Computers and Electrical Engineering 126 (2025) 110519

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

ESecMedIR: Efficient and Secure Dual-level Transformer based Medical Image Retrieval Framework in the Cloud

Arun Amaithi Rajan [a,*], Vetriselvi V. [a], Mayank Raikwar [b], Mohamed Fuzail H. [a]

[a] Security Research Lab, Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, 600025, India

[b] Department of Informatics, University of Oslo, Norway

ARTICLE INFO

Keywords:
Secure image retrieval
Medical images
Targeted encryption
Dual-level vision transformers
Hash fusion

ABSTRACT

Critical information within medical images helps in accurate diagnosis and research. Nevertheless, this increases the need for efficient storage and retrieval of information embedded in medical images. Moreover, it poses crucial challenges, such as privacy management and secure retrieval of these images. To address these challenges, this paper proposes an Efficient and Secure Medical Image Storage and Retrieval framework (ESecMedIR) leveraging cloud computing. To address privacy concerns and the challenge of balancing security with retrieval efficiency, ESecMedIR utilizes Dual-level Vision Transformer-based Hashing and Privacy Region Encryption. Sensitive regions in medical images are identified and encrypted using hyperchaos-based encryption, while hashcodes are generated and used in similar image searches for efficient retrieval. The proposed framework is tested on three standard medical datasets of Brain MRI, Chest X-ray, and Kidney CT, demonstrating a 10%–30% improvement in the retrieval accuracy over existing methods, ensuring secure and efficient image storage and retrieval management.

1. Introduction

The digitalization of healthcare has revolutionized the way medical data is generated, stored, and accessed, with medical imaging at the forefront of this transformation. From X-rays and MRIs to CT scans and ultrasounds, medical images play a crucial role in diagnosis, treatment planning, and ongoing research. However, the sensitive nature of the information embedded in these images, including identifiable patient data, necessitates robust security measures to protect against unauthorized access and data breaches [1].

The first step towards medical image information management is to create a secure storage system for the data. The storage of these image data can be in a distributed file system [2], in cloud storage, or a decentralized environment [3]. These storage systems should provide security to the stored medical images. In order to provide security, many of these systems encrypt the image and then store the encrypted images. Depending upon the employed encryption scheme, the performance of the storage systems differs. Along with the security of the image information, these secure storage systems must comply with regulations like HIPAA (Health Insurance Portability and Accountability Act) in the U.S. and GDPR (General Data Protection Regulation) in Europe, which mandate strict data protection standards. Among all the available storage options, healthcare providers increasingly adopt cloud-based storage solutions to manage the growing volume of medical images, and the challenge of ensuring secure storage and retrieval becomes more pronounced. The migration to cloud infrastructure offers scalability and accessibility benefits but introduces concerns around

* Corresponding author.
E-mail address: arunamaithirajan@gmail.com (A. Amaithi Rajan).

https://doi.org/10.1016/j.compeleceng.2025.110519
Received 2 March 2025; Received in revised form 13 May 2025; Accepted 6 June 2025
Available online 25 June 2025
0045-7906/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

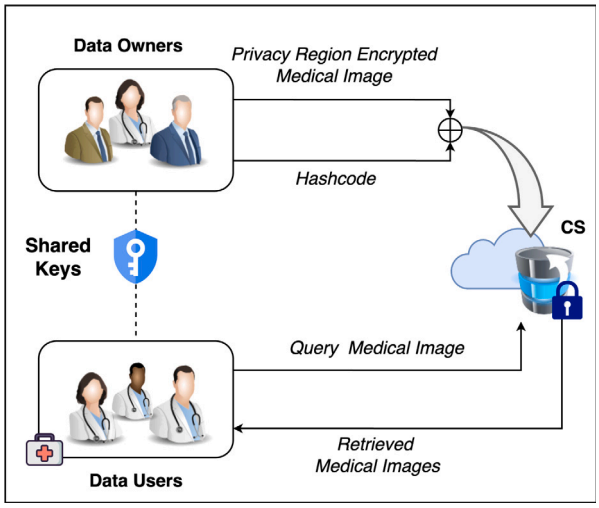


Fig. 1. An overview of the proposed framework.

data privacy, compliance with stringent healthcare regulations, and the potential for cyberattacks. These platforms need to ensure data encryption, redundancy, and compliance with healthcare regulations.

Further, the medical image data stored in a secure storage system should provide efficient retrieval of these data. Such retrieval should not leak any crucial information. As medical images are often large and stored in high volumes, there is a need for efficient compression and indexing methods that do not compromise data integrity or retrieval speed. The efficient retrieval of medical image data not only aids in diagnosing the disease swiftly but also enables researchers to find related cases for analysis from large-scale medical image databases [4,5]. This efficient retrieval of medical image data can be performed using two approaches: (1) Text-based Medical Image Retrieval (TBMIR), and (2) Content-Based Medical Image Retrieval (CBMIR). In TBMIR, images are annotated manually with associated text. However, it takes a lot of work and effort to succinctly describe the actual content of medical images using annotations [6]. However, in CBMIR, image features like shape, texture, and color without manual labeling are employed for effective retrieval [7]. Between these two approaches, CBMIR has direct diagnostic implications, influencing patient care and treatment decisions.

In a CBMIR system, challenges arise from exhibiting similarities between different medical conditions in medical images (inter-class similarities), making it essential to be able to discriminate between images that are visually similar but clinically distinct. The current research aims to address these challenges and enhance retrieval accuracy in CBMIR [8]. Moreover, due to inter-class similarities and complex features, indexing involving hashcode generation becomes a challenging process [9]. To solve this challenge, various deep hashing techniques [10], utilizing different deep learning models have been used for hashcode generation [11]. However, only a few deep hashing techniques have shown effectiveness for medical image data. In a deep hashing technique for image data, typically, the entire image is used to generate the hashcode. However, this could be advanced by adopting Vision Transformer (ViT) for hashcode generation [12]. ViTs leverage their adeptness in capturing complex visual features in images, providing enhanced accuracy in Content-based Image Retrieval (CBIR). Dual-level ViTs can be seamlessly integrated into the domain of medical images that capture the long-range dependencies and focus on fine-grained local structures, essential for distinguishing visually similar but clinically distinct cases. This hierarchical representation enhances retrieval performance by preserving both global semantic consistency and localized discriminative features in medical images.

Apart from using ViTs for hashcode generation by using only the local features of the image, a similar approach by not using the entire image for encryption can further enhance the CBMIR system. Though, conventional CBMIR systems, images undergo full encryption and are stored in conjunction with their hashcodes in database [13], resulting in a delay in the retrieval process of the image. However, not all components of medical images contribute to the crucial features for the retrieval process. Typically, in medical images, these crucial features involving sensitive information are concentrated within the Region of Interest (ROI) in medical images [14]. For example, in medical images like brain tumor MRI scans, the tumor-identified area is the ROI. The ROI encompasses not only vital information necessary for precise diagnoses but is also safeguarded against modifications, referred to as the "Region of Privacy" (ROP). Encryption of ROP enhances security and reduces retrieval time for a CBMIR system, addressing the limitations associated with the encryption of entire medical images in CBMIR systems.

In addressing challenges related to retrieval accuracy, we employ a two-fold approach: generating hashcodes independently from local and global features using dedicated ViTs, followed by their fusion for optimal outcomes. Our method, ESecMedIR, integrates Dual-level ViT-based global-local hash fusion technique and ROP-enabled encryption. To the best of our knowledge, this marks the first instance of employing ROP-enabled encryption and dual-level transformers for hashcode generation in CBMIR. An overview of our proposed framework is depicted in Fig. 1.

The key contributions of this article are summarized as follows:

Table 1
Our framework vs. Existing frameworks.

Retrieval framework	Medical dataset	Image security		Feature extraction			Hash fusion	Storage
		Image encryption	ROI image encryption	Technique	Local	Global		
Yan et al. [22]	×	✔	×	ResNet 50	×	✔	×	Linear
Du et al. [23]	×	×	×	SURF	×	✔	×	–
Hu et al. [24]	×	✔	✔	PCNN	×	✔	×	Linear
Dubey et al. [25]	×	×	×	ViT	×	✔	×	–
Guan et al. [26]	✔	×	×	DenseNet 121	✔	✔	×	–
Ozbay et al. [27]	✔	×	×	DenseNet 201	✔	✔	×	–
ESecMedIR	✔	✔	✔	Dual-level ViT	✔	✔	✔	Tree

(i) We propose an ESecMedIR framework to improve the retrieval accuracy and enhance the security of sensitive regions in CBMIR.
(ii) We present a fusion-based hashcode generation model, Dual-level Vision Transformer-based Hashing (DViTH), for optimizing the retrieval accuracy in CBMIR.
(iii) We designed a novel ROP-enabled encryption method utilizing the state-of-the-art YOLOv8 object detection framework to ensure the security of essential regions of medical images in CBMIR.
(iv) We employ a tree-based structure to enhance retrieval speed by incorporating ROP-encrypted images and DViTH-generated hashcodes.
(v) We also perform rigorous tests with three standard medical image datasets to demonstrate better retrieval performance of ESecMedIR compared to six conventional baseline methods. ESecMedIR achieves high performance and outperforms baseline models by 10%–30%.

The rest of the article is structured as follows: Section 2 presents the related work. Section 3 explains the proposed system architecture, and Section 4 describes the hashcode generation model. Section 5 gives a brief security analysis. Section 6 presents the experimental results and performance analysis. Finally, Section 7 concludes the proposed work.

2. Related work

The increasing volume of medical data generated by healthcare requires secure storage and management of the data. To fulfill this need, secure storage and processing of medical data have been an active area of research. Various frameworks and techniques have been proposed to address the dual challenges of efficiency and security in medical image storage. Similarly, many schemes and frameworks have been proposed to efficiently retrieve the stored images from the underlying storage systems. In this section, we review some of these works related to the storage and retrieval of medical images.

As stated earlier, there are various techniques suitable for storing medical image data [15]. However, cloud-based storage solutions have been widely adopted in the healthcare sector for their scalability and cost-effectiveness. To provide security for the image data in the cloud computing environment, different schemes have been proposed. Marwan et al. [16] proposed a cloud-based storage system using a secret sharing scheme to provide confidentiality of the data. Further, the authors designed and implemented a multi-cloud storage system for secure medical image storage [17]. Later, the authors proposed a storage framework [18] using segmentation and watermarking techniques and employing access control methods to enforce data security policies. Apart from using cloud-based storage for medical image data, research on new techniques has been conducted in recent years. Blockchain-based storage solutions [19,20] have been explored for medical data where the actual data is stored off-chain while the hash value and metadata are stored in the blockchain. Federated learning has also opened new avenues for secure and efficient medical image storage [21]. This approach also provides privacy to the data and gives the opportunity to analyze the medical image data without disclosing the image data.

A crucial component in the field of medical imaging and healthcare is the image retrieval framework, e.g., CBIR. Various CBIR techniques, particularly those based on deep learning and hashing, have been explored for medical images. Sharma et al. [28] conducted a review of existing CBIR mechanisms for medical images. More recently, Luo et al. [10] carried out a comprehensive survey of deep hashing methods in CBIR, categorizing them into semi-supervised deep hashing, domain adaptation deep hashing, and multi-modal deep hashing [29]. These findings and classifications offer valuable insights that can be extended to the domain of medical images [30].

Recent studies have explored various methods to enhance CBIR through unsupervised deep learning [31] and hashing techniques [32]. Yifun et al. [33] introduced Unsupervised Deep Triplet Hashing (UDTH) for CBIR, which constructs pseudo triplets by leveraging neighborhood structures in high dimensional visual feature spaces using autoencoders and binary quantization, with VGG-16 as the backbone network. Ozturk et al. [34] proposed a novel hashcode generation method for medical images utilizing ResNet-50. Ozbay et al. [35] introduced a DenseNet-201-based hashing approach for brain MRI, demonstrating superior performance compared to other deep hashing methods by employing the compare-to-learn method. Further, the challenge of error propagation in hashcode generation was addressed by Yuan et al. [36].

In the area of secure CBIR, Yan et al. [22] incorporated Software Guard Extensions (SGX) enclaves for secure similarity search in the IoT environment, relying on Intel SGX support and an encryption scheme. Later, using complex cryptographic algorithms,

Yang et al. [37] improved search efficiency and security by introducing a secure and traceable multikey image retrieval system in cloud-assisted IoT. A clustering-based privacy-preserving image retrieval system was created by Liang et al. [31] to maximize search accuracy while providing robust security assurances in cloud settings. Furthermore, Chen et al. [38] introduced a content-based image retrieval method that uses robust feature extraction techniques and privacy-preserving computations to improve security and retrieval performance in cloud-assisted IoT.

Recently, Dubey et al. [25] introduced a Vision Transformer-based Hashing approach for image retrieval, replacing the classification head with a hash layer to efficiently learn hashcodes. While Vanilla ViT [39] quickly captures global features, its effectiveness in capturing local features is not as pronounced. Contemporary techniques utilize CNN-based feature extraction to address this limitation and capture both local and global features [26]. Recognizing the significance of ROIs in medical images, ROI-based hashing has gained prominence for enhanced retrieval accuracy. Janani et al. [14] incorporated copyright protection in image retrieval by focusing on ROI separation. In general image retrieval applications with privacy concerns, encrypting specific regions (e.g., face, title card) has been explored for user privacy protection [24], though this concept has yet to be extensively applied in healthcare databases. Through our analysis, we identified a persisting need for an efficient and secure medical image retrieval system, motivating the development of our proposed system, which employs Dual-level ViT-based hashing and ROP encryption in medical image retrieval. Table 1 presents the advancement of our framework over existing ones.

3. System architecture

3.1. System entities and problem formulation

The proposed framework involves three key entities: Data Owners (*DO*), a Cloud Server (*CS*), and authorized Data Users (*DU*), as illustrated in Fig. 1. *DO* possess a collection of N medical images, denoted as $M = \{I_{mi}\}_{i=1}^{N}$, which are securely externalized to the cloud using encryption techniques. *DU* retrieves similar images by uploading a query medical image I_{mq} to the cloud. The *CS* is responsible for storing and managing medical images, as well as processing incoming queries efficiently. Given that the *CS* is assumed to be honest but curious, ensuring image privacy is a critical concern. The primary challenge lies in securely managing the storage and retrieval of medical images in the cloud. Specifically, the problem is to efficiently and securely retrieve the top-k similar medical images from an encrypted database in response to a query medical image I_{mq}. The overall system model is shown in Fig. 2. The following sections outline the proposed solution.

3.2. Proposed framework

The proposed ESecMedIR framework is depicted in Fig. 2. ESecMedIR ensures that images are secured while also being indexed for efficient and swift retrieval. In ESecMedIR, the medical images are sent to the cloud as ROP-encrypted images along with a generated hashcode. Therefore, ESecMedIR involves two processes: (1) ROP-encryption, and (2) hashcode generation. To compute the ROP encryption of a given medical image, first, the ROP is extracted using the fine-tuned state-of-the-art object detection model. Next, ROP is encrypted using a hyperchaos-based image encryption model, resulting in encrypted ROP. The encrypted ROP is then patched back into the original image, and ROP coordinates are embedded using the steganography technique. Further, for the hashcode generation, the ViT networks are employed for global and local feature extraction, producing discriminant global and local hashcodes. The fusion of these hashcodes enables efficient retrieval by creating a comprehensive representation of medical images. For the image retrieval process, the generated fused hashcode alongside the corresponding embedded image is used where the hashcode serves as the key in the tree-based cloud storage.

During the retrieval, given a query medical image, the hashcode is generated, and a search is conducted in the tree-based storage, returning the top-k embedded images based on the hamming distance for similarity measurement. Further, authorized individuals can retrieve the decrypted images relevant to the query from the received ROP-encrypted images. The proposed framework contains two phases: secure medical image storage (Section 3.3) and secure medical image retrieval (Section 3.4). Section 3.3 contains 3 sub-modules: 3.3.1 for ROP detection, 3.3.2 for similarity preserving hashcode generation, and 3.3.3 for embedded image generation. ESecMedIR framework provides the following properties compared to existing CBMIR frameworks.

- **Enhanced Retrieval Accuracy:** Our framework aims to surpass the accuracy of existing medical image retrieval models.
- **Image Security:** Crucial and sensitive regions within medical images are secured to prevent exposure or modification.
- **Optimized Retrieval Efficiency:** The search and retrieval of medical images are executed at a lower cost compared to alternative image search methods.

3.3. Secure medical image storage

In this phase, private region in the medical images are first identified and encrypted to produce ROP-encrypted images, after which hashcodes are derived globally and locally from medical images, and then fused together.

3.3.1. Region of Privacy Detection

In our proposed model, Region of Privacy (ROP) detection is crucial for both the local feature extraction in hashcode generation and the protection of sensitive regions within medical images which aids in refining retrieval based on the ROP. An accurate model

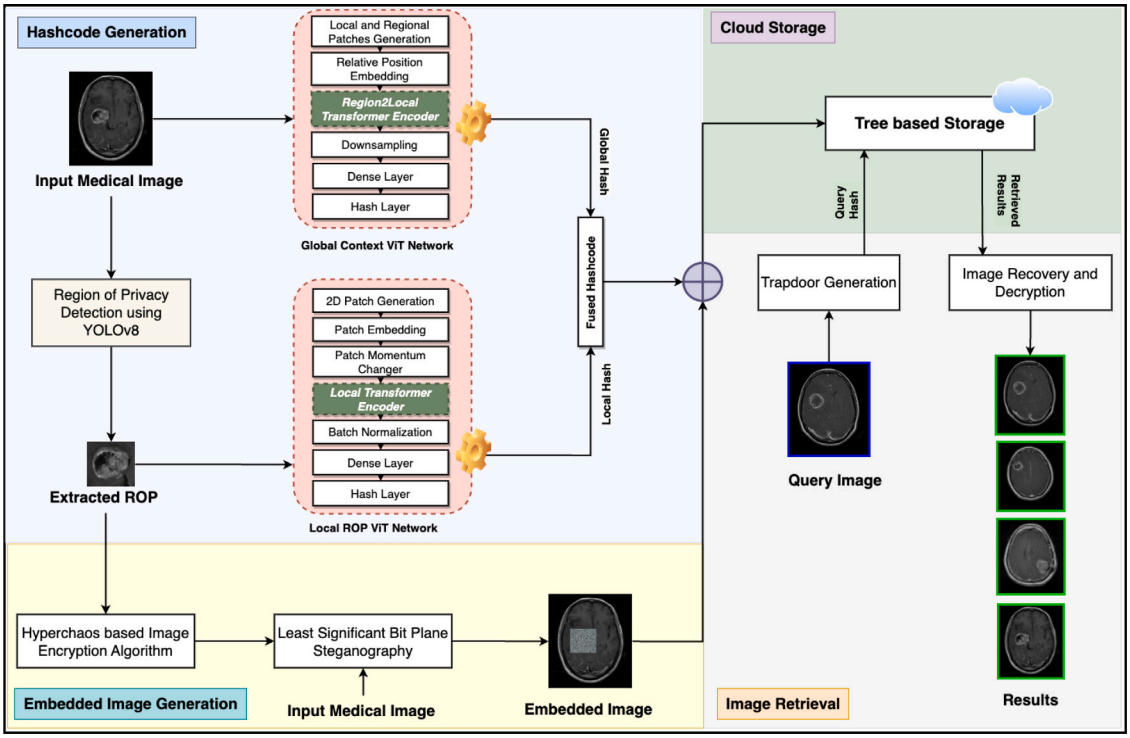


Fig. 2. Proposed ESecMedIR framework.

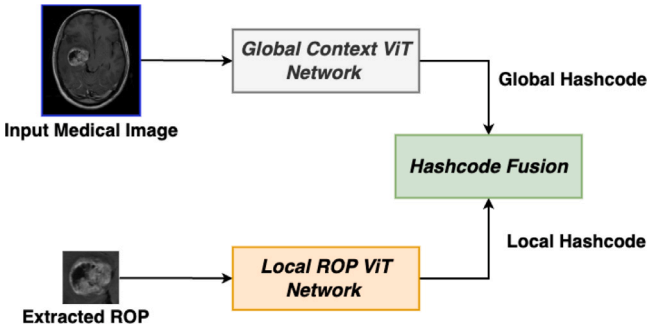


Fig. 3. DViTH: Overview of the proposed hashcode generation model.

is essential for identifying the sensitive region containing crucial diagnostic information. Among the existing models, YOLOv8 has demonstrated higher localization accuracy and faster detection [40]. Therefore, we employ the latest YOLOv8 deep learning model for ROP detection in medical images. After fine-tuning the pre-trained model with medical datasets, it becomes capable of localizing critical regions in unseen medical images, providing bounding box coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$. For instance, the YOLOv8 model trained on brain tumor MRI accurately detects tumor regions for each class based on learned features. The detected ROP can serve as input to both the hashcode generation and the embedded image generation. The deployed YOLOv8 model is detailed in Appendix A.

3.3.2. *Similarity preserving hashcode generation: DViTH*

Our novel hashcode generation model Dual-level Vision Transformer-based Hashing's (DViTH) objective is to generate a similarity-preserving hashcode that captures both overall and ROP-specific image features for accurate retrieval. This involves two essential feature extractions: global and local, contributing to corresponding hashcodes (H_{gbl}, H_{loc}) that are later fused. Fig. 3 depicts the overview of DViTH. A detailed explanation of how proposed ViT networks are employed for feature extraction and hashing, incorporating the targeted objective function, is presented in Section 4.

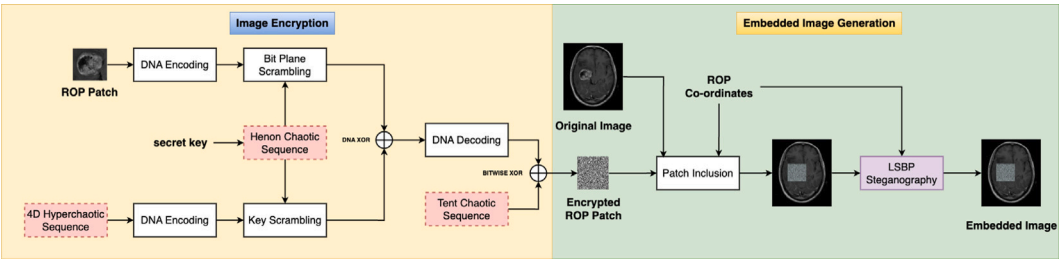


Fig. 4. Image encryption and embedded image generation flow.

3.3.3. Embedded image generation

The Region of Privacy (ROP) needs to be encrypted within the medical image to ensure confidentiality in our retrieval model. After extracting the ROP using YOLOv8, encryption is applied using the hyperchaos-based image encryption model proposed by Amaithi Rajan et al. [41]. This encryption method has been rigorously tested against various attacks and has demonstrated resilience. Security analysis of this chosen encryption model is extensively shown in Appendix B. The medical image encryption model utilizes a combination of chaotic maps applied over DNA-encoded images with bit-plane scrambling. Hyperchaotic maps generate confusion keys, while chaotic maps generate diffusion keys. The encrypted ROP is then patched back into the original image.

The ROP coordinates (bounding box coordinates) need to be embedded into the ROP-patched images to ensure correct decryption and reconstruction of the original image. To achieve this, Least Significant Bit Plane (LSBP) steganography is employed. In this process, the least significant bit plane of the encrypted ROP-patched image is selected, and the coordinate values are concealed within the bit plane values. The process of generating the embedded image is illustrated in Fig. 4.

3.4. Secure medical image retrieval

Image retrieval is a crucial process of efficiently fetching relevant medical images from a database based on certain features. For efficient and secure retrieval, a query image I_{mq} is employed as the reference point for the retrieval task. To facilitate the search, a hashcode H_q is generated for the query image, called trapdoor generation. The Hamming distance serves as the metric to quantify the similarity between the query image and the medical images stored in the B+ tree-based database. The computation of hamming distance involves XOR operations between corresponding bits of the query image hashcode H_q and the hashcodes of the medical images in the database H_m. This process is carried out for each bit from 1 to L, where L represents the length of the hashcode as shown in Eq. (1). The resulting distances are used to assess the similarity, with a higher similarity corresponding to a lower computed distance.

$$Dist(H_q, H_m) = \sum_{j=1}^{L} H_q[j] \oplus H_m[j] \tag{1}$$

Once the distances are calculated, the retrieval results are sorted based on the similarity ratio, and the top-k similar images are retrieved. Medical images stored in a tree format along with hashcodes ensure an efficient retrieval process, with a time complexity of approximately $log(N)$, where N is the number of images in the database.

4. Dual-level vision transformer-based hashcode generation

This section presents the proposed Dual-level Vision Transformer-based Hashcode Generation model, which utilizes two ViT models, each tasked with capturing global and local features separately. Table 2 provides the notations used in the model. The following subsections cover global hashcode generation in Section 4.1, local hashcode generation in Section 4.2, and the objective functions in Section 4.3.

4.1. Global hashcode generation

Leveraging a global self-attention-based vision transformer for intricate images like medical images yields superior feature maps. Challenges like uniform patch sizes in self-attention layers hinder knowledge sharing between larger and smaller patches, increasing computational demands. To overcome this and effectively capture global and local features in medical images, we integrated our hashing module with a region-based ViT architecture [42]. The model, named the Global Context ViT (GCViT) network, is designed for global hashcode generation. The GCViT network generates global hashcodes using vision transformers with a regional-to-local attention mechanism and a pyramid structure. It processes images by creating local and regional tokens, with each regional token linked to nearby local tokens. Regional self-attention gathers global information, while local self-attention focuses on specific areas, incorporating global context. This design enables GCViT to produce robust hashcodes that capture both fine details and broader medical image context, enhancing accuracy and efficiency in retrieving relevant images.

Table 2
Notations used.

Notation	Description
$M = \{I_{mi}\}_{i=1}^{N}$	N medical images
$M_r = \{R_{mi}\}_{i=1}^{N}$	Sensitive regions of N medical images
Q	Number of classes
L	Hashcode length
f_{gbl}	Global feature
h_g	Global hashcode
f_{rop_loc}	Local feature of ROP
h_l	Local hashcode
HC	Hash centers
GHM_{best}	Global hashcode generation model
LHM_{best}	Local hashcode generation model
$H_{gbl} = \{h_{gi}\}_{i=1}^{N}$	Global hashcodes of N medical images
$H_{loc} = \{h_{li}\}_{i=1}^{N}$	Local hashcodes of N medical images
$H_m = \{h_{gi} + h_{li}\}_{i=1}^{N}$	Hashcodes of N medical images
$\mathcal{L}_{csl}$	Central similarity loss
$\mathcal{L}_{quant}$	Quantization loss
$\mathcal{L}_{bit}$	Bit balance loss

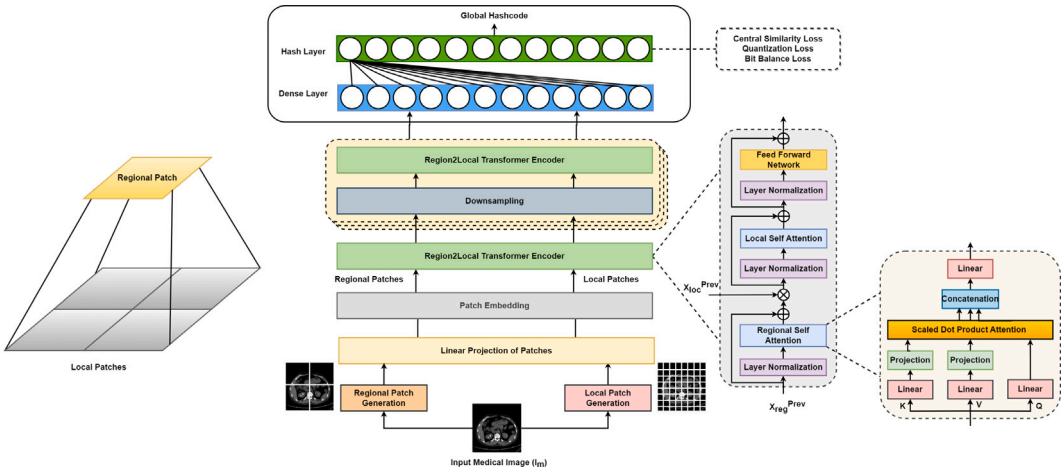


Fig. 5. Global hashcode generation: GCViT architecture.

Fig. 5 provides a detailed illustration of the GCViT architecture. The input image I_m undergoes two tokenization processes: one for regional tokens (upper path) and another for local tokens (lower path). Each tokenization converts the image into regional and local tokens. The model uses Region2Local (R2L) transformer encoders for regional-to-local attention. Initially, these encoders process two sets of tokens, and in later stages, downsampling is applied to balance computation. If RG^{s-1} and LO^{s-1} are the region and local tokens to the input layer at stage s, then the R2L transformer encoder can be expressed as shown in Eqs. (2)–(5).

$$Y^s = RG^{s-1} + \mathrm{RSA}(\mathrm{LN}(RG^{s-1})) \tag{2}$$

$$Y^s_{a,b} = Y^s_{a,b} \parallel LO^{s-1}_{a,b,m,n} \tag{3}$$

$$Z^s_{a,b} = Y^s_{a,b} + \mathrm{LSA}(\mathrm{LN}(Y^s_{a,b})) \tag{4}$$

$$RG^s_{a,b} = Z^s_{a,b} + \mathrm{FFN}(\mathrm{LN}(Z^s_{a,b})) \tag{5}$$

Here, a and b denote spatial indices with respect to regional tokens, while m and n represent indices of local tokens. LN stands for Layer Normalization, and FFN corresponds to the Feed Forward Network. The $Y^s_{a,b}$ includes one regional token and corresponding local tokens, Regional tokens are updated using RSA to capture high-level contextual features as in Eq. (2). In Eq. (3) local and regional tokens are concatenated, allowing information fusion. Local tokens are refined using LSA to ensure finer-grained feature extraction as shown in Eq. (4). Finally feature embeddings are computed using FFN for effective representation learning in Eq. (5). Thus, the information between local and regional tokens is exchanged. Both Regional Self Attention (RSA) and Local Self Attention (LSA) mirror the Multihead Self Attention (MSA) module in ViT [39]. The structures of the R2L Transformer encoder and MSA are emphasized in Fig. 5. The RSA module enables information exchange between regions, followed by parallel LSA for each region.

Algorithm 1 Global Hashcode Learning Algorithm

Input: Medical Image Dataset $M = \{I_{mi}\}_{i=1}^{N}$ along with class labels, Number of classes Q, Hashcode length L

Output: Global Hashcode Generation Model GHM_{best}

Initialization: Hadamard Matrix H_L and H_{2L}, obtained hash centers $HC_G = \{hc_g\}^N$, α, β, GHM_{best} = NULL, local patching factor $lp = 4$, regional patching factor $rp = 16$, $f_{reg} = f_{loc} = \emptyset$

Step 1: Generate Q hash centers $HC_Q = \{hc_1, hc_2, \ldots, hc_q\}$ from H_L and H_{2L}
for each epoch e **do**
  for each I_m in batch $\mathcal{B}$ **do**
    for each stage s **do**
      Step 2: Regional and local tokenization
      $RG^{s-1} \leftarrow \text{Tokenize}\left(I_m, \frac{W}{rp * 2^{s-1}} \times \frac{H}{rp * 2^{s-1}}\right)$
      $LO^{s-1} \leftarrow \text{Tokenize}\left(I_m, \frac{W}{lp * 2^{s-1}} \times \frac{H}{lp * 2^{s-1}}\right)$
      Step 3: Regional to local Transformer Encoder
      R2L_Trans(RG^{s-1}, LO^{s-1}) = (RSA, LSA, FFN) (ref. Eqs. (2)-(5))
      Update f_{reg}, $f_{loc} \leftarrow$ R2L_Trans(RG^{s-1}, LO^{s-1})
    end for
    Step 4: Hashcode Generation
    $f_{gbl} \leftarrow \text{FCN}(f_{reg}, f_{loc})$
    $h_g \leftarrow HashcodeGen(f_{gbl})$
    Step 5: Error Calculation (ref. Section 4.3)
    $\mathcal{L}_{csl} \leftarrow \frac{1}{L}\sum_{j\in L}(hc_{g,j} log(h_{g,j}) + (1 - hc_{g,j}) log(1 - h_{g,j}))$
    $\mathcal{L}_{quant} \leftarrow \sum_{j\in L}(logcosh|2h_{g,j} - 1| - 1)$
    $\mathcal{L}_{bit} \leftarrow \frac{1}{L}\sum_{j\in L} h_g(j)$
  end for
  Step 6: Back Propagation
  Calculate $\mathcal{L}_{total} \leftarrow \frac{1}{N}\sum_{1}^{N} \mathcal{L}_{csl} + \alpha\mathcal{L}_{quant} + \beta\mathcal{L}_{bit}$
  Back propagate $\mathcal{L}_{total}$
  Step 7: Best Model Selection
  if differences in $(\mathcal{L}_{total}^{e-2}, \mathcal{L}_{total}^{e-1}, \mathcal{L}_{total}^{e})$ is negligible **then**
    return GHM_{best}
  end if
end for

The FFN processes each token, creating distinct regional and local tokens. Shared RSA and LSA weights transfer local token data across regions, with RSA focusing on high-level information and LSA on local context. In the final embedding layer, regional tokens are averaged to generate the global hashcode embedding. A hash block atop the transformer encoder streamlines hashcode learning, with features projected to dimension D and passed through ReLU activation. The final global hashcode, h_g, is produced via a linear projection, with a flexible hash bit length L. The propagated losses during the learning are detailed in Section 4.3. The detailed global hashcode learning model is shown in Algorithm 1.

4.2. Local hashcode generation

Local features in medical images, especially within ROPs, are crucial for capturing fine details and identifying abnormalities. Standard ViT models may overlook these details, as their self-attention mechanism prioritizes global relationships. To address this, a CNN-ViT hybrid model is used, combining the strengths of CNNs for local feature extraction with ViTs for a broader context. This ensures the effective capture of intricate patterns within ROPs. We implemented the CNN-ViT hybrid model [43] to extract local features from ROPs and integrated a hashing module to develop the Local ROP ViT (LRViT) network. LRViT excels in addressing high inter-class similarity in medical images by generating discriminant hashcodes through an efficient convolution operation that establishes an attention mechanism. This hybrid architecture combines transformer and convolution blocks, featuring components such as the Effective Convolution Block (ECB), Local Transformer Block (LTB), and Patch Momentum Changer (PMC), as shown in Fig. 6.

The ECB, a crucial element within the hybrid architecture for multi-scale information utilization, is followed by an efficient attention-based token mixing module, comprising the Locally Feed Forward Network (LFFN) and Multi-Head Convolutional Attention (MHCA). The function of ECB is expressed through the following two equations: MHCA followed by LFFN.

$$Q^s = \text{MHCA}(P^{s-1}) + P^{s-1} \tag{6}$$

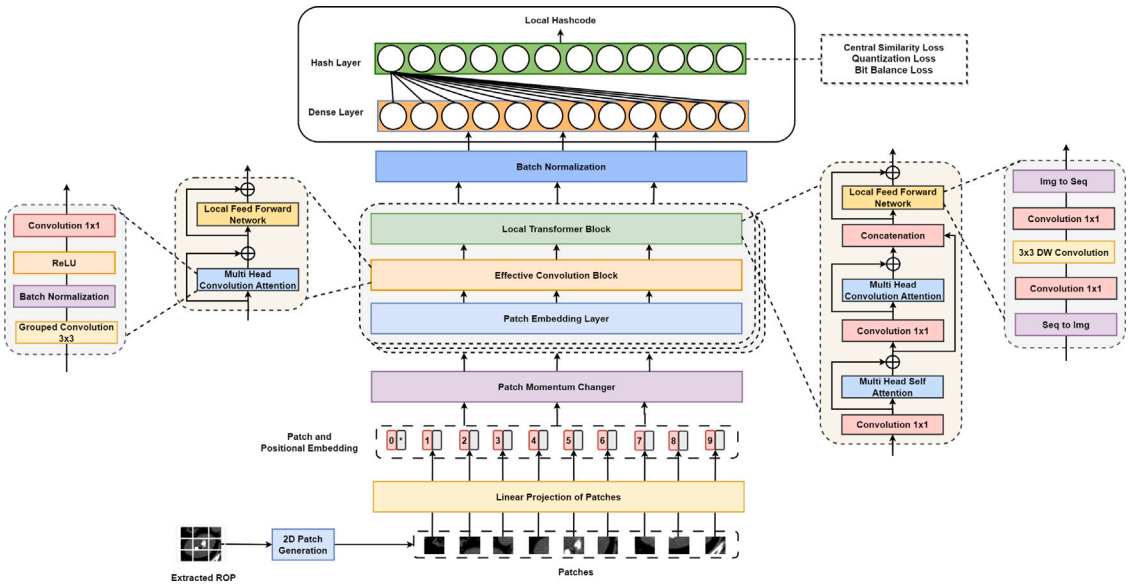


Fig. 6. Local hashcode generation: LRViT architecture.

$$R^s = \text{LFFN}(Q^s) + Q^s \tag{7}$$

P^{s-1} is input patches from layer $s-1$ to s. Q^s and R^s are outputs from MHCA and LFFN respectively. For optimal local representation learning, LRViT incorporates Convolutional Attention (CA) with a multi-head paradigm, facilitating simultaneous attention to information across various representation subspaces at different locations.

$$\text{MHCA}(X) = concat(CA(x_1), CA(x_2), \dots, CA(x_h))W^o \tag{8}$$

$$CA(x_i) = (W.T_{i,j}), \quad T_{i,j} \in X \tag{9}$$

W is trainable parameter and $T_{i,j}$ are adjacent tokens in input feature X. LFFN restores token proximity, introducing a localized perspective into the network. LFFN's local convolution operations are shown as Eqs. (10) to (12) where Q^s is the input embedding. W_a^r, W_b^r are kernels of convolution layers. $Q^{s'}, Q^s$ yields R^s.

$$Q^r = seq2img(Q^s), Q^r \in \mathbb{R}^{h \times w \times d} \tag{10}$$

$$S^r = f(Q^r * W_a^r) * W_b^r \tag{11}$$

$$Q^{s'} = img2seq(S^r) \tag{12}$$

The ECB output is directed to LTB, which efficiently captures and mixes multi-frequency signals to enhance the network's modeling capabilities. LTB integrates a Multi-Head Self-Attention (MHSA) block to specifically capture low-frequency signals, mitigating the risk of information loss in high-frequency details often observed in transformer blocks.

$$\text{MHSA}(X) = concat(SA(x_1), SA(x_2), \dots, SA(x_\psi))W^o \tag{13}$$

$$SA(x_i) = Attention(x_i.W^Q, Pl_\delta(x_i.W^K), Pl_\delta(x_i.W^V)) \tag{14}$$

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d})V \tag{15}$$

In the presented equations from (13)–(15), $X = [x_1, x_2, \dots, x_\psi]$ denotes the division of the input feature X into a multi-head form in the channel dimension. Key components include an output projection layer W^o, with ψ representing the number of heads and d as the hidden dimension of the transformer. The Pl_δ operation involves average-pooling with a stride parameter δ for downsampling. The output features from MHSA and MHCA are concatenated for a mix of high-low features called local features. An FCN layer is added after the Batch Normalization (BN) layer to extract essential features for discriminant hashcode generation. Finally, a hashing layer produces the local hashcode h_l through linear projection. The detailed algorithm is presented in Algorithm 2. The propagated losses are detailed in Section 4.3.

Algorithm 2 Local Hashcode Learning Algorithm

Input: ROP Medical Image Dataset $M_r = \{R_{mi}\}_{i=1}^{N}$ along with class labels, Number of classes Q, Hashcode length L, PMC augmented ROPs

Output: Local Hashcode Generation Model LHM_{best}

Initialization: Hadamard Matrix H_L and H_{2L}, obtained hash centers $HC_L = \{hc_l\}^N, \alpha, \beta, LHM_{best}$ = NULL, patching factor $pf = 4$, $f_{rop_loc} = f_v^e = f_v = \emptyset$

Step 1: Generate Q hash centers $HC_Q = \{hc_1, hc_2, \ldots, hc_q\}$ from H_L and H_{2L}
ECB($embedded_patches$) = (MHCA, LFFN) (ref. Eqs. (8)-(12))
LTB(f_v) = (MHSA, MHCA, LFFN) (ref. Eqs. (13)-(15) for MHSA)
for each epoch e **do**
  for each R_m in batch $\mathcal{B}$ **do**
    for each stage s **do**
      Step 2: Patch Embedding of R_m
      $P^s \leftarrow \text{Embed}\left(R_m, \frac{W}{pf * 2^s} \times \frac{H}{pf * 2^s}\right)$
      if s == 0 **then**
        Step 3: Convolution Block
        $f_v^e \leftarrow \text{ECB}(P^s)$
        Update $f_v \leftarrow f_v^e$
      else
        Step 3: Convolution and Transformer Block
        $f_v^e \leftarrow \text{ECB}(P^s)$
        Update $f_v \leftarrow \text{LTB}(f_v^e)$
      end if
    end for
    Step 4: Hashcode Generation
    $f_{rop_loc} \leftarrow \text{FCN}(\text{BN}(f_v))$
    $h_l \leftarrow HashcodeGen(f_{rop_loc})$
    Step 5: Error Calculation (ref. Section 4.3)
    $\mathcal{L}_{csl} \leftarrow \frac{1}{L}\sum_{j\in L}(hc_{l,j}log(h_{l,j}) + (1 - hc_{l,j})log(1 - h_{l,j}))$
    $\mathcal{L}_{quant} \leftarrow \sum_{j\in L}(logcosh|2h_{l,j} - 1| - 1)$
    $\mathcal{L}_{bit} \leftarrow \frac{1}{L}\sum_{j\in L} h_l(j)$
  end for
  Step 6: Back Propagation
  Calculate $\mathcal{L}_{total} \leftarrow \frac{1}{N}\sum_1^N \mathcal{L}_{csl} + \alpha\mathcal{L}_{quant} + \beta\mathcal{L}_{bit}$
  Backpropagate $\mathcal{L}_{total}$
  Step 7: Best Model Selection
  if differences in $(\mathcal{L}_{total}^{e-2}, \mathcal{L}_{total}^{e-1}, \mathcal{L}_{total}^{e})$ is negligible **then**
    return LHM_{best}
  end if
end for

4.3. Hashcode learning

In the process of similarity-preserving hashcode generation, distinct global and local hashcodes are independently learned through separate feature extractions as shown in Algorithms 1 and 2. The resulting global hashcode (h_g) and local hashcode (h_l) are then merged to form h_m, which constitutes the desired output of the DViTH model. The primary aim is to produce a semantically discriminative hashcode, enhancing the precision of image retrieval. Consequently, both local and global hashcode generation models (GHM_{best}, LHM_{best}) are learning through the propagation of cumulative loss. A non-linear hash function $f : x \rightarrow h \in \{0, 1\}^L$ is employed, transforming data into hashcode. Here, L represents the hashcode length, determining the number of nodes in the hash layer. The function $K(features)$ converts features into hashcodes for each node. The computed hashcodes, denoted as $j = 1, 2, 3, \ldots, L$ are expressed in the following equations.

$$h(j) = K(f_j) \tag{16}$$

$$K(val) = \begin{cases} 1, & \text{if } val \text{ is greater than or equal to } 0 \\ 0, & \text{if } val \text{ is less than } 0 \end{cases} \tag{17}$$

The proposed method combines Central Similarity Loss (CSL), quantization loss, and bit balance loss to improve hashcode generation. Hash centers are initially created using the Hadamard Matrix and Bernoulli distribution. CSL optimizes the central similarity of data points to their hash centers, rather than focusing on local similarity, and is based on Binary Cross Entropy (BCE) loss since hash centers are binary vectors. Suppose we have N images across Q classes. As each image belongs to each class they share the corresponding generated hash centers from HC_Q [36]. The obtained hash centers $HC = \{hc_{x1}, hc_{x2}, \ldots, hc_{xN}\}$, hashcodes $H_x = \{h_{x1}, h_{x2}, \ldots, h_{xN}\}$, and $x \in \{g, l\}$. Then the optimization objective of the central similarity loss is defined as:

$$\mathcal{L}_{csl} = \frac{1}{L}\sum_{i=1}^{N}\sum_{j\in L}[hc_{xi,j} log h_{xi,j} + (1 - hc_{xi,j}) log(1 - h_{xi,j})] \tag{18}$$

Given the inherent complexity in optimization, the current technique lacks the capability to guarantee the complete convergence of generated hashcodes to hash centers, primarily due to the binary nature of each hash center. To address this, we introduce a quantization loss $\mathcal{L}_{quant}$. By employing the computation outlined in Eq. (19), a quantization function is applied to approximate the binary code to the desired hashcode. To overcome challenges associated with calculating derivatives, we utilize the smooth function $logcosh$ as a replacement, offering improved differentiability.

$$\mathcal{L}_{quant} = \sum_{i=1}^{N}\sum_{j=1}^{L}(logcosh|2h_{xi,j} - 1| - 1) \tag{19}$$

Added to these two loss functions, bit balance loss ($\mathcal{L}_{bit}$) is also added to efficiently compute the hashcode. This means that there is a 50% chance that each hashcode will be between 0 and 1. It is possible to generate more unique hashcodes by utilizing the target function provided in Eq. (20) for L-bit hashcode generation. $h_x(j)$ is the hash layer output of jth node.

$$\mathcal{L}_{bit} = \frac{1}{L}\sum_{i=1}^{N}\sum_{j\in L} h_x(j) \tag{20}$$

Finally, the cumulative loss function for the discriminant hashcode generation is,

$$\mathcal{L}_{total_loss} = \mathcal{L}_{csl} + \alpha\mathcal{L}_{quant} + \beta\mathcal{L}_{bit} \tag{21}$$

$$\textbf{min}\ \langle\mathcal{L}_{total_loss}\rangle \tag{22}$$

All losses are merged in Eq. (21), where α and β are the weighting factors. The overall objective is to minimize this total loss. By propagating this loss in each hash generation network, efficient local and global hashcode generation models are developed. Then for a given I_m, global hashcode h_g is generated from GHM_{best} ($h_g \leftarrow GHM_{best}(I_m)$), local hashcode h_l is generated from LHM_{best} ($h_l \leftarrow GHM_{best}(R_m)$). Both hashcodes are fused to get the final hashcode $h_m \leftarrow h_g + h_l$. $\langle h_m : E(I_m)\rangle$ is being securely stored in the cloud.

5. Security of ESecMedIR framework

Definition 1 (*IND-CPA Security*)**.** An encryption scheme E is IND-CPA secure if, for any probabilistic polynomial-time (PPT) adversary $\mathcal{A}$, the advantage in distinguishing between encryptions of any two chosen plaintext images is negligible. Formally, for any $\mathcal{A}$:

$$\left|\Pr[\mathcal{A}(E(I_{m0})) = 1] - \Pr[\mathcal{A}(E(I_{m1})) = 1]\right| \leq neg(\lambda) \tag{23}$$

where I_{m0} and I_{m1} are any two plaintext images, and neg is a negligible function in the security parameter λ.

Definition 2 (*Full Encryption*)**.** The entire medical image I_m is encrypted using the hyperchaos-based encryption algorithm E, resulting in ciphertext image $C_m = E(I_m)$.

Definition 3 (*Targeted Encryption*)**.** Only the sensitive region R_m of the medical image I_m is encrypted. Let $I_m = (R_m, NS)$, where NS_m is the non-sensitive region. The image I_m after targeted encryption is $C'_m = (E(R_m), NS_m)$.

Theorem 1. *The targeted encryption of a medical image, where only the sensitive region is encrypted using a hyperchaos-based encryption algorithm, provides the same level of IND-CPA security as full encryption of the entire image, while also reducing the encryption time.*

Proof. In full encryption game, where the adversary $\mathcal{A}$ submits two images I_{m0} and I_{m1}, and receives $E(I_{mb})$ for a randomly chosen bit b. In targeted encryption game, where $\mathcal{A}$ submits two images (R_{m0}, NS_{m0}) and (R_{m1}, NS_{m1}), and receives $(E(R_{mb}), NS_{mb})$ for a randomly chosen bit b. Suppose there exists an adversary $\mathcal{A}$ that can distinguish between $(E(R_{m0}), NS_{m0})$ and $(E(R_{m1}), NS_{m1})$ with non-negligible advantage ϕ. We will construct an adversary $\mathcal{B}$ that breaks the IND-CPA security of E. Adversary $\mathcal{B}$ receives $E(R_{mb})$ for chosen plaintext regions R_{m0} and R_{m1} and constructs $(E(R_{mb}), NS_{mb})$ and provides it to $\mathcal{A}$. If $\mathcal{A}$ can distinguish $(E(R_{m0}), NS_{m0})$ from $(E(R_{m1}), NS_{m1})$ with advantage ϕ, then $\mathcal{B}$ breaks the IND-CPA security of E with the same advantage. Since E is IND-CPA secure, the advantage is negligible. Therefore, targeted encryption of the sensitive region R_m maintains IND-CPA security.

Table 3
Dataset statistics.

Dataset	Categories				Usage		
					Train	Test	Retrieval
BT-MRI	Glioma 1621	Meningioma 1645	No tumor 2000	Pituitary 1756	5700	1322	1000
C-XR	Atelectasis 3000	Cardiomegaly 3000	Infiltration 3000	Effusion 3000	9600	2400	1000
K-CT	Cyst 3709	Normal 5077	Stone 1377	Tumor 2283	10 000	2466	1000

In the full encryption method, Let the size of the medical image I_m be S_m and the time complexity for encrypting the entire image using E is $T_E(S_m)$. In the targeted encryption method, Let the size of the sensitive region R_m be S_r (where $S_r < S_m$) and the time complexity for encrypting the region R_m using E is $T_E(S_r)$. Since $S_r < S_m$, it follows that $T_E(S_r) < T_E(S_m)$.

Targeted encryption of the sensitive region achieves the same level of IND-CPA security as full encryption of the entire image because the reduction proof 1 shows that an adversary cannot distinguish between targeted and full encryption if the encryption algorithm is IND-CPA secure. Additionally, targeted encryption reduces the encryption time, as it only encrypts the sensitive region, which is smaller in size compared to the entire image.

6. Experimental results and performance analysis

This section is dedicated to analyzing the experiments conducted to showcase the outcomes of the proposed ESecMedIR, focusing on its efficient retrieval performance. The structure of this section is delineated as follows: an explanation of the dataset with details on the experimental setup and results, an analysis of ROP detection accuracy, a comparative assessment of DViTH retrieval accuracy, ablation studies, an analysis of retrieval time, and concluding with the visualization of hashcode distribution.

6.1. Datasets and experimental results

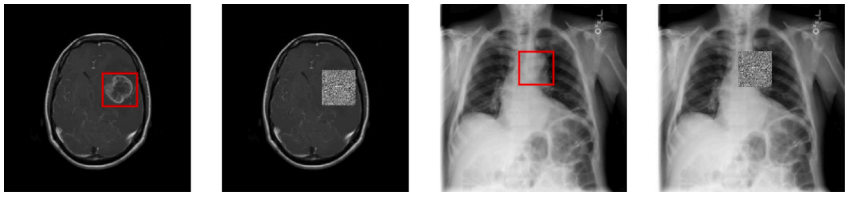
To experiment and evaluate the proposed retrieval framework, three different sensitive medical image datasets are chosen. Details of the dataset have been explained briefly here,

- **Brain Tumor MRI Dataset (BT-MRI)** [44]: The three datasets below are combined to create this dataset: figshare, SARTAJ, Br35H. There are 7023 MRI images of the human brain in this collection, divided into 4 categories: pituitary, glioma, meningioma, and no tumor. Images categorized as the 'no tumor' class were obtained from the Br35H dataset.
- **Chest X-ray Dataset (C-XR)** [45]: This dataset comes from the National Institutes of Health (NIH), which is currently the largest chest radiograph data set, containing 112,120 frontal X-ray images from 30,805 special patients. Each image was attached to the associated text disease label, which was mined from the relevant radiological reports using a natural language processing algorithm. To build the database, we selected four common and numerous thoracic disease labels, including Atelectasis, Cardiomegaly, Infiltration, and Effusion.
- **Kidney CT Dataset (K-CT)** [46]: The images were gathered from PACS (Picture Archiving and Communication System) records in several hospitals located in Dhaka, Bangladesh. These records belonged to patients who had previously received a diagnosis of kidney tumor, cyst, normal, or stone findings. The urogram and whole abdominal protocols were followed in the selection of the coronal and axial cuts from contrast and non-contrast studies. 12,446 distinct units of data make up the generated dataset.

The selection of X-ray, CT, and MRI images for the retrieval task aims to demonstrate the versatility and effectiveness of our model across various imaging modalities, showcasing its applicability and robustness in diverse clinical scenarios. Table 3 shows the statistics of the chosen datasets. The datasets are split into a training and testing ratio of 80:20, and retrieval accuracy is evaluated on a randomly selected subset of 1000 images for testing. The developed system has been executed in a computer with 64-bit Windows OS, 64 GB RAM, an NVIDIA Quadro P5000 GPU with 16 GB of memory, and an Intel Xeon processor. The overall framework has been developed with the help of Keras, and OpenCV libraries in Python.

The proposed framework is designed and tested with medical datasets, where medical images undergo ROP encryption for confidentiality, as illustrated in Fig. 7. To retrieve similar medical images, a query image is input into the system. Hashcodes are computed using the proposed DViTH and searched in the B+ tree-based storage. The search algorithm returns the top-k similar medical images based on the hamming distance similarity metric from the cloud, which are then decrypted at the client side.

The computational complexity of the proposed framework is primarily influenced by its transformer-based architectures, GCViT and LRViT, which utilize self-attention mechanisms with a complexity of $O(N^2d)$ per image, where N is the number of tokens and d is the feature dimension. In contrast, traditional CNN-based hashing methods exhibit lower complexity, typically $O(Nd)$, due to their convolutional operations, which process local receptive fields efficiently without requiring pairwise token interactions. Additionally, hashcode generation and retrieval involve efficient indexing mechanisms, while the encryption process introduces



(a) ROP detected brain MRI (b) ROP encrypted brain MRI (c) ROP detected Chest X-Ray (d) ROP encrypted Chest X-Ray

Fig. 7. Sample ROP encrypted images: The detected sensitive ROPs of (a) and (c) are encrypted in (b) and (d) respectively.

Table 4
ROP detection accuracy result analysis.

ROP detection models	Detection accuracy (%) (↑)		
	BT-MRI	C-XR	K-CT
R-FCN [47]	75.63	69.39	72.81
Faster R-CNN [48]	84.23	79.49	82.15
YOLOv5 [49]	89.86	82.60	87.94
YOLOv8 [50]	**96.65**	**90.50**	**95.61**

additional computational overhead. Despite the higher complexity of the transformer-based model, our method achieves better feature representations and retrieval accuracy, as demonstrated in our experimental results.

Sample retrieved images for each query image from different datasets are presented in Fig. 8. Column 1 indicates the dataset of the selected query image, while column 2 displays query images from each class. The remaining columns exhibit the retrieved decrypted images and the heat map of the retrieved images. Subsequent sections provide a detailed analysis of the retrieval performance of the proposed system.

6.2. ROP detection accuracy analysis

Our hashing technique incorporates both local and global features in hashcode generation, emphasizing the significance of ROP detection accuracy, which corresponds to localization accuracy, in achieving optimal retrieval accuracy. Table 4 presents the results of ROP detection using pre-trained CNN architectures for three distinct datasets, organized in ascending order based on accuracy performance metrics from the least to the highest. YOLOv8 performs better than other detection models.

The differences in YOLOv8's performance for ROP detection in Table 4 are due to the unique traits of each imaging modality. Brain MRI and kidney CT scans, which have distinct structures and precise boundaries, aid in precise detection. On the other hand, chest X-ray images present difficulties because of unclear borders and structures that overlap, leading to decreased accuracy in detection. The YOLOv8 model has been trained with medical datasets, loss graphs are shown in Appendix C. The precision has a direct influence on the creation of hashcodes for ROPs, considered as local characteristics, and therefore influences retrieval, as explained in the following subsection.

6.3. Image retrieval performance analysis

In order to evaluate our proposed DViTH which is a medical image hashing model, we have selected two metrics Mean Average Precision, and PR Curve (ROC) Analysis. The proposed method has been tested with three different medical datasets and compared with six previous state-of-the-art deep hashing methods IDHN [51], UDTH [33], DCNN [52], PMIH [26], IFFH [35] and VTH [53] as baseline models. Top-k retrieved images are used to estimate the retrieval accuracy. Image retrieval accuracy can be measured using Precision (Pr), and Recall (Re). Precision refers to the ratio of relevant retrieved images to the total number of retrieved images (k) in relation to the query image.

$$Pr@k = \frac{|\text{relevant images} \cap \text{retrieved images}|}{|\text{retrieved images}|} \tag{24}$$

Recall denotes the proportion of relevant retrieved images to the query image, considering the number of identical images in the entire dataset.

$$Re@k = \frac{|\text{relevant images} \cap \text{retrieved images}|}{|\text{total relevant images in the dataset}|} \tag{25}$$

Mean Average Precision (mAP) is the standard measure for assessing and comparing the accuracy of image retrieval. The calculation of mAP can be performed using Eq. (26) below.

$$mAP = \frac{1}{R}\sum_{m=1}^{R}\left(\frac{1}{q_m}\sum_{n=1}^{q_m} Pr_n\right) \tag{26}$$

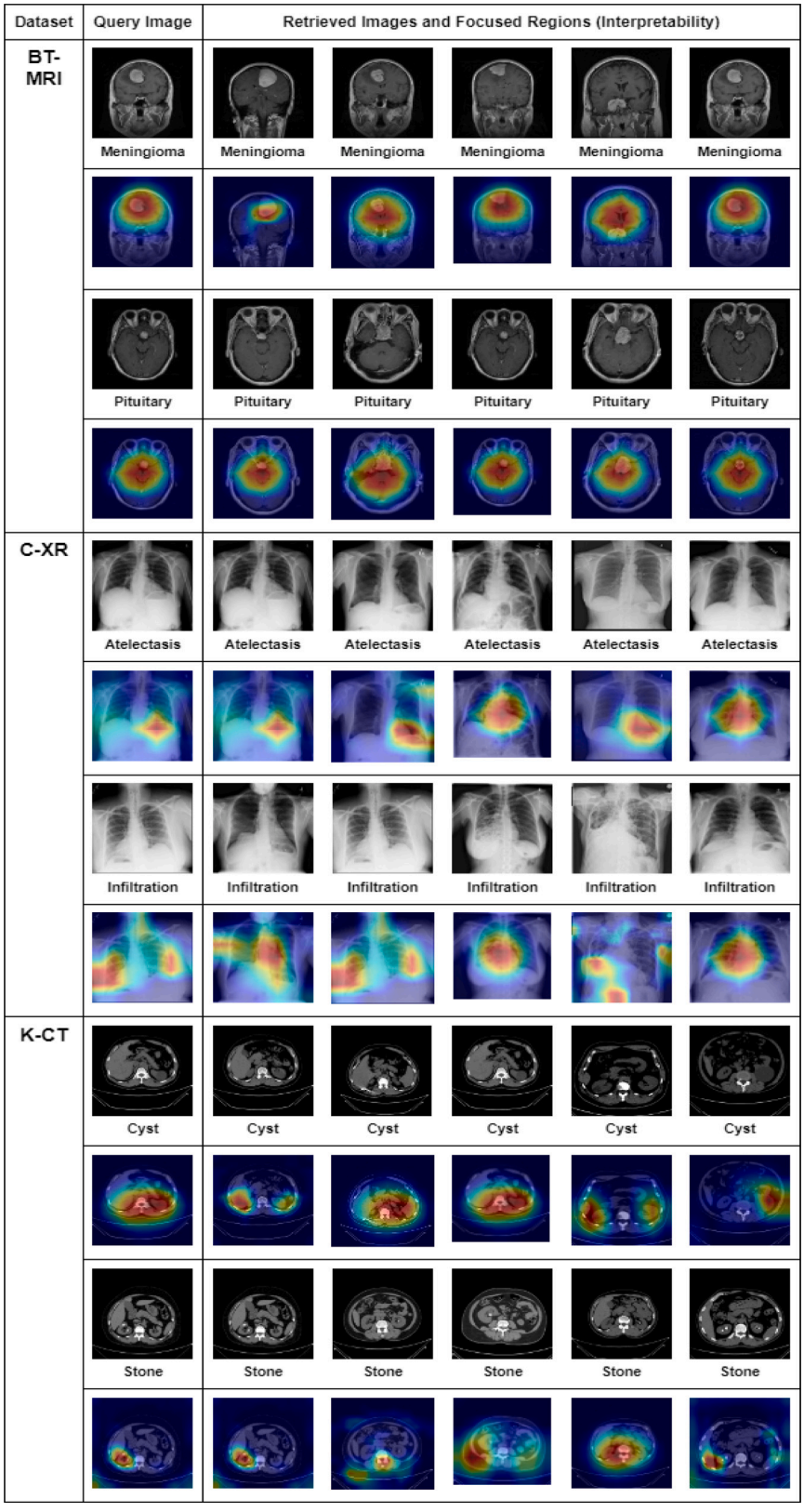
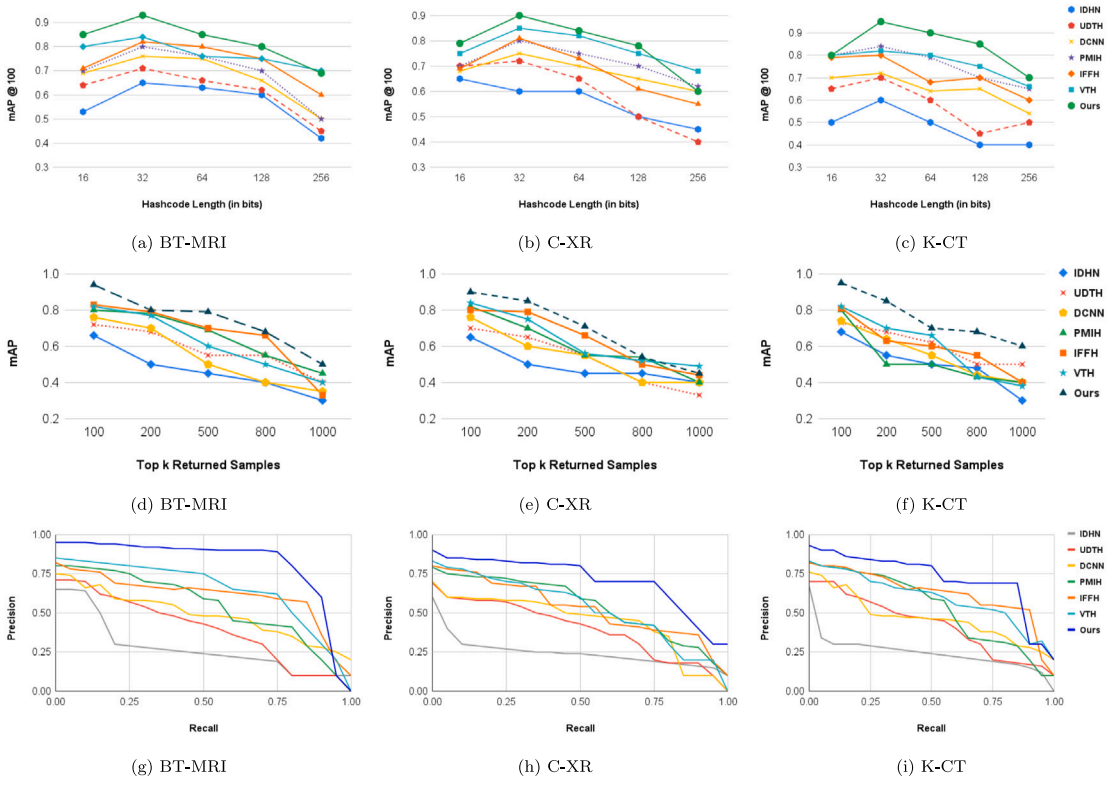


Fig. 8. Sample retrieval results: Query image is in column 2 and top-5 retrieved results for that query in subsequent columns.

Table 5
Detailed retrieval performance comparison across three datasets, evaluated by mAP under varying conditions. Baseline results were reproduced to ensure consistent benchmarking.

Hashing models		IDHN [51]			UDTH [33]			DCNN [52]			PMIH [26]			IFFH [35]			VTH [53]			Ours (↑)		
Hashcode length		16b	32b	64b	16b	32b	64b	16b	32b	64b	16b	32b	64b	16b	32b	64b	16b	32b	64b	16b	32b	64b
Medical dataset	k																					
BT-MRI	100	**0.53**	**0.65**	**0.59**	0.64	**0.73**	**0.67**	**0.68**	**0.77**	**0.74**	**0.70**	**0.78**	**0.75**	**0.76**	**0.82**	**0.79**	**0.80**	**0.85**	0.78	**0.85**	**0.94**	**0.87**
	200	0.43	0.54	0.50	**0.65**	0.68	0.62	0.63	0.71	0.67	0.67	0.73	0.70	0.73	0.79	0.74	0.75	0.78	**0.79**	0.81	0.86	0.83
	500	0.39	0.47	0.43	0.52	0.55	0.54	0.51	0.57	0.59	0.62	0.69	0.68	0.66	0.71	0.65	0.62	0.63	0.64	0.75	0.79	0.77
	800	0.38	0.41	0.39	0.43	0.53	0.41	0.43	0.49	0.51	0.54	0.56	0.54	0.62	0.66	0.64	0.59	0.59	0.53	0.69	0.71	0.72
C-XR	100	**0.65**	**0.61**	**0.62**	**0.70**	**0.72**	**0.65**	**0.69**	**0.74**	**0.71**	**0.73**	**0.81**	**0.79**	0.69	0.82	**0.74**	**0.73**	**0.83**	**0.80**	**0.81**	**0.90**	**0.84**
	200	0.57	0.53	0.51	0.62	0.63	0.57	0.62	0.63	0.64	0.65	0.76	0.77	**0.70**	**0.83**	0.68	0.67	0.75	0.72	0.78	0.85	0.79
	500	0.44	0.45	0.43	0.55	0.57	0.42	0.51	0.52	0.50	0.60	0.61	0.63	0.59	0.66	0.60	0.55	0.56	0.54	0.66	0.73	0.65
	800	0.32	0.42	0.39	0.48	0.49	0.38	0.40	0.43	0.42	0.50	0.52	0.53	0.52	0.59	0.58	0.45	0.52	0.50	0.53	0.60	0.55
K-CT	100	**0.55**	**0.63**	**0.52**	**0.65**	**0.71**	**0.69**	**0.72**	**0.75**	**0.65**	**0.76**	**0.83**	**0.79**	**0.77**	**0.81**	0.69	**0.81**	**0.84**	**0.80**	**0.84**	**0.95**	**0.89**
	200	0.43	0.59	0.51	0.60	0.68	0.62	0.64	0.68	0.67	0.68	0.70	0.72	0.64	0.68	**0.70**	0.69	0.71	0.70	0.82	0.88	0.83
	500	0.39	0.50	0.46	0.52	0.62	0.50	0.57	0.59	0.56	0.55	0.64	0.63	0.53	0.60	0.63	0.59	0.64	0.63	0.68	0.75	0.72
	800	0.34	0.48	0.39	0.50	0.58	0.45	0.44	0.43	0.43	0.53	0.55	0.57	0.50	0.53	0.48	0.55	0.59	0.57	0.59	0.69	0.63



(a) BT-MRI (b) C-XR (c) K-CT

(d) BT-MRI (e) C-XR (f) K-CT

(g) BT-MRI (h) C-XR (i) K-CT

Fig. 9. (a–c) hashcode length vs. mAP@100, (d–f) Top k retrieved results vs. mAP (mAP@k Curves), (g–i) PR Curves for all datasets.

where R is a number of queries, q_m is a number of relevant images for query n and the Pr_n is precision at nth relevant image. The implementation details of all these baseline models are detailed in Appendix D.

We randomly selected 1000 samples from each medical dataset for the analysis of retrieval accuracy. The experimental results, including mAP values across three datasets with varying hashcode lengths and different values of k, are documented in Table 5. Our method achieves improvements of 5%, 7.2%, and 8.9% for hashcode lengths of 16, 32, and 64 bits, respectively, compared to the best-performing method, VTH, on the BT-XRI dataset. On the K-CT dataset, our method outperforms other techniques, such as PMIH and IFFH, by 13% and 15%, respectively, for 32-bit hashcodes. For 16-bit binary codes, there is a notable improvement margin of 16.7%, while for 32 bits and 64 bits, the margins are 17.2% and 18.6%, respectively, when compared with all baselines. Figs.

Table 6
Retrieval accuracy mAP@100: Comparison with baseline models.

Model name	Backbone	mAP@100 (↑)		
		BT-MRI	C-XR	K-CT
IDHN [51]	AlexNet	0.65	0.61	0.63
UDTH [33]	VGG16	0.73	0.72	0.70
DCNN [52]	AlexNet	0.77	0.75	0.75
PMIH [26]	DenseNet 121	0.78	0.81	0.83
IFFH [35]	DenseNet 201	0.82	0.83	0.81
VTH [53]	Vanilla ViT	0.85	0.83	0.88
Ours	Dual ViT Hash Fusion	**0.94**	**0.90**	**0.95**

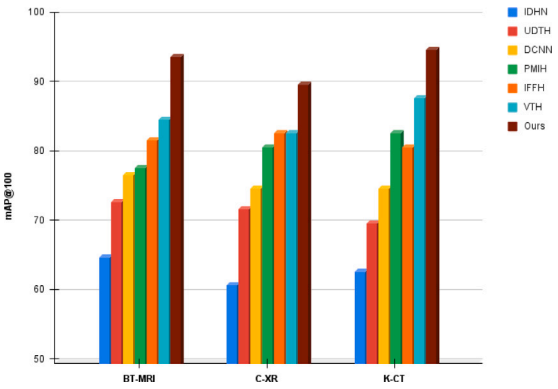


Fig. 10. Comparative analysis with existing hashing models.

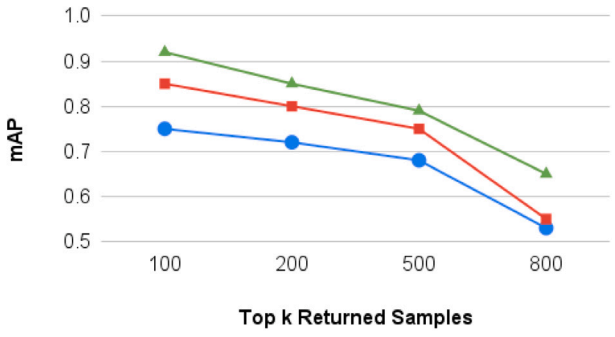
9(a)–9(c) illustrate the impact of hashcode length on image retrieval accuracy using mAP@100. Our proposed method consistently outperforms baseline models across all three medical datasets for hashcode lengths of 16, 32, 64, 128, and 256 bits. The mAP reduction with increasing bit lengths is because of a growing semantic gap between hashcodes, making hamming distances more sensitive and leading to less discriminative retrieval. This results in similar hashcodes for dissimilar images, influencing precision. Striking a balance between hashcode length and discriminative power is essential for high retrieval accuracy. Specifically, the 16-bit local and 16-bit global hashcodes, forming a 32-bit fused hashcode, achieve the best retrieval performance.

To further showcase the efficacy of our approach, we generate mAP@k and Precision-Recall (PR) curves across three datasets. The hashcode length is consistently set at 32 bits, ensuring superior retrieval accuracy across all datasets. Figs. 9(d) and 9(f) depicts the mAP@k curve for all datasets, revealing that our DViTH consistently achieves higher precision than other methods across all three datasets, particularly in the case of K-CT. In the C-XR dataset, our curves slightly surpass those of IFFH, especially when the number of retrieved images is less than 500. A decrease in mAP with higher k values in top-k retrieval indicates that as more images are retrieved, precision drops. This may be due to the spread of relevant items, challenges in distinguishing relevant from irrelevant images, or inherent traits of the medical image data.

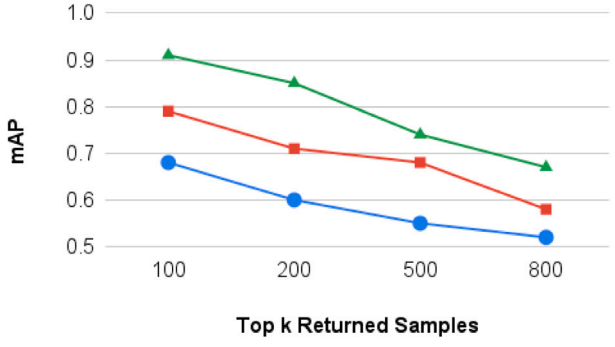
The Precision-Recall (PR) curve serves as a crucial metric for comparing the proposed methods with baseline approaches. A superior area under the PR curve typically signifies a more efficient retrieval system adept at balancing precision and recall. As depicted in Figs. 9(g)–9(i), all the PR curves of our method consistently surpass those of other methods. This positioning indicates that, at any given radius (Hamming distance), our method achieves higher precision and recall rates compared to alternative approaches. A comprehensive analysis across three medical datasets and six baseline models shows that our proposed model consistently outperforms others in retrieval accuracy, as supported by Fig. 10.

[1]The summary of the retrieval accuracy is tabulated in Table 6. The table indicates that MRI and CT images exhibit better retrieval accuracy compared to X-ray images. This discrepancy is attributed to the distinctive characteristics of X-ray, MRI, and CT medical images in terms of complex features [54]. X-ray images focus on density variations, limiting their suitability for deep hashing, while MRI provides high-resolution soft tissue details, enhancing feature extraction. CT images offer a balance between bone and tissue detail. MRI and CT outperform X-rays in deep hashing-based retrieval due to their richer, more diverse feature sets, enabling more accurate representation.

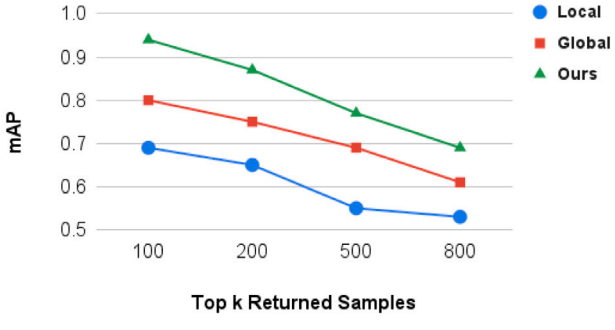
[1] Baseline results were reproduced by the authors.



(a) BT-MRI



(b) C-XR



(c) K-CT

Fig. 11. Ablation study: Contribution of local and global hashcodes - Fused hashcode provides better retrieval performance than individual local and global hashcodes.

6.4. Ablation study

We have conducted an ablation study on our proposed model to analyze how the individual contribution of each model parameter. Conducting an ablation study on the loss components, including central similarity loss (L_{csl}), quantization loss (L_{quant}), and bit balance loss (L_{bit}), is crucial to understanding their individual contributions to the overall retrieval performance. The L_{csl} ensures that similar images are projected closer to the hash center, contributing to the discriminative power of the generated hashcodes. L_{quant} enforces the hashcodes to be discrete, aligning them with the binary nature of hash functions. L_{bit}, on the other hand, maintains a balance in the number of ones and zeros in the hashcodes, further refining their discriminative nature.

This comprehensive analysis enables optimizing the trade-off between losses, achieving the best-performing hashcodes. From the study, L_{csl} significantly contributes to 75%–80% of the retrieval performance, with the remaining portion attributed to the other losses. Table 7 shows the detailed ablation study where mAP@100 and a hashcode length is 32 bits have been captured.

We investigated how localization accuracy affects retrieval performance. Hashcodes based solely on global features capture the overall characteristics but may miss intricate details and variations in medical images, making it harder to distinguish visually

Table 7
Ablation study: Contribution of loss functions.

Components	Retrieval accuracy (mAP) (↑)											
	BT-MRI				C-XR				K-CT			
	k = 100	k = 200	k = 500	k = 800	k = 100	k = 200	k = 500	k = 800	k = 100	k = 200	k = 500	k = 800
L_{csl}	0.76	0.66	0.62	0.55	0.72	0.65	0.56	0.44	0.77	0.70	0.58	0.54
$L_{csl} + L_{quant}$	0.91	0.78	0.74	0.62	0.86	0.76	0.66	0.53	0.92	0.84	0.68	0.65
$L_{csl} + L_{bit}$	0.82	0.69	0.66	0.56	0.77	0.68	0.59	0.48	0.83	0.75	0.62	0.58
$L_{csl} + L_{quant} + L_{bit}$	**0.94**	**0.82**	**0.78**	**0.65**	**0.9**	**0.81**	**0.69**	**0.56**	**0.95**	**0.88**	**0.73**	**0.68**

Table 8
Retrieval time analysis.

Database type	Dataset	Dataset size	Retrieval time (s) (↓)		
			Euclidian	*Manhattan*	*Hamming*
Linear database	BT-MRI	500	55.89	50.62	48.91
		1000	108.92	101.88	98.11
		1500	159.34	152.33	151.03
		2000	220.42	211.82	202.76
	C-XR	500	53.48	51.78	49.01
		1000	107.77	101.84	99.67
		1500	156.05	151.33	150.49
		2000	211.78	209.89	201.98
	K-CT	500	56.54	51.19	50.81
		1000	109.71	102.70	100.87
		1500	158.67	153.43	152.11
		2000	217.83	214.78	204.23
B+ tree-based database	BT-MRI	500	9.03	9.17	**8.96**
		1000	11.86	10.92	**9.95**
		1500	12.49	11.26	**10.63**
		2000	14.28	12.86	**11.05**
	C-XR	500	9.94	9.84	**9.04**
		1000	11.96	11.03	**10.11**
		1500	13.02	12.22	**11.32**
		2000	14.98	12.93	**11.89**
	K-CT	500	9.73	9.47	**8.59**
		1000	11.36	11.82	**10.53**
		1500	12.99	12.96	**10.94**
		2000	14.73	13.08	**11.09**

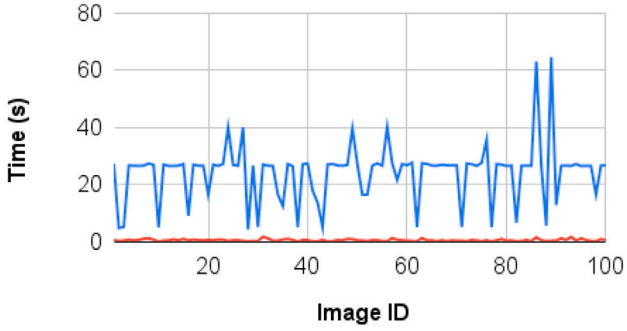
similar yet clinically different images. Conversely, hashcodes based only on local features excel in detailing specific regions but lack global context, potentially missing important patterns. Integrating both global and local features in hashcode generation provides a comprehensive representation, enhancing retrieval accuracy. Ablation studies reveal the contributions of each feature type, and experiments comparing mAP@k with global versus local features show that both contribute to overall performance. This insightful ablation study is visually presented in Fig. 11.

6.5. Retrieval time analysis

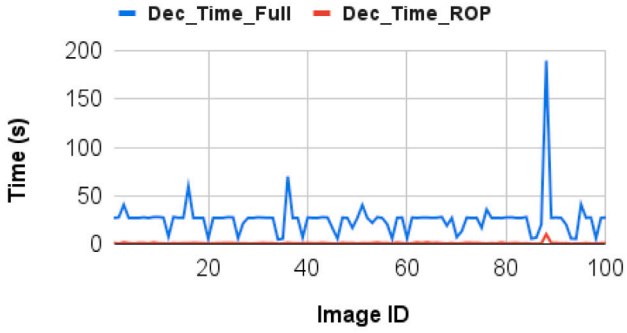
Retrieval time is crucial in secure CBMIR, balancing accuracy, security, and efficiency. Our approach reduces retrieval time by encrypting only the ROP in medical images. A comparison of fully encrypted and ROP-encrypted image databases revealed that the average decryption time for ROP-encrypted images is four times shorter. Fig. 12 shows that the maximum decryption time for ROP is lower than the minimum for fully encrypted images. We use a B+ tree-based structure for hashcodes and ROP-encrypted images, with hamming distance as the similarity metric, outperforming other methods using linear structures and Euclidean or Manhattan distances. In this experiment, with k = 100 and ROP encryption, our system achieves reduced retrieval time, ensuring both the security of medical images and optimal CBMIR performance as shown in Table 8.

6.6. Visualization

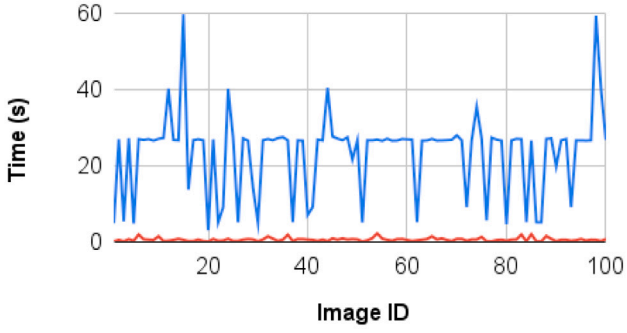
Visualizing the generated hashcodes provides valuable intuition into the hashcode generation method we employ. So, To interpret the discriminative capabilities of our model's hashcodes compared to other methods using t-SNE plots, we can visualize the clustering patterns. Our model aims to produce distinct and well-separated clusters, indicating that similar images are closer together in the hamming space. This distinctiveness implies that our deep hashing techniques effectively capture the intrinsic features of the images, leading to more discriminative hashcodes. In contrast, other baseline models exhibit overlapping or scattered clusters, reflecting a



(a) BT-MRI



(b) C-XR



(c) K-CT

Fig. 12. Decryption time analysis: ROP decryption takes less time than Full image decryption.

reduced ability to represent the inherent similarities between images. Analyzing t-SNE plots reveals our model's discriminative nature in generating hashcodes that preserve essential data structure and relationships compared to other methods. Fig. 13 displays t-SNE plots for six baseline models on all 3 datasets.

7. Conclusion

In this paper, we proposed the ESecMedIR framework that addresses the key challenges of secure and efficient medical image storage and retrieval management. To do that, we integrated Dual-level Vison Transformer-based Hashing (DViTH) with Region of Privacy (ROP) Encryption. Our framework secures the sensitive regions through hyperchaos-based image encryption. The DViTH ensures a comprehensive retrieval system by integrating global features for contextual understanding and capturing detailed

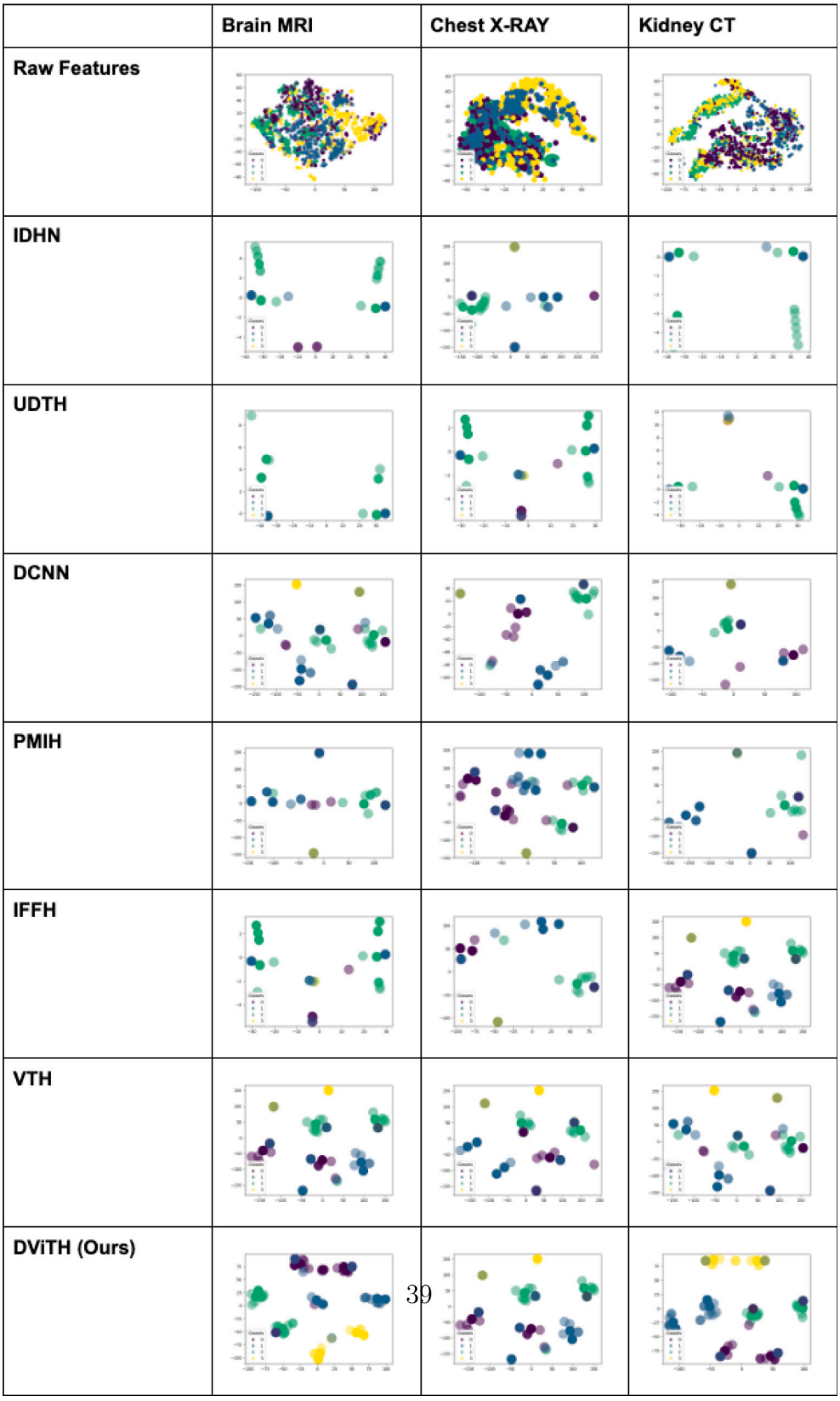


Fig. 13. Visualizing T-SNE plots reveal our model's highly discriminative hashcodes, forming distinct clusters, in contrast to less effective representations by baseline models.

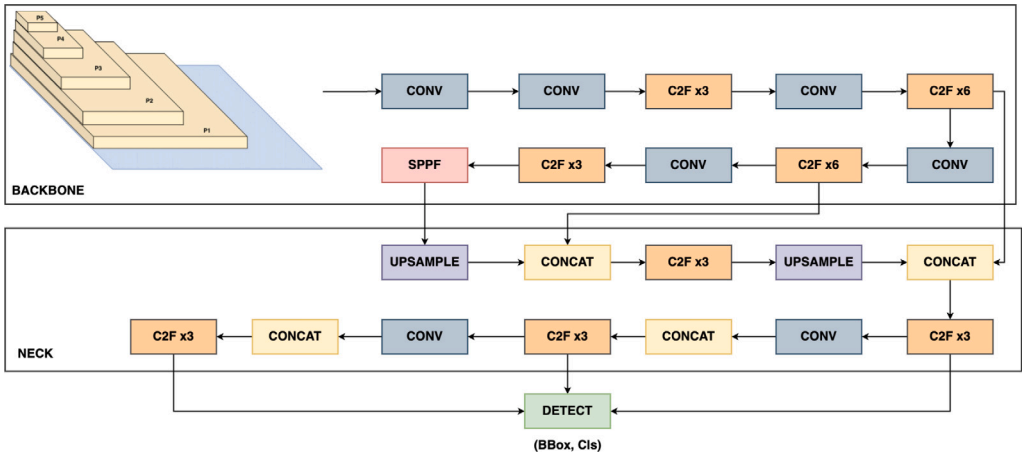


Fig. A.14. YOLOv8 architecture.

information with ROPs. Rigorous tests are conducted on three medical datasets which demonstrate that ESecMedIR achieves 10%–30% improvement in retrieval accuracy. ROP encryption contributes to reduced retrieval times, and a tree-based storage system further streamlines retrieval. Future work could explore additional encryption techniques, assess adaptability to diverse imaging modalities, and incorporate advanced learning models for continual improvement and real-world applications.

CRediT authorship contribution statement

Arun Amaithi Rajan: Conceptualization, Methodology and Development, Formal analysis and investigation, Writing – original draft, Writing – review & editing. **Vetriselvi V.:** Conceptualization, Formal analysis and investigation, Writing – original draft, Writing – review & editing, Supervision. **Mayank Raikwar:** Conceptualization, Formal analysis and investigation, Writing – original draft, Writing – review & editing. **Mohamed Fuzail H.:** Conceptualization, Methodology and Development, Formal analysis and investigation, Writing – original draft, Writing – review & editing.

Funding Information

The authors did not receive support from any organization for the submitted work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. ROP detection using YOLOv8

Fig. A.14 illustrates the YOLOv8 architecture, which includes a backbone and neck. The backbone comprises components like C2f and SPPF (Spatial Pyramid Pooling Fast) modules, optimizing feature extraction and accuracy across various scales. The C2f module enhances the capacity for feature extraction and computation efficiency through a hierarchical merge and base layer feature map division. SPPF improves context and spatial information capture by extracting features across multiple scales. The C3-CONV stage boosts accuracy and reduces computing load by integrating convolution, batch normalization, and activation functions. YOLOv8 strengthens backbone performance by integrating C2f modules with C3-CONV inspiration, promoting effective feature extraction. The neck module, incorporating PAN-FPN for feature fusion, enhances precision by segregating the head. C2f and SPPF modules elevate detection accuracy through contextual information and broaden the network's receptive field with pooling operations. The distinctive C3-CONV operation and mosaic augmentation, provide superior detection accuracy and generalizability compared to other models prompting us to choose the YOLOv8 model for ROP detection in medical images [50]. The YOLOv8 architecture combines feature extraction, feature fusion, and prediction stages to enhance region identification. Fig. A.15 shows the ROP detection results for each class within the three distinct medical datasets.

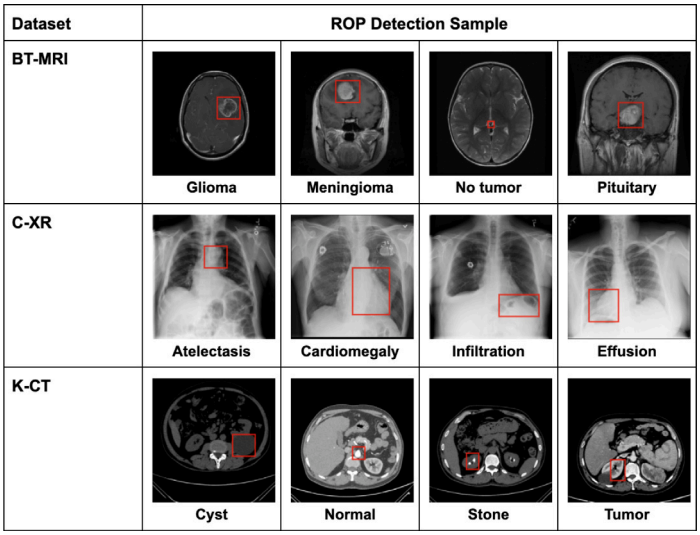


Fig. A.15. ROP Detection: Every row illustrates sample ROP detection instances from the specified medical dataset classes presented in column 1.

Appendix B. Security analysis

Amaithi Rajan et al. [41] proposed a novel image encryption model for medical images using hyperchaos and bitplane scrambling and did extensive attack analysis. There are multiple cryptographic attack resistance analyses such as chosen-plaintext attacks, differential attacks, histogram attacks, statistical attacks, and brute-force attacks done in that article, which has been carried out on our selected 3 samples from medical datasets. Then it has been used in our ESecMedIR. We have taken 3 medical images M1, M2, M3 which are BT-MRI, C-XR, and K-CT respectively. They are the ROP extracted and then encrypted using hyperchaos-based image encryption. Fig. B.16 shows the selected medical images and their ROP encrypted images.

B.1. Histogram examination

The histogram showed the image's statistical properties. The histogram of an image was flat when the values of each pixel were almost identical. This demonstrated how strong its defenses against statistical attacks were. The plaintext and ciphertext histograms of sample medical images are shown in Fig. B.17. It shields the data against attackers using histograms.

B.2. Correlation analysis

Image correlation is an indicator for evaluating the randomness of an image. The plaintext images adjacent pixels had a strong correlation. Randomness between the pixels could be increased by encrypting the plaintext image. The following Eqs. (B.1) to (B.4) help to derive the correlation coefficient of an image. There are 3 correlation coefficients Horizontal, Vertical, and Diagonal in image analysis.

$$E(p) = \frac{1}{n}\Sigma_{i=1}^{n} p_i \tag{B.1}$$

$$D(p) = \frac{1}{n}\Sigma_{i=1}^{n} (p_i - E(p))^2 \tag{B.2}$$

$$cov(p,q) = \frac{1}{n}\Sigma_{i=1}^{n} (p_i - E(p))(q_i - E(q)) \tag{B.3}$$

$$\rho_{pq} = \frac{cov(p,q)}{D(p)D(q)} \tag{B.4}$$

Table B.9 shows the H, V, and D correlation values of the selected images. Horiz_Corr, Vert_Corr, Diag_Corr represent horizontal, vertical, and diagonal correlation coefficients, respectively. Sub figures in Table B.10 clearly explain the encrypted image M1, M2, and M3's randomness in 3 directions.

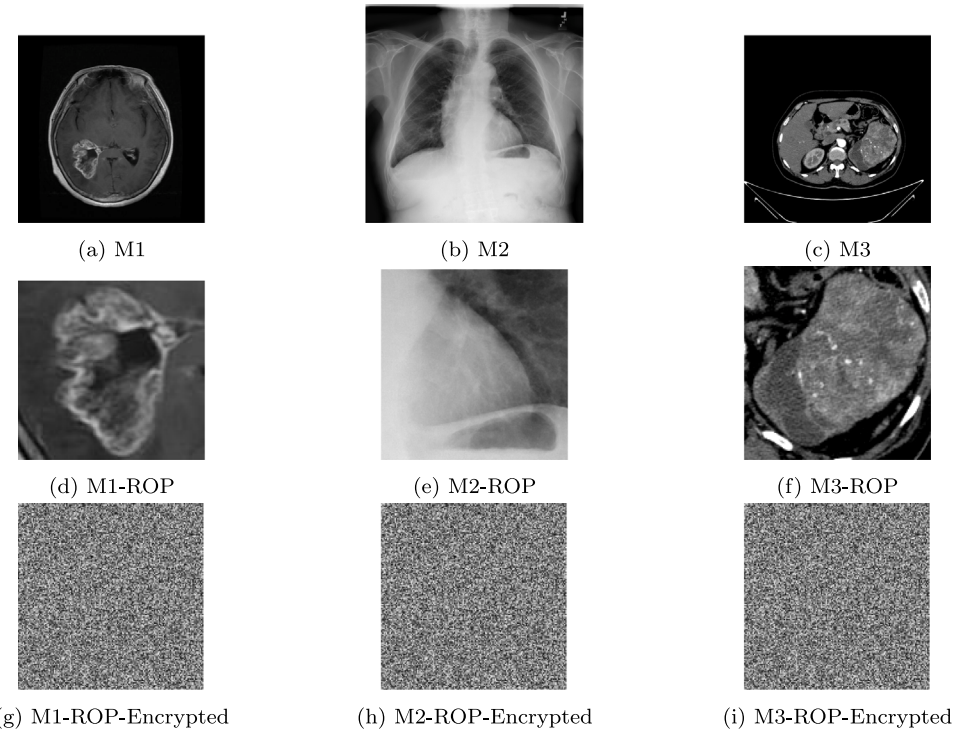


Fig. B.16. Sample Images and Encryption: (a)–(c) are the selected medical image from each dataset, (d)–(f) are the ROPs of selected images, (g)–(i) are the encrypted ROPs.

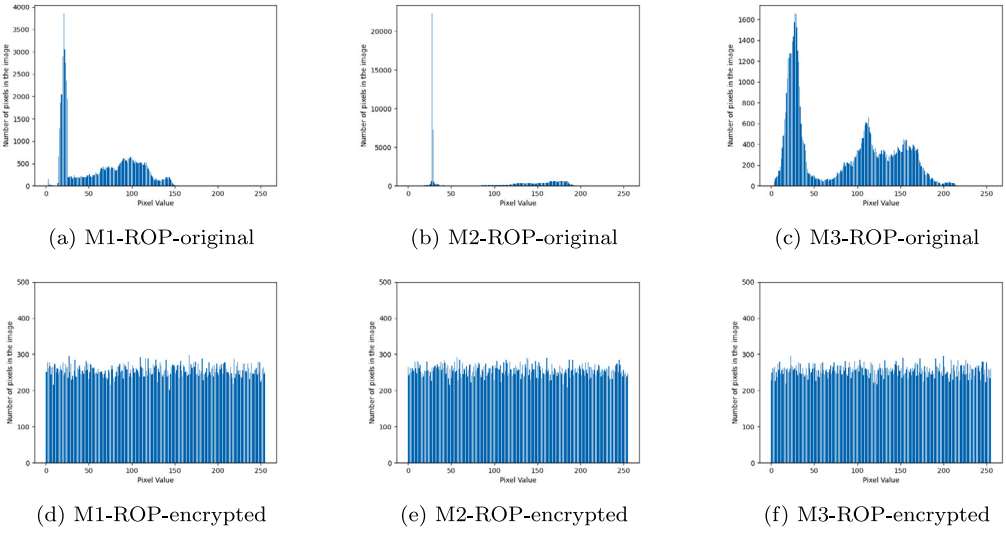


Fig. B.17. Histogram analysis.

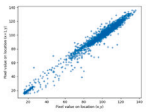
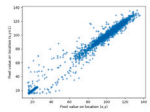
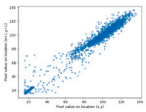
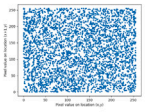
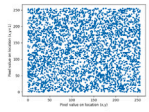
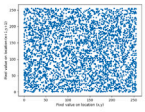
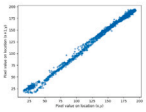
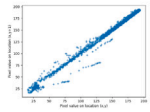
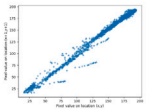
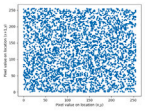
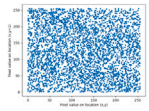
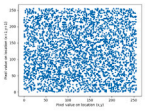
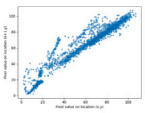
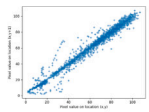
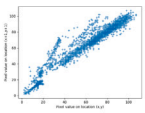
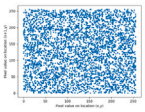
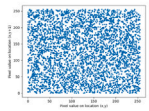
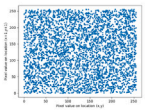
B.3. Differential attack analysis

Utilizing a differential attack, it is possible to determine how sensitive the encryption technique is to even the smallest alterations in the plain image. The Unified Average Change in Intensity (UACI) and Number of Pixel Change Rate (NPCR) are the parameters utilized to assess how well the proposed technique thwarts differential attacks. Using the following Eqs. (B.5) and (B.6), we can

Table B.9
Correlation coefficient and entropy analysis.

Image	Horiz_Corr		Vert_Corr		Diag_Corr		Entropy	
	Original	Encrypted	Original	Encrypted	Original	Encrypted	Original	Encrypted
M1	0.9353	−0.0146	0.7821	−0.0324	0.8952	−0.0045	5.9183	7.9869
M2	0.9919	−0.0033	0.9952	−0.0098	0.9539	0.0034	4.4784	7.9951
M3	0.9903	0.0079	0.9587	−0.0087	0.8873	0.0056	5.6754	7.9981

Table B.10
Correlation analysis of M1, M2, M3.

Image	Horizontal_Corr	Vertical_Corr	Diagonal_Corr
 M1-ROP			
 M1-ROP-Encrypted			
 M2-ROP			
 M2-ROP-Encrypted			
 M3-ROP			
 M3-ROP-Encrypted			

determine NPCR.

$$NPCR = \frac{\Sigma_{l,m} Diff(l,m)}{w \times h} \times 100\% \tag{B.5}$$

Here,

$$Diff(l,m) = \begin{cases} 1, & \text{if M(l,m) equals C(l,m).} \\ 0, & \text{if M(l,m) not equals C(l,m).} \end{cases} \tag{B.6}$$

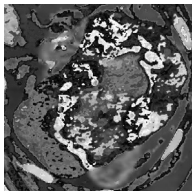
The letters w and h stand for the image's width and height, respectively. The value of $Diff(l,m)$ represents the difference between the corresponding pixels of the original image ($M(l,m)$) and the encrypted image ($C(l,m)$). UACI is the average difference in pixel intensity between the test and the encrypted image. It is yet another often-employed efficiency indicator for evaluating the capacity to resist a differential attack. The UACI value is determined using Eq. (B.7).

$$UACI = \frac{\Sigma_{u,v} M(u,v) - C(u,v)}{255 \times w \times h} \times 100 \tag{B.7}$$

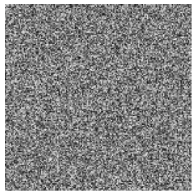
Table B.11 lists the UACI and NPCR of the selected medical images in the proposed system.

Table B.11
Differential attack analysis.

Image	UACI	NPCR
M1	33.72	99.49
M2	33.93	99.87
M3	32.95	99.53

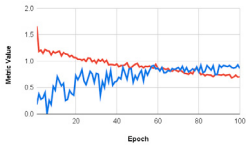


(a) $M_1(l,m) \oplus M_2(l,m)$



(b) $Enc_1(l,m) \oplus Enc_2(l,m)$

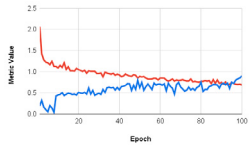
Fig. B.18. CP attack simulation.



(a) BT-MRI



(b) C-XR



(c) K-CT

Fig. C.19. ROP Detection: Training loss and detection accuracy - Brain tumor detection has 96.65% detection accuracy. The chest X-ray dataset gives 90.50% and the kidney dataset gives 95.61% accuracies.

B.4. Chosen Plaintext (CP) attack simulation

The suggested approach uses an XOR operation to diffuse the values of the pixels. Eq. (B.8) is utilized to do the CP Attack analysis.

$$M_1(l,m) \oplus M_2(l,m) = Enc_1(l,m) \oplus Enc_2(l,m) \tag{B.8}$$

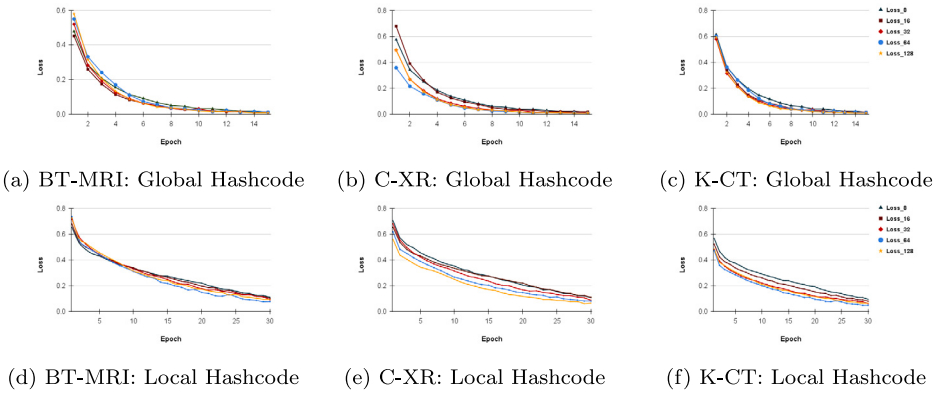
where M_1 and M_2 are two test images and the corresponding encrypted images are Enc_1 and Enc_2. If Eq. (B.8) is satisfied, then the CP attack is possible. The suggested approach is protected from a chosen-plaintext attack, as shown by Fig. B.18, which demonstrates that Eq. (B.8) above is not justified.

Appendix C. Training loss analysis

The YOLOv8 model is trained with our medical dataset to improve the ROP detection accuracy. Fig. C.19 illustrates the qualitative results of training loss. The x-axis represents the iteration times of training (Epoch), while the y-axis represents both training loss and detection accuracy. The sub-figures a–c depict the training loss and ROP detection accuracy concerning the medical dataset, demonstrating the network's fast convergence and eventual stable fitting.

Given the direct influence of the hashcode on image retrieval, it is imperative to closely monitor the process of hashcode learning. After extracting the ROPs, they are categorized by classes and fed into the LRViT for local hashcode generation, as illustrated in Fig. 6.

Simultaneously, the entire images are inputted into GCViT for the global hashcode generation with the help of the architecture shown in Fig. 5. The ViT networks are designed to propagate cumulative loss, combining central similarity loss, quantization loss, and bit balance loss. The objective is to generate a discriminative hashcode while minimizing the loss defined in Eq. (21). The network undergoes training for hashcodes of lengths 8, 16, 32, 64, and 128 bits for each dataset separately. Over epochs, the training loss, depicted in Fig. C.20(a–f), decreases for both local and global networks. From the graph, it is observable that, global features are quickly captured in GCViT as it learns sparse features with R2L encoder (R2L communication). For the ROP images, complex local features are learned in more iterations with modified convolution attention of LRViT. Once the model has been trained, both models are used to generate global and local hashcodes of each image in the dataset. The produced hashcodes are fused and stored alongside the embedded image in a B+tree-based storage system. Since we generate hashcodes in lengths of 8, 16, 32, 64, and 128 bits, our fused hashcodes will have lengths of 16, 32, 64, 128, and 256 bits. This database serves to evaluate the retrieval accuracy of the system.



(a) BT-MRI: Global Hashcode (b) C-XR: Global Hashcode (c) K-CT: Global Hashcode

(d) BT-MRI: Local Hashcode (e) C-XR: Local Hashcode (f) K-CT: Local Hashcode

Fig. C.20. Hashcode learning: Training loss.

Table D.12
Summary of baseline models: Backbone, Datasets, Loss functions, and Implementation handling.

Model name	Backbone	Original dataset(s)	Loss functions	Code availability and handling in our work
IDHN [51]	AlexNet	NUS-WIDE, Flickr, VOC2012, IAPRTC12	Pairwise similarity loss, soft similarity loss	Code available on GitHub. So, used official TensorFlow code; trained on our datasets without architecture change.
UDTH [33]	Pre-trained VGG-16	CIFAR-10, NUS-WIDE, MIRFLICKR-25K	Triplet loss, reconstruction loss, binary quantization loss	Exact code: Not available. So, re-implemented in PyTorch using architecture and hyperparameters from paper. Finetuned on BT-MRI, C-XR, and K-CT.
DCNN [52]	AlexNet	MESSIDOR, OASIS-MRI, I-ELCAP CT	Classification and hashing loss (SGD optimizer)	Exact code: Not available. So, reproduced using AlexNet; followed original training settings (20 epochs, lr = 0.0001, batch = 64).
PMIH [26]	DenseNet-121 (Global + Local)	CheXpert, MIMIC-CXR	Cross-entropy, quantization, bit balance loss	Exact code: Not available. Re-implemented architecture with multi-branch DenseNet. Used same loss and parameters as paper (50 epochs, lr = 0.0001, batch = 64).
IFFH [35]	DenseNet-201 (Global + Local)	BT-MRI	Cross-entropy, quantization, bit balance loss	Exact code: Not available. Reproduced per paper. Since same BT-MRI dataset was used, we obtained nearly identical performance for that dataset.
VTH [53]	Vision Transformer (ViT)	COVID-19, Breast Cancer	Self-supervision loss, contrastive loss	Code available on GitHub. So, used official codebase; adapted directly to BT-MRI, C-XR, and K-CT datasets.

Appendix D. Discussion on baseline models

To ensure a fair and consistent evaluation, we reproduced or re-executed all relevant baseline models on the same three medical datasets used in our study: BT-MRI, C-XR, and K-CT. Many of the original papers used either non-medical datasets or entirely different medical domains, which necessitated a re-implementation or fine-tuning process to ensure comparability. Table D.12 provides a detailed comparison of the baseline methods.

We emphasize that none of the baseline models (except IFFH) originally evaluated their methods on the exact datasets we used which areBT-MRI, C-XR, and K-CT. Due to domain-specific variations across medical datasets, direct comparison using reported results from prior works would be invalid. Thus, we reproduced or retrained all models in a consistent environment. For models with publicly available code (IDHN, VTH), we directly used the authors' implementations and adapted them to our datasets. For others (UDTH, DCNN, PMIH, IFFH), we followed their published architecture, loss formulations, and training settings. Most of these models build upon standard deep learning backbones like AlexNet, VGG-16, and DenseNet-121/201 and employ customized loss functions to improve retrieval performance.

Our reproduction was carefully designed to align with the original work, ensuring that hyperparameters (learning rate, batch size, epoch count) and optimization methods matched exactly. As most papers used pre-trained models with minor architectural tuning and loss-based innovations, our reimplementations provide a faithful reproduction. For IFFH, which used the same BT-MRI dataset, our results closely match their reported performance, indicating correctness in implementation. This approach ensures a rigorous and fair comparison under identical conditions. It also highlights how our proposed model generalizes better across varying medical imaging modalities while maintaining security and retrieval efficiency.

Data availability

Publicly available datasets are used for modeling which have been mentioned in the manuscript.

References

[1] Raghupathi W, Raghupathi V, Saharia A. Analyzing health data breaches: a visual analytics approach. AppliedMath 2023;3(1):175–99.
[2] Li W, Feng C, Yu K, Zhao D. MISS-D: A fast and scalable framework of medical image storage service based on distributed file system. Comput Methods Programs Biomed 2020;186:105189.
[3] Gayathri, Gowri. Securing medical image privacy in cloud using deep learning network. J Cloud Comput Adv Syst Appl 2023;12(1).
[4] Duan Y, Li Y, Lu L, Ding Y. A faster outsourced medical image retrieval scheme with privacy preservation. J Syst Archit 2022;122. http://dx.doi.org/10.1016/j.sysarc.2021.102356.
[5] Tabatabaei Z, Wang Y, Colomer A, Oliver Moll J, Zhao Z, Naranjo V. WWFedCBMIR: World-wide federated content-based medical image retrieval. Bioengineering 2023;10(10). http://dx.doi.org/10.3390/bioengineering10101144.
[6] Galbusera F, Cina A. Image annotation and curation in radiology: an overview for machine learning practitioners. Eur Radiol Exp 2024;8(1):11.
[7] Guo JM, Prayuda AWH, Prasetyo H, Seshathiri S. Deep Learning based Image Retrieval with Unsupervised Double Bit Hashing. IEEE Trans Circuits Syst Video Technol 2023. http://dx.doi.org/10.1109/TCSVT.2023.3268091.
[8] Abid MH, Ashraf R, Mahmood T, Faisal CMN. Optimized transfer learning based multi-modal medical image retrieval. Multimedia Tools Appl 2023;83(15):44069–100.
[9] Kumar S, Pal AK, Islam SH, Hammoudeh M. Secure and efficient image retrieval through invariant features selection in insecure cloud environments. Neural Comput Appl 2021. http://dx.doi.org/10.1007/s00521-021-06054-y.
[10] Luo X, Wang H, Wu D, Chen C, Deng M, Huang J, Hua XS. A survey on deep hashing methods. ACM Trans Knowl Discov Data 2023;17(1). http://dx.doi.org/10.1145/3532624.
[11] Li Y, Wei X, Cai G, Wang H. Privacy-preserving multi-source content-based image retrieval for IoT in cloud computing. Ann Des Telecommun/Ann Telecommun 2023. http://dx.doi.org/10.1007/s12243-023-00946-x.
[12] Chen Y, Zhang S, Liu F, Chang Z, Ye M, Qi Z. TransHash: Transformer-based hamming hashing for efficient image retrieval. In: ICMR 2022 - Proceedings of the 2022 international conference on multimedia retrieval. Association for Computing Machinery, Inc; 2022, p. 127–36. http://dx.doi.org/10.1145/3512527.3531405.
[13] Janani T, Brindha M. Secure similar image matching (SESIM): An improved privacy preserving image retrieval protocol over encrypted cloud database. IEEE Trans Multimed 2022;24:3794–806. http://dx.doi.org/10.1109/TMM.2021.3107681.
[14] Janani T, Darak Y, Brindha M. Secure similar image search and copyright protection over encrypted medical image databases. IRBM 2021;42(2):83–93. http://dx.doi.org/10.1016/j.irbm.2020.02.005.
[15] Prasad S, Singh AK. Survey on medical image encryption: From classical to deep learning-based approaches. Comput Electr Eng 2025;123:110011. http://dx.doi.org/10.1016/j.compeleceng.2024.110011, URL https://www.sciencedirect.com/science/article/pii/S0045790624009364.
[16] Marwan M, Kartit A, Ouahmane H. Secure cloud-based medical image storage using secret share scheme. In: 2016 5th international conference on multimedia computing and systems. ICMCS, IEEE; 2016, p. 366–71.
[17] Marwan M, Kartit A, Ouahmane H. Design a secure framework for cloud-based medical image storage. In: Proceedings of the 2nd international conference on big data, cloud and applications. 2017, p. 1–6.
[18] Marwan M, Kartit A, Ouahmane H. A framework to secure medical image storage in cloud computing environment. J Electron Commer Organ (JECO) 2018;16(1):1–16.
[19] Kumar R, Kumar J, Khan AA, Ali H, Bernard CM, Khan RU, Zeng S, et al. Blockchain and homomorphic encryption based privacy-preserving model aggregation for medical images. Comput Med Imaging Graph 2022;102:102139.
[20] Gan C, Yang H, Zhu Q, Zhang Y, Saini A. An encrypted medical blockchain data search method with access control mechanism. Inf Process Manage 2023;60(6):103499. http://dx.doi.org/10.1016/j.ipm.2023.103499, URL https://www.sciencedirect.com/science/article/pii/S0306457323002364.
[21] Sohan MF, Basalamah A. A systematic review on federated learning in medical image analysis. IEEE Access 2023;11:28628–44.
[22] Yan H, Chen Z, Jia C. SSIR: Secure similarity image retrieval in IoT. Inform Sci 2019;479:153–63. http://dx.doi.org/10.1016/j.ins.2018.11.046.
[23] Du A, Wang L, Cheng S, Ao N. A privacy-protected image retrieval scheme for fast and secure image search. Symmetry 2020;12(2). http://dx.doi.org/10.3390/sym12020282.
[24] Hu L, Xiang T, Guo S. SensIR: Towards privacy-sensitive image retrieval in the cloud. Signal Process, Image Commun 2020;84. http://dx.doi.org/10.1016/j.image.2020.115837.
[25] Dubey SR, Singh SK, Chu W-T. Vision transformer hashing for image retrieval. 2021, URL http://arxiv.org/abs/2109.12564.
[26] Guan A, Liu L, Fu X, Liu L. Precision medical image hash retrieval by interpretability and feature fusion. Comput Methods Programs Biomed 2022;222. http://dx.doi.org/10.1016/j.cmpb.2022.106945.
[27] Özbay E, Özbay FA. Interpretable pap-smear image retrieval for cervical cancer detection with rotation invariance mask generation deep hashing. Comput Biol Med 2023;154. http://dx.doi.org/10.1016/j.compbiomed.2023.106574.
[28] Sharma AK, Nandal A, Dhaka A, Dixit R. A survey on machine learning based brain retrieval algorithms in medical image analysis. Heal Technol (Berl) 2020;10(6):1359–73.
[29] Zhang Y, Ou W, Zhang J, Deng J. Category supervised cross-modal hashing retrieval for chest X-ray and radiology reports. Comput Electr Eng 2022;98:107673. http://dx.doi.org/10.1016/j.compeleceng.2021.107673, URL https://www.sciencedirect.com/science/article/pii/S0045790621005942.
[30] Zhou W, Wang N, Liu Z, Fu J, Deng L, Wu Q. Privacy-preserving IoT data retrieval scheme with lightweight fine-grained access control in cloud computing. IEEE Internet Things J 2025. http://dx.doi.org/10.1109/jiot.2024.3519578, 1–1.
[31] Liang J, Li P, Liu Z, He H. Privacy-preserving clustering-based image retrieval in cloud-assisted internet of things. IEEE Internet Things J 2025. http://dx.doi.org/10.1109/JIOT.2025.3546416, 1–1.
[32] Hussain A, li H-C, Hussain M, Ali M, Abbas S, Ali D, Rehman A. A gradual approach to knowledge distillation in deep supervised hashing for large-scale image retrieval. Comput Electr Eng 2024;120:109799. http://dx.doi.org/10.1016/j.compeleceng.2024.109799, URL https://www.sciencedirect.com/science/article/pii/S0045790624007262.
[33] Gu Y, Zhang H, Zhang Z, Ye Q. Unsupervised deep triplet hashing with pseudo triplets for scalable image retrieval. Multimedia Tools Appl 2020;79(47–48):35253–74. http://dx.doi.org/10.1007/s11042-019-7687-0.
[34] Öztürk Ş. Class-driven content-based medical image retrieval using hash codes of deep features. Biomed Signal Process Control 2021;68:102601.
[35] Özbay E, Altunbey Özbay F. Interpretable features fusion with precision MRI images deep hashing for brain tumor detection. Comput Methods Programs Biomed 2023;231. http://dx.doi.org/10.1016/j.cmpb.2023.107387.
[36] Yuan L, Wang T, Zhang X, Tay FE, Jie Z, Liu W, Feng J. Central similarity quantization for efficient image and video retrieval. 2019, URL http://arxiv.org/abs/1908.00347.

[37] Yang T, Li Y, He J, Liu Z, Ren F, Wang T, Hou G. Secure and traceable multikey image retrieval in cloud-assisted internet of things. IEEE Internet Things J 2024;11(24):40875–87. http://dx.doi.org/10.1109/JIOT.2024.3457017.
[38] Chen L, Yang Y, Yang L, Xu C, Miao Y, Liu Z, Liu X, Choo K-KR, Liang Z, Hong C. Efficient and secure content-based image retrieval in cloud-assisted internet of things. IEEE Internet Things J 2025;12(5):6001–13. http://dx.doi.org/10.1109/JIOT.2024.3489957.
[39] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. 2020, URL http://arxiv.org/abs/2010.11929.
[40] Mercaldo F, Brunese L, Martinelli F, Santone A, Cesarelli M. Object detection for brain cancer detection and localization. Appl Sci (Switzerland) 2023;13(16). http://dx.doi.org/10.3390/app13169158.
[41] Amaithi Rajan A, Vetrian V, Gladys A. Secure image encryption model for cloud-based healthcare storage using hyperchaos and DNA encoding. 2023, p. 89–103. http://dx.doi.org/10.1007/978-3-031-38296-3{_}8, URL https://link.springer.com/10.1007/978-3-031-38296-3_8.
[42] Chen C-F, Panda R, Fan Q. RegionViT: Regional-to-local attention for vision transformers. 2021, URL http://arxiv.org/abs/2106.02689.
[43] Manzari ON, Ahmadabadi H, Kashiani H, Shokouhi SB, Ayatollahi A. MedViT: A robust vision transformer for generalized medical image classification. 2023, http://dx.doi.org/10.1016/j.compbiomed.2023.106791, URL http://arxiv.org/abs/2302.09462 http://dx.doi.org/10.1016/j.compbiomed.2023.106791.
[44] Nickparvar M. Brain tumor MRI dataset. 2021.
[45] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Tech. rep., URL https://uts.nlm.nih.gov/metathesaurus.html.
[46] Islam MN, Hasan M, Hossain MK, Alam MGR, Uddin MZ, Soylu A. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. Sci Rep 2022;12(1). http://dx.doi.org/10.1038/s41598-022-15634-4.
[47] Dai J, Li Y, He K, Sun J. R-FCN: Object detection via region-based fully convolutional networks. 2016, URL http://arxiv.org/abs/1605.06409.
[48] Gavrilescu R, Fosalau C, Zet C, Skoczylas M, Cotovanu D. Faster R-CNN: an approach to real-time object detection. In: 10th international conference and exposition on electrical and power engineering. 2018, p. 165–8.
[49] Jung HK, Choi GS. Improved YOLOv5: Efficient object detection using drone images under various conditions. Appl Sci (Switzerland) 2022;12(14). http://dx.doi.org/10.3390/app12147255.
[50] Satila Passa R, Nurmaini S, Rini DP. YOLOv8 based on data augmentation for MRI brain tumor detection. Sci J Inform 2023;10(3):363. http://dx.doi.org/10.15294/sji.v10i3.45361, URL http://journal.unnes.ac.id/nju/index.php/sji.
[51] Zhang Z, Zou Q, Lin Y, Chen L, Wang S. Improved deep hashing with soft pairwise similarity for multi-label image retrieval. IEEE Trans Multimed 2020;22(2):540–53. http://dx.doi.org/10.1109/TMM.2019.2929957.
[52] Mohite NB, Gonde AB. Deep features based medical image retrieval. Multimedia Tools Appl 2022;81(8):11379–92. http://dx.doi.org/10.1007/s11042-022-12085-x.
[53] Kumar M, Singh R, Mukherjee P. VTHSC-MIR: Vision transformer hashing with supervised contrastive learning based medical image retrieval. Pattern Recognit Lett 2024;184:28–36. http://dx.doi.org/10.1016/j.patrec.2024.06.003.
[54] Xiong C, Xu X, Zhang H, Zeng B. An analysis of clinical values of MRI, CT and X-ray in differentiating benign and malignant bone metastases. Tech. rep., 2021, URL www.ajtr.org.