

Review of Computer Engineering Research

2026 Vol. 13, No. 4, pp. 49-64

ISSN(e): 2410-9142

ISSN(p): 2412-4281

DOI: 10.18488/76.v13i4.5201

© 2026 Conscientia Beam. All Rights Reserved.



GAST: Geometry-aware self-supervised transformer for efficient chest x-ray image retrieval

 Sannapureddy
Vijaya Lakshmi^{1*}

 K. Vijayalakshmi²

^{1,2}Department Computer Science and Engineering, Annamalai University,
India.

¹Email: svijayalakshmi@kitsguntur.ac.in

²Email: kv11898@annamalaiuniversity.ac.in



(+ Corresponding author)

ABSTRACT

Article History

Received: 19 March 2026

Revised: 14 May 2026

Accepted: 8 June 2026

Published: 29 September 2026

Keywords

Attention mechanism

Chest X-ray

Deep learning

FAISS

Image retrieval

Medical imaging

Self-supervised learning

Vision transformer.

Chest radiography is currently one of the most widely used diagnostic techniques in the medical field. However, rapid and accurate retrieval of clinically correct cases from large-sized image databases is still a major challenge. The existing Content-Based Medical Image Retrieval (CBMIR) methods are mainly based on handcrafted features or hybrid optimization approaches. They face problems such as being weak in generalization, having high computational cost, and not being able to correctly identify important anatomical structures. To overcome these challenges, this study introduces a new framework called Geometry-Aware Self-Supervised Transformer (GAST). The recommended approach is based on a Vision Transformer backbone that is first trained using self-supervised methods such as masked autoencoding and contrastive learning. In addition, a geometry-aware attention module is introduced to guide the model toward important anatomical regions, particularly the lungs and heart. The proposed procedure was tested on a chest X-ray dataset available to the public. As per the results, GAST showed better performance compared to conventional deep learning models and achieved 98.85% accuracy. Also, there has been a significant improvement in retrieval criteria like Recall@K and Mean Average Precision. Faster response time was also achieved by using FAISS-based indexing. Overall, the research demonstrated that both interpretability and efficiency of the model were improved by the combination of geometry-aware attention and sparse feature learning. The proposed system will provide a reliable and beneficial solution in medical image retrieval and is likely to help in real-life clinical decision-making in the future.

Contribution/Originality: This study contributes to the existing literature by introducing a geometry-aware transformer framework for medical image retrieval. The study integrates a dual-head architecture that jointly supports medical image classification and retrieval within a unified framework.

1. INTRODUCTION

Medical Imaging is a cornerstone of modern medical diagnosis. Chest Radiography is the most widely used test. It is increasingly performed worldwide due to its accessibility and high efficacy [1-3]. The rapidly growing digital imaging data has further increased the need for advanced computational methods to assist physicians. This will help reduce errors in diagnosis and improve workflow efficiency. In this context, Content-Based Medical Image Retrieval (CBMIR) has emerged as a key technology [4, 5]. It can query a database based on a patient image and retrieve visually and pathologically similar cases with their diagnoses and treatment outcomes. This increases the confidence of physicians and provides a strong basis for treatment planning and radiology education. The development of the

field of Content-Based Image Retrieval (CBIR) has been mainly driven by the advancement of Deep Learning. It overcomes the limitations of traditional handcrafted descriptors and provides semantic understanding [6, 7]. A survey by Dubey [6] highlighted the importance of deep learning in CBIR. However, the difficulties have been compounded by the complexity and heterogeneity of clinical images used in medical imaging [4, 5]. For chest X-rays, methods based on handcrafted feature descriptors such as GLCM and PCA [8] or query expansion methods [9, 10] were initially popular in CBMIR. Later, hybrid meta-heuristic optimization [11] or hybrid feature engineering [5] based on deep learning was used, but these methods still rely on handcrafted descriptors and have limited generalization on large-scale heterogeneous datasets [12].

Recently, the advent of CNNs [13-15] and Vision Transformers (ViTs) [16, 17] has made it possible to learn automated representations. Self-Supervised Learning (SSL)-based pretrained models on large natural image datasets [18] have yielded robust and semantic features. However, they have some limitations for direct application in the medical field: (i) computationally expensive non-end-to-end feature selection [19]. (ii) failure to model anatomical structural information (lung fields, heart silhouette) [13, 20]. (iii) inability to effectively address the class imbalance problem [21, 22]. These limitations indicate the need for an end-to-end, self-supervised, geometry-aware framework. To this end, we propose a new approach called Geometry-Aware Self-Supervised Transformer (GAST).

Despite the development of deep learning and Vision Transformer-based approaches in recent times, there still remain some challenges in medical image retrieval. Most CBMIR systems cannot effectively use anatomical information and, therefore, have limitations in correctly identifying vital components such as the lungs and heart. In addition, the use of non-differentiable feature selection methods increases computational cost and reduces efficiency.

To address these issues, a framework called Geometry-Aware Self-Supervised Transformer (GAST) was proposed in this research. This model combines attentional anatomical information and self-supervised learning into a single structure with differentiable feature selection. This improves interpretability, efficiency, and overall retrieval performance.

The main contributions of this work are summarized as follows:

1. Integration of Self-Supervised Learning and Feature Selection

This framework combines self-supervised pretraining with differentiable feature selection, providing an end-to-end approach to ensure that the model learns meaningful features and automatically selects the essential features needed.

2. Improved Retrieval Performance with Practical Efficiency

The proposed approach also retains real-time efficiency while showing improved performance in classification and image retrieval tasks. Therefore, it is suitable for clinical decision support systems.

3. End-to-End Self-Supervised Transformer Backbone

This research pretrained the Vision Transformer (ViTS/MAE) with masked image modelling and contrastive learning. It eliminates the need for handcrafted descriptors and learns semantic and frequency-aware representations.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed methodology. Section 4 discusses experimental results and analysis. Section 5 outlines the paper and future research directions.

2. RELATED WORK

Rapid evolution in deep learning techniques has revolutionized the field of medical imaging analysis, particularly in the domain of Content-Based Image Retrieval. A CBIR system finds an image that is similar to a query image from a large database [6, 7]. This is helpful for medical training, computer-aided diagnosis, and physician decision-making. The related work section reviews the relevant literature on medical image analysis, retrieval optimization methods, and their applications in X-ray image interpretation.

2.1. Deep Learning for Medical Image Analysis and Public Datasets

The progress of deep learning models is closely tied to the availability of large-scale and well-annotated datasets. Several publicly available datasets have enabled significant research progress, particularly in the field of chest X-ray imaging analysis. Popularly used benchmarks include the Chest X-ray Pneumonia dataset [1]. Datasets incorporating COVID-19 cases, such as the collection presented in Patel [2], have supported pandemic-related diagnostic studies. The NIH Chest X-ray dataset [3], which contains metadata for multiple pathologies, is one of the largest data resources. A survey conducted by Çallı et al. [23] comprehensively reviewed the use of deep learning techniques for chest X-ray image analysis. In this, the development from traditional approaches to modern deep neural network structures is clearly explained. Currently, deep learning with CNN constructs has become a standard in medical image analysis. They are widely used to detect pneumonia from chest X-ray images [13] and to detect COVID-19 from X-ray images [14].

2.2. Feature Extraction and Representation Learning

Extracting relevant features from images is a fundamental step in CBIR architectures. Early-stage CBIR systems relied on handcrafted features to describe the information in the image. Texture analysis by Gray-Level Co-occurrence Matrix (GLCM) [8] has been used primarily. An extensive review of the use of deep learning for feature extraction was provided in a survey conducted by Dubey [6]. The survey explains that deep learning approaches are replacing handcrafted descriptors by learning image representations directly from the data itself. Convolutional Neural Networks (CNN) models that are pre-trained on large datasets such as ImageNet are widely used as powerful feature extractors [12]. Recently, transformers, which gained prominence in NLP, have been adapted to computer vision tasks. Their application in medical imaging has demonstrated significant potential, which is highlighted by Shamshad et al. [16]. Models like Med-Former [17] exemplify the adaptation of transformers to medical image analysis and signify their ability to represent long-range interactions in visual data.

With recent standard architectures, researchers are investigating methods to enhance the quality and Interpretability of learned representations. Lee et al. [18] proposed a new approach, area-based contrastive pretraining, to improve retrieval performance in anatomical queries. Also, studies have suggested that binary encoding of high-dimensional features can provide an effective solution to large-scale image retrieval [24, 25].

2.3. Optimization and Hybrid Technologies in CBIR

In recent research, hybrid models combining deep learning with heuristic optimization algorithms have been examined to improve retrieval accuracy and efficiency. Shetty et al. [4] proposed a hybrid model that uses meta-heuristic optimization for feature selection, which led to improved retrieval results. Similarly, a study conducted by Issaoui et al. [11] also described the effect of similar hybrid approaches. The Archimedes optimization algorithm was incorporated into a deep learning framework for the healthcare sector CBIR. Other nature-inspired techniques, including Artificial Rabbits Optimization [19] and Harris Hawks Optimization [26, 27], have demonstrated notable improvement in retrieval outcomes. Query expansion is another strategy to improve retrieval performance that refines the query based on knowledge extracted from top results. Methods such as RbQE [9] and those relying on top-ranked images [10] have achieved promising outcomes between low-level features and higher-level clinical concepts.

2.4. Application to Chest X-ray and Multi-Label Classification

Given their widespread availability and diagnostic utility, Chest x-ray has attracted significant attention in medical CBIR research. Agrawal et al. [13] conducted a study and proposed a CBIS system for lung diseases using a deep convolutional neural network. A key challenge in this domain is that chest x-ray frequently exhibits multiple co-occurring comorbidities requiring systems that extend beyond single disease classification. This challenge is

addressed by Chakravarty et al. [28] and Chen et al. [29] using a graph-based neural network combined with deep hierarchical architectures for multi-label classification of chest X-rays, ensuring retrievals align with a patient's full clinical profile.

2.5. Comparison with Existing CBMIR Methods

Although existing research has made significant strides, several unsolved challenges persist. Many systems concentrated on performance optimization but failed to address privacy issues, as highlighted by Sadik et al. [30]. Hybrid models indicate potential, yet there is a need for more advanced architectures that integrate multiple deep learning paradigms (CNNs, LSTMs, Transformers) to handle both spatial and sequential context modelling, as demonstrated in CNN – LSTM models for COVID-19 [14] and croup cough detection [15].

Table 1. Comparative analysis of existing CBMIR methods and proposed GAST.

Ref.	Method / Study	Position Relative to GAST	Limitation
Issaoui, et al. [11]	Archimedes optimization with deep learning	GAST avoids separate optimization and learns features directly during training	The optimization stage adds overhead and limits scalability
Agrawal, et al. [13]	CNN-based CBMIR	GAST uses transformer-based global modelling with anatomical guidance	Limited ability to capture long-range dependencies and a lack of anatomical awareness
Mahbod, et al. [12]	Pretrained CNN feature extractors	GAST introduces geometry-aware attention for clinically relevant focus	Generic features do not incorporate anatomical priors
Shamshad, et al. [16]	Transformer-based medical imaging approaches	GAST integrates anatomical priors into the attention mechanism	Standard transformers do not explicitly model anatomical structures
Lee, et al. [18]	Region-based contrastive pretraining	GAST integrates self-supervised learning with a unified retrieval framework	Focus is limited to region-based learning, not full retrieval optimization
Proposed	GAST	Provides an end-to-end, geometry-aware, and efficient CBMIR framework	-

Table 1 shows the Comparative analysis of existing CBMIR Methods and the proposed GAST Framework. Extending prior research studies, our research introduces a novel framework for content-based chest x-ray retrieval, integrating a hybrid deep learning technique with a nature-inspired optimization strategy. The proposed framework focuses on multi-label chest X-ray retrieval while embedding efficient hashing and privacy-preserving measures to mitigate the gaps in accuracy and real-world applicability.

3. METHODOLOGY

The proposed GAST framework is designed to improve the retrieval of chest X-ray images. It combines self-supervised learning, geometry-aware attention, and differentiable feature selection into a single unified architecture. The entire process includes image preprocessing, transformer-based feature extraction, geometry-aware attention enhancement, feature selection, and dual-head classification and retrieval modules.

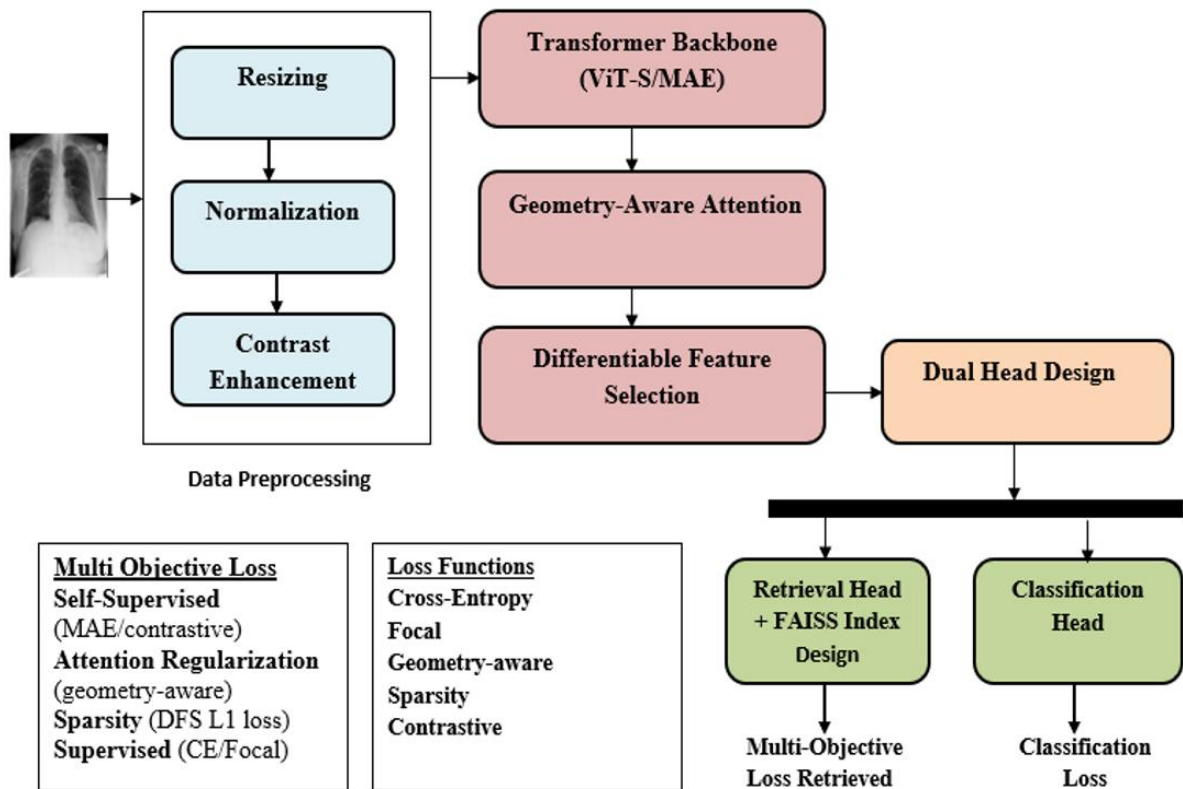


Figure 1. Proposed framework of the geometry-Aware transformer with dual-head design for classification and retrieval.

To address the challenges of efficient and accurate chest X-ray retrieval and classification, we propose the novel framework termed Geometry-Aware Self-Supervised Transformer (GAST). The overall workflow of the architecture is shown in Figure 1, combines a transformer backbone (ViT-S/MAE), geometry-aware attention, and differentiable feature selection into a dual-head architecture. This pipeline starts with standard preprocessing actions like resizing, normalization, and contrast enhancement to retain uniformity in the image quality. The processed images are subsequently fed into the transformer backbone for feature extraction, which are then enhanced by a geometry-aware attention mechanism to capture spatial dependencies within anatomical structures. Differentiable feature selection is applied to reduce redundancy and promote sparsity ahead of the dual-head design.

One head focuses on disease category classification, whereas the other incorporates a FAISS-based retrieval index for efficient similarity search. The model is optimized by integrating supervised, self-supervised, and regularization-based multi-task losses, ensuring robust representation learning for diagnostic classification as well as clinically meaningful case retrieval. Table 2 shows the notations used in the GAST Framework.

Table 2. Notations used in the proposed GAST framework.

Symbol	Meaning
I	Input image
I'	Preprocessed image
H, W, C	Height, Width, Channels
p	Patch size
d	Feature dimension
N	Number of patches

3.1. Image Preprocessing

Medical images undergo a preprocessing pipeline to ensure uniform resolution, stability, and improved visibility of diagnostic details before feature extraction. Let the raw medical image be denoted as I . A preprocessing function $P(\cdot)$ is applied to obtain a normalized and enhanced input I' .

$$I' = P(I), I' \in \mathbb{R}^{(H \times W \times C)} \quad (1)$$

Where H, W, and C represent the standardized height, width and number of channels, respectively. The pre-processing phase consists of three main steps:

- A. Resizing: To provide uniformity to the transformer backbone, all images are scaled to a fixed resolution (for example, 224×224).

$$I_{resized} = \text{Resize}(I, H, W) \quad (2)$$

- B. Normalization: Mapping or standardizing pixel intensity values to the same general range is done in this way:

$$I'_{norm} = (I - \mu) / \sigma \quad (3)$$

Where μ and σ are the dataset mean and standard deviation. These help in stable training convergence.

- C. Contrast Enhancement: local histogram equalization (e.g., CLAHE) is applied to highlight subtle anatomical regions such as lung boundaries and lesions.

$$I' = \text{CLAHE}(I_{norm}) \quad (4)$$

The resulting I' is then fed into the transformer backbone for feature extraction, ensuring consistent quality and enhanced visibility of clinically relevant regions. Preprocessing is performed to ensure consistent input quality and improve model performance. Resizing is done so that all the images have the same size, which is compatible with the transformer backbone. Normalization standardizes the pixel intensity values and keeps the training process consistent. Also, contrast enhancement using CLAHE helps in better feature extraction, making important anatomical components such as lungs and lesions clearly visible.

3.2. Transformer Backbone

Once preprocessing is complete, I' is given as input to the transformer backbone for feature extraction. Now, image $I' \in \mathbb{R}^{(H \times W \times C)}$ is split into a sequence of fixed-size non-overlapping patches. Each patch of a $p \times p$ image is converted to a flattened and linearly embedded d -dimensional feature vector.

$$x_i = W_p \cdot \text{Flatten}(I'_i) + b_p, \quad i = 1, \dots, N \quad (5)$$

Where $W_p \in \mathbb{R}^{((p^2 \cdot C) \times d)}$ is the learnable projection matrix, b_p is the bias term, and $N = (H \cdot W) / p^2$ is the total number of patches. The patch embeddings are then enriched with positional encodings to retain spatial structure.

$$z_0 = [x_1 + e_1, x_2 + e_2, \dots, x_n + e_n] \quad (6)$$

Where x_i are the patch embeddings, and e_i are the corresponding positional encodings. These embeddings are processed by multi-head self-attention (MHSA) layers; they are also able to understand and model distal relationships between different patches in the image.

$$\text{Attention}(Q, K, V) = \text{Softmax}((QK^T) / \sqrt{d_k}) V \quad (7)$$

With

$$Q = zW_Q, K = zW_K, V = zW_V \quad (8)$$

Where W_Q, W_K, W_V are learnable projection matrices.

The output from successive MHSA and feed-forward layers forms a hierarchical feature representation.

$$z^l = \text{TransformerBlock}(z^{l-1}), \quad l = 1, 2, \dots, L \quad (9)$$

Where L is the number of transformer layers.

Because traditional CNN backbones rely on local receptive fields, they limit their ability to capture comprehensive (Global) information in the image. In contrast, the MHSA method in the proposed Geometry-Aware Transformer (GAT) backbone effectively detects the interactions of local and global space boundaries. Moreover, its geometry-aware attention enables it to be more sensitive to architecture-related cues, such as anatomical boundaries. Existing CNN-based pipelines require deep convolution layers or handcrafted features. But our backbone learns hierarchical representations directly from the enhanced images. This approach provides better generalization over different datasets.

3.3. Differentiable Feature Selection (DFS)

The embeddings produced by transformer backbones are high-dimensional, where redundant or weakly salient features limit efficiency and clarity in representation. The transformer backbone produces a high-dimensional embedding matrix.

$$Z = [z_1, z_2, \dots, z_n], Z \in \mathbb{R}^{(N \times d)} \quad (10)$$

Where N is the number of patches and d is the embedding dimension.

To overcome the limitation mentioned above, it needs a mechanism that selectively highlights informative features while suppressing redundancies, facilitating compact and robust representations. For this, a differentiable feature selection (DFS) layer is applied. Each feature dimension is assigned a learnable gating weight w_j .

$$\hat{z}_i = [w_1 z_{i1}, w_2 z_{i2}, \dots, w_d z_{id}], w_j \in [0, 1] \quad (11)$$

Where the weights w_j are optimized during training. To restrict sparsity, the L1-regularization term is introduced.

$$L_{DFS} = \lambda \sum_j |w_j| \quad (12)$$

With λ controlling the degree of sparsity. The output is a compact embedding matrix.

$$\hat{Z} = [\hat{z}_1, \hat{z}_2, \dots, \hat{z}_n], \hat{Z} \in \mathbb{R}^{(N \times d')} \text{ with } d' < d \quad (13)$$

Conventional pipelines typically use manual feature selection or dimensionality reduction (e.g., PCA, t-SNE). They are suitable for specific tasks and are not differentiable. In contrast, the proposed DFS is a fully end-to-end trainable approach. This allows the network to function on its own and retain the features that are most useful for classification and retrieval. This approach improves computational efficiency, reduces unnecessary features, and provides more robust performance compared to current CNN + human-directed feature pipelines.

3.4. Geometry-Aware Attention Module (GAA)

While the transformer backbone efficiently captures local and global dependencies by self-attention, it remains agnostic to explicit structural cues such as shape, orientation, and anatomical boundaries. In medical imaging, these geometric attributes are critical for distinguishing subtle variations across regions of interest. To bridge this gap, we introduce the Geometry-Aware Attention (GAA) module, which integrates spatial priors into the attention mechanism, thereby enhancing the model's sensitivity to structural context. Let the feature embeddings from the transformer backbone be.

$$Z = [z_1, z_2, \dots, z_n], Z \in \mathbb{R}^{(N \times d)} \quad (14)$$

Where N denotes the number of image patches, and d is the feature dimension. Standard self-attention computes contextual weights between patches using query-key similarity.

$$\alpha_{ij} = \exp((q_i^T k_j) / \sqrt{d}) / \sum_j' \exp((q_i^T k_j') / \sqrt{d}) \quad (15)$$

with $q_i = W_q z_i$ and $k_j = W_k z_j$.

However, in medical images, purely content-based similarity may overlook anatomical priors such as lung boundaries or heart regions. GAA integrates the geometric attitude term β_{ij} , obtained from the relative spatial positions.

$$\tilde{\alpha}_{ij} = \exp((q_i^T k_j) / \sqrt{d} + \beta_{ij}) / \sum_j' \exp((q_i^T k_j') / \sqrt{d} + \beta_{ij}') \quad (16)$$

Where β_{ij} encodes spatial relationships (e.g., patch distance, anatomical alignment). The attended representation is then.

$$z_i' = \sum_j \tilde{\alpha}_{ij} v_j, v_j = W_v z_j \quad (17)$$

Thus, attention weights are both content-aware and geometry-aware, ensuring focus on clinically relevant regions. Conventional attention in ViT or CNN relies solely on appearance similarity, which can cause spurious focus (e.g., noise, artifacts). Existing methods may use post-hoc region masks, but they are non-trainable and dataset-dependent. The proposed GAA is differentiable and trainable, directly integrating anatomical priors into the attention

mechanism. This will help make the model easier to understand, reduce incorrect activations, and increase trust in clinical use.

3.5. Dual-Head Architecture

The compact embedding matrix produced by the differentiable feature selection (DFS) serves as a distilled representation of the input data. To maximize the utility of the compact embeddings, we propose a Dual-Head Architecture. This architecture has two specialized heads: one is for classification, and the other is for retrieval and similarity search. The output of the Differentiable Feature Selection (DFS) layer is the compact embedding matrix.

$$\hat{Z} = [\hat{z}_1, \hat{z}_2, \dots, \hat{z}_n], \quad \hat{Z} \in \mathbb{R}^{(N \times d')} \quad (18)$$

This embedding is fed into a dual-head design consisting of.

1. Classification Head

The classification branch applies a fully connected layer followed by a softmax activation to predict class probabilities.

$$\hat{y}_i = \text{softmax}(W_c \hat{Z}_i + b_c) \quad (19)$$

Where $W_c \in \mathbb{R}^{(C \times d')}$ is the learnable weight matrix, b_c is the bias, and C is the number of classes. The classification loss is defined using focal loss to handle class imbalance.

$$L_{cls} = -\alpha_t (1 - \hat{y}_t)^\gamma \log(\hat{y}_t) \quad (20)$$

Where α_t balances class frequencies and γ emphasizes hard-to-classify samples.

2. Retrieval Head

The retrieval head generates embeddings for similarity-based image retrieval using FAISS. Given a query embedding $q \in \mathbb{R}^{(d')}$ and a stored database embedding $x \in \mathbb{R}^{(d')}$, similarity is measured by.

$$S(q, x) = (q \cdot x) / (||q|| ||x||) \quad (\text{cosine similarity}) \quad (21)$$

$$D(q, x) = ||q - x||^2 \quad (L2 \text{ distance}) \quad (22)$$

The retrieval loss is computed using contrastive learning.

$$L_{retrieval} = -\log [\exp(S(q, x^+)/\tau) / \Sigma_x - \exp(S(q, x^-)/\tau)] \quad (23)$$

where x^+ is a positive sample, x^- is a negative sample, and τ is a temperature parameter. This ensures embeddings of similar images are close in latent space. Existing methods rely on handcrafted features and predefined similarity measures for both classification and retrieval. The proposed dual-head architecture is end-to-end trainable, allowing the network to learn both retrieval and classification simultaneously from the same feature embeddings. FAISS-based embedding retrieval enables queries to be retrieved in real-time (< 100 ms). At the same time, the classification head utilizes spars and geometry-aware features, providing higher accuracy and better reliability compared to conventional pipelines.

3.6. Comparison with Existing Methods

Compared to traditional CNN-based and simple transformer approaches, the proposed GAST framework identifies anatomical structures more efficiently using geometry-aware attention. In addition, it effectively manages end-to-end learning by using differentiable feature selection in place of traditional optimization methods. Moreover, the dual-head design further improves the overall performance by optimizing the classification and retrieval tasks together in a single structure.

4. RESULTS AND DISCUSSIONS

The proposed GAST framework provides two major improvements over existing deep learning methods, such as RNN, CNN, Bi-LSTM, and DenseNet, for X-ray image analysis. These improvements were systematically examined using the COVID-19 Chest X-ray dataset available to the public.



Figure 2. Sample visualizations from the COVID-19 Chest X-ray image dataset.

Table 3. Sample counts for training and testing in the COVID-19 Chest X-ray dataset.

Dataset: COVID-19 Chest X-ray			
Class	Training Samples	Testing Samples	Total
Normal	1266	317	1583
Pneumonia	3418	855	4273
COVID-19	460	116	576
Overall	5144	1288	6432

4.1. Description of the Dataset

We used the COVID-19 chest X-ray image dataset for experimental validation. This dataset is organized into two main folders: train and test. Each folder has three sub-folders, which are related to the following clinical categories: COVID-19, Pneumonia, and Normal. This dataset contains 6,432 chest X-ray images in total. Amongst these, 80% are for training. 20% is reserved for testing. The distribution of samples in the three categories is shown in Table 3. This balanced dataset structure enables the model to be evaluated under actual clinical conditions. It covers both binary and multi-class disease classifications. Disease variation in the dataset to the GAST framework, Geometry-based attention patterns, Intuitive features' ability to learn effectively. In addition, the retrieval head provides case reference and clinical decision support by performing content-based image retrieval. Representative examples of sample images in the dataset are shown in Figure 2. Table 4 shows Hyperparameter details used in the proposed model.

Table 4. Hyperparameter details used in the proposed model.

Hyperparameter	Value
Model	Vision Transformer (ViT-Base- Patch16-224-in21k) with GAA + DFS
Learning rate	2e-5
Epochs	20
Batch size	16
Loss function	Cross-Entropy Loss
Optimizer	AdamW
Training data	80%
Testing data	20%

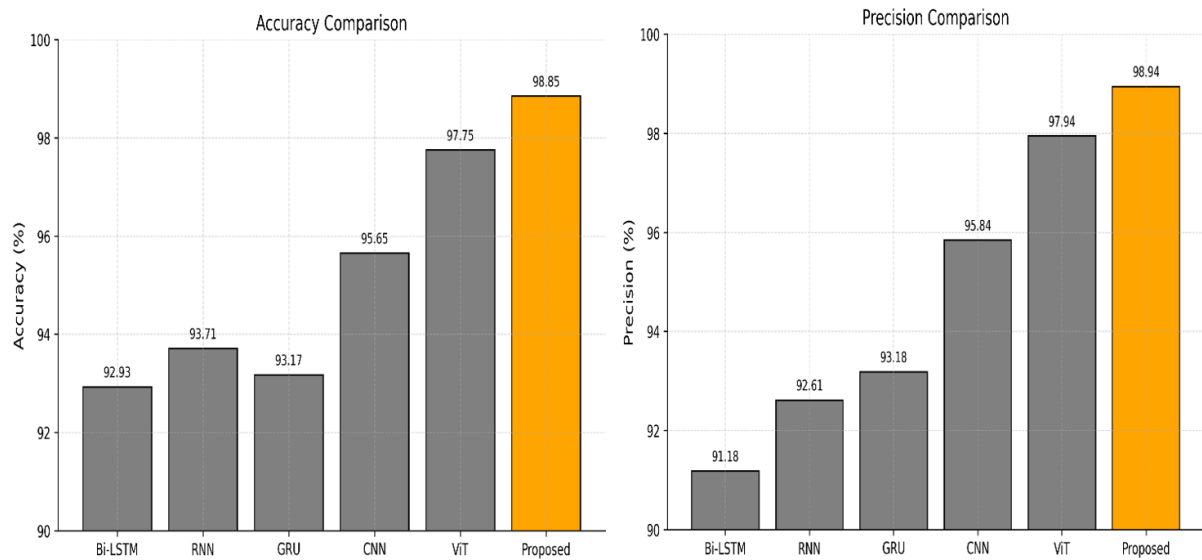
4.2. Classification Head

The classification head effectively increases the confirmation accuracy by utilizing Geometry-Aware Attention and Dynamic Feature Sparsity. Compared to traditional models, this framework shows significantly better performance in terms of Accuracy, Precision, Recall, and F1-score, indicating its ability to accurately differentiate disease categories. As indicated in Table 5, the proposed approach provides better results compared to older methods.

Table 5. Classification performance (%) of the proposed framework compared with existing methods on the chest X-ray image dataset.

Model	Accuracy	Precision	Recall	F1-Score
Bi-LSTM	92.93	91.18	91.07	91.13
RNN	93.71	92.61	92.37	92.47
GRU	93.17	93.18	90.63	91.85
CNN	95.65	95.84	95.55	95.68
ViT	97.75	97.94	97.69	97.82
Proposed	98.85	98.94	98.69	98.91

As shown in Figure 3, the proposed GAST framework demonstrates better performance than traditional models such as Bi-LSTM, RNN, GRU, CNN, and ViT, especially in terms of accuracy and precision. This improvement is even more apparent compared to the recurrent-based models (Bi-LSTM, RNN, GRU). This ensures that the model can more effectively capture the subtle confirmation signals in the chest X-ray by using approaches such as Geometry-Aware Attention and Sparse Feature Learning. In addition, the increased precision score shows that this framework reduces false positive conclusions, which is the most important thing. Mistakenly identifying a patient who does not have COVID-19 in clinical conditions can lead to unnecessary treatments and panic attacks.

**Figure 3.** Assessment of the proposed model's performance through accuracy and precision metrics.

The recall and F1-score in Figure 4 also reinforce this advantage. The high recall achieved by the GAST framework shows the ability to correctly identify COVID-19 and pneumonia cases, thereby reducing the number of false negatives. This is the most crucial of medical diagnoses - because leaving a patient undiagnosed can lead to serious consequences. An improved F1-score indicates a balanced benefit to both precision and recall. At the same time, it indicates that the model performs strongly despite the category mismatch in the dataset. Overall, Figure 3 and Figure 4 clearly demonstrate that the proposed GAST framework provides not only linear accuracy but also a comprehensive improvement in all key performance parameters. So, it can be said that it is a reliably usable approach in the resolution of clinical decisions.

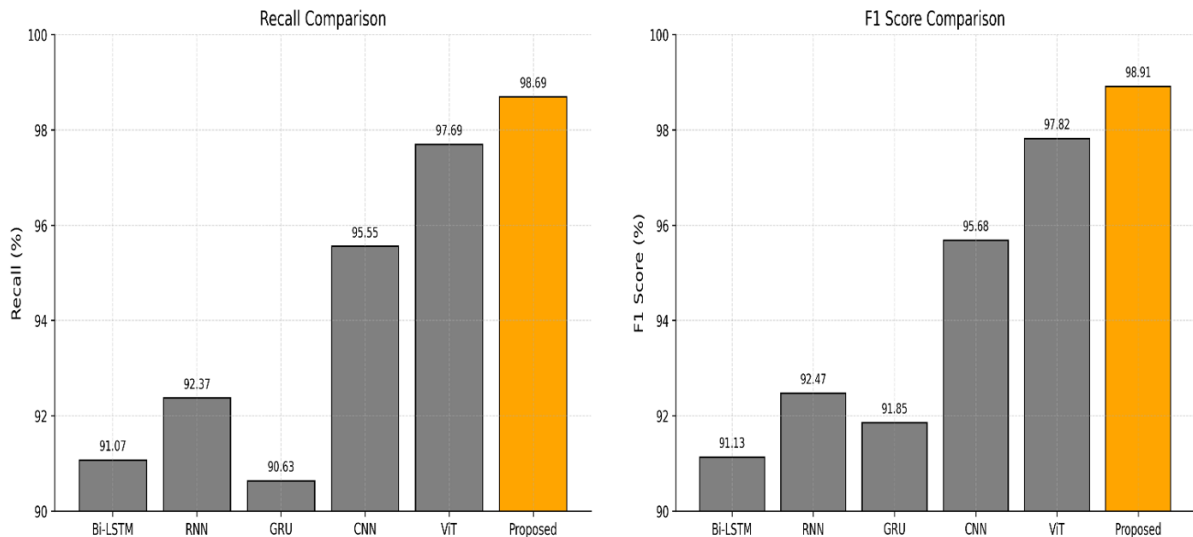


Figure 4. Visualization of model performance with respect to Recall and F1-score.

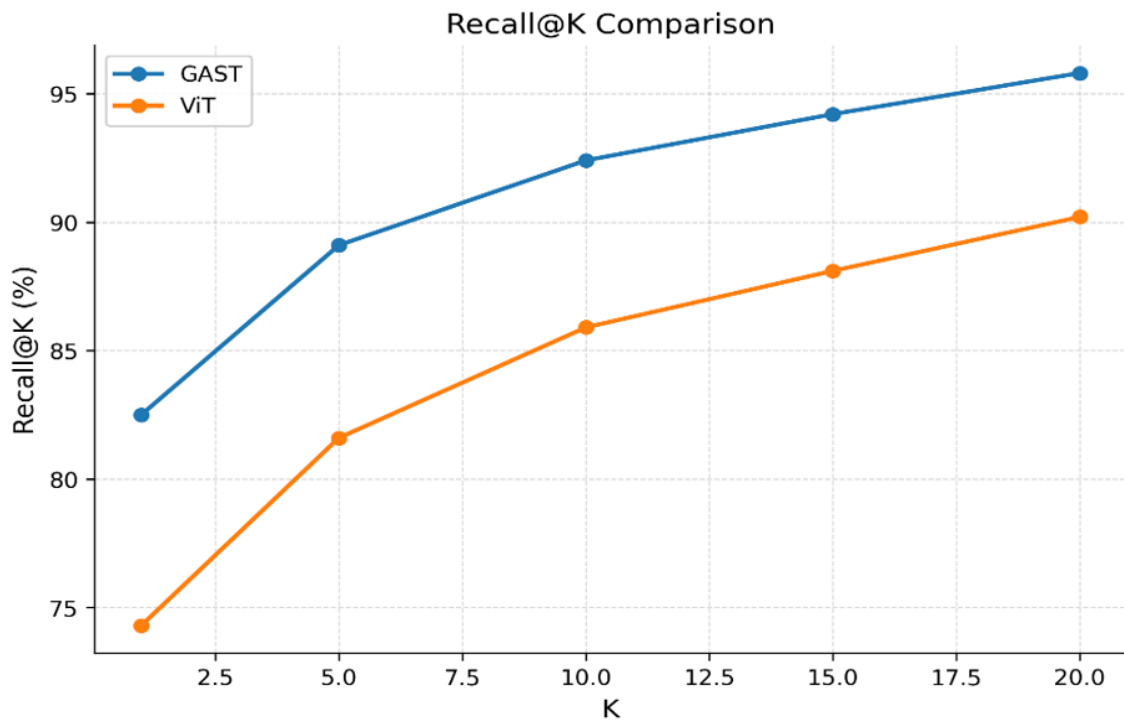


Figure 5. Recall@K comparison between the proposed GAST framework and the ViT baseline.

4.3. Retrieval Head

Since the retrieval head in the proposed GAST framework is integrated with the FAISS indexing, the content-based medical image shows significant improvement in the retrieval performance. As shown in Figure 5, the proposed model achieved a Recall@K score consistently higher than the ViT baseline at all K values. This indicates that the correct confirmation category is more often found in top-K similar images. This allows clinicians to access relevant patient records in a shorter time. The progress achieved, especially at low K values, demonstrates that the GAST model provides clinically meaningful cases.

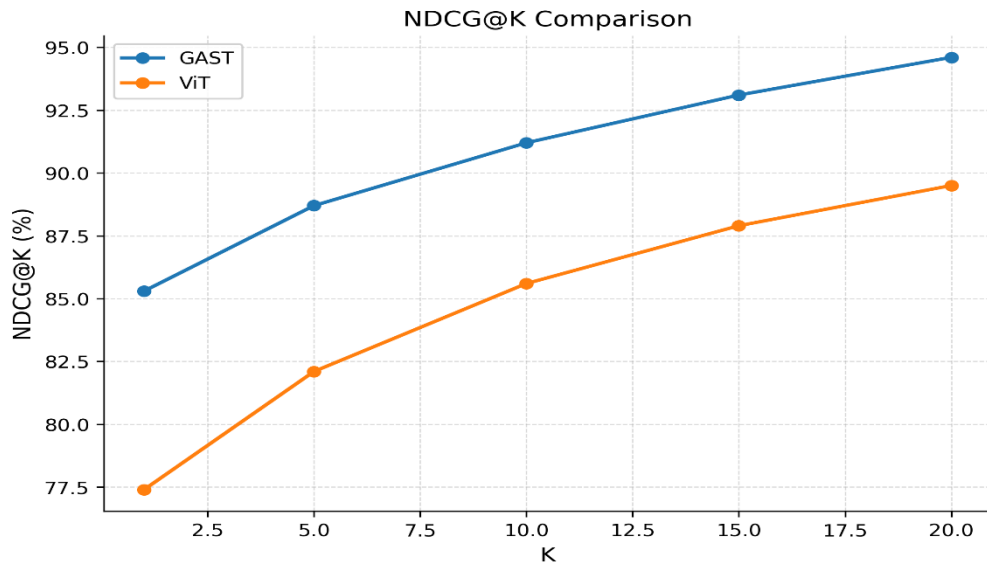


Figure 6. NDCG@K comparison of the proposed GAST framework with ViT and CNN baselines.

The NDCG@K comparison between the models in Figure 6 shows the ability of GAST not only to identify pure cases but also to rank them in the correct order of priority. The proposed approach achieved the best possible ranking relative scores, surpassing the ViT and CNN baselines. This suggests that geometry-aware attention and sparse feature learning play a key role in placing clinically relevant images in order of priority - which is highly beneficial for case-based analysis in the diagnostic process. Finally, the mAP comparison is shown at cut-off values 5 and 10 in Figure 7. The GAST framework achieves the best mAP values: $\text{mAP}@5 = 91.2\%$, $\text{mAP}@10 = 93.0\%$. These are clearly better than the Bi-LSTM, RNN, GRU, CNN, and ViT models. These results ensure that the retrieval head maintains precision and stability even on large sets.

In Figures 5 to 7, retrieval is not only the function of head classification, but it has also been confirmed to provide robust, accurate, and clinically valuable case retrieval. This makes it an extremely useful tool in real-world medical diagnosis environments.

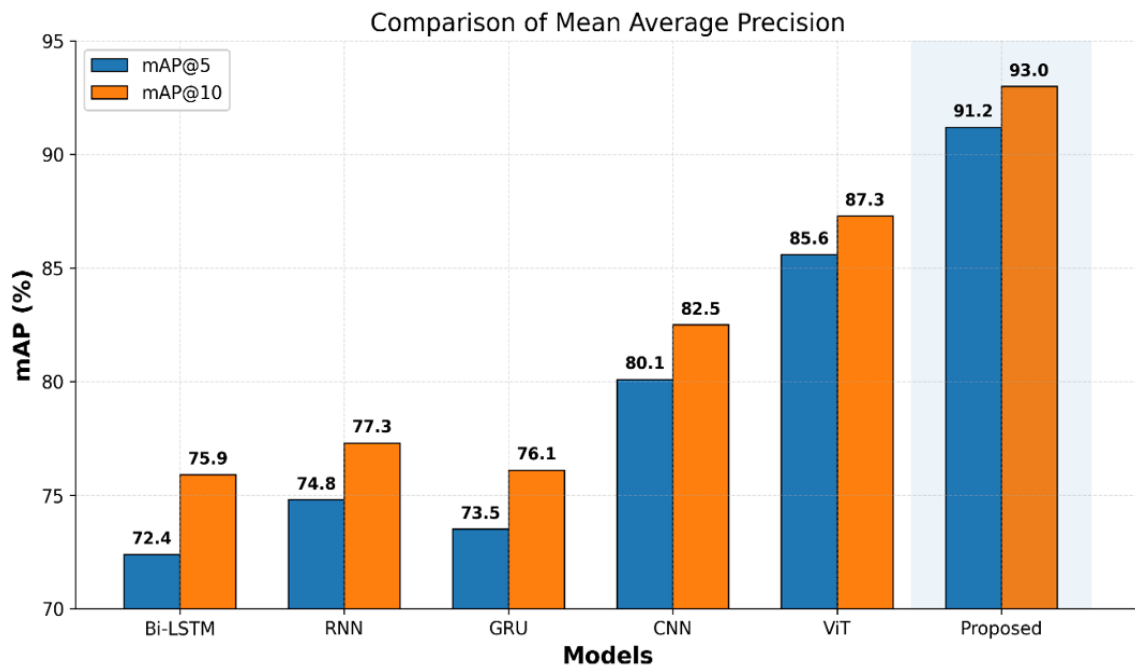


Figure 7. Mean average precision (mAP) comparison of GAST and baseline models.

4.4. Ablation Study

Table 6 shows the ablation analysis of the proposed GAST framework. As per the results, the performance decreases when the geometry-aware attention module or the differentiable feature selection module is removed. The full GAST model confirms the importance of these two components by achieving the highest accuracy.

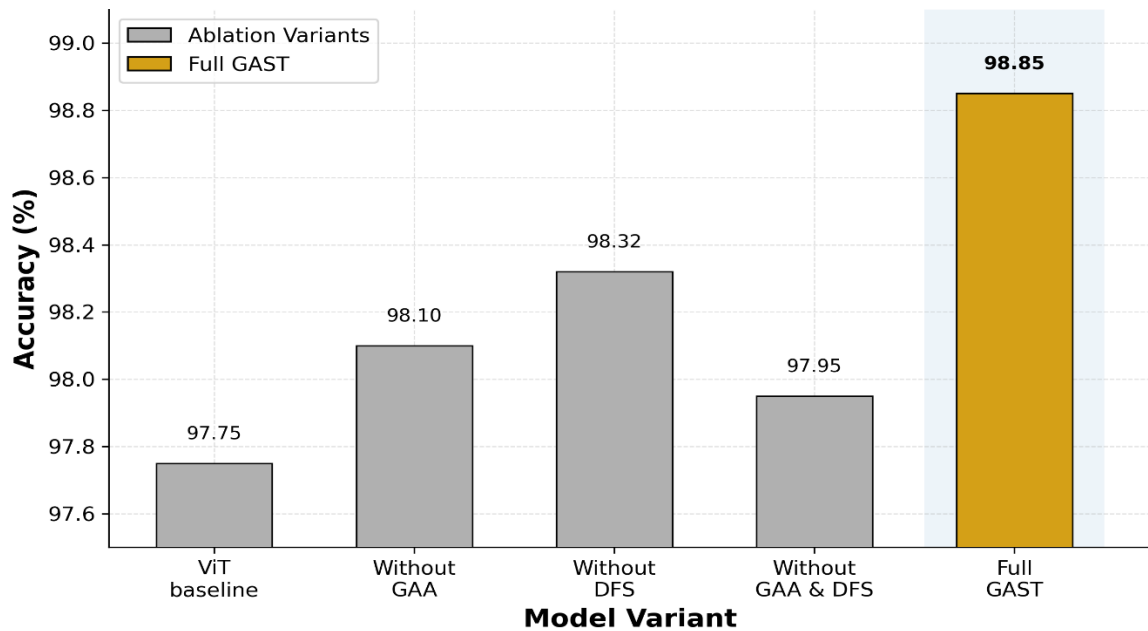


Figure 8. Performance comparison of GAST ablation variants.

The ablation analysis in Figure 8 illustrates the effect on classification accuracy when GAA and DFS are removed. By the appearance of a significant performance decrease when these two components are removed simultaneously, their combined importance in the proposed framework becomes clear.

Table 6. Ablation analysis of the proposed GAST framework.

Model variant	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ViT baseline	97.75	97.94	97.69	97.82
GAST without GAA	98.10	98.21	98.03	98.11
GAST without DFS	98.32	98.40	98.25	98.31
GAST without GAA & DFS	97.95	98.05	97.88	97.96
Full GAST	98.85	98.94	98.69	98.91

4.5. Sensitivity Analysis

Sensitivity analysis was conducted to evaluate the effect of λ parameter on the differentiable feature selection module. Higher values of λ tend to cause more sparsity and reduce unnecessary features, but placing too high a value can also remove useful information. The moderate λ value provided the best balance between compactness and performance. In addition, the geometry-aware attention module has further improved performance by increasing focus on areas of anatomical importance.

4.6. Error Analysis and Discussion

Many errors have been seen between COVID-19 and pneumonia images, especially in cases of low contrast, because they look similar. These results are consistent with the latest studies on anatomy-aware retrieval and powerful pretrained feature extractors. However, many existing methods still rely on generic representations or hybrid optimization, making the use of explicit geometry-aware modeling less visible.

5. CONCLUSION AND FUTURE WORK

The proposed Geometry-Aware Self-Supervised Transformer (GAST) framework demonstrated powerful performance in chest X-ray analysis by combining geometry-aware attention and differentiable feature selection into a single model. The results showed superior classification performance in parameters such as Accuracy, Precision, Recall, and F1-score, as well as improved retrieval performance in Recall@K, NDCG@K, and Mean Average Precision. These results show how effective medical image analysis can be by incorporating anatomical information and selecting relevant features. Being able to do simultaneous classification and case-based retrieval through a dual-head design will help doctors identify similar cases. It will increase trust in diagnosis and promote the use of intelligent medical imaging systems in health services.

5.1. Limitations

This model, being a transformer-based structure, requires more computational resources. Also, it was tested only on a single-modality chest X-ray dataset. So there is a possibility that the generalization ability on different types of images may be limited.

5.2. The Future Directions of Research

Future research will focus on extending this framework to multi-modal medical datasets. Also, lightweight models are developed for real-time applications that operate in resource-constrained environments. Further, research also continues on improving the efficiency and robustness of the model for a wide range of clinical applications.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable.

Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.

Competing Interests: The authors declare that they have no competing interests.

Authors' Contributions: Both authors contributed equally to the conception and design of the study. Both authors have read and agreed to the published version of the manuscript.

REFERENCES

- [1] P. Mooney, "Chest X-ray (Pneumonia), Kaggle," Retrieved: <https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data>. 2018.
- [2] P. Patel, "Chest X-ray (COVID-19 & Pneumonia), Kaggle," Retrieved: <https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia>. 2020.
- [3] NIH, "Chest X-ray dataset, Kaggle," Retrieved: <https://www.kaggle.com/datasets/nih-chest-xrays/data>. 2017.
- [4] R. Shetty, V. S. Bhat, and J. Pujari, "Content-based medical image retrieval using deep learning-based features and hybrid meta-heuristic optimization," *Biomedical Signal Processing and Control*, vol. 92, p. 106069, 2024. <https://doi.org/10.1016/j.bspc.2024.106069>
- [5] E. Saranya and M. Chinnadurai, "Medical application driven content based medical image retrieval system for enhanced analysis of X-ray images," *Scientific Reports*, vol. 15, no. 1, p. 29115, 2025. <https://doi.org/10.1038/s41598-025-14282-8>
- [6] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687-2704, 2022. <https://doi.org/10.1109/TCSVT.2021.3080920>
- [7] D. Srivastava, S. S. Singh, B. Rajitha, M. Verma, M. Kaur, and H.-N. Lee, "Content-based image retrieval: A survey on local and global features selection, extraction, representation, and evaluation parameters," *IEEE Access*, vol. 11, pp. 95410-95431, 2023. <https://doi.org/10.1109/ACCESS.2023.3308911>

- [8] Priyanka and D. Kumar, "Feature extraction and selection of kidney ultrasound images using GLCM and PCA," *Procedia Computer Science*, vol. 167, pp. 1722-1731, 2020. <https://doi.org/10.1016/j.procs.2020.03.382>
- [9] M. Rashad, I. Afifi, and M. Abdelfatah, "RbQE: An efficient method for content-based medical image retrieval based on query expansion," *Journal of Digital Imaging*, vol. 36, no. 3, pp. 1248-1261, 2023. <https://doi.org/10.1007/s10278-022-00769-7>
- [10] A. Ahmed and S. J. Malebary, "Query expansion based on top-ranked images for content-based medical image retrieval," *IEEE Access*, vol. 8, pp. 194541-194550, 2020. <https://doi.org/10.1109/ACCESS.2020.3033504>
- [11] I. Issaoui, M. A. Alohal, W. Mtouaa, F. A. Alotaibi, A. Mahmud, and M. Assiri, "Archimedes optimization algorithm with deep learning assisted content-based image retrieval in healthcare sector," *IEEE Access*, vol. 12, pp. 29768-29777, 2024. <https://doi.org/10.1109/ACCESS.2024.3367430>
- [12] A. Mahbod, N. Saeidi, S. Hatamikia, and R. Woitek, "Evaluating pre-trained convolutional neural networks and foundation models as feature extractors for content-based medical image retrieval," *Engineering Applications of Artificial Intelligence*, vol. 150, p. 110571, 2025. <https://doi.org/10.1016/j.engappai.2025.110571>
- [13] S. Agrawal, A. Chowdhary, S. Agarwala, V. Mayya, and S. Kamath S, "Content-based medical image retrieval system for lung diseases using deep CNNs," *International Journal of Information Technology*, vol. 14, no. 7, pp. 3619-3627, 2022. <https://doi.org/10.1007/s41870-022-01007-7>
- [14] M. Z. Islam, M. M. Islam, and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," *Informatics in Medicine Unlocked*, vol. 20, p. 100412, 2020. <https://doi.org/10.1016/j.imu.2020.100412>
- [15] E. Vetrimani, M. Arulselvi, and G. Ramesh, "A new hybrid CNN-LSTM model for the X-ray image-based detection of Paediatric croup cough disease," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 7s, pp. 398-405, 2023.
- [16] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023. <https://doi.org/10.1016/j.media.2023.102802>
- [17] G. J. Chowdary and Z. Yin, "Med-former: A transformer based architecture for medical image classification," presented at the International Conference on Medical image Computing and Computer-Assisted Intervention, Springer, 2024.
- [18] H. H. Lee *et al.*, "Region-based contrastive pretraining for medical image retrieval with anatomic query," *arXiv preprint arXiv:2305.05598*, 2023. <https://doi.org/10.48550/arXiv.2305.05598>
- [19] L. Wang, Q. Cao, Z. Zhang, S. Mirjalili, and W. Zhao, "Artificial rabbits optimization: A new bio-inspired meta-heuristic algorithm for solving engineering optimization problems," *Engineering Applications of Artificial Intelligence*, vol. 114, p. 105082, 2022. <https://doi.org/10.1016/j.engappai.2022.105082>
- [20] Y. Bao, X. Liu, Y. Qi, R. Liu, and H. Li, "Adaptive information fusion network for multi-modal personality recognition," *Computer Animation and Virtual Worlds*, vol. 35, no. 3, p. e2268, 2024. <https://doi.org/10.1002/cav.2268>
- [21] N. Arora, A. Kakde, and S. C. Sharma, "An optimal approach for content-based image retrieval using deep learning on COVID-19 and pneumonia X-ray images," *International Journal of System Assurance Engineering and Management*, vol. 14, no. Suppl. 1, pp. 246-255, 2023. <https://doi.org/10.1007/s13198-022-01846-4>
- [22] Y. Xu and S. Liu, "Botanical-based simulation of color change in fruit ripening: Taking tomato as an example," *Computer Animation and Virtual Worlds*, vol. 35, no. 1, p. e2225, 2024. <https://doi.org/10.1002/cav.2225>
- [23] E. Çalli, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest X-ray analysis: A survey," *Medical Image Analysis*, vol. 72, p. 102125, 2021. <https://doi.org/10.1016/j.media.2021.102125>
- [24] L. Cui and M. Liu, "An intelligent deep hash coding network for content-based medical image retrieval for healthcare applications," *Egyptian Informatics Journal*, vol. 27, p. 100499, 2024. <https://doi.org/10.1016/j.eij.2024.100499>
- [25] Y. Chen *et al.*, "Performance evaluation of attention-deep hashing based medical image retrieval in brain MRI datasets," *Journal of Radiation Research and Applied Sciences*, vol. 17, no. 3, p. 100968, 2024. <https://doi.org/10.1016/j.jrras.2024.100968>

- [26] I. D. Apostolopoulos and T. A. Mpesiana, "COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, vol. 43, no. 2, pp. 635–640, 2020. <https://doi.org/10.1007/s13246-020-00865-4>
- [27] R. Padate, A. Gupta, P. Chakrabarti, and A. Sharma, "An enhanced approach to content-based image retrieval using Harris Hawks optimization and automated image captioning," presented at the 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, 2024.
- [28] A. Chakravarty, T. Sarkar, N. Ghosh, R. Sethuraman, and D. Sheet, "Learning decision ensemble using a graph neural network for comorbidity aware chest radiograph screening," presented at the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020.
- [29] H. Chen, S. Miao, D. Xu, G. D. Hager, and A. P. Harrison, "Deep hierarchical multi-label classification of chest X-ray images," presented at the International Conference on Medical Imaging with Deep Learning, PMLR, 2019.
- [30] M. J. Sadik, N. A. Samsudin, and E. F. Bin Ahmad, "Balancing privacy and performance: Exploring encryption and quantization in content-based image retrieval systems," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 10, pp. 903–920, 2024. <https://doi.org/10.14569/IJACSA.2024.0151092>

Views and opinions expressed in this article are the views and opinions of the author(s). Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.