SSNMLRGKSR at ImageCLEFmedical Caption 2024: Medical Concept Detection using DenseNet-121 with MultiLabelBinarizer

Notebook for the ImageCLEF Lab at CLEF 2024

Ramyaa Dhinagaran[1,*,†], Sheerin Sitara Noor Mohamed[1,†] and Kavitha Srinivasan[1,†]

[1]*Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603110, Chennai, India*

Abstract

Concept detection in the medical domain is important for early diagnosis and decision-making to improve overall efficiency and quality of healthcare. It also helps as a prerequisite to generate accurate and contextually relevant captions for medical images that helps timely interventions and to improve patient health. In this paper, the study, analysis and implementation of concept detection is carried out for the dataset given by ImageCLEFmedical forum 2024 task. It has 80,080 radiology images of different modalities such as X-Ray, Computer Tomography (CT), and Magnetic Resonance Imaging (MRI) for development. To develop an optimized concept detection model, a fine-tuned DensetNet-121 architecture has been used (Model1) that efficiently handles complex features and a DenseNet-121 combined with MultiLabelBinarizer (MLB) for preprocessing (Model2) to handle Multi-labeled data, despite with limited data. The proposed Model2 outperforms Model1 with an improved F1-score and secondary F1-score of 0.600 and 0.905 respectively. This score is 5.4 and 10.9 percentage points (0.546 and 0.796) higher than Model1. Also, Model2 achieved fourth rank in ImageCLEFmedical 2024 concept detection task.

Keywords

Concept detection, DenseNet-121, MultiLabelBinarizer, Multi-label classification, Medical domain, Radiology image

1. Introduction

The rapid technological growth and advancement in healthcare domain prompts the need to enhance both the quality and quantity of medical data. As technology evolves, accurate diagnosis and early detection becomes more challenging due to the factors such as ensuring data privacy and security, managing large volumes of information, and addressing disparities in healthcare access. Also, automatic medical image captioning becomes essential due to the increasing availability of medical images across various types of data such as image, text and clinical report or its combinations for different diseases. In this context, medical concept detection plays a crucial role.

Concept detection in medical domain involves the process of identifying and labeling the important features or findings within a medical image for various applications that may be a specific organ or disease [1] [2]. In [3] Lin et al., implemented a concept detection model in the field of Medical Visual Question Answering (VQA) to predict plausible answers for the questions based on the corresponding medical images. The study in [4] shows that concept detection from Electronic Health Records (EHRs) can be used effectively to predict future medical concepts for patients with high precision and recall. Apart from this, it can be used in Natural Language Processing (NLP) and Medical Decision Support Systems (MDSS) [5]. Experts highlight the gap between practical use of concept detection from NLP in healthcare, making them more applicable in real-world applications [6]. Also, MDSS based on NLP [5]

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

[†] These authors contributed equally.

ramyaad@ssn.edu.in (Ramyaa Dhinagaran); sheerinsitaran@ssn.edu.in (Sheerin Sitara Noor Mohamed); kavithas@ssn.edu.in (Kavitha Srinivasan)

0009-0006-6741-0663 (Ramyaa Dhinagaran); 0000-0003-1752-2107 (Sheerin Sitara Noor Mohamed); 0000-0003-3439-2383 (Kavitha Srinivasan)

shows improved acceptance rate among clinicians to expand clinical concepts and integrating more closely with EHR systems.

The above stated concept detection applications are implemented using various Machine Learning (ML) and Deep Learning (DL) approaches such as transfer learning, multi-modal learning, ensemble learning, active learning and Convolutional Neural Network (CNN). In ImageCLEFmedical 2020 concept detection task, Long Short-Term Memory (LSTM) and CNN image encoders [7] are commonly used to develop a better model. With the same dataset, multi-modal learning [8] technique is proposed to developed a reliable model using CNN, NLP, and clustering techniques. Even though CNN is used widely for medical image concept detection, its full potential is often hindered by limited amount of labeled data. This can be addressed by using transfer learning approach [9], [10] where the Deep Convolution Neural Networks (DCNN) improves accuracy even with sparse labeled data. In addition, ensemble learning algorithms and other pre-trained DL models implemented with mammogram images in [11] achieved promising results with limited labeled data. Similarly, in [12] a sparse ensemble learning approach is developed for concept detection from videos. Apart from the limitations of sparse data, problem with class imbalance can be solved using active learning methods. In [13] an active learning model is generated across four datasets which surpasses state-of-the-art methods even with imbalanced data.

Medical concept detection models can be optimized further by choosing appropriate image and label preprocessing techniques. A quality image can be achieved by various techniques such as normalization, resizing and denoising. In addition, label preprocessing techniques will convert the labels into a format compatible with ML algorithms, ensuring consistency and generalization. It also addresses the issues of handling imbalanced data with resampling techniques and multi-label classification using MultiLabelBinarizer.

In general, DL outperforms traditional methods for concept detection [14], indicating its effectiveness in medical domain with sparse data. To advance the field of concept detection, ImageCLEF started conducting concept detection task annually from 2008 to promote open science principles [15] and focused on identifying the presence and location of relevant concepts in a large corpus of medical images. Over the years, datasets and tasks have evolved, becoming larger and more challenging, with a focus on multimodal data access and realistic evaluations. This initiative also aims to improve the accuracy and efficiency of concept detection in medical imaging, thereby enhancing various medical applications and decision support systems. In this paper, two models are developed using DensetNet-121 architecture (Model1) [16] and DenseNet-121 combined with MultiLabelBinarizer (MLB) for preprocessing (Model2) [17] to handle multi-label concept detection dataset.

Remaining sections of this paper are structured as follows: In Section 2, the dataset given by ImageCLEFmedical concept detection task forum and its inferences are discussed. Section 3 describes the proposed system design for concept detection task in detail. A brief summary about the experimental setup and results are given in Section 4, followed by the conclusion and future work in Section 5.

2. Dataset

The ImageCLEFmedical 2024 concept detection dataset consists of 80,080 multi-modality radiology images for development with 70,108 images are allocated for training, 9,972 for validation and 17,237 for testing which is represented in Table 1. In this dataset, each image has several concept IDs (CUIs) that varies from minimum 1 to maximum 27 concepts in train set and 1 to 24 concepts in validation set and is given in Table 2. Each CUIs in turn represents a single concept name (Canonical name). The top 10 frequent CUIs with its canonical names are given in Table 3.

Table 1
Dataset description for concept detection [18]

Dataset	Training Set	Validation Set	Test Set
No. of images	70108	9972	17237
No. of concept IDs (CUIs)	1945	1751	-
No. of concept name (Canonical name)	1945	1751	-

Table 2
Inference from concept detection dataset [18]

Dataset	Concept Per Image	Count	Concept IDs (CUIs)
Training	Minimum	1	7587 images have only one concept
	Maximum	27	C1306645, C0817096, C0205297, C0225756, C0225758, C0006826, C0009450, C1145640, C0877248, C0036525, C0040053, C0085590, C0021102, C0034065, C0919267, C0226550, C0042459, C0225844, C0028259, C0030747, C0225754, C0013922, C0042449, C0740422, C1541923, C0521108, C0524702
Validation	Minimum	1	980 images have only one concept
	Maximum	24	C0041618, C0042149, C0030797, C0032961, C1260954, C0205207, C1288329, C0226378, C0002940, C1510412, C0012359, C0034052, C0225342, C0507850, C0332234, C0040300, C0205321, C0009450, C0877248, C0003842, C0042591, C0229664, C0443294, C0040053

Table 3
Top 10 concept ID with its frequency and its canonical name

S.No.	CUIs	Frequency	Canonical Name
1	C0040405	24227	X-Ray Computed Tomography
2	C1306645	19363	Plain x-ray
3	C0024485	11296	Magnetic Resonance Imaging
4	C0041618	9870	Ultrasonography
5	C0817096	8912	Chest
6	C0002978	4387	angiogram
7	C0037303	3737	Bone structure of cranium
8	C0000726	3726	Abdomen
9	C0030797	3154	Pelvis
10	C0023216	2791	Lower Extremity

3. System Design

The aim of the proposed system is to develop a concept detection model using DL approach. Because, medical image concept detection is significantly more accurate using DL models due to their superior ability of automatically extracting complex features from images as stated in [14]. Hence in the proposed work, two models such as Model1 and Model2 are developed using DenseNet-121 architecture and MLB preprocessing technique and the respective code is available in Github [https://github.com/Sheerin786/Concept-Detection.git]. Model1 is developed using DensetNet-121 architecture [16] to promote better feature extraction from medical images with its dense connections. Model2 is a combination of DenseNet-121 with MLB for preprocessing [17] to effectively handle multiple concept in a single radiology image simultaneously. The system design proposed for the concept detection task is shown in Figure 1. Initially, Model1 and Model2 are developed based on radiology images, concept IDs and concept names from training set and, validation set is used to measure the performance of Model1 and Model2. The parameters used to create Model1 and Model2 are given in Table 4. Since the performance of these models with the given parameters are comparatively higher during development phase, both training

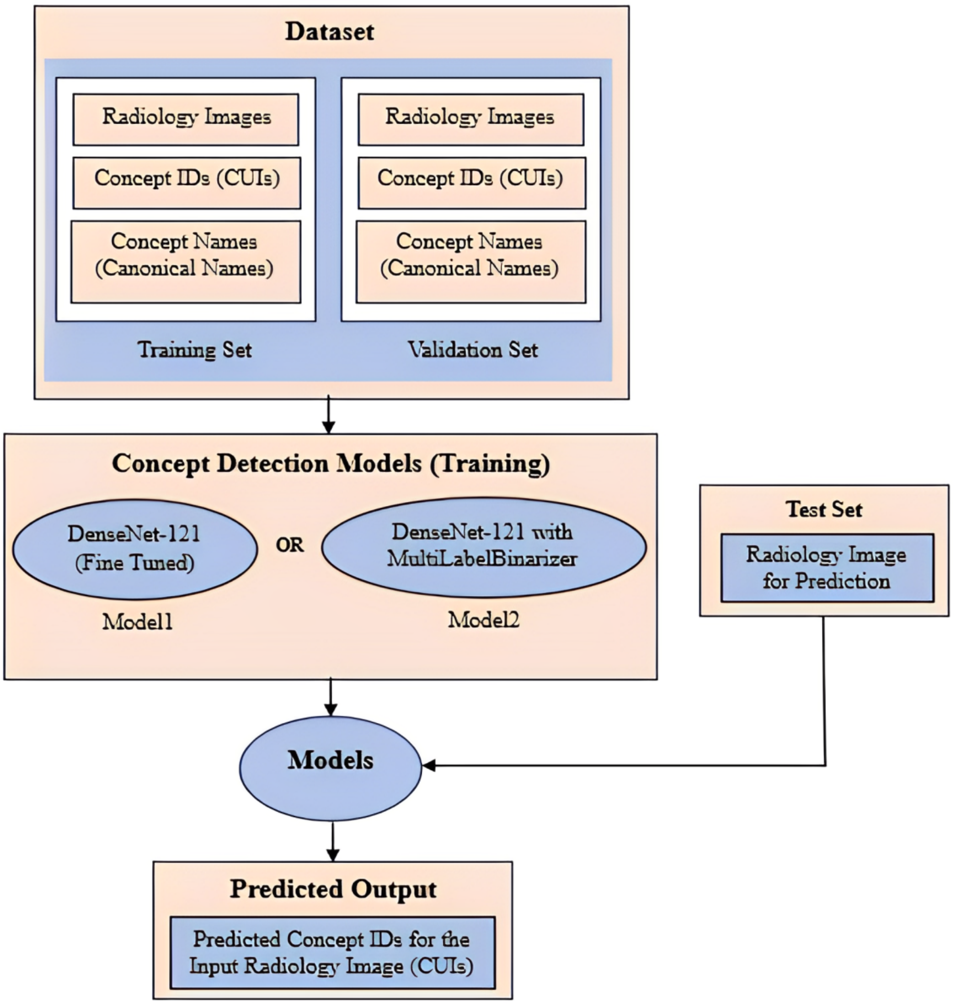

Dataset
Radiology Images
Concept IDs (CUIs)
Concept Names (Canonical Names)
Training Set
Radiology Images
Concept IDs (CUIs)
Concept Names (Canonical Names)
Validation Set
Concept Detection Models (Training)
DenseNet-121 (Fine Tuned)
OR
DenseNet-121 with MultiLabelBinarizer
Model1
Model2
Test Set
Radiology Image for Prediction
Models
Predicted Output
Predicted Concept IDs for the Input Radiology Image (CUIs)

Figure 1: System design proposed for concept detection task

Table 4
Parameters used in Model1 and Model2

Parameters	Value
Loss function	Binary Cross Entropy
Pooling	Global Average Pooling
Activation function	Sigmoid
Optimizer	Adam
Learning rate	0.001
Batch size	32
Epochs	15

and validation dataset are used together for model generation during testing phase.

3.1. Model1: Fine-tuned DenseNet-121

In concept detection tasks, DenseNet-121 architecture allows to extract rich and hierarchical features from medical radiology images [16] [19] irrespective of its modality. These features are essential for identifying complex patterns associated with various medical concepts. Also, DenseNet-121 requires fewer parameters than traditional CNN. Since it is a pretrained model, relearning of redundant feature map is not required and leveraging existing knowledge for faster development and potentially improved performance. In Model1, the pre-trained DenseNet-121 architecture is fine-tuned by freezing its last layer to retain the pre-trained weights and only the weights from added layers are updated. The added layers then extracts each image features and is then mapped with its respective concepts in the training phase. Moreover, the proposed Model1 is further fine-tuned by extracting the features from layer 277 to preserve the lower-level features. Because, Model1 performs comparatively better while learning the features from this layer. During testing, the developed model predicts probability value for each concept with respect to the test images. Further, the concepts with its probability value greater than the threshold value are considered as the predicted concepts.

3.2. Model2: DenseNet-121 with MultiLabelBinarizer

Medical images often shows multiple conditions. The MultiLabelBinarizer (MLB) preprocessing technique effectively handles the multi-label nature of medical annotations [17] and allows better comprehensive labeling of multiple concepts even from a single image. Also, different diseases may have similar features that can be recognized and classified by MLB simultaneously. Hence, before training Model1, MLB transforms the list of concepts and images in the training data as a binary matrix by representing each unique concept as a column and each image as a row. The presence of a concept in the corresponding image is marked as 1 and its absence as 0 in the matrix. This MLB preprocessed data helps Model2 to learn all the suitable concepts better than Model1. During testing, Model2 computes the probability scores for each concept with respect to each test radiology image in the matrix. Further, the relevant concepts for the test images are predicted with the help of two threshold values (0.3 and 0.5).

4. Implementation and Results

In this section, the performance of the concept detection models are analyzed and measured using F1-score and secondary F1-score metrics. F1-score measures the balance between precision and recall. Whereas, secondary F1-score provides additional insight into the model's performance under specific conditions for particular subsets.

4.1. System Specification

The hardware and software required to implement the concept detection models include, (i) Intel i5 processor with NVIDIA graphics card, 4800M at 4.3GHZ clock speed, 16GB RAM, Graphical Processing Unit and 2TB disk space, (ii) Linux – Ubuntu 20.04 operating system, Python 3.7 package with required libraries namely tensorflow 2.7.0, sklearn, tqdm, pickle, pandas, numpy and os.

4.2. Results of Concept Detection Models

The performance of Model1 and Model2 for the test set are measured in terms of F1-score and secondary F1-score as given in Table 5. The F1 and secondary F1-score of submitted Run ID 421 shows the performance of Model1. Here, the feature vectors for the training images of size 255 x 255 pixels are extracted using fine-tuned DenseNet-121 with a learning rate of 0.001 and a batch_size of 32. Apart from learning rate, Model1 is also implemented with a learning rate decay of 90% for every 100 steps. This implies, the learning rate is multiplied by 0.9 for every 100 steps so that the learning rate decreases to 90% from its previous value at each specified interval. This helps in gradually reducing the learning rate by improving the stability and convergence of the training process which in turn makes the model to learn more precise. Further, the model is fine-tuned by frozen the first 276 layers of the DenseNet-121 and the weights from 277 layers are added to improved accuracy. This fine-tuned Model1 gives an F1-score of 0.546 and secondary F1-score of 0.796.

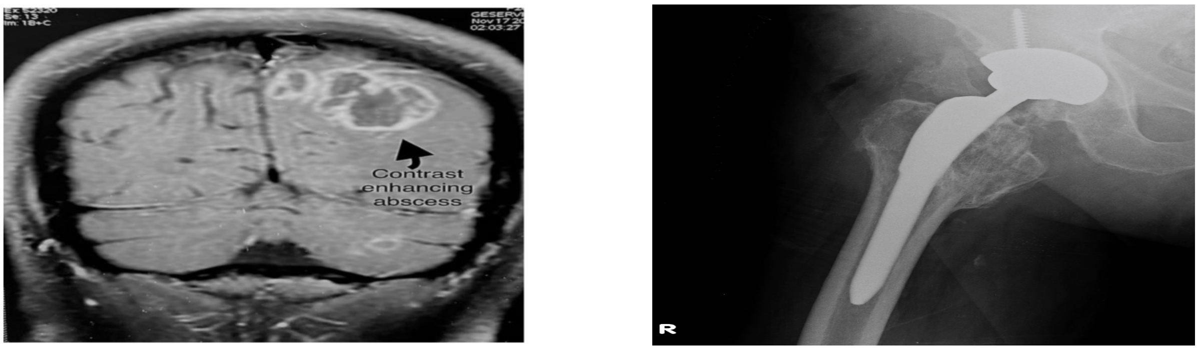
To optimize the model further, Model2 is generated where the input data is preprocessed with MLB and converted to a binary matrix to handle multi-labeled data. The scores of Model2 are given as submitted Run ID 422 and 425 in Table 5. The architecture and design of Model2 with both Run IDs are

Table 5
Results of each run

Run number	Submitted Run ID	Approach/Techniques	F1 Score	Secondary F1 Score
1	421	Model1	0.546	0.796
2	***422***	***Model2 (Threshold=0.3)***	***0.600***	***0.905***
3	425	Model2 (Threshold=0.5)	0.600	0.905

similar, except the threshold value set to detect the relevant concept based on the probability values. The threshold value is set to 0.3 for the model in Run ID 422 and is then changed to 0.5 in Run ID 425 for comparison and they gave the same result. The obtained F1-score and secondary F1-score of both the RunIDs of Model2 are 0.600 and 0.905 respectively. This indicates that MLB is balanced between precision and recall, and remains robust and stable across both the thresholds. Also, the decision boundary set by the Model2 is well-defined for concept detection regardless of specific threshold value within the range of 0.3 and 0.5.

In addition, the F1-score of Model2 is better than Model1 (0.600 vs. 0.546) indicates Model2 performs the concept detection task 5.4 percentage points better than Model1. This indicates a better balance between precision and recall, reducing both false positives and false negatives by Model2. Also, Model2 shows 0.9 percentage points improved performance with secondary F1-score of 0.905 compared to Model1 0.796 irrespective of the threshold values. This shows that, the use of MLB in Model2 significantly enhances its ability to manage multiple labels simultaneously. For instance, a sample of correctly and incorrectly detected samples by Model2 are given in Figure 2. Model2 correctly detects the concept of the image in Figure 2a as C0024485 that implies Magnetic Resonance Imaging. For comparison, in [20] this image is captioned as Magnetic Resonance Imaging scan showing multiple ring enhancement lesions. In contrast, for another image shown in Figure 2b, Model2 wrongly detects the concept as C0024485 (Magnetic Resonance Imaging) instead of X-Ray image. This may be due to feature overlap, since both X-rays and MRIs has similar patterns. Moreover, Model2 may not learned distinguished features. As the model is balck-box in nature the specific reason for wrong prediction is unknown. This can be addressed by using suitable Intrepretable AI techniques like Explainable AI, Decision Tree or Rule-based systems.



(a) Model2 correctly detects the concept C0024485: Magnetic Resonance Imaging, CC BY [Sagar et al. (2022)]

(b) Model2 incorrectly detects the concept as C0024485: Magnetic Resonance Imaging, instead of X-Ray image. CC BY [Muacevic et al. (2022) (Incorrect prediction)]

Figure 2: Correct and incorrect prediction by Model2 [21]

In ImageCLEF 2024 Concept Detection task, 56 teams registered and 10 participated. Among them, our team SSNMLRGKSR made three successful submissions and achieved fourth rank in the concept detection task using Model2. The overall ranking achieved by top 9 teams are listed in Table 6. This improved results are due to an increase in the number of dataset samples under each modality and improved system configuration to create models over years.

Table 6
Top 9 ranking of ImageCLEFmedical 2024 concept detection task

Rank	Team Name	Run ID	No. of Runs Submitted	F1 Score	Secondary F1 Score
1	DBS-HHU	601	8	0.637	0.953
2	auebnlpgroup	644	10	0.631	0.939
3	DS@BioMed	653	5	0.619	0.931
4	***SSNMLRGKSR***	***425***	***3***	***0.600***	***0.905***
5	UACH-VisionLab	235	2	0.598	0.936
6	MICLabNM	681	3	0.579	0.883
7	Kaprov	558	1	0.460	0.730
8	VIT_Conceptz	233	4	0.181	0.264
9	CS_Morgan	530	1	0.107	0.210

5. Conclusion and Future Works

In the context of medical radiology images, DenseNet-121 and MLB techniques have different purposes in different stages during model generation. In this paper, two different concept detection models were developed namely Model1 (fine-tuned DenseNet-121) and Model2 (DenseNet-121 with MultiLableBinarizer) to detect the concepts from radiology images. In addition, F1-score and secondary F1-score of Model2 was analyzed using two different threshold values 0.3 and 0.5. Among Model1 and Model2, the later performs better with an F1-score of 0.600 and secondary F1-score of 0.905, irrespective of the fixed threshold values. This consistent behavior of Model2 suggests that, it is comparatively better than Model1 to handle the concept detection task under various conditions for the given multi-modality, multi-labeled radiology images. In future, these models can be optimized further by using different preprocessing techniques. Also, behaviour of these models can be analyzed to reduce the error. This can be achieved by identifying the important contributing features and the reason for detection using suitable Explainable AI techniques like LIME, SHAP, GradCAM, etc.

Acknowledgments

Our profound gratitude to Sri Sivasubramaniya Nadar College of Engineering, Department of CSE, for allowing us to utilize the High Performance Computing Laboratory and GPU Server, enabling us to contribute to the ImageCLEF 2024 concept detection task successfully.

References

[1] D. Miranda, V. Thenkanidiyoor, D. A. Dinesh, Review on approaches to concept detection in medical images, Biocybernetics and Biomedical Engineering 42 (2022) 453–462. URL: https://www.sciencedirect.com/science/article/pii/S0208521622000158. doi:https://doi.org/10.1016/j.bbe.2022.02.012.

[2] M. Palaniappan, H. Bharathi, E. A. Chodisetty, A. Bhaskar, K. Desingu, Concept detection and caption prediction from medical images using gradient boosted ensembles and deep learning (2023).

[3] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, Z. Ge, Medical visual question answering: A survey, Artificial Intelligence in Medicine (2023) 102611.

[4] N. Le, M. Wiley, A. Loza, V. Hristidis, R. El-Kareh, et al., Prediction of medical concepts in electronic health records: similar patient analysis, JMIR Medical Informatics 8 (2020) e16008.

[5] G. T. Berge, O.-C. Granmo, T. O. Tveit, B. E. Munkvold, A. Ruthjersen, J. Sharma, Machine learning-driven clinical decision support system for concept-based searching: a field trial in a norwegian hospital, BMC Medical Informatics and Decision Making 23 (2023) 5.

[6] S. Tamang, M. Humbert-Droz, M. Gianfrancesco, Z. Izadi, G. Schmajuk, J. Yazdany, et al., Practical

considerations for developing clinical natural language processing systems for population health management and measurement, JMIR Medical Informatics 11 (2023) e37805.

[7] O. Pelka, C. M. Friedrich, A. García Seco de Herrera, H. Müller, Overview of the imageclefmed 2020 concept prediction task: Medical image understanding, CLEF2020 Working Notes 2696 (2020).

[8] R. Sonker, A. Mishra, P. Bansal, A. Pattnaik, Techniques for medical concept detection from multi-modal images., in: CLEF (Working Notes), 2020.

[9] S. Singh, K. Ho-Shon, S. Karimi, L. Hamey, Modality classification and concept detection in medical images using deep transfer learning, in: 2018 International conference on image and vision computing New Zealand (IVCNZ), IEEE, 2018, pp. 1–9.

[10] S. S. N. Mohamed, K. Srinivasan, Ssn mlrg at imageclefmedical caption 2022: Medical concept detection and caption prediction using transfer learning and transformer based learning approaches., in: CLEF (Working Notes), 2022, pp. 1505–1515.

[11] M. C. Berrones-Reyes, M. A. Salazar-Aguilar, C. Castillo-Olea, Use of ensemble learning to improve performance of known convolutional neural networks for mammography classification, Applied Sciences 13 (2023) 9639.

[12] S. Tang, Y.-T. Zheng, Y. Wang, T.-S. Chua, Sparse ensemble learning for concept detection, IEEE Transactions on Multimedia 14 (2011) 43–54.

[13] M. Khaleel, A. Idris, W. Tavanapong, J. R. Pratt, J. Oh, P. C. de Groen, Visactive: Visual-concept-based active learning for image classification under class imbalance, ACM Transactions on Multimedia Computing, Communications and Applications 20 (2023) 1–21.

[14] S. Iqbal, A. N. Qureshi, J. Li, T. Mahmood, On the analyses of medical images using traditional machine learning techniques and convolutional neural networks, Archives of Computational Methods in Engineering 30 (2023) 3173–3233.

[15] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, B. Bracke, H. Damm, T. M. G. Pakull, C. S. Schmidt, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[16] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[17] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, International Journal of Data Warehousing and Mining (IJDWM) 3 (2007) 1–13.

[18] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. Garcıa Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[19] S. S. N. Mohamed, K. Srinivasan, Ssn mlrg at imageclefmedical caption 2023: Automatic concept detection and caption prediction using conceptnet and vision transformer (2023).

[20] T. Sagar, L. Sheoran, A. Prajapati, P. P. Jana, P. Pandey, S. Saxena, et al., Aspergillus fumigatus cerebral abscess following hemodialysis: A case report, Current Medical Mycology 8 (2022) 32.

[21] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology Objects in COntext version 2, an updated multimodal image dataset, Scientific Data (2024). URL: https://arxiv.org/abs/2405.10004v1. doi:10.1038/s41597-024-03496-6.