J. Innovative Eng. Nat. Sci., vol. 5, no. 2, pp. 477-489, 2025.
http://doi.org/10.61112/jiens.1529215

Original Paper

Journal of Innovative Engineering and Natural Science
(Yenilikçi Mühendislik ve Doğa Bilimleri Dergisi)
https://dergipark.org.tr/en/pub/jiens

Generating informative chest X-ray captions with LSTM architecture

Omer Faruk Guzel[a], Harun Tanriverdi[a], and Gokhan Bakal[b,*]

[a]*Electical and Computer Engineering Division, Abdullah Gul University, Barbaros, Kayseri, 38080, Turkiye.*
[b]*Department of Computer Engineering, Abdullah Gul University, Barbaros, Kayseri, 38080, Turkiye.*

ARTICLE INFO

Article history:
Received 7 August 2024
Received in revised form 24 December 2024
Accepted 13 March 2025

Available online

Keywords:
Biomedikal imaging
Text mining
Deep learning
Medical informatics

ABSTRACT

Biomedical imaging is the most effective medical screening procedure for medical specialists. Specifically, X-ray images are intensively used as a reference point for medical diagnostic purposes. However, understanding the underlying matters from the X-ray images requires significant radiological knowledge. In this study, a deep learning model, which employs the DenseNet121 neural network architecture as an encoder module and textual data (captions) items as word embedding layers, is trained to predict the corresponding title/caption information of the given X-ray images. The generated model is a typical sequence-to-sequence model used particularly for neural machine translation tasks. In the experiments, the Open-i database curated by Indiana University is used for the training and testing phases. The dataset consists of 7,470 X-ray images and 3,955 patient reports stored in XML format, composed by a domain expert. The textual reports contain four specific captions, including impressions, findings, comparisons, and indications. During the model development, the textual data under the impression captions was exploited in the training and testing steps. To measure the model's performance, the Bilingual Evaluation Understudy Score (BLUE) was calculated and utilized as the primary performance evaluation metric. Based on the BLUE scores, the best performance score was achieved when four words (four grams) were predicted with the BLUE score of 0.38368 compared to other n-gram sets (where n: 1, 2, and 3). This research effort demonstrates the power of sequence-to-sequence models on the text generation task in medical image datasets for automatic diagnosing purposes.

I. INTRODUCTION

Since the considerable advances in the Artificial Intelligence (AI) field, researchers have intensively been conducting numerous studies utilizing AI methodologies. In this regard, Machine Learning (ML) has considered one widely used subfield of the AI area, whereby the ML algorithms construct intelligent models using known data examples, also called training datasets. One of the prominent ML experiments is a classification task of data examples by the trained models regardless of the application domain [1, 2]. Over the past two decades, the biomedical informatics field, which principally exploits AI techniques, including classification and prediction methods, dramatically emerged and became an indispensable research area [3]. On this basis, visual data elements (e.g., X-Ray images) and textual data representations (e.g., diagnosis reports) are reasonably relevant elements in constructing successful ML models. Considering these recent directions, we can summarize the leading biomedical research efforts as follows:

- Automatic medical diagnosis [4-6],
- Potential disease predictions [7-9],
- In silico drug discovery and drug repositioning [10-13],
- Precision medicine applications [14-16].

*Corresponding author. Tel.: +90-544-563-8710; e-mail: gokhan.bakal@agu.edu.tr

As a subfield of medical diagnosis, biomedical imaging is one of the most investigated research efforts for disease diagnosis in healthcare [17, 18, 19]. Building a successful diagnosis model depends on various factors affecting the model's accuracy, such as the quality of medical images, filtering operations, and tuning brightness and contrast features. X-Ray is a biomedical imaging method that is broadly operated in all healthcare institutions as well as medical emergency units [20]. Nowadays, having an X-Ray taken is widespread and more accessible for people. Usually, patients' conditions are evaluated by chest X-Ray images unless their health problems are not a fracture or a crack in a body part. The reason is that it helps the healthcare specialist (called radiologists) see the internal organs and nearby structures. Evaluating biomedical images requires domain experts to write reports, interpret the imaging, and diagnose diseases. To gain this expertise, medical doctors must train with numerous biomedical images and handle many actual cases. Even if medical doctors obtain this expertise, it is impractical to appoint radiologists with the same level of knowledge in all healthcare institutions around the world.

Recently, any potential data elements have become valuable; thus, biomedical imaging data and methods have also earned even more importance correspondingly [21]. Therefore, relevant research topics, such as classification, biomedical image segmentation, and abnormality detection, have been successfully examined lately. Also, these days, the captioning operation of image processing has been widely performed by researchers. Basically, this process can be defined as producing subtitles of the image by extracting discriminative contextual features from the image content. In this sense, Pavlopoulos et al. [22] demonstrated that this captioning process can also be applied to biomedical images successfully.

In this research effort, the primary aim is to construct consistent captions for the given chest X-ray images. In the dataset having radiology reports and chest X-ray images, comparison, indication, findings, and impression data have been acquired using RegEx pattern matching approaches. Next, the chest image features have been extracted using the DenseNet121 architecture [23]. Using the obtained visual and textual context information, a combined ML model has been built and trained to test the model's performance. The rest of the paper is organized as follows. Section 2 elaborates more on the recent related works. Section 3 describes the details of the dataset used in the proposed study, while Section 4.1 presents the designed deep learning model in more detail. Section 5 presents the results and evaluates them in a brief discussion. Finally, the overall conclusion is given in Section 6, and potential future directions are mentioned in Section 7.

II. BACKGROUND & RELATED EFFORTS

Generally, there are various in-use imaging modalities, such as X-ray, Computerized Tomography (CT) scan, Positron Emission Tomography (PET) scan, and Magnetic Resonance Imaging, for distinct purposes [24]. To this end, beyond medical use, X-Ray imaging is also used for objectives, such as security screening. Biomedical imaging plays a critical role in medical diagnostic processes due to supplying beneficial insights to medical specialists. One of the practical applications in the medical field is to classify the target entities utilizing X-Ray images [25, 26]. During the recent COVID-19 pandemic period, the value of biomedical imaging techniques to diagnose COVID-19 contamination by analyzing the chest X-Rays is well-understood globally [27-29]. Plus, non-radiological image processing-based machine learning models are also intensively examined by many researchers [30, 31]. Another functional research direction in biomedical imaging is to predict the evaluation reports composed

by medical domain experts. Ayesha et al. (2021) [32] showed that medical image analysis and interpretation by combining image processing and natural language processing techniques can be successfully performed. In addition to captioning tasks, computational studies on medical imaging report generation have also been widely investigated by bioinformatics researchers [33]. Despite significant progress in automated medical image analysis, existing approaches to chest X-ray captioning often struggle to generate comprehensive and clinically relevant reports. These models typically focus on identifying individual findings rather than providing a holistic image interpretation. In order to address this limitation, our study proposes a novel sequence-to-sequence deep learning model that integrates DenseNet121 for image feature extraction, GloVe embeddings for textual representation, and an LSTM network for capturing long-range dependencies in the radiology report. By combining these powerful techniques, we aim to develop a model that can automatically generate accurate and comprehensive chest X-ray reports, potentially reducing the workload of radiologists and improving diagnostic efficiency.

The principal contribution of the paper is that we proved the combination of both image and text contextual information works smoothly. In this regard, we built a sequence-to-sequence deep learning model with GloVe word pre-embeddings from the textual reports.

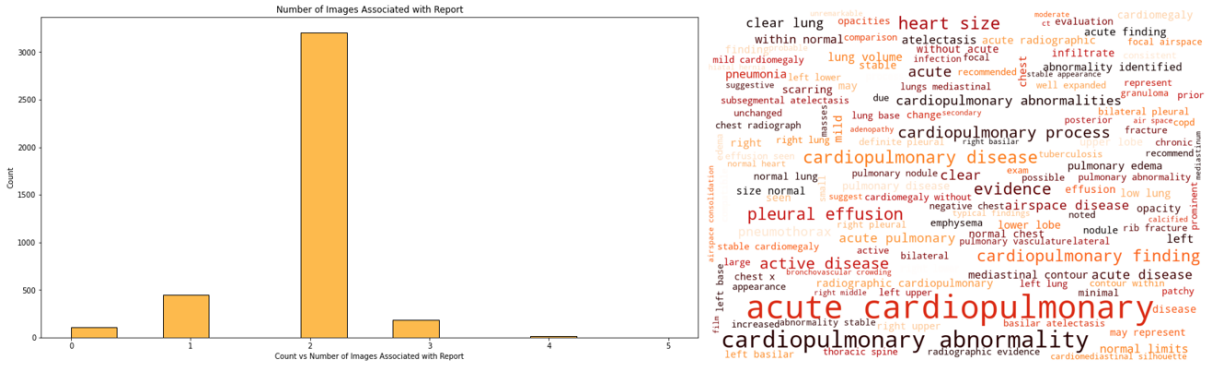
III. DATASET DETAILS

To build the model, we used the Indiana University Chest X-Ray dataset, which is publicly available and includes chest X-Ray images and reports in the XML format. Each image has been paired with four captions providing clear medical descriptions in the XML reports: Impressions, Findings, Comparison, and Indication as shown in Figure 1.

Figure 1. An X-Ray image and report content example [34]

In the dataset, there are 7,471 X-ray images in the ".png" file format (containing lateral and frontal views for each patient) and 3,955 textual patient reports formed in the ".XML" format. Naturally, some patients may have multiple, max of 5 X-ray images, and this frequency distribution is presented as a graph in Figure 2a. To understand

and visualize the textual content, we generated the word cloud based on the image reports, as demonstrated in Figure 2b.



(a) (b)

Figure 2. Distributions of both X-Ray images and the word cloud representation, (**a**) The frequency distribution of X-Ray images, (**b**) The word cloud of textual reports

IV. METHODOLOGY

This study aims to develop a deep-learning model capable of generating accurate and informative captions for chest X-ray images. To achieve this, we employ sequence-to-sequence architecture inspired by its success in neural machine translation. This architecture effectively captures the relationship between visual features extracted from the images and the corresponding textual descriptions. Technically, our model consists of three main components: an image encoder, a text encoder, and a decoder. The image encoder utilizes a pre-trained DenseNet121 network to extract relevant features from the X-ray images. The text encoder employs GloVe word embeddings to represent the textual information from the associated radiology reports. These encoded representations are fed into a Long Short-Term Memory (LSTM) network, which learns the long-range dependencies between the visual and textual information. Finally, a decoder, implemented using a greedy search algorithm, generates the output captions based on the learned representations.

4.1 Data Pre-Processing

The raw data obtained from the Indiana University Chest X-Ray dataset required careful pre-processing to ensure optimal model training. This involved several steps applied to both the textual and image data:

Textual Data:

- Extraction: Radiology reports, stored in XML format, were parsed to extract the relevant textual captions (impressions, findings, comparisons, and indications) using regular expression (RegEx) pattern matching.
- Cleaning: Extracted text was cleaned by removing non-alphanumeric characters, punctuation marks, and extra whitespaces. This step aimed to standardize the text and reduce noise.
- Contraction Expansion: Contracted word forms (e.g., "won't", "don't") were expanded to their full forms (e.g., "will not", "do not") to ensure consistency and improve tokenization.

- Deduplication: To mitigate overfitting, duplicate captions associated with more than two images were removed. This step prioritized unique and diverse data samples for training.

Image Data:

- Resizing: To accommodate the input requirements of the pre-trained DenseNet121 model, all images were resized to a uniform dimension of 512x512 pixels. This involved adjusting the height of the images while maintaining their aspect ratio.

- Data Equalization: Given the uneven distribution of images per patient (Figure 2a), the dataset was balanced by replicating images for those with a single X-ray and selecting the first two images for those with more than two. This ensured equal representation and prevented bias during training.

This comprehensive pre-processing ensured that both textual and image data were properly formatted, cleaned, and standardized for effective model training. It also addressed potential overfitting issues and ensured balanced data representation.

4.2 Proposed Model Details

The proposed deep learning model for chest X-ray captioning is designed as a sequence-to-sequence architecture, leveraging the power of both image and text processing techniques. This architecture, commonly used in neural machine translation, effectively learns the mapping between visual features and corresponding textual descriptions. The model comprises five core components:

4.2.1 Image Encoder

A pre-trained DenseNet121 architecture serves as the image encoder. DenseNet121 is a convolutional neural network (CNN) known for its dense connections between layers, allowing for efficient feature extraction and representation learning. This encoder takes the pre-processed X-ray images as input and outputs a 1024-dimensional feature vector, capturing essential visual information.

4.2.2 Text Encoder

The text encoder processes the corresponding textual captions from the radiology reports. We utilize pre-trained GloVe word embeddings to represent each word in captions as a dense vector. GloVe captures semantic relationships between words based on their co-occurrence statistics in a large corpus. This allows the model to understand the meaning and context of the textual descriptions. The embedded captions are then fed into an LSTM network.

4.2.3 LSTM Network

LSTM networks are a type of recurrent neural network (RNN) particularly well-suited for sequential data like text. They effectively capture long-range dependencies within the input sequence. A standard LSTM cell structure is displayed in Figure 3.

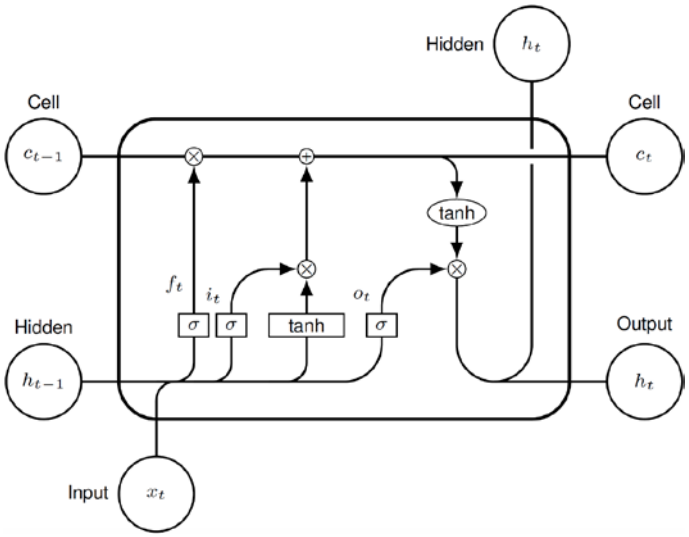


Figure 3. An example of LSTM cell structure [35]

Here, the cell state unit behaves as a memory area in the network and holds characteristic information about the data to be transmitted to other cells. In this manner, short-term information can be stored and transported throughout the network. The Forget Gate erases the unnecessary information from memory depending on the sigmoid function (activation function generating value within [0, 1]) result calculated with the information coming from the previous cell and the information in the current cell state. The Input Gate updates the data in the cell state unit according to the result of the sigmoid operation with the data elements from the previous and the current cell. Finally, the output gate determines which information will be transmitted to other cells. First, the sigmoid operation is executed with the previous and the current cell data. Then, the sigmoid function result and tanh function (activation function generating value within [-1, 1]) of the data in the cell state unit are multiplicated. According to this multiplication result, the crucial information is identified and forwarded as previous cell information. Thus, LSTM architecture shows much higher classification and prediction performance with this advanced gating structure simulating the memory functionality compared to a standard RNN model. For a more simplified definition, imagine reading a long sentence; we need to remember the earlier words to understand the meaning of the whole sentence. LSTMs work similarly, using a 'memory cell' to store important information from previous words in the caption. This ability helps the model capture the context and relationships between words, which is essential for generating accurate and meaningful chest X-ray descriptions.

In our model, the LSTM network takes the GloVe embeddings as input and learns the relationships between the words in the captions. This enables the model to understand the grammatical structure and contextual information of the textual descriptions.

4.2.4 Decoder

The decoder is responsible for generating the output caption based on the learned representations from the image and text encoders. We employ a greedy search algorithm as the decoder. This algorithm iteratively selects the word with the highest probability at each time step, conditioned on the previously generated words and the encoded image and text features. The process continues until a complete caption is generated.

4.2.5 Model Integration

The output from the DenseNet121 image encoder is concatenated with the output from the LSTM text encoder. This combined representation is then passed through a dense layer to further integrate visual and textual information. The final output of this dense layer is then fed to the decoder for caption generation. This model architecture effectively combines the strengths of CNNs for image feature extraction, GloVe embeddings for textual representation, and LSTM networks for sequence learning to generate informative captions for chest X-ray images.

4.3 Model Architecture

The architecture of our deep learning model is designed to effectively capture and integrate visual and textual information to generate accurate chest X-ray captions. As visualized in Figure 4, the model begins with two distinct input layers: one for the pre-processed X-ray images and another for the corresponding textual captions. The image input is fed into a pre-trained DenseNet121 architecture, serving as our image encoder. DenseNet121 extracts a rich set of visual features from the input images due to its dense connectivity pattern. Simultaneously, the textual captions are processed by the text encoder. This operation involves embedding each word in the captions using pre-trained GloVe word vectors, which capture semantic relationships between words. These embeddings are then sequentially fed into an LSTM network to learn the long-range dependencies and context within the caption information. The outputs from the DenseNet121 image encoder and the LSTM text encoder are concatenated by merging the visual and textual representations. This combined representation is further processed by a dense layer to learn higher-level features that integrate both modalities. Finally, a decoder, implemented using a greedy search algorithm, generates the output caption word by word. This algorithm iteratively selects the word with the highest probability, conditioned on the previously generated words and the encoded image and text features.

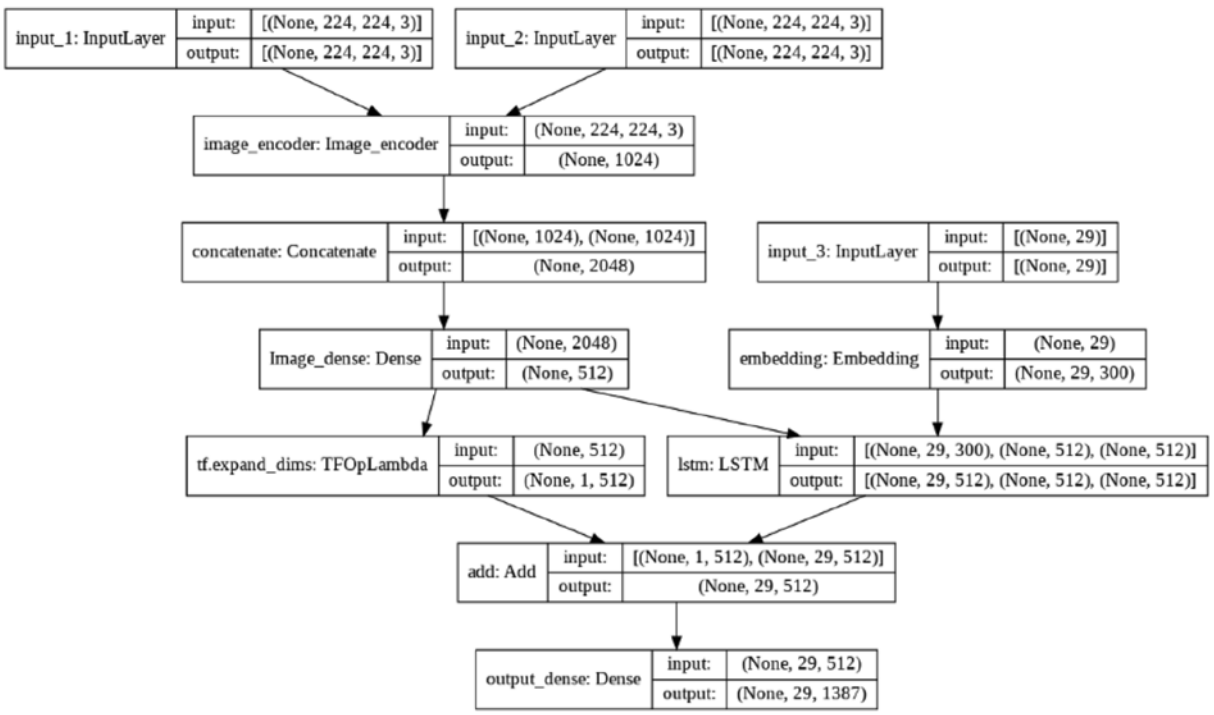


Figure 4. Overall neural model architecture

4.3.1 Dense Neural Network

In our model architecture, we incorporate a Dense Neural Network (DNN) layer to further process the combined visual and textual features extracted by the encoders. A DNN, also known as a fully connected layer, is characterized by its dense connections, where each neuron in the layer receives input from every neuron in the preceding layer. This dense connectivity allows the network to learn complex non-linear relationships between the features. In our study, the DNN takes the concatenated output from the image encoder (DenseNet121) and the text encoder (LSTM) as input. By applying a series of matrix-vector multiplications and activation functions, the DNN effectively integrates and transforms the visual and textual information into a higher-level representation. This refined representation is then passed to the decoder for caption generation. Incorporating the DNN layer enhances the model's capacity to learn intricate interactions between the image and text modalities, leading to more accurate and contextually relevant captions.

4.3.2 Global Vectors for Word Representation (GloVe)

To accurately represent the textual information in our model, we employ Global Vectors for Word Representation (GloVe), an open-source project developed at Stanford University [36]. GloVe is a distributed word representation technique that utilizes an unsupervised learning approach to learn meaningful vector representations for words. This approach maps words onto a vector space where the distance between words reflects their semantic similarity [37]. GloVe achieves this by leveraging global word-word co-occurrence statistics derived from a large text corpus. By analyzing how often words appear together in different contexts, the model learns to capture semantic relationships. Essentially, GloVe combines the strengths of global matrix factorization and local context window methods, operating as a log-bilinear regression model to learn word representations in an unsupervised manner [38]. While effective in capturing semantic relationships, including synonyms and related concepts, GloVe has limitations in distinguishing homographs, words with identical spelling but different meanings. This stems from the unsupervised nature of the learning process, which assigns a single vector to entities with the same morphological structure. Despite this limitation, GloVe provides valuable semantic representations for the textual captions in our chest X-ray analysis.

4.3.3 Greedy Search Algorithm

In our caption generation process, we use a Greedy Search Algorithm to create the final caption from our model's output. The algorithm works by choosing the word with the highest probability at each step of the process. Starting with an initial token, the algorithm predicts the next word based on the model's output and the previous word and continues this step-by-step process until an "end" token is reached or a maximum length is reached. While the algorithm may not always find the best possible solution, it strikes a good balance between efficiency and performance, making it suitable for our caption generation task. After generating the word sequence, we use the BLEU score as our primary metric to assess the quality of the captions.

4.4 Model Training Process

To facilitate the training process, we utilized an Ubuntu server hosted on a virtual machine within Amazon Web Services (AWS). This cloud-based infrastructure provided the necessary computational resources, including 8 vCPUs and 32 GB of RAM, to efficiently train our deep learning model. We employed a Jupyter Notebook environment for code development and execution. Within this environment, we installed the necessary Python packages, such as TensorFlow, Keras, and NLTK, to support model development and training.

The pre-processed dataset, as described in Section 4.1, was uploaded to the AWS server and subsequently split into training and validation sets with an 80:20 ratio. This partitioning allowed us to assess the model's performance on unseen data during training, helping to prevent overfitting and ensure generalizability. The overall model architecture flow is illustrated in Figure 5.

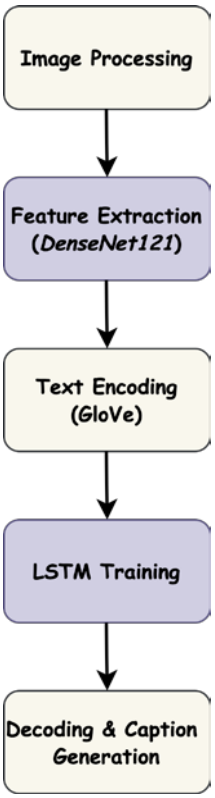


Figure 5. Proposed model architecture flow

We trained the model for 10 epochs using the Adam optimizer, which is known for its efficiency and effectiveness in deep learning applications. The following hyperparameters were carefully selected based on preliminary experiments and best practices:

- embedding_dim: 300
- dense_dim: 512
- lstm_units: 512
- dropout_rate: 0.2

A dropout rate of 0.2 was applied to regularize the model and mitigate the risk of overfitting. Each training epoch took approximately 4 minutes to complete, resulting in a total training time of 40 minutes. Upon completion of the training process, the model achieved an accuracy of 70% on the validation set, demonstrating its ability to learn and generalize from the provided data.

V. RESULTS AND DISCUSSIONS

To evaluate the performance of our chest X-ray captioning model, we employed the Bilingual Evaluation Understudy (BLEU) score as our primary metric. BLEU is a widely used metric for evaluating the quality of machine-generated text, particularly in machine translation tasks. It measures the similarity between the generated text and a set of reference translations by comparing the presence and order of n-grams (sequences of n words) in both the generated and reference texts. In our effort, we used the Greedy Decoder Algorithm, as described in Section 4.2.5, to generate the captions from the model's output. This algorithm selects the word with the highest probability at each time step, which results in a complete caption. We then calculated the BLEU scores for different n-gram orders (n=1, 2, 3, and 4) to assess the model's performance at different levels of granularity. The results of our evaluation are presented in Table 1.

Table 1. BLEU scores of n-gram models

BLEU-1	BLEU-2	BLEU-3	BLEU-4
0.306819	0.302596	0.339031	0.383689

As shown in Table 1, the model achieved the highest BLEU score of 0.38368 for the 4-grams configuration, which indicates its ability to generate captions that are not only fluent but also capture longer-range dependencies in the radiology report, like human-written reports. This outcome suggests that the model could be used to assist radiologists by providing initial drafts of reports, potentially reducing their workload and improving reporting efficiency. However, the lower BLUE scores for 1-gram and 2-gram indicate that the model may struggle with capturing specific medical terms or local word choices. These conclusions highlight both the potential of our approach for automated chest X-ray reporting and the need for further refinement to improve the model's sensitivity to fine-grained details. Furthermore, it is worth noting that the BLEU scores gradually decrease as the n-gram order decreases, which suggests that the model' s performance is better at capturing local word sequences than longer-range dependencies.

VI. CONCLUSIONS

This study successfully demonstrated the feasibility of automatically generating descriptive captions for chest X-ray images using a deep learning model. Our evaluation, using the BLEU score as the primary metric, demonstrated the effectiveness of our model in generating relevant captions for chest X-ray images. Our sequence-to-sequence architecture effectively combined image and text information to produce informative captions, achieving the highest BLEU score with 4-grams. This research contributes to the growing field of medical image analysis and has the potential to assist radiologists in their diagnostic workflow. By automating the captioning process, our model can improve efficiency, reduce workload, and potentially enhance the accuracy of interpretations. Although

the BLEU score is helpful in providing a quantitative measure of how well the model performs, there are several limitations. The algorithm used for decoding is a greedy search heuristic; while efficient, it will not necessarily find the best word sequence, which may affect BLEU again. Beyond this, BLEU mainly stresses n-gram overlap and fails to capture most semantic nuances and clinical accuracy in the generated captions. Trying more advanced decoding methods, such as beam search, will be interesting sub-studies in future work. Moreover, metrics that can better reflect the clinical relevance of the generated reports will be considered.

VII. FUTURE DIRECTIONS

In the context of future directions, we mentioned potential research direction below:

- Incorporating Clinical Knowledge: Integrating clinical knowledge, such as disease ontologies or patient history, into the caption generation process could improve the accuracy and relevance of the generated reports.
- Multi-Modal Analysis: Combining chest X-ray images with other data modalities, such as electronic health records or clinical notes, could enhance the model's ability to provide a comprehensive and personalized interpretation.
- Explainable AI: Developing methods for explaining the model's predictions, for example, by highlighting the image regions that influenced the generated caption, would improve trust and transparency, facilitating clinical adoption.
- Longitudinal Analysis: Analyzing temporal changes in chest X-ray images and generating captions that reflect disease progression or treatment response could provide valuable insights for patient management.

ACKNOWLEDGMENT

We gratefully acknowledge the insightful feedback and valuable suggestions provided by the paper reviewers.

REFERENCES

1. Lu D, Weng Q (2007) A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing 28(5):823-870.
2. Bakal G, Talari P, Kakani EV, Kavuluru R (2018) Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. Journal of Biomedical Informatics 82:189-199.
3. Bernstam EV, Smith JW, Johnson TR (2010) What is biomedical informatics? Journal of Biomedical Informatics 43(1):104-110.
4. Kampouraki A, Vassis D, Belsis P, Skourlas C (2013) e-Doctor: A web based support vector machine for automatic medical diagnosis. Procedia - Social and Behavioral Sciences 73:467-474.
5. Ma F, Sun T, Liu L, Jing H (2020) Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. Future Generation Computer Systems 111:17-26.
6. Shanthi T, Sabeenian RS, Anand R (2020) Automatic diagnosis of skin diseases using convolution neural network. Microprocessors and Microsystems 76, 103074.
7. Islam MM, Haque MR, Iqbal H, Hasan MM, Hasan M, Kabir MN (2020) Breast cancer prediction: A comparative study using machine learning techniques. SN Computer Science 1:1-14.
8. Xie S, Yu Z, Lv Z (2021) Multi-disease prediction based on deep learning: A survey. Computer Modeling in Engineering and Sciences 128(2): 489-522.

9. Thieme AH, Zheng Y, Machiraju G, Sadee C, Mittermaier M, Gertler M, et al (2023) A deep-learning algorithm to classify skin lesions from mpox virus infection. Nature Medicine 29(3):738-747.
10. Bakal G, Kilicoglu H, Kavuluru R (2019) Non-negative matrix factorization for drug repositioning: Experiments with the repoDB dataset. In AMIA Annual Symposium Proceedings. American Medical Informatics Association, pp 238.
11. Shaker B, Ahmad S, Lee J, Jung C, Na D (2021) In silico methods and tools for drug discovery. Computers in Biology and Medicine 137, 104851.
12. Akkaya A, Bakal G (2023) A computational drug repositioning effort using patients' reviews dataset. In 2023 International Conference on Smart Applications, Communications and Networking (SmartNets). IEEE, pp 1–6.
13. Park JH, Cho YR (2024) Computational drug repositioning with attention walking. Scientific Reports 14(1):10072.
14. Yang SR, Schultheis AM, Yu H, Mandelker D, Ladanyi M, Büttner R (2022) Precision medicine in non-small cell lung cancer: Current applications and future directions. Seminars in Cancer Biology 84:184–198.
15. MacEachern SJ, Forkert ND (2021) Machine learning for precision medicine. Genome 64(4):416–425.
16. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al (2021) Precision medicine, AI, and the future of personalized health care. Clinical and Translational Science 14(1):86–93.
17. Chan HP, Hadjiiski LM, Samala RK (2020) Computer-aided diagnosis in the era of deep learning. Medical Physics 47(5):e218–e227.
18. Guler Ayyildiz B, Karakis R, Terzioglu B, Ozdemir D (2024) Comparison of deep learning methods for the radiographic detection of patients with different periodontitis stages. Dentomaxillofacial Radiology 53(1):32–42.
19. Şahin E, Özdemir D, Temurtaş H (2024) Multi-objective optimization of ViT architecture for efficient brain tumor classification. Biomedical Signal Processing and Control 91:105938.
20. Özdemir D, Arslan NN (2022) Analysis of deep transfer learning methods for early diagnosis of the Covid-19 disease with Chest X-ray images. Düzce Üniversitesi Bilim ve Teknoloji Dergisi 10(2):628–640.
21. Arslan NN, Ozdemir D (2024) Analysis of CNN models in classifying Alzheimer's stages: Comparison and explainability examination of the proposed separable convolution-based neural network and transfer learning models. Signal, Image and Video Processing:1–15.
22. Pavlopoulos J, Kougia V, Androutsopoulos I (2019) A survey on biomedical image captioning. In Proceedings of the Second Workshop on Shortcomings in Vision and Language. pp 26–36.
23. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp 4700–4708.
24. Kasban H, El-Bendary MAM, Salama DH (2015) A comparative study of medical imaging techniques. International Journal of Information Science and Intelligent System 4(2):37–58.
25. Sharma A, Raju D, Ranjan S (2017) Detection of pneumonia clouds in chest X-ray using image processing approach. In 2017 Nirma University International Conference on Engineering (NUiCONE). IEEE, pp 1–4.
26. Matsui T, Kamata T, Koseki S, Koyama K (2022) Development of automatic detection model for stem-end rots of 'Hass' avocado fruit using X-ray imaging and image processing. Postharvest Biology and Technology 192:111996.
27. Civit-Masot J, Luna-Perejón F, Domínguez Morales M, Civit A (2020) Deep learning system for COVID-19 diagnosis aid using X-ray pulmonary images. Applied Sciences 10(13):4640.
28. Tabik S, Gómez-Ríos A, Martín-Rodríguez JL, Sevillano-García I, Rey-Area M, Charte D, et al (2020) COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images. IEEE Journal of Biomedical and Health Informatics 24(12):3595–3605.
29. Jain R, Gupta M, Taneja S, Hemanth DJ (2021) Deep learning based detection and analysis of COVID-19 on chest X-ray images. Applied Intelligence 51:1690–1700.
30. Mishra R, Daescu O (2017) Deep learning for skin lesion segmentation. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp 1189–1194.
31. Harangi B (2018) Skin lesion classification with ensembles of deep convolutional neural networks. Journal of Biomedical Informatics 86:25–32.
32. Ayesha H, Iqbal S, Tariq M, Abrar M, Sanaullah M, Abbas I, et al (2021) Automatic medical image interpretation: State of the art and future directions. Pattern Recognition 114:107856.
33. Yin C, Qian B, Wei J, Li X, Zhang X, Li Y, Zheng Q (2019) Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In 2019 IEEE International Conference on Data Mining (ICDM). IEEE, pp 728–737.
34. National Institutes of Health. Open-i: Biomedical image search engine. Chest X-ray Collection. Retrieved 08.05.2025 from https://openi.nlm.nih.gov/gridquery?sub=x&it=xg&coll=cxr&m=1.
35. Erkantarci B, Bakal G (2024) An empirical study of sentiment analysis utilizing machine learning and deep learning algorithms. Journal of Computational Social Science 7(1):241–257.

36. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp 1532–1543.
37. Abad A, Ortega A, Teixeira A, Mateo CG, Hinarejos CDM, Perdigão F, et al (2016) Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH 2016, Lisbon, Portugal, November 23–25, 2016, Proceedings. Springer, Vol. 10077.
38. Kalajdziski S, Ackovska N (2018) ICT Innovations 2018. Engineering and Life Sciences: 10th International Conference, ICT Innovations 2018, Ohrid, Macedonia, September 17–19, 2018, Proceedings. Springer, Vol. 940.