Journal of Computer Science

Research Article

Automatic Skin Lesion Diagnosis and Medical Report Generation Based on Image Captioning

Abdelouahed Sabri, Chaimae Zouitni, Hamza El Medhoune and Abdellah Aarab

Department of Computer Science, Faculty of Sciences Dhar el Mahraz, Sidi Mohammed Ben Abdellah University, Fez, Morocco

Article history
Received: 08-10-2024
Revised: 25-11-2024
Accepted: 20-03-2025

Corresponding Author:
Abdelouahed Sabri
Department of Computer Science, Faculty of Sciences Dhar el Mahraz, Sidi Mohammed Ben Abdellah University, Fez, Morocco
Email:
abdellah.aarab@usmba.ac.ma

Abstract: Captioning or textual description of the visual content of images involves generating meaningful words and sentences to describe the content of an image. This work lies at the crossroads of Natural Language Processing (NLP) and computer vision. When dealing with medical images and especially skin lesions, it goes beyond simple classification to generate detailed textual reports describing the skin lesion's condition comprehensively. Such reports are crucial for supporting clinical diagnosis and decision-making. The novelty of this study lies in the creation of the first dataset specifically designed for skin lesion captioning, generated using expert-validated descriptions based on the ABCDE rules. Our approach integrates the VGG16 architecture for feature extraction and LSTM for textual description generation. The proposed method was evaluated on the PH2 dataset and achieved a BLEU-1 score of 0.50, demonstrating its promise for aiding dermatological diagnosis.

Keywords: Image Captioning, Skin Lesion, Deep Learning, VGG16, NLP, PH2, BLEU Score

Introduction

Artificial Intelligence (AI) has had a crucial impact on the development of medical diagnostic tools. It has enabled the design and development of powerful tools for medical image classification. Image analysis and classification is a highly coveted field of research that has enabled and will undoubtedly enable the development of very powerful systems. Captioning or textual description of medical images is a recent research area that has attracted the attention of several researchers. A textual description system of medical images allows, in addition to classification, to generate reports and thus helps specialists to make decisions.

Image captioning, as a crossroad of Natural Language Processing (NLP) and computer vision, involves describing the visual content of images with meaningful textual descriptions (Hossain *et al.*, 2019). In the context of medical images, NLP plays a crucial role in structuring these descriptions into coherent sentences that align with professional diagnostic language. This dual-domain approach ensures that the generated captions are not only visually descriptive but also semantically and clinically relevant. In our study, NLP facilitates the transformation of extracted visual features into textual medical reports, bridging the gap between image analysis and practical diagnostic support.

Captioning systems are based on supervised classification models, which require a labeled learning database. The main issue encountered in medical image captioning systems is the availability of datasets containing both the images and their captions. Medical image datasets with textual descriptions or captions are very rare and obstruct the development of this type of system (Tian *et al.*, 2020). Image datasets of skin lesions with captions are not yet available.

The focus of this work is on skin lesion captioning, a domain with unique challenges and high clinical relevance. Unlike existing methods that primarily classify lesions, our approach aims to generate "almost complete textual reports," capturing critical diagnostic features of skin lesions. These reports provide a more nuanced understanding of lesion characteristics, aligning closely with how dermatologists assess and document lesions in practice.

One of the main novelties of this study is the creation of the first dataset specifically tailored for skin lesion captioning. Leveraging the PH2 dataset, we generated captions based on the ABCDE rules of dermatological diagnosis, validated by domain experts (Mendonca *et al.* 2013, Filali *et al.*, 2020). This dataset bridges a significant gap in the field, enabling the application of advanced deep learning models to a previously unexplored domain.

Our proposed method combines the benefits of computer vision and NLP. To extract visual features, we employ the VGG16 network, and utilize a Long Short-Term Memory (LSTM) model for the generation of corresponding textual descriptions. This integration ensures the captions are both clinically meaningful and linguistically coherent. The evaluation results, including a BLEU-1 score of 0.50, demonstrate the feasibility and potential of this approach in aiding dermatological diagnostics.

Medical Image Captioning

Image captioning is a prevalent challenge in contemporary Artificial Intelligence (AI), focused on generating descriptive text that conveys the content of a given image (Figure 1). This task lies at the crossroads of Natural Language Processing (NLP) and computer vision. In the case of medical images, the challenge is big, because medical image captioning will allow the automatic generation of diagnosis reports.

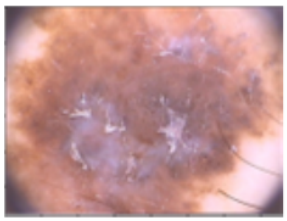
Long-term studies have focused on how low-level visual elements like color, texture, and shape can be used to perceive and comprehend high-level semantic information found in images, such as scenes, objects, and relationships. This was carried out through multiple stages, including segmentation, object detection, classification, and semantic interpretation (Filali *et al.*, 2022; 2019). Advancements in each of these stages have paved the way for generating comprehensive textual interpretations of images, a task commonly referred to as semantic description or image captioning. Developing an effective image captioning system typically involves three key components that must be taken into account (Sharma *et al.*, 2017).

1. The dataset to use since we need image datasets with at least one caption for each image
2. The approach to use to extract relevant features from images
3. The mechanism to generate textual description (caption)
4. The evaluation and validation metrics

The number of published works dealing with captioning in the medical field is very low when compared to the number of published articles for medical image classification. This is certainly due to the shortage of datasets containing both medical images and their textual descriptions. In a recently published survey, the authors raised this problem while emphasizing the problems encountered (Beddiar *et al.*, 2022).

As already mentioned, the major problem encountered is the scarcity of medical image datasets with textual descriptions compared to natural image datasets. A range of image captioning datasets exists, varying in image count, object diversity, and associated attribute complexity (Luo *et al.* 2022). These include MS COCO Dataset, Flickr30K and Flickr8K datasets, and Visual Genome Dataset. These datasets contain many images with a fairly large number of descriptions per image. Within the domain of medical images, we can cite the Indiana University (IU) Chest X-Ray dataset that contains more than 7000 chest X-Ray images, and each image has multiple annotations (Demner-Fushman *et al.*, 2016). CheXpert is another Chest X-ray dataset used in medical image captioning. This dataset contains almost 225000 multi-view chest radiographs (Irvin *et al.*, 2019). The ImageCLEF dataset used in the ImageCLEF competition in 2017 contains 184614 biomedical images (Eickhoff *et al.*, 2017).

A skier performing a jump against some snow



A fully asymmetric lesion that contains two colors with An irregular pigment network, along with unusual dots and globules, as well as the presence of a blue whitish veil

Fig. 1: Example of images with captions

To date, there is no dataset available for skin lesion captioning. And we had to create our own dataset to be able to develop a system for the generation of diagnostic reports for this type of cancer.

Related Work

Medical images captioning is attracting more and more attention from the scientific research community. A fairly consistent number of scientific research works have been published in recent years. Singh *et al.* (2022) presented an approach to medical captioning based on the Attend and Tell (ATM) model. They propose to refine the model using the Strength Pareto Evolutionary Algorithm (SPEA). The proposed approach has been tested and evaluated using the VQA-Med medical image captioning dataset (Abacha *et al.*, 2019). The proposed approach resulted in an improvement in terms of the kappa measure to 0.9382. An attention based GRU language generator in combination with a CNN encoder model was proposed by Beddiar *et al.* (2019) to generate medical image captions. The proposed approach was able to achieve a BLEU score value equal to 0.243 on the Image CLEF challenge. Harzig *et al.* (2019) proposed to use a Deep Convolutional Neural Network (DCNN) to classify diseases from gastrointestinal images. They also propose to use Class Activations Maps (CAM) to generate textual rapport. A Semantic Fusion Network (SFNet) that involves a model for the detection of lesion

area and for the diagnostic generation is proposed by Zeng *et al.* (2020). The proposed approach has been tested and evaluated using the Open-i X-ray Image Dataset. AMAnet is an Adaptive Multimodal Attention network proposed in the paper published by Yang *et al.* (2021). The main idea is to generate medical diagnosis reports. The proposed approach is based on the prediction of multi-labels using local proprieties. To generate the final report, a semantic word embedding vector is used.

Medical image captioning has gained significant attention in recent years, with various approaches leveraging deep learning for automatic diagnosis report generation. For example, CNN-GRU-based models have been applied to the ImageCLEF dataset with BLEU score evaluations, and attention-based networks like AMAnet have shown promising results on chest X-ray datasets such as CheXpert and IU X-Ray (Harzig *et al.*, 2019). However, these methods rely heavily on existing datasets tailored for specific modalities like radiology, which include pre-labeled captions. In contrast, the domain of dermatology lacks a publicly available dataset for captioning skin lesion images.

Our work addresses this critical gap by creating a novel dataset based on the PH2 database, where captions were generated from expert-validated features derived from the ABCDE dermatological diagnostic rules. Unlike existing studies, which primarily compare algorithmic performance on pre-existing datasets, our contribution lies in establishing a foundation for skin lesion captioning research. This novel dataset enables the application of state-of-the-art techniques to a previously unexplored area, paving the way for advancements in automatic dermatological report generation.

Materials and Methods

Medical images captioning has become a very attractive topic in recent years. Its purpose is to assist physicians and health professionals in report writing and diagnosis. The scarcity of image databases with descriptions has slowed down the development of this research area. In this paper, we proposed an automatic captioning approach for skin lesion images. We proposed to use an approach by merging the visual features of the skin lesion images with the prefixes of the descriptions. The general scheme of the proposed approach is presented in Figure (2).

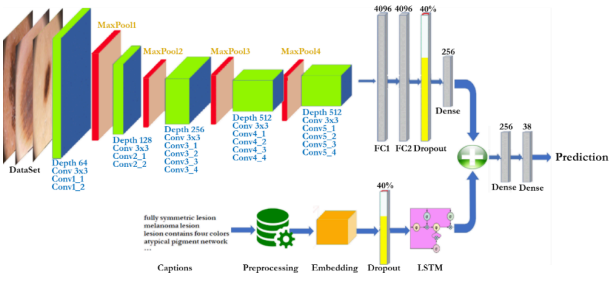


Fig. 2: General scheme of the proposed skin lesion captioning approach

Building the Training Dataset

As previously mentioned, there is not yet a dataset for skin lesions with captions, so we have proposed to create our own image base. We used the PH2 dataset, and we used the features provided with the base to create the training descriptions (Teresa *et al.*, 2013). The PH2 database is freely available for research and benchmarking purposes through the National Library of Medicine (Teresa *et al.*, 2013). These features are derived from the criteria used in clinical diagnosis. They are based on the ABCDE (Asymmetry, Border, Color, Diameter, and Evolving) rules used by dermatologists. The PH2 dataset contains annotation of medical diagnosis for the 200 images which are:

- Colors Present: May include black, blue-gray, light brown, dark brown, red, and/or white
- Asymmetry: Indicates whether the lesion is asymmetrical or not
- Pigment Network: Categorized as typical or atypical
- Dots/Globules: Classified as absent, typical, or atypical
- Streaks: Either present or absent
- Regression Areas: Either present or absent
- Blue-Whitish Veil: Either present or absent

These features will be used to create and generate descriptions for each of the images in the dataset. These descriptions are verified and validated by professionals. Table (1) presents an example of features of some images of the PH2 database. The PH2 database is composed of 200 dermoscopy images, where 120 are non-melanoma and 80 are melanoma.

Captions were validated by a dermatology professional. Table (2) shows examples of captions generated based on the features presented in Table (1).

Visual Features Extraction

In our approach, we use the VGG16 architecture, a widely recognized convolutional neural network pre-trained on the ImageNet dataset, for feature extraction. VGG16 was chosen due to its proven effectiveness in extracting robust feature representations for skin lesion classification tasks, including prior applications to the PH2 dataset. Its deep yet simple architecture ensures high-quality visual feature extraction, which is further refined in our approach by adding a dense layer to reduce the dimensionality of extracted features, tailoring them for the caption generation process (Sabri *et al.*, 2020).

In our proposed approach, as illustrated in Figure (2), 4096 features were extracted from each of the images in the PH2 dataset, and a dense layer is used to create only 256 features by combining the previously extracted 4096 features.

Captions Preprocessing and Cleaning

The captions were cleaned up and pre-processed by removing unnecessary special characters and spaces and adding caption delimiters. We used a dropout layer by

removing 40% of the entries to reduce the effect of overfitting. The captions are used by the LSTM architecture to create textual features.

Design of the Caption Generation Model

The caption generation process involves integrating visual features extracted using the VGG16 architecture with textual representations generated using the LSTM model. NLP is pivotal in this step, as it ensures that the generated captions are linguistically coherent and clinically meaningful. By leveraging the sequential modeling capabilities of LSTM, the system translates numerical features into structured natural language descriptions that adhere to dermatological reporting standards. This integration of NLP enhances the utility of the captions as diagnostic aids.

To characterize in a single component the textual and visual representation, the separately extracted visual and textual features are merged. Subsequently, to create new hybrid features, we applied two "Dense" layers and thus reduce the size of the descriptor vector to 38.

In this work, we utilize the Long Short-Term Memory (LSTM) architecture, a widely recognized deep learning model introduced by Hochreiter and Schmidhuber in 1997, this architecture was specifically designed to capture long-range dependencies in sequential data. For a comprehensive understanding of the LSTM mechanism, including its gating architecture and training process, readers are referred to (Hochreiter and Schmidhuber (1997).

Evaluation Metrics

Various evaluation metrics have been introduced in the literature for assessing image captioning methods. Many of these are adapted from Natural Language Processing (NLP), including BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), and ROUGE (Recall-Oriented Understudy for Gisting Evaluation). In addition, other measures have been proposed specifically for the case of image captioning approaches such as; Semantic Propositional Image Caption Evaluation (SPICE), and Consensus-based Image Description Evaluation (CIDEr). These measures are not yet well-developed and do not yet allow properly evaluate the quality of the generated descriptions (Beddiar *et al.*, 2022).

Medical reports of skin lesions are mainly based on the visual aspect of the lesions and are established based on the ABCDE rules (Asymmetry, Border, Color, Diameter, and Evolving). Therefore, we adopted as an evaluation measure of the proposed approach the BLEU score, which allows us to analyze the frequency of co-occurrence of n-grams in the predicted and original description.

The BLUE score can be defined as follows:

$$BLEU_N = BP * e^{\sum_{n=1}^{N} w_n \log(p_n)} \quad (1)$$

where, n is the number of the N-grams considered, the $w_n$is the weight of the n-gram, and BP is the brevity penalty used to penalize short sentences, defined by:

$$BP = e^{\min\left(1-\frac{len(reference)}{len(prediction)},0\right)} \quad (2)$$

The BLUE score can be measured for different lengths of N-grams to consider, varying from 1 to 4-grams.

Table 1: Examples of features of some images from the PH2 dataset

Image	Diagnosis[1]	Asymmetry	Pigment Network	Dots/Globules	Streaks	Regression Areas	Blue-Whitish Veil	Colors[2]
IMD022	CN	Fully Symmetric	Typical	Absent	Absent	Absent	Absent	LB
IMD024	CN	Fully Symmetric	Typical	Absent	Absent	Absent	Absent	LB, DB
IMD002	CN	Symmetric in 1 axis	Atypical	Absent	Absent	Absent	Absent	LB, DB
IMD226	CN	Fully Asymmetric	Atypical	Atypical	Present	Absent	Present	LB, DB, BG, B
IMD434	AN	Fully Asymmetric	Atypical	Atypical	Present	Absent	Present	LB, DB, BG
IMD058	Me	Fully Symmetric	Atypical	Absent	Absent	Present	Present	W, R, DB, BG
IMD061	Me	Fully Asymmetric	Atypical	Absent	Absent	Present	Present	LB, BG

[1]*Me = Melanoma, CN = Common Nevus, AN = Atypical Nevus*
[2]*W= White, R = Red, LB = Light-Brown, DB = Dark-Brown, BG = Blue-Gray, B = Black*

Table 2: Captions created from the features presented in the Table 1

Image	Caption
IMD022	Fully symmetric lesion that contains one color with typical pigment network
IMD024	Fully symmetric lesion that contains two colors with typical pigment network
IMD002	Asymmetry in one axis lesion that contains two colors with atypical pigment network
IMD226	Asymmetry in one axis lesion that contains two colors with atypical features observed in both the pigment network and in the distribution of dots and globules
IMD434	Fully asymmetric lesion that contains three colors with atypical features observed in both the pigment network and in the distribution of dots and globules and presence of streaks and blue whitish veil
IMD058	Fully symmetric lesion that contains four colors with atypical pigment network and presence of regression areas and blue whitish veil
IMD061	Fully asymmetric lesion that contains two colors with atypical pigment network and presence of regression areas and blue whitish veil

Results and Discussion

In this section, we will present the simulation results of the proposed approach. For the evaluation of the implemented architecture, the dataset is partitioned into 80% for training and 20% for validation. The evaluation metrics used are BLEU1, BLEU2, BLEU3, and BLEU4.

Figure (3) shows the captions results using our proposed approach of three skin lesions images.

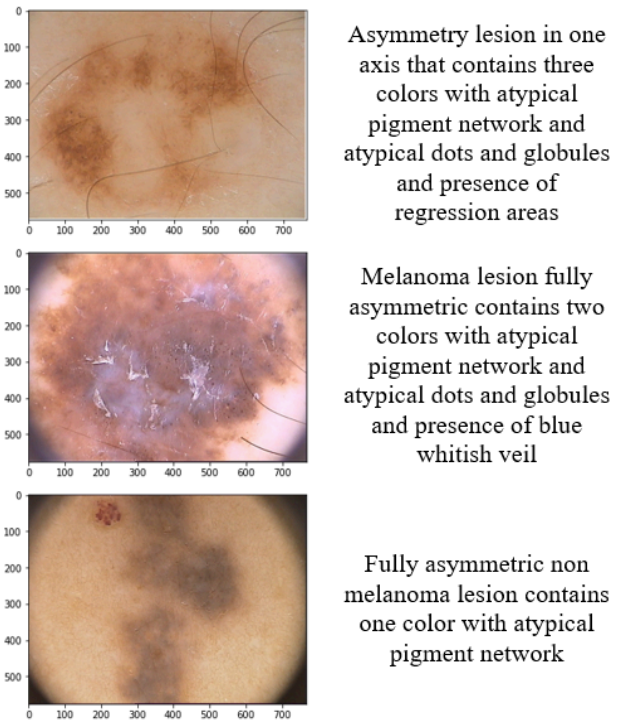


Fig. 3: An example of caption prediction results of 3 images of skin lesions by the proposed approach

It can be seen from the results presented in Figure (3) that the proposed approach was able to generate complete captions (reports) of lesion images. Certainly, these cations are not 100% correct, but they are very meaningful and can be used as a diagnostic aid.

To measure and evaluate the performance of the proposed captioning approach, Table (3) shows the BLEU-1, 2, 3, and 4 scores measured on the validation dataset. It is important to note that the values N=1, 2, 3, and 4 represent the length of the N-grams used (see Eqs. 1-2).

Table 3: Evaluation measures of captions generated by our proposed approach

	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Proposed approach	0.51	0.33	0.28	0.18

We achieved a BLEU-1 score of 0.50, indicating that the unigrams in the generated captions closely match those in the expert-validated ground truth. This score highlights the model's effectiveness in producing meaningful and clinically relevant textual descriptions of skin lesions. While this score may seem modest compared to benchmarks in natural image captioning, it is highly significant given the novelty of our dataset and the limited size of the training data. These results demonstrate the feasibility of using our proposed approach as a diagnostic aid, paving the way for future improvements through larger datasets and enhanced models.

The results also validate the integration of VGG16 for feature extraction and LSTM for sequential modeling in this domain, emphasizing the potential for further advancements by incorporating attention mechanisms or pre-trained transformers.

Conclusion

Captioning or textual description of images is a fast-growing research topic. The goal is to be able to describe the visual content of images by words or sentences. It is at the crossroads of Natural Language Processing (NLP) and Computer Vision. In the case of medical images, it is not yet very developed as in the case of natural images. This is due to the rarity of medical image datasets with captions. Most of the existing are chest X-ray image datasets. In this paper, we have proposed an approach to generate medical captions and reports for skin lesion diagnosis. We used the PH2 dataset, which is a well-known used dataset for skin lesion classification. The initial captions of all the images in the dataset were created manually using the features provided with the dataset. Thus, we proposed a captions generation model based on a fusion process. The visual features were extracted using the VGG16 Deep learning architecture, and the textual features were extracted using the LSTM architecture. Simulation results as well as the BLEU scores showed great interest in our proposed approach. As perspectives, it is necessary, firstly, to use a large base of skin lesions as the ISIC challenge dataset and secondly to use the attention mechanism for the extraction of visual features as well as the generation of captions.

Acknowledgment

The authors thank God Almighty for granting us the strength, health, courage, and patience to accomplish the work presented in this article. We also extend our sincere gratitude to the Faculty of Sciences, Sidi Mohamed Ben Abdellah University, Fez, Morocco, and to everyone who contributed directly or indirectly to this research.

Funding Information

The authors received no financial support for the research, authorship, or publication of this article.

Author's Contributions

Abdelouahed Sabri: Developpement, expermimentations testing, validation and writing the manuscript.

Chaimae Zouitni: Expermimentations validation and proofreading.

Hamza El Medhoune: Concept developpement, experimentations testing, validation and proof reading.

Abdellah Aarab: Concept developpement, experimentations validation and proof reading.

Ethics

This paper is original with unpublished material. The corresponding author confirms that this manuscript has not been published elsewhere and that no ethical issues are involved.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Abacha, A. B., Hasan, S. A., Datla, V. V., Liu, J., Demner-Fushman, D., & Müller, H. (2019). Vqa-Med: Overview of the Medical Visual Question Answering Task at Imageclef 2019. *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*, *2*.

Abdelouahed, S., Filali, Y., & Aarab, A. (2019). An Improved Segementation Approach for Skin Lesion Classification. *Statistics, Optimization & Information Computing*, *7*(2), 456-467. https://doi.org/10.19139/soic.v7i2.533

Beddiar, D. R., Oussalah, M., & Seppänen, T. (2021). Attention-Based CNN-GRU Model for Automatic Medical Images Captioning: Imageclef 2021. *Proceedings of the Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum*, 1160.

Beddiar, D.-R., Oussalah, M., & Seppänen, T. (2023). Automatic Captioning for Medical Imaging (MIC): A Rapid Review of Literature. *Artificial Intelligence Review*, *56*(5), 4019-4076. https://doi.org/10.1007/s10462-022-10270-w

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., & McDonald, C. J. (2016). Preparing a Collection of Radiology Examinations for Distribution and Retrieval. *Journal of the American Medical Informatics Association*, *23*(2), 304-310. https://doi.org/10.1093/jamia/ocv080

Eickhoff, C., Schwall, I., & Herrera A, G. S. (2017). Overview of Imageclefcaption 2017-Image Caption Prediction and Concept Detection for Biomedical Images. *Proceedings of the CLEF 2017 Working Notes*. CEUR workshop proceedings.

Filali, Y., Khoukhi, H. E., Sabri, M. A., & Aarab, A. (2022). Analysis and Classification of Skin Cancer Based on Deep Learning Approach. *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 1-6. https://doi.org/10.1109/iscv54655.2022.9806087

Filali, Y., Sabri, M. A., & Aarab, A. (2021). Efficient Skin Cancer Diagnosis Based on Deep Learning Approach Using Lesions Skeleton. *International Journal of Cloud Computing*, *10*(5/6), 565. https://doi.org/10.1504/ijcc.2021.120395

Harzig, P., Einfalt, M., & Lienhart, R. (2019). Automatic Disease Detection and Report Generation for Gastrointestinal Tract Examination. *Proceedings of the 27th ACM International Conference on Multimedia*, 2573-2577. https://doi.org/10.1145/3343031.3356066

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hossain, MD. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Computing Surveys*, *51*(6), 1-36. https://doi.org/10.1145/3295748

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 590-597. https://doi.org/10.1609/aaai.v33i01.3301590

Luo, G., Cheng, L., Jing, C., Zhao, C., & Song, G. (2022). A Thorough Review of Models, Evaluation Metrics and Datasets on Image Captioning. *IET Image Processing*, *16*(2), 311-332. https://doi.org/10.1049/ipr2.12367

Mendonca, T., Ferreira, P. M., Marques, J. S., Marcal, A. R. S., & Rozeira, J. (2013). PH2 - A Dermoscopic Image Database for Research and Benchmarking. *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka. https://doi.org/10.1109/embc.2013.6610779

Sabri, M. A., Filali, Y., El Khoukhi, H., & Aarab, A. (2020). Skin Cancer Diagnosis Using an Improved Ensemble Machine Learning model. *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 1-5. https://doi.org/10.1109/iscv49265.2020.9204324

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556-2565. https://doi.org/10.18653/v1/p18-1238

Singh, A., Krishna Raguru, J., Prasad, G., Chauhan, S., Tiwari, P. K., Zaguia, A., & Ullah, M. A. (2022). Medical Image Captioning Using Optimized Deep Learning Model. *Computational Intelligence and Neuroscience*, *2022*, 1-9. https://doi.org/10.1155/2022/9638438

Tian, J., Zhong, C., & Zhongchao, S. (2020). Towards Automatic Diagnosis From Multi-Modal Medical Data. *Interpretability Mach Intell Med Image Comput Multimodal Learn Decis Support*, *11797*, 67-74.

Yang, S., Niu, J., Wu, J., Wang, Y., Liu, X., & Li, Q. (2021). Automatic Ultrasound Image Report Generation with Adaptive Multimodal Attention Mechanism. *Neurocomputing*, *427*, 40-49. https://doi.org/10.1016/j.neucom.2020.09.084

Zeng, X., Wen, L., Xu, Y., & Ji, C. (2020). Generating Diagnostic Report for Medical Image by High-Middle-Level Visual Information Incorporation on Double Deep Learning Models. *Computer Methods and Programs in Biomedicine*, *197*, 105700. https://doi.org/10.1016/j.cmpb.2020.105700