Computational and Structural Biotechnology Journal 24 (2024) 434–450

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

Instance-level medical image classification for text-based retrieval in a medical data integration center

Ka Yung Cheng [a,*], Markus Lange-Hegermann [b], Jan-Bernd Hövener [c], Björn Schreiweis [a]

[a] Institute for Medical Informatics and Statistics, Kiel University and University Hospital Schleswig-Holstein, Kiel, Germany
[b] Institute Industrial IT, OWL University of Applied Sciences and Arts, Lemgo, Germany
[c] Department of Radiology and Neuroradiology, Section Biomedical Imaging, Kiel University and University Hospital Schleswig-Holstein, Kiel, Germany

ARTICLE INFO

Keywords:
DICOM images
Medical image captioning
Medical image interchange
SNOMED CT body structure

ABSTRACT

A medical data integration center integrates a large volume of medical images from clinical departments, including X-rays, CT scans, and MRI scans. Ideally, all images should be indexed appropriately with standard clinical terms. However, some images have incorrect or missing annotations, which creates challenges in searching and integrating data centrally. To address this issue, accurate and meaningful descriptors are needed for indexing fields, enabling users to efficiently search for desired images and integrate them with international standards.

This paper aims to provide concise annotation for missing or incorrectly indexed fields, incorporating essential instance-level information such as radiology modalities (e.g., X-rays), anatomical regions (e.g., chest), and body orientations (e.g., lateral) using a Deep Learning classification model - ResNet50. To demonstrate the capabilities of our algorithm in generating annotations for indexing fields, we conducted three experiments using two open-source datasets, the ROCO dataset, and the IRMA dataset, along with a custom dataset featuring SNOMED CT labels. While the outcomes of these experiments are satisfactory (Precision of >75%) for less critical tasks and serve as a valuable testing ground for image retrieval, they also underscore the need for further exploration of potential challenges. This essay elaborates on the identified issues and presents well-founded recommendations for refining and advancing our proposed approach.

1. Introduction

Medical imaging plays a pivotal role in modern healthcare, contributing significantly to diagnostic accuracy and clinical decision-making. Currently, medical imaging accounts for 70% of non-invasive clinical diagnoses and comprises 80% to 90% of all hospital data [1]. However, the escalating volume and usage have raised concerns regarding retrieval performance [2], data interoperability, and integration [3].

Other than dealing with the ever-growing amount of medical imaging data, incorrect or incomplete image annotations complicate the retrieval of medical imaging data. Addressing these and other healthcare data-sharing challenges, Medical Data Integration Center (MeDIC) was established [4], aiming to streamline data accessibility and facilitate secondary use. Our MeDIC should be able to access or reference medical imaging data from existing Picture Archiving and Communication Systems (PACS) using the Digital Imaging and Communication in Medicine (DICOM) standard [5] and handling non-DICOM formats like-Joint Photographic Experts Group (JPEG), Portable Network Graphics (PNG), and others. Therefore, there is a need for an automatic image annotation process in MeDIC, and it should accommodate both DICOM and non-DICOM formats.

The image files in DICOM format, the preferred format in radiology for storing medical images, are supposed to have complete annotation information in their header section; however, we discovered that legacy data in our PACS at University Hospital Schleswig-Holstein (UKSH) miss or contain erroneous annotations and local clinical terms. These images can still benefit from incorporating an automated annotation system for retrieval. A DICOM file often bundles up to hundreds or even thousands of image slices. Each DICOM file can be generated by different manufacturers and modalities such as CT, MRI, and X-ray [6]. While DICOM may contain valuable information, its DICOM tags may not always align with the search queries for specific images. Besides, research has indicated that approximately 15% of DICOM images contain erroneous tag entries during automated examination processes [7]. For

* Corresponding author.
E-mail address: kayung.cheng@uksh.de (K.Y. Cheng).

https://doi.org/10.1016/j.csbj.2024.06.006
Received 29 February 2024; Received in revised form 5 June 2024; Accepted 6 June 2024
Available online 12 June 2024

instance, improper image registrations often occur in multiple imaging exams associated with a single physician order. Furthermore, older DICOM images frequently lack crucial information such as body parts and image modalities. Compounding these challenges, clinic departments tend to document DICOM imaging data with local codes rather than standardized keywords or international terminologies. These poor DICOM practices degenerate effective communication and understanding among healthcare systems and institutions worldwide. All these challenges accentuate the need for automated annotate solutions for DICOM image files in medical data platforms.

Non-DICOM data can also benefit from automatic annotation to enable efficient searching. A study conducted by ISH-VNA in 2017 [8] revealed that over 75% of medical images consist of non-DICOM imaging data, including biosignals like Electrocardiograms (ECG) and Electroencephalograms (EEG). Even though DICOM supports these biosignals in various Information Object Definitions (IODs) [9], outdated medical imaging equipment and diagnostic devices often restrict the availability of DICOM for legacy data [10]. Moreover, DICOM files are often converted to JPEG [11]. This is predominantly driven by the smaller file sizes offered by JPEG and its compatibility across diverse computer platforms. However, this conversion process eliminates DICOM headers and often leads to a loss of annotative description in the resulting non-DICOM files. Additionally, most non-DICOM data types lack standardized descriptive annotations or possess incomplete annotations related to image modalities and anatomical parts. This poses challenges for search engines attempting to retrieve these data. Therefore, automatic annotation is essential for managing non-DICOM data, especially in scenarios involving outdated medical devices, loss of DICOM headers during non-DICOM conversion, and the overall absence of standard annotations in non-DICOM data types.

Due to the complex anatomy structures, constraints, and inherent properties of medical image data sets, there were only a handful of studies and review papers in medical image captioning tasks [12] [13]. Their coverage tended to be specialized rather than mapping all image objects in a generic image captioning system. Chiang et al. [14] focused on four classes using CNN, “CT of the abdomen”, “CT of the brain”, “MRI of the brain” and “MRI of the lumbar spine”. Wasserthal et al. [15] segmented CT 104 anatomical structures, and Zhang et al. [16] trained their CNN models using Empirical Mode Decomposition (EMD) to retrieve X-ray images, i.e. IRMA dataset, with a Content-based approach. [14]. [17–19] trained with the IRMA 2009 dataset to retrieve top-ranked hits. Shamna et al. [17] implement the Topic and Location Model, which led to the retrieval of content-based images with a precision rate of 97.5% for the top ten images. Srinivas et al. [18] employed dictionary learning and achieved a precision of 93.7% for the top ten images. Meanwhile, [19] uses local binary patterns (LBP) and support vector machines (SVM), resulting in a precision of 87.2% for the top 20 hits.

1.1. Objective

Our primary objective was to explore the potential and challenges of utilizing AI for automatically annotating in retrieving and integrating medical imaging data within the MeDIC platform [20]. This work focused on generating instance-level keywords using existing Deep Learning (DL) methods and datasets. These instance-level keywords were terminologies related to imaging modality, orientation, and anatomical parts depicted in the images. Integrating these standardized annotations into MeDIC enables efficient Text-based Image Retrieval (TBIR) through Elasticsearch full-text search queries.

Furthermore, our approach emphasized using standardized clinical terminologies, such as SNOMED Clinical Terms (SNOMED CT), to enhance data sharing and collaboration across healthcare institutions. While our project primarily targets non-DICOM images in JPEG and PNG formats, it also extended to handling DICOM images, ensuring comprehensive coverage of medical imaging data.

2. Methods and materials

Image captioning involves generating human-readable textual descriptions for images [21], interchangeably referred to as “automatic image annotation”, “image indexing”, or “image tagging”. Approaches to generating textual descriptions of images can be categorized as: 1) classifying images into predefined classes [14], or 2) describing (captioning) objects or scenes in images. To accomplish our goal of generating captions using instance-level keywords, we focus on classifying medical images into predefined categories. We use “classify” instead of “caption” when a classifier is applied in the remainder of this paper.

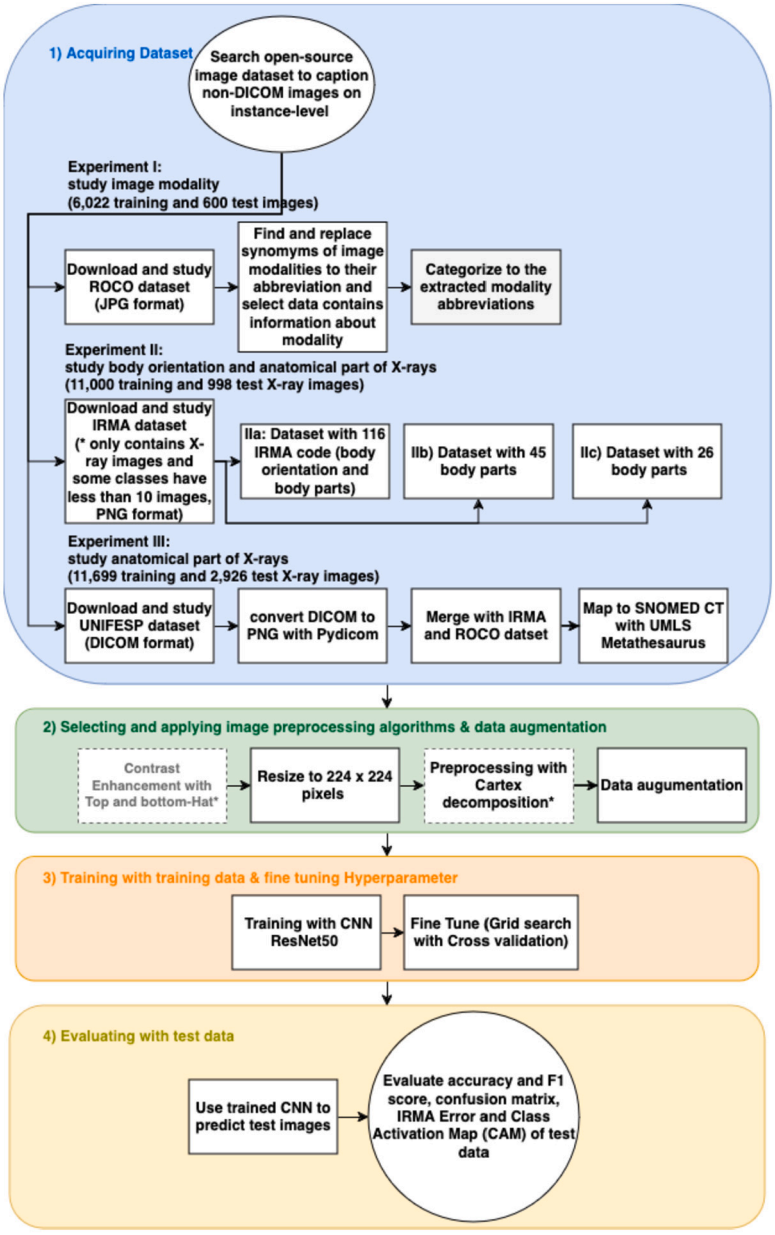
As depicted in Fig. 1, our DL classification is structured according to four basic Machine Learning steps [22]. ***1) Data curation*** Initially, we acquired three publicly available datasets enriched with instance-level labels: ROCO [23], IRMA [24], and UNIFESP [25], containing keywords of image modalities, body directions, anatomical parts, and relevant descriptions. Three distinct training datasets, labeled as Dataset I, II, and III, were chosen as the foundation of our research, comprising i) ROCO, ii) IRMA, and iii) a custom-defined SNOMED CT dataset created by aggregating and refining data from the aforementioned open-source datasets. ***2) Image Preprocessing*** We conducted an experiment for each dataset, denoted them as Experiment I, II, and III. In each experiment, the corresponding dataset underwent different image preprocessing techniques, such as Top-Bottom Hat Morphological Operators [26] and Cartoon+Texture decomposition [27]. We evaluated the impact of these techniques across shorter training epochs and determined the best preprocessing procedures that improved the training model. ***3) Training and fine tuning*** In each experiment training, we utilized a pretrained Convolutional Neural Network (CNN), specifically ResNet50 (IMAGENET1K V2) [28] from torchvision. Hyperparameter fine-tuning was conducted through grid search methodology aided by 5-fold stratified cross-validation. ***4) Evaluation*** Ultimately, we comprehensively evaluated the optimally trained models using diverse metrics, including Confusion Matrix [29] and Class Activation Map [30], to generate instance-level keywords for medical imaging data through our DL classifier. In the following sections, each step will be described in greater detail.

2.1. Three instance-level datasets with labels of modalities, IRMA codes and SNOMED CT body parts

We chose two publicly available datasets with instance-level labels, along with a custom-defined dataset aggregated from three publicly available datasets, ROCO, IRMA, and UNIFESP. We selected these datasets due to their ample training data, particularly given the constrained availability of publicly accessible datasets in the medical domain.

2.1.1. ROCO dataset - 81,000 radiology images with captions including various image modalities

One of the datasets we employed was the **R**adiology **O**bjects in **CO**ntext (ROCO) dataset [23], which consists of images obtained from an open-access biomedical literature database - PubMedCentral. It contains over 81,000 radiology images with medical imaging modalities, including Computer Tomography (CT), Ultrasound, X-ray, Fluoroscopy, Positron Emission Tomography (PET), Mammography, Magnetic Resonance Imaging (MRI), etc. Each image within the ROCO dataset is accompanied by a caption, relevant keywords, Unified Medical Language Systems Concept Unique Identifiers (UMLS CUIs) by the National Library of Medicine (NLM) and Semantic Types (SemTypes). Two deep CNN-based binary classifiers were trained to categorize images into radiology and non-compound figures. The textual annotations per image were extracted from the captions and converted into keywords, which were then transformed into CUIs and SemTypes using QuickUMLS. The images were reviewed, and any false positives were manually identified



To generate instance-level keywords for medical imaging data, the DL workflow has four primary components: 1) acquiring data, 2) preprocessing and augmentation, 3) training and tuning hyperparameter and 4) evaluation. The "" symbols denote as optional steps; refer to Section 2.2 for detailed descriptions.*

Fig. 1. DL workflow for instance-level medical image annotation.

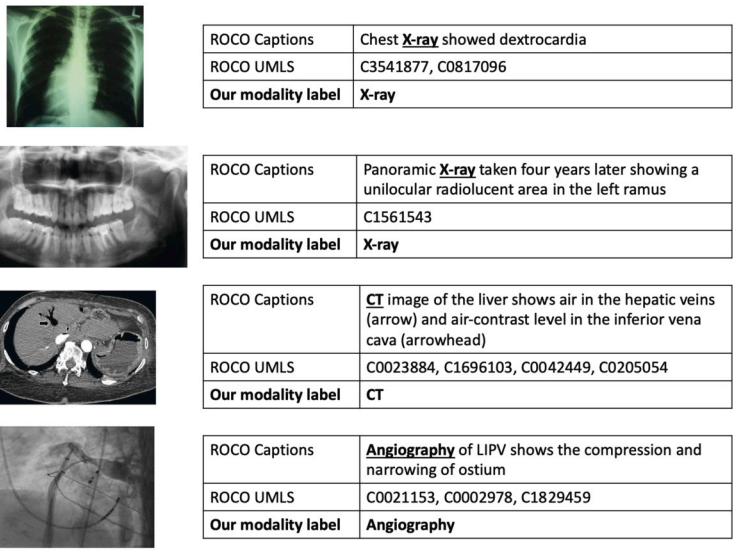
and corrected [23]. Fig. 2 depicts four sample images from the ROCO dataset along with their corresponding captions.

To simplify the image captioning task, we approached it as a classification problem. Before training the classifier, it is essential to note that the ROCO dataset has an imbalanced data distribution. The predominant class "CT" contains approximately 3500 images, whereas the minority classes, such as "mammography", consist of fewer than 50 images each. The uneven distribution of data does not accurately reflect the characteristics of routine clinical datasets, as evidenced by fewer X-ray images compared to CT and MRI images [31]. This is due to ROCO relying on medical publications rather than clinical records, resulting in captions lacking proper modalities description.

As depicted in Fig. 3, we chose the top six modalities and grouped other minority (tail) categories in ROCO dataset as "Unknown" in this experiment to handle the imbalanced data problem of ROCO dataset. We utilized the scikit-learn Library for the random stratified train-test split and then loaded the chosen images as Python Imaging Library (PIL) images into our DL models. Our resultant dataset comprised 8,430 training images and 2,108 test images, comprising the top six modalities and one augmented class.

2.1.2. IRMA dataset - body parts and orientation of X-rays

The IRMA dataset [24] is another publicly available medical image dataset from the Department for Radiological Diagnostics at Radiology, University Hospital RWTH Aachen. The dataset contains 12,000 anonymized X-ray images divided into 116 classes, 11,000 radiology images for training, and the remaining 998 radiology images as the test set. Fig. 4 illustrates three sample images accompanied by an IRMA



ROCO Captions	Chest **X-ray** showed dextrocardia
ROCO UMLS	C3541877, C0817096
Our modality label	**X-ray**

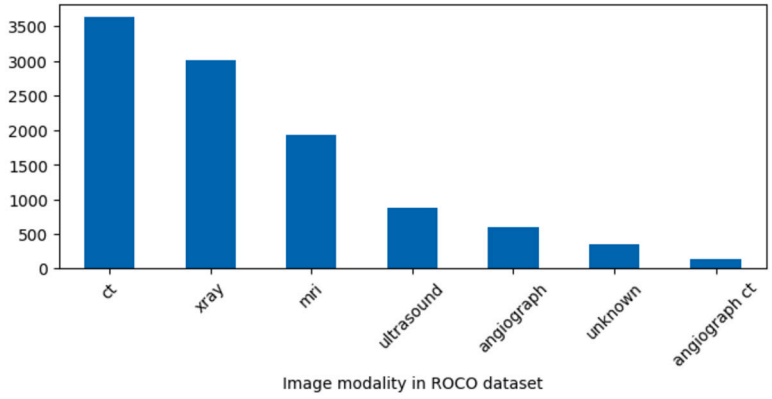
ROCO Captions	Panoramic **X-ray** taken four years later showing a unilocular radiolucent area in the left ramus
ROCO UMLS	C1561543
Our modality label	**X-ray**

ROCO Captions	**CT** image of the liver shows air in the hepatic veins (arrow) and air-contrast level in the inferior vena cava (arrowhead)
ROCO UMLS	C0023884, C1696103, C0042449, C0205054
Our modality label	**CT**

ROCO Captions	**Angiography** of LIPV shows the compression and narrowing of ostium
ROCO UMLS	C0021153, C0002978, C1829459
Our modality label	**Angiography**

UMLS is a comprehensive terminology system to facilitate the integration and retrieval of biomedical information. It assigns specific identifiers, CUIs, to each concept, allowing for standardized representation and linking of concepts across different terminologies and vocabularies.

Fig. 2. Four sample images of ROCO dataset with corresponding captions and UMLS.



The "unknown" class contains uncertain images and tails classes, e.g., fluoroscopy, PET, mammography and SPECT/CT.

Fig. 3. Imbalance distribution of image modality in ROCO dataset.

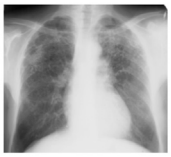
code (IRMA notation: TTTT – DDD – AAA – BBB). Each IRMA code has four segments, where T, D, A, and B denote a coding or sub-coding digit of the technical (image modality), directional, anatomical, and biological axis, respectively. This dataset covers various anatomical and directional codes but exclusively includes X-ray images. Other modalities mentioned in the IRMA T-code [24] or the ROCO dataset are not represented.

In addition, the IRMA dataset has an imbalanced data distribution, as shown in Fig. A.12 and Fig. A.13. The major class "chest" (IRMA anatomical code: 500) has over 2000 training images, while minority classes such as "lower lumbar spine" (IRMA anatomical code: 333) have only ten training images. The imbalance distribution of IRMA dataset aligns with the distribution observed in other publicly available materials and real clinical data. However, this can lead to model over-fitting. To address the risk of an over-trained model resulting from the imbalanced distribution of IRMA dataset, we employed data augmentation in the following steps and closely observed the validation loss while training. This improved the overfitting with synthetic samples, and no information was lost.

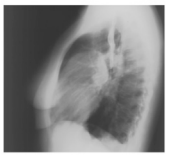
2.1.3. Custom dataset with SNOMED CT labels - body parts of X-rays

In the creation of Dataset III, we enhanced our previous datasets by incorporating the UNIFESP dataset into the ROCO dataset [23] and the IRMA dataset [24] used in Dataset I and Dataset II. The UNIFESP dataset [25], sourced from the Federal University of São Paulo in Brazil, includes 2,481 X-ray images in DICOM format. 1,880 images are annotated by radiologists with multi-label annotations, covering 20 distinct body parts and one augmented (unknown) class. We converted the DICOM images into NumPy arrays using the Pydicom library and loaded them as PIL images into Dataset III. Subsequently, we combined the converted UNIFESP images with 1,210 ROCO X-ray images from Dataset I and 11,998 IRMA images from Dataset II into a single dataset.

Our previous datasets utilized the ROCO and IRMA datasets with their respective keywords and labeling systems. In Dataset III, we relabeled all images from the three datasets (32 ROCO CUIs, which have the semantic type "Anatomical Structure or Region", 40 IRMA code, and 22 UNIFESP annotations) to one or multiple SNOMED CT [32] labels using the UMLS Metathesaurus [33]. SNOMED CT can be easily cross-mapped to other international terminologies. Therefore, incorporating SNOMED



IRMA code	1123-110-500-000
Technical code	X-ray, plain radiography, analog, high beam energy
Directional code	Coronal, posteroanterior (PA), unspecified
Anatomical code	Chest, unspecified
Biological code	Unspecified

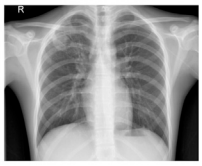


IRMA code	1123-211-500-000
Technical code	X-ray, plain radiography, analog, high beam energy
Directional code	Sagittal, lateral, right-left, inspiration
Anatomical code	Chest, unspecified
Biological code	Unspecified

IRMA code	1121-120-200-700
Technical code	X-ray, plain radiography, analog, overview image
Directional code	Coronal, anteroposterior (AP, coronal), unspecified
Anatomical code	Cranium, unspecified
Biological code	Musculosceletal system

TTTT in IRMA notation represents for technical imaging modality, DDD represents for the direction of the body, AAA represents for the anatomical part examined, and BBB represents for the biological system studied.

Fig. 4. Three sample images (two chest images and a cranium image) with the corresponding IRMA code (TTTT-DDD-AAA-BBB) [24].



Original dataset	ROCO dataset
Original label	ROCO UMLS: *C0015726, C1306645, C1962945, C1548003, C1522577, C1261259, C0423899, C0751438, C0043299, C0817096, C0796494, C1561540, C1561538* ROCO captions: *Follow-up chest radiograph obtained 2 weeks after (day 37) showing a ball-shaped cavitating mass with crescentic cavitation in the posterior segment of the right upper lobe.*
Original format	JPEG
Our SNOMED label	**51185008 (Chest)**

Original dataset	IRMA dataset
Original label	IRMA code: 1121-230-943-700 (X-ray,plain radiography, analog, overview image - sagittal,mediolateral, lower extremity / leg,knee - musculosceletal system)
Original format	PNG
Our SNOMED label	**61685007 (Lower limb), 72696002 (Knee-region-structure)**

Original dataset	UNIFESP dataset
Original label	21 (Wrist)
Original format	DICOM
Our SNOMED label	**120574008 (Upper extremity part), 8556200 (Hand structure), 8205005 (Wrist)**

Fig. 5. Three sample images in custom dataset, sourced from ROCO, IRMA and UNIFESP respectively.

CT codes will enhance precision and compatibility, facilitating precise access to medical images in MeDIC across diverse stakeholders and institutions.

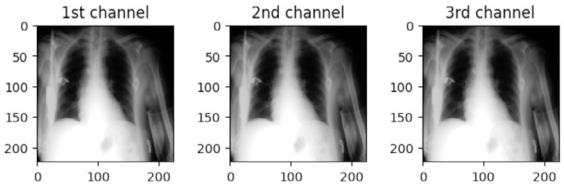
Another advancement was addressing the issue of limited data for certain classes within Dataset I and II and their integration into the multi-labeled UNIFESP dataset and handling hierarchical anatomical concepts. We transformed the classification challenge into a multi-label classification problem. This approach enables each image to be associated with multiple labels and allows for detecting more than one body part for each sample. For example, an upper leg image is annotated as “lower extremity” (SNOMED CT code: 61685007) and “thigh” (SNOMED CT code: 7569003), instead of just “upper leg” (IRMA anatomical code: 953) as in the IRMA dataset. Refer to Fig. 5 for a visual representation of three sample images in Dataset III.

We performed a stratified split on the dataset with a split factor of 0.2 using the scikit-learn Library, creating randomized training and testing subsets. The resulting Dataset III comprises 11,700 training images and 2,922 test images, encompassing 45 SNOMED CT codes and an augmented (unknown) class.

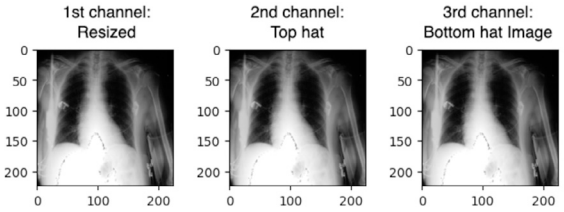
2.2. Image preprocessing with grayscale images

Grayscale images, characterized by their single channel, pose compatibility challenges with prevalent pre-trained models designed for RGB images with three dimensions [34]. To address this issue, fine-tuning pre-trained models on grayscale images can be achieved through either replicating the single channel to construct a pseudo-color image with three channels [35] or by modifying the initial model layer [36]. In our study, we converted all images to three channels because it is more common to alter the images than the pre-trained model itself, particularly when dealing with grayscale medical images [37].

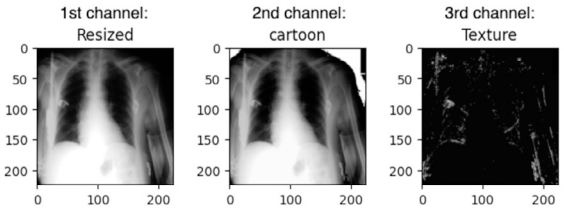
Thoughtful preprocessing enhances the performance of DL models [38]. To optimize image quality and refine edge detection, we designed four distinct image preprocessing combinations (A, B, C, and D) as shown in Fig. 6, each utilizing four fundamental transformations (resizing, Top-Bottom Hat, Cartex, and data augmentation) or fewer. Within each upcoming experiment, we started with determining the optimal image preprocessing combination among the four options. Details of each basic transform are outlined below:



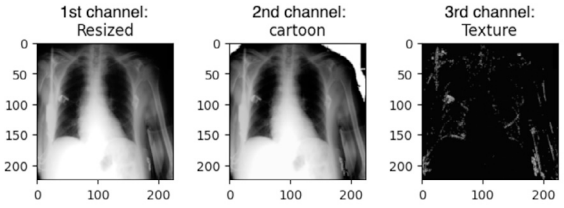
(a) Combination A (minimally processed)- Resizing + data augumentation



(b) Combination B - Resizing + Top-Bottom Hat + data augumentation



(c) Combination C - Resizing + Cartex + data augumentation



(d) Combination D - Resizing + Top-Bottom Hat + Cartex + data augumentation

We analyzed the optimal image preprocessing combination from among the four available options.

Fig. 6. Preprocessing a sample image.

1) Resize We performed bicubic interpolation over a 4×4 (16 pixels) neighborhood to resize images to 224×224. The resized images allow DL models to train faster on small images with mini-batch training and overcome the computation constraints [39] ***2) Top- and Bottom-Hat (Top-Bottom Hat) Transform*** We enhanced the contrast of the images using Top-Bottom Hat morphological operators [40]. This was achieved by adding the bright regions (obtained after applying morphological opening to the original images) and subsequently subtracting the dark regions (obtained after applying morphological closing to the original images) to the original images. ***3) Cartoon+Texture (Cartex) decomposition*** We decomposed the global structure image information and the locally-patterned image information with the Cartex algorithm [27] by decomposing the image into contrasted shapes and high oscillating patterns as Algorithm 1. ***4) Data augmentation*** Data augmentation improves the model performance by adding new artificial data derived from existing training data [41]. Common techniques include resizing, flipping, rotating, cropping, padding, contrast and brightness transformations. Before each batch of images was loaded into the model, we applied a list of predefined transforms from the fast.ai Library, with parameters as outlined in [41]. The list of transformations includes as follows: we applied Dihedral transform [42] with a probability of 0.5. With a probability of 0.75, we applied a random rotation of an additional 10 degrees. With a probability of 0.75, a random zoom between a scale of 0.9 and 1.1, and a perspective warping of 0.2. With a probability of 0.75, we applied a change in brightness and contrast of a maximum scale of 0.2. The random resize crop was picked as a random scale in the range from 0.9 to 1, and then the resize was done. Since the task involves classifying direction, we skipped the flipping step for Dataset II (IRMA). This ensured that the number of samples remained unchanged while allowing the model to be trained with more generalized data.

2.3. Choosing DL model for training

Convolutional Neural Networks (CNNs) have been widely used in supervised learning, such as image classification, object detection, semantic segmentation, etc. VGG, ResNet, GoogLeNet, Inception-ResNet, and MobileNet are common CNN architectures [28]. “Residential network (ResNet)” [28] by Kaiming He et al. is usually recommended for the early training stage because its stacked residual blocks with “skip connections (or short-cuts)” lead to high accuracy with ease and avoid overfitting. “ResNet-50” is a variant with 50 layers, comprising convolutional, pooling, and fully connected layers, renowned for its state-of-the-art performance in visual recognition tasks. It processes input images of size $224 \times 224 \times 3$, generating predictions for object classes within the image. Hence, we selected “ResNet-50” to continue our training.

Table 1
Set-up for DL.

PC Hardware & Operating System	
CPU	AMD Ryzen™ 9 5900X
RAM	Skill DIMM 32 GB DDR4-3200 Kit
GPU	NVIDIA GeForce RTX 960
Operating System	Ubuntu 20.04
DL settings and hyperparameters	
Model	ResNet50 (IMAGENET1K V2)
Max Epoch	50 (Early Stopping patience = 8)
Batch	64
Loss	Cross-Entropy
Shuffle	True
Fine tuning	5-fold grid search cross validation

2.4. Optimizing and assessing experiment performance

The "search terms generator" was implemented as a DL classification model using the Fastai Python Machine Learning Library (Version 2.7.12). Based on the open-source dataset we found, our image captioning task is divided into three experiments: **Experiment I)** Classifying the top five image modality classes of the ROCO dataset. **Experiment II)** Classifying 116 classes of the IRMA dataset and **Experiment III)** Classifying 45 anatomical classes of our custom SNOMED CT dataset.

All DL training occurred under consistent environments, as outlined in Table 1. In each experiment, we determined the optimal image preprocessing combination among four options with 20 training epochs. These combinations included: Combination A) resizing only, Combination B) resizing with Top-Bottom Hat, Combination C) resizing with Cartex, and Combination D) preprocessing with Top-Bottom Hat and Cartex. We aimed to optimize the training process and improve model performance and generalization by selecting the most suitable preprocessing algorithm.

After identifying the optimal preprocessing combination, we employed hyperparameter tuning with cross-validation to refine our models further. This strategy allowed us to handle diverse and previously unseen data and led to improved predictive accuracy and real-world performance. We conducted a Grid Search to fine-tune the optimizer, with the optimizer space defined between Stochastic Gradient Descent (SGD), Adam, and Adagrad [43]. The learning rate of each optimizer was specified by a Learning Rate Finder with a Cyclical Learning Rate policy [44]. Training processes were halted automatically to avoid overtraining through an Early-Stopping function with the patience of 8 epochs when a monitored metric - validation loss - stopped improving. Through comprehensive exploration of the space of optimizer, learning rate, and epoch, we identified the optimal model configuration with the best performance.

We evaluated with standard classification metrics [45], such as Average Precision, Average Recall, Recall, and F1 score. Average Precision measured the model ability to identify prediction instances of particular classes. In contrast, the F1 score, calculated as the harmonic mean of precision and recall, provided a more robust measure for scenarios involving imbalanced class distributions. These metrics range from 0 to 1, with 0 indicating poor performance and 1 representing flawless accuracy. Additionally, we enriched our analysis using Confusion Matrices [26] and Class Activation Map (CAM) [30]. Confusion Matrices provide graphical insights into the model effectiveness, by comparing actual target values with model predictions. Class Activation Maps (CAMs) visualize the decision-making process of CNNs, highlighting distinct image regions that significantly contribute to the classification results.

3. Results

3.1. Experiment I: classifying modality with ROCO dataset

3.1.1. Choosing image preprocessing algorithm

The data presented in Table 2 illustrates the reason for the preference for image preprocessing Combination B for the ROCO dataset over three alternative algorithms, Combinations A, C, and D. Our comprehensive evaluation of each combination involved the assessment of test set with key performance metrics, including accuracy, average precision score, F1 score, and average recall score. These evaluations were carried out after a maximum of 20 epochs, utilizing a 5-fold cross-validation methodology.

In terms of accuracy, Combination B and D outperformed the other two combinations, achieving a noteworthy accuracy of 0.925. Combinations A and C exhibited an accuracy of 0.920 and 0.924, respectively. Combination B and D demonstrated a 1% improvement in accuracy over the other two combinations. This improvement can be significant in large clinical datasets.

Combination B and C excelled with an Average Precision of 0.747 and 0.752, respectively. Combinations A and D resulted in F1 scores of 0.733 and 0.728, respectively.

Average Recall also favored Combination B, which registered a precision score of 0.823. In contrast, Combinations A, C, and D achieved slightly lower precision scores of 0.803, 0.788, and 0.744, respectively. Combination B displayed a 2% higher precision than the other three combinations.

Given Combination B consistently overall outperformed the alternative combinations across the four evaluation criteria, it solidifies the decision to employ the Combination B approach, resizing and applying Top-Bottom Hat preprocessing method, before feeding the images into the model for this specific Experiment I.

3.1.2. Performance

After applying grid search with 5-fold cross-validation for fine-tuning, we achieved with our best-trained model a precision of 0.840 and an F1 score of 0.854, indicating the model ability to classify instances accurately. The confusion matrix in Fig. 7 reveals that the model classified most classes with an accuracy exceeding 75%. The figure displays prominent dark diagonal elements, representing true positives, while the small number of off-diagonal elements indicates low confusion between the classes.

Additionally, we conducted experiments by specifically training with five modalities. We combined the "angiography CT" category into "angiography" and excluded the "unknown" class. This approach resulted in a slight improvement, achieving a precision of 0.91 and an F1 score of 0.93.

3.2. Experiment II: classification IRMA code of IRMA dataset

3.2.1. Choosing image preprocessing algorithms

As in Experiment I, we tested all four image preprocessing combinations to determine the optimal preprocessing algorithms. The outcomes presented in Table 3 reveal our findings with the IRMA dataset. In terms of accuracy, Combination D demonstrated an accuracy of 0.918. In comparison, Combinations A, B, and C displayed 0.829, 0.902, and 0.904 accuracy, respectively. Combination D showcased an improvement ranging from 2% to 7% in accuracy over the other combinations.

For Average Precision, Combination D achieved a precision score of 0.789, higher than Combinations A, B, and C, which recorded precision scores of 0.714, 0.760, and 0.753, respectively. Combination D exhibited a 3% to 8% higher precision than the other three combinations.

In terms of the F1 score, Combination D excelled with a score of 0.769, surpassing Combinations A, B, and C, which yielded F1 scores of 0.680, 0.732, and 0.743, respectively.

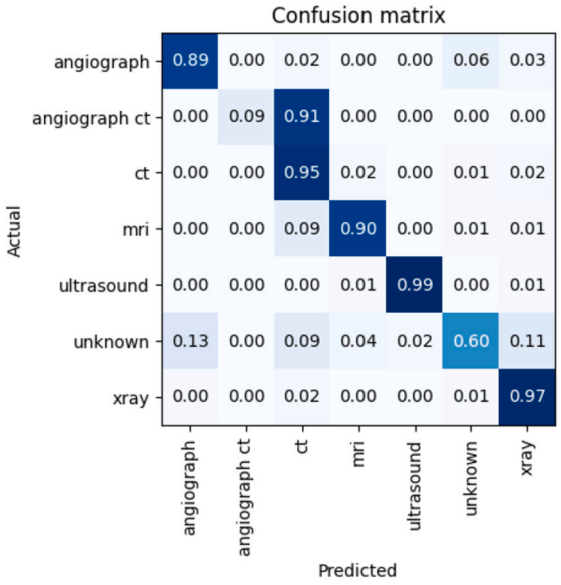
Table 2
Comparing four preprocessing combinations for classifying ROCO modalities after 20 epochs model training.

Image Preprocessing Combination	A	**B**	C	D
Brief description	Resize	**Top-Bottom Hat**	Cartex	B and C
Accuracy	0.920	**0.925**	0.924	0.925
Average Precision	0.733	**0.747**	0.752	0.728
Average Recall	0.803	**0.823**	0.788	0.744
F1 score	0.753	**0.763**	0.763	0.816

We opted for Combination B to proceed with Experiment I.

Table 3
Comparison of four preprocessing combinations with IRMA dataset after 20 epochs.

Image Preprocessing Combination	A	B	C	**D**
Brief description	Resize	Top-Bottom Hat	Cartex	**B and C**
Accuracy	0.849	0.902	0.904	**0.918**
Average Precision	0.714	0.760	0.753	**0.789**
Average Recall	0.680	0.732	0.759	**0.767**
F1 score	0.680	0.734	0.743	**0.769**



It involved the top six modalities in ROCO dataset. The classifier exhibited high consistency with the ground truth labels but struggled in distinguishing between "angiography CT" and "CT" scans due to the variety of angiography subtypes like CT-, X-ray, and MRI angiography.

Fig. 7. Confusion matrix of Experiment I.

Combination D showed an improvement of approximately 7% in accuracy and precision compared to the other combinations. Furthermore, it is easily implementable and offers relatively fast processing. As a result, Combination D is the preferred choice for further model training in Experiment II.

3.2.2. Performance

Using the Combination D steps for the whole IRMA dataset, we achieved classifying the 116 IRMA classes with an average precision of 0.789 (F1-score of 0.771). However, some minority classes were misclassified in Fig. 8. Three test images, representing 100% test images of the class of 1121-127-700-400 (abdomen, spine), were predicted as 1121-115-700-400 (abdomen, upright). Additionally, one image, representing 100% of the test images for class 1127-430-215-700 (mandible, other direction), was predicted as 1127-120-310-700 (cervical spine, anteroposterior).

Furthermore, we conducted experiments by explicitly focusing on training with two different sorting of the dataset: anatomical parts (45 body parts) and orientation parts. When training with labels only encompassing anatomical parts, we achieved a precision of 0.802 and an F1-score of 0.789. Similarly, we focused on only training with the orientation parts. We achieved a precision of 0.858 and an F1-score of 0.851. These findings suggested that the model performance did not significantly benefit from separating the classification tasks based on anatomical parts or orientation. Thus, combining all the relevant information during training seems to be more effective in achieving accurate predictions.

3.3. Experiment III: classifying SNOMED CT code of the custom dataset

The resultant dataset in Experiment III consists of 11,700 training images and 2,925 test images, encompassing 45 SNOMED CT codes and one augmented (unknown) class as classes.

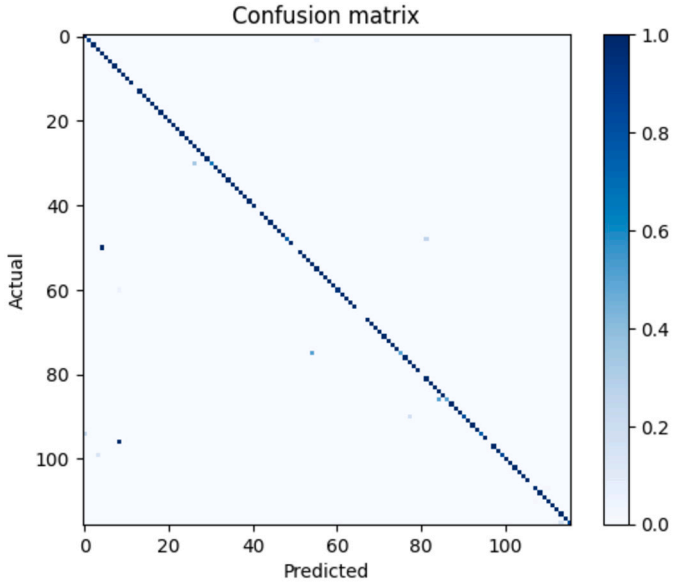
The multi-label approach allows classes with broader anatomical concepts, such as lower extremity, to have more images for model training, thereby providing the model with more data to learn these classes. However, applying this multi-label method and integrating multiple datasets exacerbates the long-tailed (imbalanced) distribution issue. For instance, the class labeled as "Chest" (SNOMED CT code: 51185008) is represented in approximately 6,000 images, while classes like "Trachea" (SNOMED CT code: 44567001), "Colon" (SNOMED CT code: 71854001), and "Esophageal structure" (SNOMED CT code: 32849002) have fewer than 10 images each, as shown in Table 4.

3.3.1. Choosing image preprocessing algorithm

The results in Table 5 outline the outcomes obtained through the four image preprocessing combinations.

It is noteworthy that precision holds greater significance than recall in our case, given our preference for accurate predictions over the comprehensiveness of positive predictions. Combinations C and D achieved a precision score of 0.652 and 0.645, slightly surpassing Combinations A and B, which recorded precision scores of 0.622 and 0.636.

Image preprocessing Combination D (Preprocessing with Top-Bottom Hat and Cartex) exhibited slightly better overall performance, so we decided to preprocess custom SNOMED CT dataset with Combination D for Experiment III. This decision allows us to build upon the



Test data was from IRMA dataset involving 116 IRMA classes. The classifier demonstrated high consistency with most ground truth labels, achieving accuracy exceeding 0.9 for 100 classes and over 0.7 for 105 classes. However, some minority classes were misclassified.

Fig. 8. Confusion matrix of Experiment II.

positive outcomes with X-ray images observed in previous experiment II, further improving the overall performance of our classifier.

3.3.2. Performance

After fine-tuning, our best model with combination D achieved a satisfactory accuracy of 0.936 on the test set, with a precision score of 0.750 and an F1 score of 0.779. In Fig. 9, we displayed a multi-label confusion matrix class-wise to visualize the classification accuracy for each class. Each class presented a binary confusion matrix with four groups of sample counts, comprising True Negative (top left box), False Positive (top right box), False Negative (bottom left box), and True Positive (bottom left box).

The classifier demonstrated high consistency (achieving a Precision exceeding 99%) in distinguishing between the majority classes, “Chest, Spine, Upper limb, and Lower limb”. It displayed a low occurrence of misclassifications, with minimal entries in both the False Positive and False Negative categories. However, the model struggled to achieve precise classifications in the following classes: upper arm (Precision of 0.6), thigh (precision of 0.69), and Tibia (Precision of 0.5).

3.3.3. Class activation map (CAM)

To gain a deeper understanding of our classifier, we utilized CAMs to generate a heatmap within all test images. The light region of the heatmap represents the crucial Region of Interest (ROI) for our CNN classification. Our CAMs, featuring a plasma color map, illustrated the significance of different image regions in classifying anatomical SNOMED CT codes. Image regions with a higher contribution were represented in yellow, while those with a lower contribution were displayed in purple.

Four typical scenarios are depicted in the following figures. The classifier successfully identified the key anatomical parts in three scenarios, as illustrated in Fig. 10. However, the classifier failed to highlight the relevant anatomical features in one scenario in Fig. 11.

In Fig. 10a, our classification network effectively highlights anatomical parts; even the two X-ray images involved complex osteosynthetic materials on the hip and ankle, respectively. However, there were instances where the network made incorrect predictions due to similar visual contents, as observed in Fig. 10b, or when the predicted region was nearby but not accurately aligned with the anatomical location, as shown in Fig. 10c. Nevertheless, Fig. 11 also shows that the CNN occasionally identified incorrect image features, particularly when markings were present. These misleading optical features could lead to erroneous predictions by the classifier.

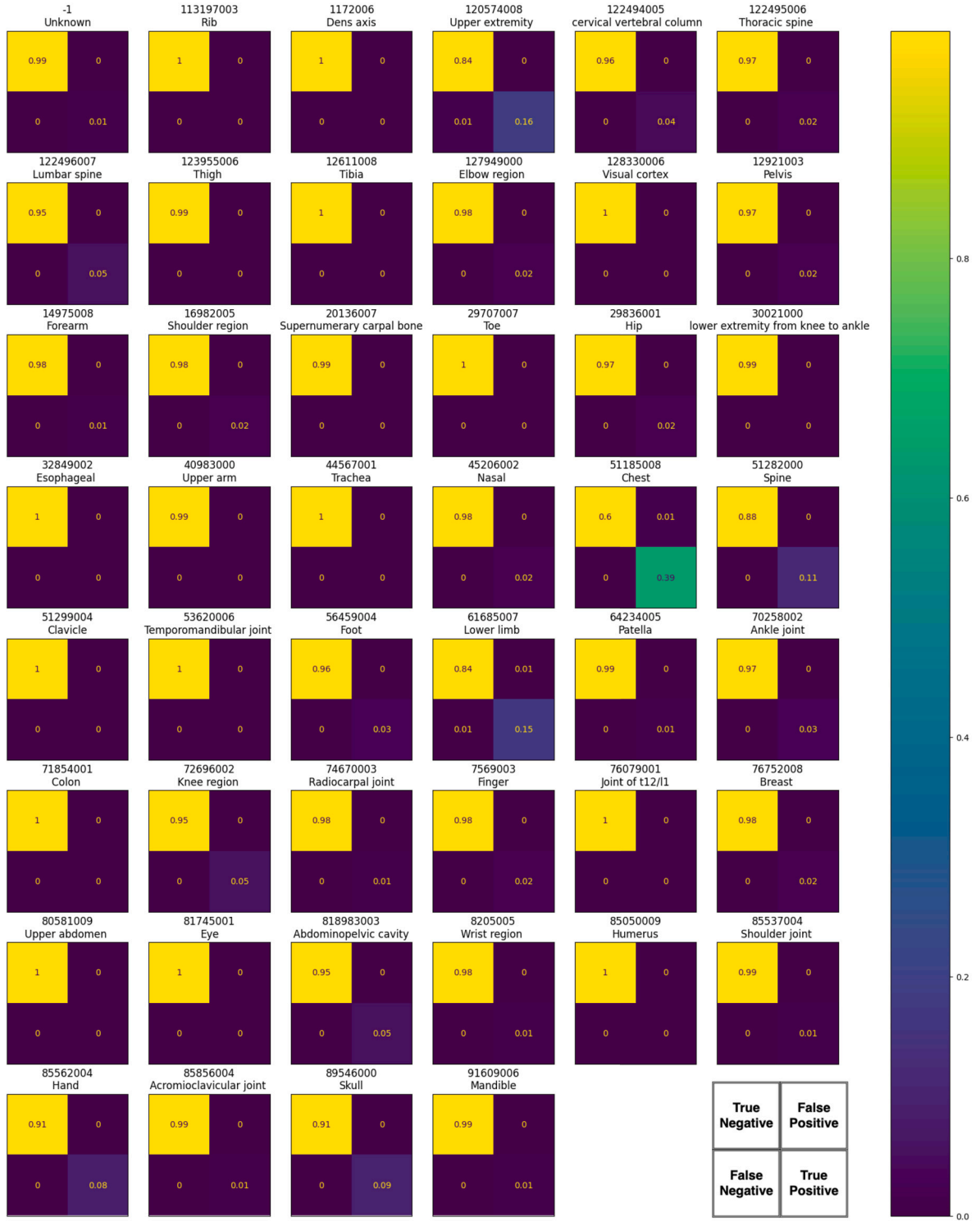
4. Discussion

Our study introduced using DL classification for captioning medical images with alphabetic metadata or SNOMED CT code, showcasing its ability to improve annotation precision for DICOM and non-DICOM images, thereby simplifying medical data integration and retrieval.

We assessed the proposed approach with three experiments: 1) Six medical image modalities from the ROCO dataset were chosen and could be classified with a precision of 0.840 (F1 score of 0.854). 2) 116 IRMA classes of X-ray images from the IRMA dataset with a precision of 0.789 (F1 score of 0.771). 3) ROCO, IRMA and UNIFESP images were fused and refined into 45 SNOMED CT body parts and the classifier achieved a precision of 0.750 (F1 score of 0.779). While 15% of DICOM images contains 15% erroneous tags [7] and even more in non-DICOM images, the performance of our proposed DL algorithm demonstrated its capability to address our real-world problem, achieving an accuracy surpassing 0.936. Consequently, the integration of our approach into comprehensive medical data integration platforms, such as MeDIC, holds promise for enhancing the accuracy of current annotations in both DICOM and non-DICOM images.

4.1. Outliers: image similarity, imprecise annotations, and data quality

The results obtained from the three datasets indicated that a significant portion of our CNN predictions effectively aligned with the actual values. However, it is essential to analyze instances where misclassification occurred. This discrepancy could attributed to several factors that impact the precision of our model.



The classifier is accurate in many cases (45 classes achieved accuracy exceeding 99% and 39 classes achieved precision exceeding 70%).

Fig. 9. Multi-label confusion matrix of 46 SNOMED CT code in Dataset III.

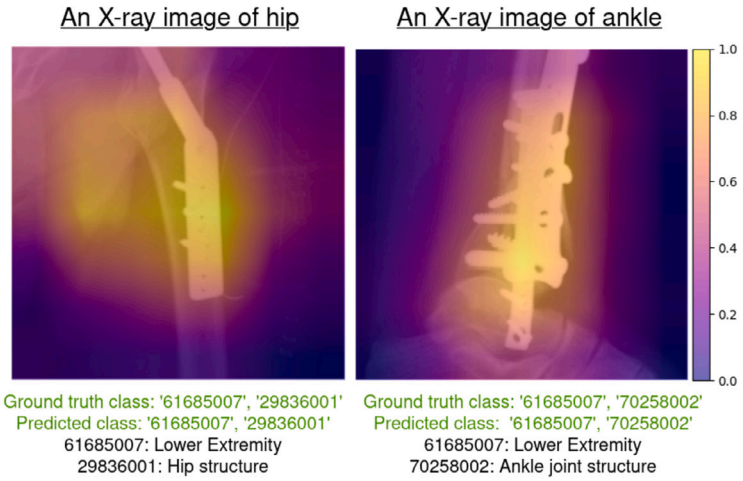
Table 4
Label distribution of SNOMED CT Dataset.

SNOMED CT code	Anatomical meaning	Count	Semantic Types
113197003	Rib	80	Body Part, Organ, or Organ Component
1172006	Dens axis	37	Body Part, Organ, or Organ Component
120574008	Upper extremity	2418	Body Part, Organ, or Organ Component
122494005	cervical vertebral column	622	Body Part, Organ, or Organ Component
122495006	Thoracic spine	341	Body Part, Organ, or Organ Component
122496007	Lumbar spine	637	Body Location or Region
123955006	Thigh	67	Body Location or Region
12611008	Tibia	33	Body Part, Organ, or Organ Component
127949000	Elbow	256	Body Location or Region
128330006	Visual cortex	45	Body Part, Organ, or Organ Component
12921003	Pelvis	441	Body Part, Organ, or Organ Component
14975008	Forearm	207	Body Part, Organ, or Organ Component
16982005	Shoulder region	349	Body Location or Region
20020131	Carpus	56	Body Part, Organ, or Organ Component
29707007	Toe	90	Body Part, Organ, or Organ Component
29836001	Hip	329	Body Part, Organ, or Organ Component
30021000	Low extremity	152	Body Part, Organ, or Organ Component
32849002	Esophageal	3	Body Part, Organ, or Organ Component
40983000	Upper arm	92	Body Part, Organ, or Organ Component
44567001	Trachea	9	Body Part, Organ, or Organ Component
45206002	Nasal	264	Body Part, Organ, or Organ Component
51185008	Chest	5846	Body Location or Region
51282000	Spine	1564	Body Part, Organ, or Organ Component
51299004	Clavicle	46	Body Part, Organ, or Organ Component
53620006	Temporomandibular joint	19	Body Space or Junction
56459004	Foot	485	Body Part, Organ, or Organ Component
61685007	Lower limb	2251	Body Part, Organ, or Organ Component
64234005	Patella	117	Tissue
70258002	Ankle joint	482	Body Space or Junction
71854001	Colon	5	Body Part, Organ, or Organ Component
72696002	Knee region	674	Body Location or Region
74670003	Radiocarpal joint	218	Body Space or Junction
7569003	Finger	345	Body Part, Organ, or Organ Component
76079001	Joint of T12/L1	25	Body Space or Junction
76752008	Breast	326	Body Part, Organ, or Organ Component
80581009	Upper abdomen	37	Body Part, Organ, or Organ Component
81745001	Eye	50	Body Part, Organ, or Organ Component
818983003	Abdominopelvic cavity	709	Body Location or Region
8205005	Wrist region	194	Body Location or Region
85050009	Humerus	12	Body Part, Organ, or Organ Component
85537004	Shoulder joint	134	Body Space or Junction
85562004	Hand	1303	Body Part, Organ, or Organ Component
85856004	Acromioclavicular joint	147	Body Space or Junction
89546000	Skull	1291	Body Part, Organ, or Organ Component
91609006	Mandible	99	Body Part, Organ, or Organ Component
-1	Unknown	102	-

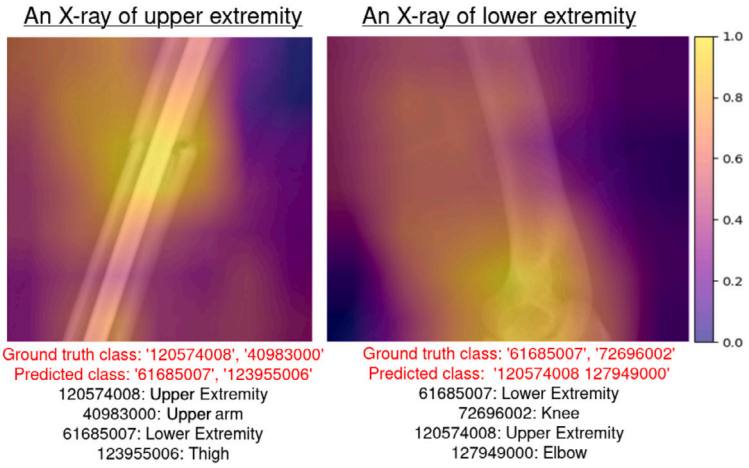
Table 5
Comparison of four preprocessing combinations with custom SNOMED CT dataset.

Image Preprocessing Combination	A	B	C	**D**
Brief description	Resize	Top-Bottom Hat	Cartex	**B and C**
Accuracy	0.897	0.900	0.890	**0.904**
Average Precision	0.622	0.636	0.652	**0.645**
Average Recall	0.747	0.747	0.741	**0.742**
F1 score	0.657	0.664	0.668	**0.673**

One main reason for the misclassification was the visual similarity among specific imaging data, particularly noticeable when differentiating between the upper and lower limbs. In these cases, images lacked distinct visual information beyond bones, as shown in Fig. 11. This ambiguity challenged preprocessing algorithms to highlight distinctive features effectively [46]. Another significant discrepancy stemmed



(a) **CAM overlays of two X-ray images with implants** Two observations emerged from the CAMs and predictions: i) The classifier accurately identified and highlighted the hip and ankle as key anatomical regions. ii) The predicted SNOMED CT code aligned with the ground truth labels, showcasing the capability of the model when complex materials were involved.



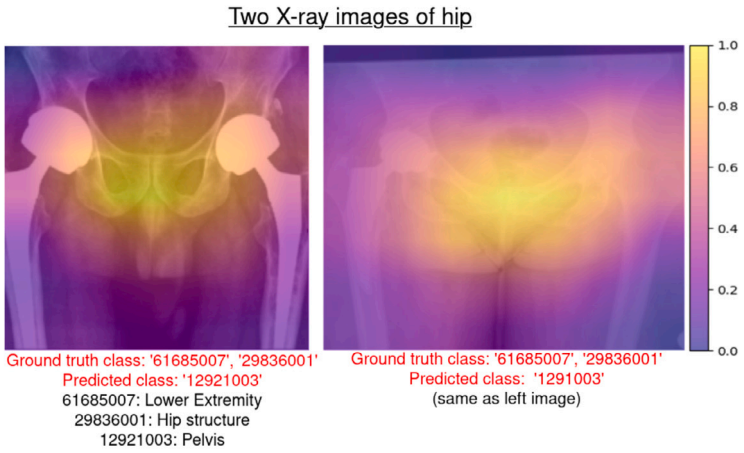
(b) **CAM overlays of two X-ray images with similar visual content** The CAMs and predictions provided two observations: i) CAMs correctly highlighted the upper and lower extremity bones as key anatomical regions. ii) The predicted SNOMED CT codes deviated from the ground truth. The limited depiction of a single bone in both sample images did not provide sufficient information for our classifier, which led to the potential for misclassification.

Fig. 10. Exemplary CAMs overlays in Experiment III highlighted key anatomical regions in grayscale X-ray images.

from the imprecise annotations in the dataset, particularly regarding hierarchical terminology and organ locations. This imprecision led to inaccuracies and negatively impacted the training and evaluation process.

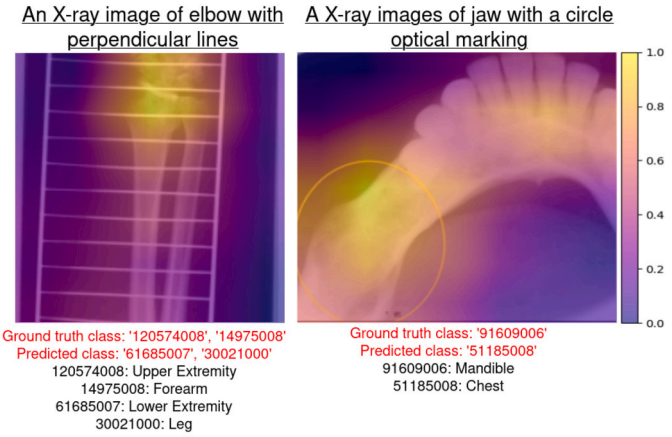
The first key factor leading to unfair evaluation was the imprecision in most medical terms related to radiological modality and human anatomy, which were often organized into complex hierarchies. For instance, our modality classifier in Fig. 7 showed difficulties in distinguishing “CT” and “angiography CT”. Since angiography is a commonly used medical examination method that allows healthcare professionals to visualize and study the blood vessels within the body, medical professionals, depending on the procedure requirements, perform angiography using various modalities, including X-rays, CT, or MRI scans. In other words, angiography can be further classified as “X-ray angiography”, “CT angiography” and “MRI angiography”. This ambiguity contributed to the difficulty in defining these classes. Anatomical terms are also ordered in a specific hierarchical structure. A single-class label was often insufficient to represent anatomical terms in images comprehensively. For instance, various abdominal structures may be present in a chest X-ray, highlighting the need for more nuanced annotations in the training datasets to accurately capture the complexity of anatomical features.

The second imprecise factor stemmed from the intricate details of body parts in the publicly available datasets. Our anatomical classifier occasionally outperformed the dataset by predicting more specific body parts. For example, some abdomen images lacked annotations indicat-



(c) **CAM overlays of two X-ray images with adjacent anatomical regions** Our observations reveal two findings: i) CAMs highlighted the pelvis or hip area in images as the key anatomical region. ii) Unfortunately, the predicted SNOMED CT code did not match the ground truth due to the adjacency of the acquisition area. Although clinical definitions distinguish "pelvis" and "hip" as separate body parts, their physical connection results in their joint appearance in photos, as evident in these two sample images. Consequently, the prediction failed to align with the annotation provided in the test set. Indeed, both predictions should be considered correct, and further recommendations will be provided in our discussion section.

Fig. 10. (*continued*)



Two scanning images with optical markings are shown, and two observations were identified: i) CAM overlays highlighted the optical markings as key anatomical regions, and not our target regions. ii) The predicted SNOMED CT code also failed to align with the ground truth. Our classifiers were distracted by the additional optical markings unrelated to the human body.

Fig. 11. Exemplary CAMs overlays in Experiment III that failed to highlight key anatomical regions.

ing the presence of the spine in the background. Another example is when our model analyzed an X-ray image of the cranium; our classifier might accurately identify the presence of the cranium and nose, showcasing a higher level of accuracy and detail in our predictions than the original dataset. However, the dataset annotations did not encompass such specific predictions. This annotation discrepancy between our model capabilities and the dataset led to unfair evaluation. These imprecise annotations increased false positives during evaluation [47], resulting in lower prediction performance than expected. Given that our annotation process was primarily designed for TBIR, the presence of these false positives can provide valuable insights for user search rather than being viewed as a shortcoming [48].

The third imprecision was that the dataset annotations failed to represent relationships and locations of body parts accurately. Many body parts were located near or within other target areas, but these specific relationships and locations must be precisely annotated. As a result, our classifier struggled to distinguish between certain groups of body parts, such as the hip versus pelvis and the chest versus upper abdomen, because they were usually captured together in medical imaging due to their anatomical positions.

Additionally, dataset biases and quality issues can significantly impact model performance, particularly in sensitive domains such as healthcare. Dataset biases, including non-real-world population representative sample selection, inconsistent annotation processes, and overrepresenting incorrect features, can cause lower overall accuracy and reliability. We have only identified three open-source datasets suitable for our instance-level classification purpose. The quantity and size of clinical data collection were restricted due to research ethics consent and data processing consent [49]. Collecting sufficient annotated images while addressing data imbalance presented a significant challenge to our DL training. These publicly available datasets may exhibit different data distributions compared to authentic clinical images and limit the validity of the practical integration of our models. Data quality issues from the collected datasets, such as noise, missing values, and low resolution, can also adversely affect model performance. These biases can lead to skewed model performance, where the model may perform well on specific subgroups of data but poorly on others.

In the paper, we implemented data preprocessing and augmentation, stratified sampling, and standardized labels with SNOMED CT code to address the quality issues and bias. More data cleaning and preprocessing, bias detection and correction, and multi-source training can be applied in future work to enhance model performance and ensure equitable outcomes. Our next step must also involve using clinical data as the training and test set. The routine patient dataset will improve our precision and validate the robustness and generalization of our current models in real-world scenarios. Other than training with clinical

routine data, we can generate synthetic data to provide more data for training [50]. Furthermore, incorporating domain-specific knowledge and expertise in the future can ensure more precise and detailed annotations that capture the hierarchical structure of the imaging modality and the human anatomy and fairer evaluation. The correctly annotated dataset can further enhance the performance of image captioning models in the medical sector. By accurately representing the relationships and locations of body parts within the dataset, we can improve the ability of the classifier to differentiate between similar groups and reduce false positives. These improvements in the dataset annotations will lead to a more reliable evaluation process and facilitate better model performance.

Our paper addressed quality issues and biases by implementing data preprocessing and augmentation, stratified sampling, and standardized labels using SNOMED CT codes. For future work, additional data cleaning and preprocessing, bias detection and correction, and multi-source training will be applied to enhance model performance further and ensure equitable outcomes. An essential next step involves using clinical data for both training and testing. Utilizing routine patient datasets will improve precision and validate the robustness and generalization of our models in real-world scenarios. Besides training with real patient data, generating synthetic data can augment the training dataset, providing more diverse examples [50]. Furthermore, incorporating domain-specific knowledge and expertise ensures more precise and detailed annotations, especially when it involves the complex hierarchical structure of the imaging modality and human anatomy. By accurately representing the relationships and locations of body parts, the ability of the classifier to differentiate between similar groups will improve, reducing false positives, improving model performance, and ensuring a more reliable evaluation process. Correctly annotated datasets can enhance the performance of image captioning models in the medical sector.

In summary, the primary reasons behind misclassification and unfair evaluation were the visual similarity among images, imprecise annotations in the source dataset—particularly concerning hierarchical details and organ locations—and data bias and quality issues in the collected dataset.

4.2. Other limitations and future work

Beyond unexpected results caused by the dataset, we encountered limitations during training and evaluation, including the challenges of applied preprocessing methods and network architecture, training parameters, and integration complexities for annotating PACS images within MeDIC, highlighting areas requiring further development.

Image preprocessing and augmentation Because of the limitation of current hardware, we resized the images into 224×224 to capture image features for DL classification. However, this could lead to the loss of crucial information in the model. In the future, this practice can be improved based on the future hardware settings.

While our experiments demonstrated improved performance through preprocessing with the proposed techniques, further refinement is required for their application in the context of MeDIC for future extensions. Top-Bottom Hat may remove crucial information or introduce blurring effects, potentially affecting the accuracy. Similarly, Cartex decomposition risks losing significant details in the texture components. Additionally, the efficacy of Cartex depends on parameter selection, including the decomposition method, thresholding, and filtering techniques. These parameters must be thoroughly fine-tuned to effectively integrate these preprocessing techniques into the MeDIC framework.

The applied augmentation algorithm with suggested values implemented in the fast.ai library provides suitable parameters in [41]. Determining the optimal augmentation method for specific data types remains an ongoing challenge [51]. Our augmentation scheme can be further tailored to address long-tailed distribution challenges. Combining class-generic features from head classes with those from tail classes can achieve performance enhancements for the tail classes, as suggested by Chu et al. [52].

DL model and other training parameters ResNet50 has a deep hierarchy and strong representation capability. However, the performance of the proposed pre-trained ResNet50-based classifier can be further improved with other deep learning methods, like a Vision Transformer or a Capsule Neural Network (CapsNet) [53]. CapsNet is more robust to the rotation, translation, and other transformations of objects in the images, and thus, they have been employed for image classifications [53]. More comparing experiments with different models for our classification task is needed.

The cross-entropy loss function used in this work is commonly used [54]. Hybrid loss functions have been used recently to obtain higher performances from different network models [55]. Additionally, a better representation of the hierarchical penalty of human anatomy can be customized in future work.

Integration in MeDIC Integrating automatic annotation within a medical data integration platform like MeDIC can significantly enhance the platform's functionality and efficiency. This enhancement spans improving data interoperability and quality to accelerating advanced analytics without manual data integration. While the benefits are substantial, integrating automatic annotation into MeDIC also presents challenges and works: choosing suitable classification approaches, extending support to diverse image formats, and ensuring accuracy with advanced algorithms and continuous validation.

This study addresses the annotation of non-DICOM data. For DICOM data, we introduced a straightforward conversion method to PIL image objects using PyDicom. In the case of multi-layer DICOM images, our plan involves converting them into multiple non-DICOM images to benefit from our algorithm. Further exploration through additional data and experiments is necessary to validate these approaches thoroughly.

Future integration into MeDIC requires selecting suitable classification approaches. Two classification tasks for anatomical parts can be categorized: pathological classification and anatomical classification. We applied pathological classification in Experiment II and anatomical classification in Experiment III. Pathological classification is often treated as a single-label problem [17], often with one primary diagnosis per image. This approach simplifies annotation, prediction, and model training, aligning with the interests of clinicians. On the other hand, anatomical classification is often treated as a multi-label problem, as images usually present multiple body parts. For instance, a chest X-ray may show the lungs, ribs, and spine, each needing separate identification and analysis. This multi-label approach provides comprehensive understanding, aiding diagnostics, treatment planning and identifying the relevant data for specific research questions. Choosing the suitable method is crucial for effective integration into MeDIC.

Our proposed methodology, based on accuracy with found open-source datasets, offers a viable solution for annotating medical imaging data from PACS into MeDIC. It accurately classifies instance-level keywords into standardized terms, achieving a precision of more than 0.750. Our next steps include extending the PACS image formats, training and testing with routine clinical data, enhancing annotations with expert collaboration, testing other DL models and parameters, and implementing continuous validation mechanisms.

4.3. Comparison with previous studies

While previous studies focus on specific modalities or anatomical classes [12] [13], our research tackled a broader spectrum of challenges. Specifically, we aimed to caption image objects within a generic platform, such as MeDIC, using instance-level standardized terminologies. In this context, our average precision cannot be directly compared with the precision of other studies.

Classification-based approaches in medical imaging are highly efficient and provide clear diagnostic categories, making them well-suited

for straightforward clinical decisions. In contrast, medical image captioning offers comprehensive descriptions that can enhance understanding of complex cases but require more computational resources, extensively annotated data, and complex preprocessing algorithms to limit the data quality and diversity [13]. In addition to classifying with pretrained ResNet50, we explored alternative captioning methods with Encoder-Decoder-based image caption generator [56]. Unlike the classification model, the Encoder-Decoder-based image caption generator handles captions, including complex or ambiguous medical language. The obtained Precision (< 0.6) and BLEU score (< 0.2) fell below satisfactory levels. The suboptimal results can be due to the inherent challenges posed by the source of the dataset. ROCO is a dataset of images collected from publications in PubmedCentral and does not reach standard medical quality and resolution. The unprocessed dataset obtains medical images from diverse medical devices, which are often heterogeneous and low-quality and strongly affect DL performance [13].

Our classifier is aimed to generate standardized clinical terms for TBIR approach [5]. However, one limitation of TBIR was its dependency on manual annotation for image categorization and language-based queries. This constraint can be overcome with an inference processor or alternative retrieval strategies, such as Content-based Image Retrieval (CBIR), as demonstrated in several studies [17–19]. While the CBIR method enhances the reliability of top-ranked results, unfortunately, it cannot support textual queries within our search engine [20], and it demands more memory and computational resources in comparison to TBIR. Hence, it is crucial to carefully consider the trade-offs between different retrieval strategies and choose the most appropriate approach based on the specific requirements of the image captioning task. In future research, we will explore integrating multiple retrieval strategies, e.g., hybrid CBIR [57], to achieve more comprehensive and accurate captioning results.

We need to enhance the quality of our training set to elevate our overall performance. Notably, the proposed classifier exhibited superior efficacy when confronted with heterogeneous data as opposed to the captioning models. The choice between a Classification-based or Caption-based (Encoder-Decoder) approach depends on the specific clinical needs and available resources, each offering unique advantages in different scenarios.

5. Conclusion

We showed the potential of DL classification methods for classifying instance-level (modality, orientation, and anatomy in SNOMED CT) classes while addressing the challenges in future research. All our training sets were publicly available: i) ROCO, ii) IRMA, and iii) custom SNOMED CT dataset. We utilized a ResNet50 CNN to train each set. Our proposed traditional image preprocessing algorithms with Top-Bottom Hat algorithm and Cartex decomposition slightly improve the X-ray classification.

With a precision exceeding 0.75, our method offers a viable option for annotating medical imaging data from PACS in MeDIC. It enables users to search images for secondary use in the platform using the TBIR approach. However, the quality of annotations and the imbalance of images strongly affect the evaluation. Therefore, our next steps will be collecting, cleaning, training, and evaluating productive clinical image data.

CRediT authorship contribution statement

Ka Yung Cheng: Writing – original draft, Validation, Software, Methodology. **Markus Lange-Hegermann:** Writing – review & editing, Methodology. **Jan-Bernd Hövener:** Writing – review & editing. **Björn Schreiweis:** Writing – review & editing, Supervision.

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.

Data availability

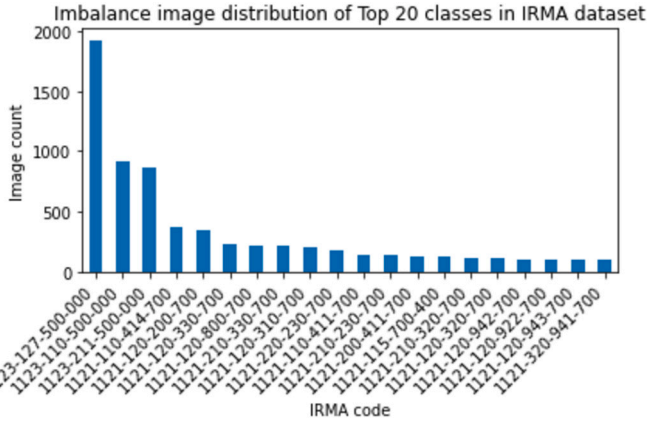
To support the outcomes of this study, the ROCO dataset in Experiment I can be obtained from a public repository at https://github.com/razorx89/roco-dataset, while the IRMA2007 dataset in Experiment II is available at http://publications.rwth-aachen.de/. Additionally, the customized SNOMED dataset utilized in Experiment III is accessible through the following URL: https://opendata.uni-kiel.de/receive/fdr_mods_00000029?accesskey=mCrZLQ4sd5albYQxiFW3hfZOcpPiZYUn and the UNIFESP dataset can be accessed through Kaggle at https://www.kaggle.com/datasets/felipekitamura/unifesp-xray-bodypart-classification. The models are publicly available (https://github.com/KYCheng-Ahoi/IMBODY.git)

Acknowledgement

This project was conducted within the UKSH HiGHmed and IMPETUS projects, funded by the German Federal Ministry of Education and Research [Grant number: FKZ 01ZZ1802T and FKZ 01ZZ2011].

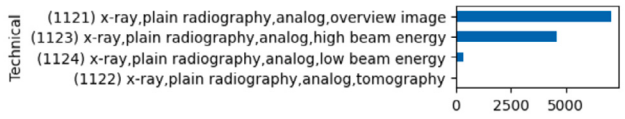
Appendix A. Data and segment distribution of IRMA dataset

Fig. A.12 illustrates the distribution of the 116 classes within the IRMA Dataset (Dataset II). Fig. A.13 analyzes the distribution of each IRMA segment. Fig. A.13(a) depicts a dataset consisting exclusively of X-ray images. Fig. A.13(b) shows the majority of images are labeled as “coronal” and “sagittal”. Figure (c) highlights that “chest” related images are predominant, with over 4000 instances. In contrast, the second most common class, “spine”, and the third class, “hand”, contain fewer than 500 images each. Lastly, in Fig. A.13(d), we observe that the biological segments in the dataset do not provide significant valuable information for our classification task.

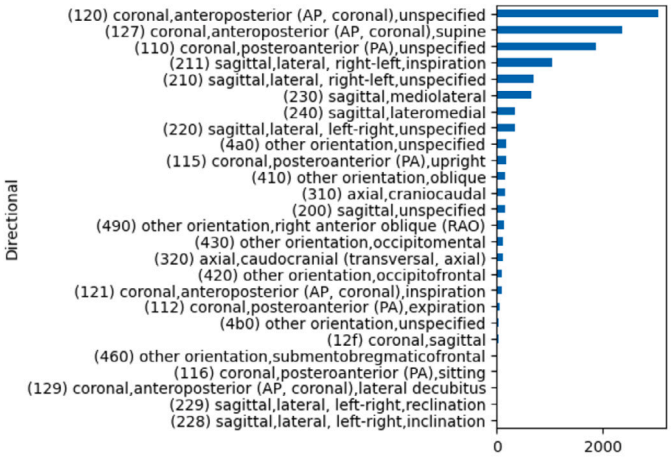


The distribution of images among 116 classes in the IRMA dataset is imbalanced. For more information about the distribution of each IRMA segment, please refer to Fig. A.13. The table of IRMA codes can be found in [24].

Fig. A.12. Imbalance image distribution of Top 20 classes in IRMA dataset.



(a) Technical segment (TTTT)



(b) Directional segment (DDD)

Fig. A.13. Segment distribution of IRMA Dataset.

Appendix B. Cartoon+texture decomposition

In addition to the original algorithm, we introduced a condition to avoid exceptions during the relative reduction rate (rrr) calculation of Local Total Variation (LTV). If a pixel with a value of zero is encountered, the quotient is returned as zero. Furthermore, we adjusted the size of the Gaussian kernel to 7 and reduced the Low Pass Filter (LPF) iteration to 2 — these modifications aimed at capturing more precise textual details.

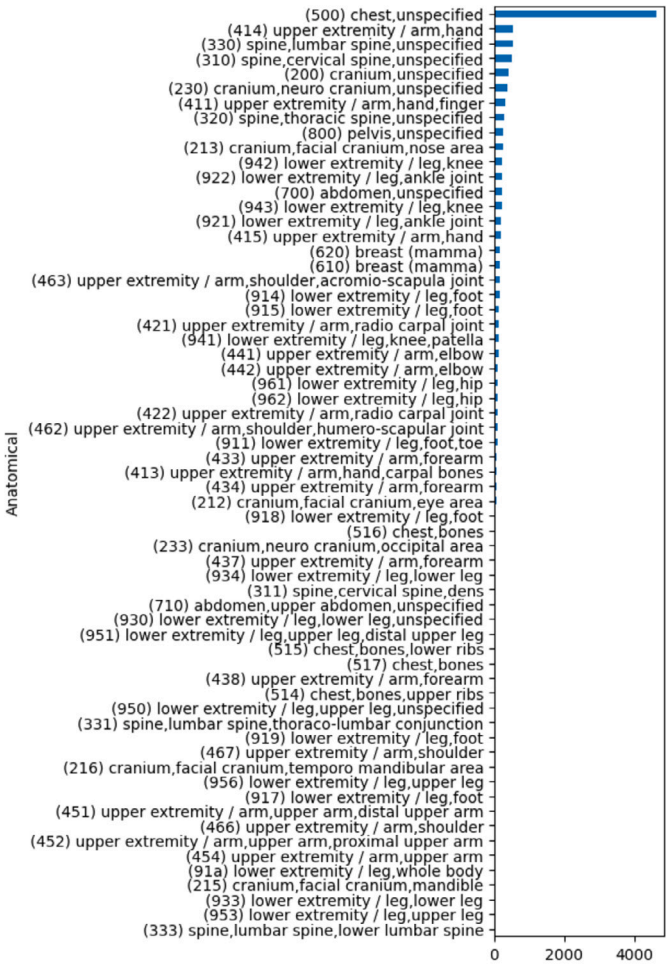
Algorithm 1 Cartoon+Texture decomposition.

Input A grayscale image f
Output (1) the cartoon part U, and (2) the texture part V

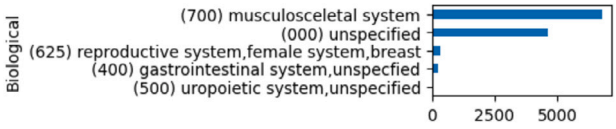
1: Apply a simple low pass filter (LPF) L_σ to the initial image f using *Gaussian blur (smoothing)* iteratively to reduce the high-frequency components of image
2: Compute the Local Total Variation (LTV) of the original and low pass filter images $LTV_\sigma = G_\sigma * |\Delta f|(x)$ to measure the relationship between pixels which are next/close to each other, by convolving G_σ (a Gaussian kernel with standard deviation $\sigma = 2$) with the approximated image gradients (combined the horizontal and vertical changes with Sobel-filter by using Euclidean norm $|\Delta u| = \sqrt{u_x^2 + u_y^2}$) at each point of the initial image f and the low pass filtered image $L_\sigma * f$.
3: Deduce the relative reduction rate (rrr) at each point $\lambda(x) = \frac{LTV_\sigma(x)(f) - LTV_\sigma(x)(L_\sigma * f)}{LTV_\sigma(x)(f)}$ in the image.
4: Compute the value of the cartoon image U as a weighted average of the initial image and the low pass filtered image by $u(x) = (1 - \omega(\lambda(x)))f + \omega(\lambda(x))L_\sigma * f$. If $\lambda(x)$ is small, the function f is non-oscillatory around x, thus $u(x) = f(x)$; if $\lambda(x)$ is large, the function f is locally oscillatory and locally replaced by $L_\sigma * f$.

$$\omega(x)_{a_1=0.25, a_2=0.5} = \begin{cases} 0 & a_1 < x \\ \frac{1}{a_2 - a_1}(x - a_1) & a_1 \le x \le a_2 \\ 1 & x < a_2 \end{cases}$$

5: Compute the texture V as the difference $f - U$



(c) Anatomical segment (AAA)



(d) Biological segment (BBB)

Fig. A.13. (*continued*)

References

[1] Papp L, Spielvogel CP, Rausch I, Hacker M, Beyer T. Personalizing medicine through hybrid imaging and medical big data analysis. Front Phys 2018;6. https://doi.org/10.3389/fphy.2018.00051.

[2] Chen W, Liu Y, Wang W, Bakker EM, Georgiou T, Fieguth PW, et al. Deep image retrieval: a survey. CoRR. arXiv:2101.11282 [abs], 2021. https://arxiv.org/abs/2101.11282.

[3] Mantri M, Taran S, Sunder G. DICOM integration libraries for medical image interoperability: a technical review. IEEE Rev Biomed Eng 2022;15:247–59. https://doi.org/10.1109/rbme.2020.3042642.

[4] Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, et al. HiGHmed – an open platform approach to enhance care and research across institutional boundaries. Methods Inf Med 2018;57(S 01):66–81. https://doi.org/10.3414/me18-02-0002.

[5] Cheng KY, Pazmino S, Bergh B, Lange-Hegermann M, Schreiweis B. An image retrieval pipeline in a medical data integration center. Studies in health technology and informatics. IOS Press; 2024. https://mi-ki.eu/wp-content/uploads/2022/06/Image_Retrieve_Poster_Cheng.pdf.

[6] Tehrani MS. What's the difference between all the different head scans (xray, ct, mri, mra, pet scan)? And what do they show in the head? https://sdbif.org/index/wp-content/uploads/2020/02/Differences-Between-Different-Head-Scan-Types.pdf.

[7] Gueld MO, Kohnen M, Keysers D, Schubert H, Wein BB, Bredno J, et al. Quality of dicom header information for image categorization. In: Medical imaging 2002: PACS and integrated medical information systems: design and evaluation, vol. 4685. SPIE; 2002. p. 280–7.

[8] Marx E. Voices of innovation: fulfilling the promise of information technology in healthcare. HIMSS book series. Taylor & Francis; 2019. https://books.google.de/books?id=9vSDDwAAQBAJ.

[9] Pianykh O. Digital imaging and communications in medicine (DICOM): a practical introduction and survival guide. Springer Berlin Heidelberg; 2009. https://books.google.de/books?id=GpQmSXqhDcMC.

[10] Desjardins B, Mirsky Y, Ortiz M, Glozman Z, Tarbox L, Horn R, et al. Dicom images have been hacked! Now what? Am J Roentgenol 2019;214:1–9. https://doi.org/10.2214/AJR.19.21958.

[11] Varma DR. Managing dicom images: tips and tricks for the radiologist. Indian Jo Radiol Imag 2012;22(01):4–13. https://doi.org/10.4103/0971-3026.95396.

[12] Hossain MZ, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. ACM Comput Surv 2019;51(6):1–36.

[13] Beddiar D-R, Oussalah M, Seppänen T. Automatic captioning for medical imaging (mic): a rapid review of literature. Artif Intell Rev 2022. https://doi.org/10.1007/s10462-022-10270-w.

[14] Chiang C-H, Weng C-L, Chiu H-W. Automatic classification of medical image modality and anatomical location using convolutional neural network. PLoS ONE 2021;16(6). https://doi.org/10.1371/journal.pone.0253205.

[15] Wasserthal J, Breit H-C, Meyer M, Pradella M, Hinck D, Sauter A, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiol Artif Intell 2023;5. https://doi.org/10.1148/ryai.230024.

[16] Zhang S, Zhi L, Zhou T. Medical image retrieval using empirical mode decomposition with deep convolutional neural network. BioMed Res Int 2020;2020:1–12. https://doi.org/10.1155/2020/6687733.

[17] Shamna P, Govindan V, Abdul Nazeer K. Content based medical image retrieval using topic and location model. J Biomed Inform 2019;91:103112. https://doi.org/10.1016/j.jbi.2019.103112.

[18] Srinivas M, Naidu RR, Sastry C, Mohan CK. Content based medical image retrieval using dictionary learning. Neurocomputing 2015;168:880–95. https://doi.org/10.1016/j.neucom.2015.05.036. https://www.sciencedirect.com/science/article/pii/S0925231215006967.

[19] Camlica Z, Tizhoosh HR, Khalvati F. Autoencoding the retrieval relevance of medical images. In: 2015 international conference on image processing theory, tools and applications (IPTA). IEEE; 2015. p. 550–5.

[20] Cheng KY, Pazmino S, Schreiweis B. Etl processes for integrating healthcare data - tools and architecture patterns. Stud Health Technol Inform 2022. https://doi.org/10.3233/shti220974.

[21] Ankit Kumar KK, Garg M. Image captioning and image retrieval. Int J Innov Sci Res Technol 2019;4:909–12.

[22] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[23] Pelka O, Koitka S, Rückert J, Nensa F, Friedrich CM. Radiology objects in COntext (ROCO): a multimodal image dataset. In: Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis. Lecture notes in computer science. Springer International Publishing; 2018. p. 180–9.

[24] Lehmann T, Schubert H, Keysers D, Kohnen M, Wein B. The irma code for unique classification of medical image. Proc SPIE Int Soc Opt Eng 2003;5033. https://doi.org/10.1117/12.480677.

[25] The unifesp x-ray body part classification dataset. https://www.kaggle.com/datasets/felipekitamura/unifesp-xray-bodypart-classification, 2024.

[26] Jalba A, Roerdink J, Wilkinson M. Morphological hat-transform scale spaces and their use in texture classification. In: Proceedings 2003 international conference on image processing (cat. no. 03CH37429), vol. 1; 2003. p. 329–32.

[27] Chang L, Ma W, Yu J, Xu L. An image decomposition fusion method for medical images. Math Probl Eng 2020;2020. https://www.proquest.com/scholarly-journals/image-decomposition-fusion-method-medical-images/docview/2431753227/se-2.

[28] El Ouazzani R, Fattah M, Benamar N. AI applications for disease diagnosis and treatment. In: Advances in medical diagnosis, treatment, and care. IGI Global; 2022. https://books.google.de/books?id=H2V2EAAAQBAJ.

[29] Heydarian M, Doyle TE, Samavi R. Mlcm: multi-label confusion matrix. IEEE Access 2022;10:19083–95. https://doi.org/10.1109/ACCESS.2022.3151048.

[30] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 2921–9.

[31] Abumalloh R, Nilashi M, Yousoof M, Alhargan A, Alghamdi A, Alzahrani A, et al. Medical image processing and covid-19: a literature review and bibliometric analysis. J Infect Public Health 2021;15. https://doi.org/10.1016/j.jiph.2021.11.013.

[32] Bfarm - snomed ct. https://www.bfarm.de/EN/Code-systems/Terminologies/SNOMED-CT/_node.html, 2024.

[33] Selden C, Humphreys B. Unified medical language system: current bibliographies in medicine (jan. '86-dec. '96). Current bibliographies in medicine. DIANE Publishing Company; 1997. https://books.google.de/books?id=m3l4JhmuaaIC.

[34] Mascarenhas S, Agarwal M. A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification. In: 2021 international conference on disruptive technologies for multi-disciplinary research and applications (CENTCON), vol. 1; 2021. p. 96–9.

[35] Xie Y, Richmond D. Pre-training on grayscale ImageNet improves medical image classification. Lecture notes in computer science. Springer International Publishing; 2019. p. 476–84.

[36] Ahmad I, Shin S. An approach to run pre-trained deep learning models on grayscale images. In: 2021 international conference on artificial intelligence in information and communication (ICAIIC); 2021. p. 177–80.

[37] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–324. https://doi.org/10.1109/5.726791.

[38] Yousif M. Enhancing the accuracy of image classification using deep learning and preprocessing methods. Artif Intell Robot Dev J 2024;01. https://doi.org/10.52098/airdj.2023348.

[39] Luke J, Joseph R, Balaji M. Impact of image size on accuracy and generalization of convolutional neural networks. Int J Res Anal Rev 2019;6:70–80.

[40] Hassanpour H, Samadiani N, Mahdi Salehi S. Using morphological transforms to enhance the contrast of medical images. Egypt J Radiol Nucl Med 2015;46(2):481–9. https://doi.org/10.1016/j.ejrnm.2015.01.004. https://www.sciencedirect.com/science/article/pii/S0378603X15000054.

[41] Oronowicz-Jaśkowiak W, Wasilewski P, Kowaluk M. Empirical verification of the suggested hyperparameters for data augmentation using the fast.ai library. CSRN; 2022. https://api.semanticscholar.org/CorpusID:251486323.

[42] Higgins I, Racanière S, Rezende D. Symmetry-based representations for artificial and biological general intelligence. Front Comput Neurosci 2022;16:836498. https://doi.org/10.3389/fncom.2022.836498.

[43] Ruder S. An overview of gradient descent optimization algorithms. arXiv:1609.04747 [abs], 2016. https://api.semanticscholar.org/CorpusID:17485266.

[44] Smith LN, Topin N. Super-convergence: very fast training of neural networks using large learning rates. In: Pham T, editor. Artificial intelligence and machine learning for multi-domain operations applications. SPIE; 2019.

[45] Powers D. Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. J Mach Learn Technol 2011;2:2229–3981. https://doi.org/10.9735/2229-3981.

[46] Abraham A, Rodríguez J, González S, de Paz Santana J. International symposium on distributed computing and artificial intelligence. Advances in intelligent and soft computing. Springer Berlin Heidelberg; 2011. https://books.google.de/books?id=664jzyWwdbsC.

[47] Suganya R, Rajaram S, Abdullah A. Big data in medical image processing. CRC Press; 2018. https://books.google.de/books?id=ZkoPEAAAQBAJ.

[48] Li H, Bin Y, Liao J, Yang Y, Shen HT. Your negative may not be true negative: boosting image-text matching with false negative elimination. In: Proceedings of the 31st ACM international conference on multimedia. New York, NY, USA: ACM; 2023. p. 924.

[49] Singleton P, Wadsworth M. Consent for the use of personal medical data in research. BMJ 2006;333(7561):255–8. https://doi.org/10.1136/bmj.333.7561.255.

[50] Yang W, Nam W. Data synthesis method preserving correlation of features. Pattern Recognit 2022;122:108241. https://doi.org/10.1016/j.patcog.2021.108241. https://www.sciencedirect.com/science/article/pii/S0031320321004222.

[51] Goceri E. Medical image data augmentation: techniques, comparisons and interpretations. Artif Intell Rev 2023;56(11):12561–605. https://doi.org/10.1007/s10462-023-10453-z.

[52] Chu P, Bian X, Liu S, Ling H. Feature space augmentation for long-tailed data. In: Computer vision – ECCV 2020. Lecture notes in computer science. Springer International Publishing; 2020. p. 694–710.

[53] Pawan SJ, Rajan J. Capsule networks for image classification: a review. Neurocomputing 2022;509. https://doi.org/10.1016/j.neucom.2022.08.073.

[54] Bishop CM. Pattern recognition and machine learning. Information science and statistics. Springer; 2007. https://www.worldcat.org/oclc/71008143.

[55] Goceri E. Polyp segmentation using a hybrid vision transformer and a hybrid loss function. J Imag Inf Med Jan 2024. https://doi.org/10.1007/s10278-023-00954-2.

[56] Kinghorn P, Zhang L, Shao L. A region-based image caption generator with refined descriptions. Neurocomputing 2018;272:416–24. https://doi.org/10.1016/j.neucom.2017.07.014. https://www.sciencedirect.com/science/article/pii/S0925231217312511.

[57] Pavithra L, Sharmila TS. An efficient framework for image retrieval using color, texture and edge features. Comput Electr Eng 2018;70:580–93. https://doi.org/10.1016/j.compeleceng.2017.08.030. https://www.sciencedirect.com/science/article/pii/S0045790616308229.