ImageCLEF-Medical 2025: MedCLIP Model for Medical Caption Prediction and Concept Detection*

Notebook for the sakthiii at ImageCLEFmedical 2025

Sakthi Mukesh Thanga Mariappan[1,*,†], Beulah Arul[1,†] and Muthulakshmi Ramasamy[2,†]

[1]Rajalakshmi Engineering College, Chennai, Tamil Nadu, India
[2]Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India

Abstract

This paper presents our contribution to the ImageCLEF-Medical 2025 challenge, which focused on three core tasks: concept detection, caption prediction and Explainability task for radiological images. This paper focuses solely on the approaches and outcomes related to the Concept Detection and Caption Prediction subtasks undertaken as part of our participation in the ImageCLEF-Medical 2025 challenge. The concept detection task aimed to identify relevant UMLS concepts from biomedical images, while the caption prediction task required systems to generate clinically accurate textual descriptions. We employed a MedCLIP-based transformer model, fine-tuned in a staged manner—first on concept detection, then adapted for caption generation using the learned weights to retain semantic understanding. Our best model achieved an F1 score of 0.4003 for concept detection, with a secondary F1 of 0.9082. For caption prediction, evaluation metrics yielded scores of 0.7957 (Similarity), 0.5553 (BERTScore Recall), 0.1607 (ROUGE-1), and 0.2806 (BLEURT). These results highlight the effectiveness of our transformer based approach in capturing clinical semantics across both tasks.

Keywords

ImageCLEF, Computer Vision, Concept Detection, Multi Label Classification, Image Captioning, Image Understanding, Radiology, MedClip, Deep learning

1. Introduction

ImageCLEF Medical[1] is a prominent initiative within the ImageCLEF framework, aimed at advancing the field of medical image retrieval and analysis. This initiative provides a platform for researchers and practitioners to evaluate and compare various methodologies in medical image processing, including tasks such as image captioning and concept detection. By facilitating the sharing of datasets, benchmarks, and evaluation metrics, ImageCLEF Medical fosters collaboration and innovation among the global research community. The initiative has evolved over the years, adapting to the rapid advancements in technology and the increasing complexity of medical imaging data. It serves as a critical resource for developing automated systems that can assist healthcare professionals in interpreting and utilizing medical images effectively.

Concept detection and captioning are critical components of medical image analysis as they bridge the gap between complex visual data and clinical decision-making by enabling automated interpretation and contextual understanding of radiological images (Shin et al., 2016)[2]. Concept detection focuses on identifying standardized medical terms such as, concepts of the Unified Medical Language System (UMLS)[3], which supports structured indexing, retrieval, and integration with electronic health records, thus enhancing diagnostic precision and interoperability. Captioning, on the other hand, generates coherent and clinically relevant textual descriptions that mimic radiologist reports, offering interpretability, aiding non expert users, and serving as decision support in scenarios with limited radiology expertise. Together, these tasks facilitate efficient, scalable, and explainable AI-driven healthcare systems.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

sakthiimukesh@gmail.com (S. M. T. Mariappan); beulah.a@rajalakshmi.edu.in (B. Arul); muthulakshmiiir@gmail.com (M. Ramasamy)

0009-0005-3044-6841 (S. M. T. Mariappan); 0000-0002-3891-0806 (B. Arul); 0009-0009-9911-5812 (M. Ramasamy)

The domain of medical image analysis has seen advancements driven by deep learning techniques, with particular focus on tasks such as image captioning and concept detection. One of the foundational challenges in the deployment of machine learning models in medical environments is the phenomenon of concept drift, which can adversely affect model performance over time. Huggard et. al.[4] addressed this issue by proposing a calibrated drift detection method[5](CDDM) tailored for medical triage systems. Their work emphasizes the importance of detecting changes in data distributions to ensure models remain accurate and reliable, a consideration that is crucial for maintaining the integrity of captioning and concept detection systems in dynamic clinical settings.

In the broader context of medical image analysis, deep learning has been extensively explored for various pattern recognition tasks[6]. Rehman et. al.[7] conducted a comprehensive survey highlighting the improvements achieved through deep learning approaches in multiple applications, including lesion classification, segmentation, and disease detection. Their review underscores the versatility of deep learning models, such as convolutional neural networks (CNNs)[8], in extracting meaningful features from complex medical images, which is directly relevant to concept detection tasks. These models facilitate the identification of salient features that can be mapped to clinical concepts, thereby supporting automated annotation and captioning. The application of generative models, particularly Generative Adversarial Networks (GANs)[9], has also been explored for medical image analysis. In another research work, Kazeminia et. al. [10] provided a broad overview of GANs, discussing their potential to generate realistic medical images and augment datasets, which can be beneficial for training captioning and concept detection systems, especially in scenarios with limited annotated data. GANs' ability to synthesize diverse and high-quality images can help improve the robustness of models tasked with recognizing and describing medical concepts.

Transformers[11] have emerged as a powerful architecture in medical computer vision. Parvaiz et. al. [12] reviewed the integration of Vision Transformers (ViTs)[13] in medical imaging, emphasizing their effectiveness in tasks such as disease classification, segmentation, and report generation. Their analysis suggests that transformer-based models excel in capturing long-range dependencies and contextual information, which are essential for accurate captioning and concept detection. Similarly, another researcher Shamshad et. al. [14] provided a comprehensive survey of transformers in medical imaging, highlighting their applications across various tasks, including detection and classification, which are directly relevant to the development of automated captioning systems. The importance of explainability and interpretability in medical AI has been recognized as a critical factor for clinical adoption. Bie et. al. [15], proposed a multi-level image-concept alignment framework that enhances explainability by semantically aligning medical images with clinical concepts at multiple levels. This approach addresses a key challenge in concept detection and captioning: ensuring that models not only perform accurately but also provide interpretable outputs that clinicians can trust. Their work underscores the necessity of aligning visual features with semantic concepts to improve the transparency of AI systems.

Recent advancements also include multimodel approaches that combine visual and linguistic data to improve understanding and explanation of medical images. Pham et. al. [16], introduced Silvar-Med, a speech-driven visual language model that integrates speech interaction with medical image analysis. Although primarily focused on abnormality detection, this multimodel framework exemplifies the potential for integrating natural language processing with image analysis, which could be extended to captioning tasks. Such models can facilitate more natural and interpretable communication between AI systems and clinicians, enhancing the usability of automated captioning and concept detection tools. Despite these advancements, challenges remain in the domain of medical captioning and concept detection, particularly in the context of limited annotated datasets and the need for models to generalize across diverse medical conditions and imaging modelities [17]. The literature indicates a trend toward leveraging transfer learning, as demonstrated by Hassan et. al. [15], who used pre-trained ResNet50 models combined with linear discriminant analysis for medical image modelity classification. Transfer learning can be instrumental in addressing data scarcity issues and improving model performance in captioning and concept detection tasks.

Furthermore, the integration of attention mechanisms, as discussed by Liu et. al.[18], has shown promise in enhancing the performance of transformer-based models. Their review highlights that attention mechanisms enable models to focus on relevant regions within images, which is crucial for accurate concept detection and generating meaningful captions. This aligns with the broader movement toward more interpretable and context-aware AI systems in medical imaging. The importance of detecting and adapting to concept drift, as well as integrating multimodel data, are emerging themes that will likely shape future developments in medical captioning and concept detection. As the field progresses, combining these approaches with domain-specific knowledge and clinical validation will be essential to develop robust, accurate, and interpretable AI systems capable of supporting medical decision-making more effective.

2. Dataset

This paper targets two tasks. They are Concept Detection and Caption Prediction for medical images. Both tasks make use of the datasets supplied by the ImageCLEFmedical 2025 challenge.

The dataset for the caption prediction task in the ImageCLEFmedical 2025 challenge consists of carefully selected medical images sourced from biomedical literature, each accompanied by expert-annotated captions and manually curated UMLS terms provided as metadata [19]. For implementation purpose, Radiology Objects in Context Version 2 (ROCOv2) [20] is used. The dataset is split into three portions: a training set consisting of 80,091 radiology images to develop the model, a validation set containing 17,277 radiology images used for tuning and optimization of the model, and a test set formed by 19,267 radiology images that is mainly used for final evaluation of performance. The goal of concept detection is to identify others that are medically relevant from a predetermined set of 2022 UMLS concepts for each image. Performance of concept detection is judged in terms of F1, and F1 secondary scores. The second task, caption prediction, attempts to develop systems that can generate captions that are contextually accurate and meaningful in a medical sense for radiology images. Each image has a caption with human annotations as a reference, and the goal is to produce ultimate text with clinical contents underrepresented in the image. The performance of caption prediction is evaluated by Similarity, BERTScore(Recall)[21], ROUGE-1 [22], BLEURT[23], Relevance average[24], UMLS concept F1[25], AlignScore [26], Factuality average[27].

Most of the images in the dataset are in grayscale, with common modelities being Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and X-ray. The dataset emphasizes diagnostic imaging and anatomical localization, often highlighting key regions such as the head, neck, thoracic cavity, and abdomen. Many images deal with soft-tissue differentiation and structural mapping. CT scans bring axial, coronal, and sagittal anatomical views to human anatomy using soft tissue window settings. These window settings emphasize the difference in tissue density, vastly assisting radiologists in spotting lesions, invasions, and pathological transformations clearly. Among the most usual cuts in this dataset, axial forms provide coarse pictures of the supposed structure-the larynx, piriform fossae, and thyroid cartilage-which are necessary for staging head and neck cancers and assessing anatomical invasion.

To illustrate this, one such example from the dataset is an axial contrast-enhanced CT image of the neck. The Figure 1 shows a soft tissue window and demonstrates the heterogeneously enhancing lesion involving the two sides of the supraglottis, extending into the right piriform sinus, and invading the thyroid cartilage. This anatomical clarity in the image facilitates precise concept annotation and clinically meaningful captioning.

Table 1 represents the UMLS concepts associated with the Figure 1. This example highlights the dataset's rich multimodel structure—combining diagnostic imagery, anatomical precision, and textual descriptions to reflect typical clinical documentation. The annotations enable systems to understand not just object presence, but their clinical relevance, spatial relationships, and contextual importance.

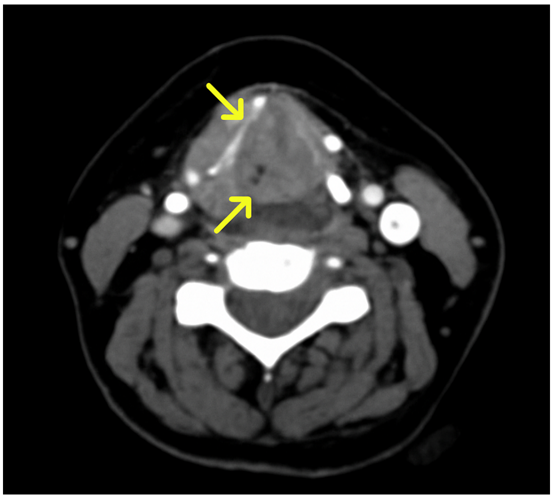

Figure 1: An example image from the dataset. CC BY [Muacevic et al. (2023)]. Caption - Radiological image (axial cuts) Axial cut, soft tissue window contrast computed tomography of the neck showing a heterogeneously enhancing lesion of both sides of the supraglottis extending to the right pyriform sinus (lower arrow) invading the thyroid cartilage (upper arrow).

Table 1
UMLS Concepts and Descriptions

UMLS Code	Concept Description
C0040405	CT
C0225317	Soft Tissues, NOS
C0027530	Neck Structure
C2239273	Supraglottic Part of Larynx
C0227170	Piriform Fossa
C0040126	Thyroid Cartilage

3. Approach

Our methodology, developed and implemented by team sakthiii, was centered around the MedCLIP model [28], a multimodel transformer architecture specifically designed to align medical images with their corresponding textual descriptions. This model, pre-trained on medical image-caption pairs, provided a strong foundation for both concept detection and caption generation tasks. The workflow of the proposed concept detection and caption prediction is shown in Figure 2.

In the first stage, fine-tuning the MedCLIP model for the concept detection task has performed. The training process spanned 11 epochs using a batch size of 32. The optimizer selected was Adam algorithm [29], with a learning rate of 1e-5, which provided effective convergence while maintaining generalization. The dataset consisted of radiology images paired with UMLS concepts, allowing the model to learn mappings between visual features and structured medical vocabulary. After training, the model was validated on a separate validation dataset and then evaluated on the test set released by the organizers. Several iterations were performed to optimize hyperparameters and ensure stability in detection accuracy across multiple runs.

Following the completion of concept detection training, we transitioned to the caption prediction task by reusing the same model weights. The goal was to utilize the semantic representations already learned during concept identification to aid in generating context-aware textual descriptions. Each image was pre-processed and converted to RGB format, then paired with the corresponding concepts from the dataset. Using the MedCLIP processor and tokenization pipeline based on the Transformers library, we prepared both models for their respective tasks (Caption prediction and Concept detection). During model training, the model utilized Cross-Entropy Loss for the caption prediction module, enabling

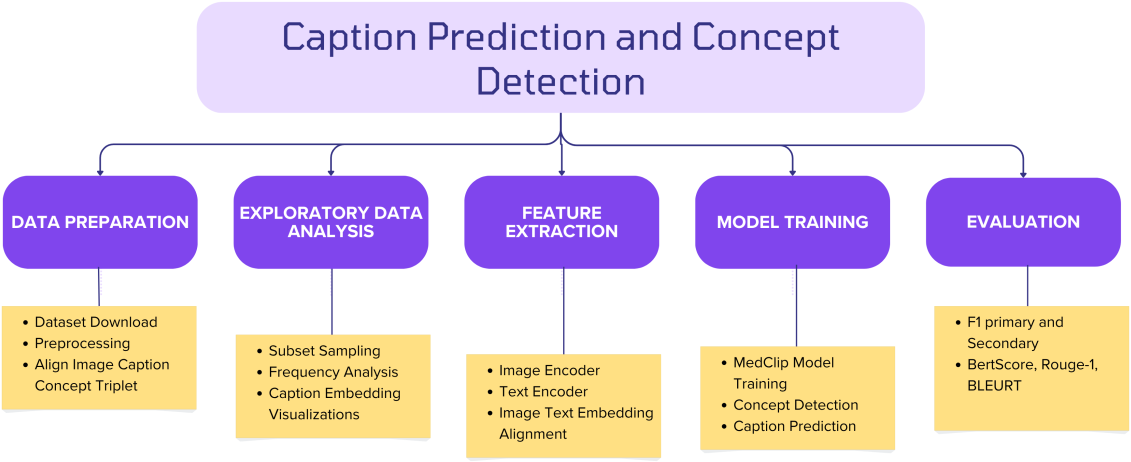
Caption Prediction and Concept Detection
DATA PREPARATION
• Dataset Download
• Preprocessing
• Align Image Caption Concept Triplet
EXPLORATORY DATA ANALYSIS
• Subset Sampling
• Frequency Analysis
• Caption Embedding Visualizations
FEATURE EXTRACTION
• Image Encoder
• Text Encoder
• Image Text Embedding Alignment
MODEL TRAINING
• MedClip Model Training
• Concept Detection
• Caption Prediction
EVALUATION
• F1 primary and Secondary
• BertScore, Rouge-1, BLEURT

Figure 2: Workflow of the proposed Concept detection and Caption prediction.

it to learn accurate word sequences from the given image features. For the concept detection task, Binary Cross-Entropy Loss[30] is employed to handle the multi-label nature of the medical concepts (CUIs), allowing the model to independently assess the presence of each concept in a given input. This seamless transition between tasks helped us to retain shared knowledge and reduce redundant training overhead. This approach demonstrated efficient reuse of learned embeddings and resulted in consistent performance on both tasks. Before finalizing this approach, we did a lot of trail and error with multiple other pre-trained models, which then did not perform well. The codes and models that have used to train the dataset for the tasks, including the models that yielded poor training results, are given in the repository link

- GitHub

. The model's training and evaluation for both the concept detection and caption prediction tasks were performed on a local machine equipped with an NVIDIA RTX 4050 GPU. This setup was sufficient to handle the computational requirements of the training within a reasonable timeframe.

4. Result and Analysis

The results and analysis are discussed in this section. Tables 2 and 3 present the top-performing results achieved by each participating team for the concept detection and caption prediction subtasks, respectively.

4.1. Results for the Concept Detection Sub-task

The Concept Detection Model used exhibits strong performance in accurately identifying relevant concepts within radiological images. Evaluation metrics such as F1-score and F1 secondary score highlight the model's effectiveness. F1 Secondary Score is a variant of the F1 score that measures model performance using an expanded or relaxed version of the ground truth labels, allowing for partial matches, synonyms, or semantically related concepts. It helps capture a model's ability to generalize concept detection beyond exact matches. Among the teams that are participated in the ImageCLEFmedical 2025 Concept detection challenge, our team sakthiii achieved eighth rank in the concept detection task. Table 2 summarizes the best-performing submission from each team.

Table 2
Top Rankings of ImageCLEFmedical 2025 Concept Detection Task

Team Name	F1 Score	F1 Secondary Score
auebnlpgroup	0.588788	0.948442
bahareh0281	0.576579	0.929936
thesalimi	0.576579	0.929936
mapan	0.565985	0.929801
oggyds312	0.561317	0.910382
ds4dh	0.522459	0.867173
oggysashimi	0.454259	0.719997
sakthiii	0.400278	0.908151
jaimage	0.398163	0.832920
ronghaopan	0.239768	0.537660
lekshmiscopevit	0.149379	0.229757

Table 3
ImageCLEFmedical 2025 Captioning - Core Metrics

Owner	Overall	Similarity	BERTScore (Recall)	ROUGE-1	BLEURT
UMUTeam	0.3432	0.9271	0.5977	0.2594	0.3230
DS4DH	0.3362	0.9016	0.6067	0.2516	0.3096
AI Stat Lab	0.3229	0.8919	0.5823	0.2440	0.3173
UIT-Oggy	0.3211	0.8798	0.5951	0.2535	0.3020
AUEB NLP Group	0.3068	0.7947	0.5884	0.2176	0.3030
JJ-VMed	0.3043	0.8251	0.5953	0.2389	0.3094
sakthiii	0.2746	0.7957	0.5553	0.1607	0.2806
csmorgan	0.2315	0.5704	0.5180	0.1598	0.2385

Table 4
ImageCLEFmedical 2025 Captioning - Conceptual and Factual Metrics

Owner	Relevance Avg	UMLS Concept F1	AlignScore	Factuality Avg
UMUTeam	0.5268	0.1816	0.1375	0.1596
DS4DH	0.5174	0.1682	0.1417	0.1549
AI Stat Lab	0.5089	0.1524	0.1213	0.1369
UIT-Oggy	0.5076	0.1672	0.1021	0.1346
AUEB NLP Group	0.4759	0.1429	0.1325	0.1377
JJ-VMed	0.4922	0.1366	0.0964	0.1165
sakthiii	0.4481	0.1094	0.0928	0.1011
csmorgan	0.3717	0.0741	0.1087	0.0914

4.2. Results for the Caption Prediction Subtask

The Caption Prediction Model shows powerful results in building coherent and context-aware captions for medical images. Evaluation metrics such as Similarity, BERTScore (Recall), ROUGE-1, BLEURT, Relevance Average, UMLS Concept F1, AlignScore, Factuality Average were applied to check linguistic quality and semantic alignment between generated caption and reference annotations. The results showed that the model mostly generated relevant and rich captions, comparable to human-written descriptions. The Caption Prediction subtask attracted several participants submitting number of runs that were graded in the ImageCLEF-medical 2025 challenge. Among them, our team sakthiii has achieved the eighth rank in the caption prediction task. Table 3 and Table 4 present the results of the submission.

5. Conclusion

This paper presents a solution for the ImageCLEFmedical 2025 challenge, addressing both caption prediction and concept detection using a two-phase training approach built on the MedCLIP model. In the first phase, MedCLIP was fine-tuned for concept detection by optimizing contrastive and concept-specific losses like Binary-Cross-Entropy to ensure accurate output. In the second phase, the weights of the concept detection MedClip model is used to extract clinical entities from concepts, allowing a BERT-based classifier to predict medical captions directly from visual features. Our system showed strong results across metrics like BLEURT, ROUGE-1, BERTScore and overall, highlighting the effectiveness of combining multimodel embeddings with domain-specific knowledge. While the approach had limitations, such as dependence on static thresholds and occasional generic captions, future improvements will explore dynamic thresholding, prompt-based generation, and more advanced encoders. In general, our work demonstrates the potential of AI-driven frameworks in clinical decision support and medical image understanding.

6. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly to: Grammar and spelling check. After using these services, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

References

[1] H. Damm, T. M. G. Pakull, H. Becker, B. Bracke, B. Eryilmaz, L. Bloch, R. Brüngel, C. S. Schmidt, J. Rückert, O. Pelka, H. Schäfer, A. Idrissi-Yaghir, A. B. Abacha, A. G. S. de Herrera, H. Müller, C. M. Friedrich, Overview of imageclefmedical 2025 – medical concept detection and interpretable caption generation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[2] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning, IEEE Transactions on Medical Imaging 35 (2016) 1285–1298. doi:10.1109/TMI.2016.2528162.

[3] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.

[4] H. Huggard, Y. S. Koh, G. Dobbie, E. Zhang, Detecting concept drift in medical triage, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1733–1736. URL: https://doi.org/10.1145/3397271.3401228. doi:10.1145/3397271.3401228.

[5] S. E. Davis, R. A. Greevy Jr, T. A. Lasko, C. G. Walsh, M. E. Matheny, Detection of calibration drift in clinical prediction models to inform model updating, Journal of biomedical informatics 112 (2020) 103611.

[6] P. RAVISANKAR, A. Beulah, S. ELANGO, Cell classification in microscopic images for anemia detection, Romanian Journal of Information Technology and Automatic Control 34 (2024) 7–12.

[7] A. Rehman, M. Ahmed Butt, M. Zaman, A survey of medical image analysis using deep learning approaches, in: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1334–1342. doi:10.1109/ICCMC51019.2021.9418385.

[8] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: Analysis, applications, and prospects, IEEE Transactions on Neural Networks and Learning Systems 33 (2022) 6999–7019. doi:10.1109/TNNLS.2021.3084827.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, Generative

adversarial networks: An overview, IEEE Signal Processing Magazine 35 (2018) 53–65. doi:10.1109/MSP.2017.2765202.

[10] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, A. Mukhopadhyay, Gans for medical image analysis, Artificial Intelligence in Medicine 109 (2020) 101938. URL: https://www.sciencedirect.com/science/article/pii/S0933365719311510. doi:https://doi.org/10.1016/j.artmed.2020.101938.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[12] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, M. M. Fraz, Vision transformers in medical computer vision—a contemplative retrospection, Engineering Applications of Artificial Intelligence 122 (2023) 106126. URL: https://www.sciencedirect.com/science/article/pii/S095219762300310X. doi:https://doi.org/10.1016/j.engappai.2023.106126.

[13] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., A survey on vision transformer, IEEE transactions on pattern analysis and machine intelligence 45 (2022) 87–110.

[14] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, H. Fu, Transformers in medical imaging: A survey, Medical Image Analysis 88 (2023) 102802. URL: https://www.sciencedirect.com/science/article/pii/S1361841523000634. doi:https://doi.org/10.1016/j.media.2023.102802.

[15] M. Hassan, S. Ali, H. Alquhayz, K. Safdar, Developing intelligent medical image modality classification system using deep transfer learning and lda, Scientific reports 10 (2020) 12868.

[16] T.-H. Pham, T.-D. Bui, M. L. Quang, T. H. Pham, C. Ngo, T. S. Hy, Silvar-med: A speech-driven visual language model for explainable abnormality detection in medical imaging, in: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops, 2025, pp. 2984–2994.

[17] A. Beulah, T. S. Sharmila, V. Pramod, Degenerative disc disease diagnosis from lumbar mr images using hybrid features, The Visual Computer 38 (2022) 2771–2783.

[18] Z. Liu, Q. Lv, Z. Yang, Y. Li, C. H. Lee, L. Shen, Recent progress in transformer-based medical image analysis, Computers in Biology and Medicine 164 (2023) 107268.

[19] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, L.-D. Ştefan, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science (LNCS), Madrid, Spain, 2025.

[20] 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. Horn, F. Nensa, C. M. Friedrich, Rocov2: Radiology objects in context version 2, an updated multimodal image dataset, Scientific Data 11 (2024). doi:10.1038/s41597-024-03496-6.

[21] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[22] J.-P. Ng, V. Abrecht, Better summarization evaluation with word embeddings for ROUGE, in: Conference on Empirical Methods in Natural Language Processing, volume abs/1508.06034, 2015. doi:10.18653/v1/d15-1222.

[23] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: D. Jurafsky,

J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704/. doi:10.18653/v1/2020.acl-main.704.

[24] S. E. Robertson, E. Kanoulas, E. Yilmaz, Extending average precision to graded relevance judgments, in: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, pp. 603–610.

[25] A. Abbas, M. Afzal, J. Hussain, T. Ali, H. S. M. Bilal, S. Lee, S. Jeon, Clinical concept extraction with lexical semantics to support automatic annotation, International Journal of Environmental Research and Public Health 18 (2021) 10564.

[26] Y. Zha, Y. Yang, R. Li, Z. Hu, AlignScore: Evaluating factual consistency with a unified alignment function, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11328–11348. URL: https://aclanthology.org/2023.acl-long.634/. doi:10.18653/v1/2023.acl-long.634.

[27] R. Saurí, J. Pustejovsky, Are you sure that this happened? assessing the factuality degree of events in text, Computational linguistics 38 (2012) 261–299.

[28] A. U. R. Hashmi, D. Mahapatra, M. Yaqub, Envisioning medclip: a deep dive into explainability for medical vision-language models, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE, 2024, pp. 1–5.

[29] M. Reyad, A. M. Sarhan, M. Arafa, A modified adam algorithm for deep neural network optimization, Neural Computing and Applications 35 (2023) 17095–17112.

[30] U. Ruby, V. Yendapalli, et al., Binary cross entropy with deep learning technique for image classification, Int. J. Adv. Trends Comput. Sci. Eng 9 (2020).