neural network for classification and localization of cardiomegaly on chest X-rays,” Scientific Reports, vol. 13, p. 6247, 2023.

[22] D. Demner-Fushman , M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez , S. Antani, G. R. Thoma dan C. J. McDonald, “Preparing a collection of radiology examinations for distribution and retrieval,” Journal of the American Medical Informatics Association,, vol. 23, no. 2, p. 304–310, March 2016.

[23] J. Ramos , “Using TF-IDF to Determine Word Relevance in Document Queries,” dalam Proceedings of the first instructional conference on machine learning, 2003.

[24] D. M. Blei, A. Y. Ng dan M. I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol. 3 , pp. 993-1022, January 2003.

[25] T. Baltrušaitis, C. Ahuja dan L.-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 2018.

[26] L. McInnes, J. Healy dan J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” arXiv preprint arXiv:1802.03426, 9 February 2018.

[27] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, . I. W. H. Kwok, L. G. Ng, F. Ginhoux dan E. W. Newell, “Dimensionality reduction for visualizing single-cell data using UMAP,” Nature Biotechnology, vol. 37, no. 1, pp. 38-44, January 2019.