

Medical Image Captioning in Kurdish Language Using AI

Sangin Jamal Hamza

Department of Software, College of Engineering, Salahaddin University - Erbil,
Kurdistan Region, Iraq

Email: sangin.hamza@su.edu.krd

Marden Abdullah Anwer

Department of Software, College of Engineering, Salahaddin University - Erbil,
Kurdistan Region, Iraq

Email: mardin.anwer@su.edu.krd

ARTICLE INFO

Article History:

Received: 9/7/2025

Accepted: 10/8/2025

Published: Autumn 2026

Keywords:

*Deep Learning; LSTM;
Bahdanau Attention;
Disease-aware Model;
Natural Language
Generation; Low-
resource Languages*

Doi:

10.25212/lfu.qzj.11.4.51

ABSTRACT

The diagnosis of thoracic disorders such as cardiomegaly, effusion, and pneumothorax depends heavily on medical imaging, especially chest X-rays. However, skilled radiologists are needed to interpret these images, and their shortage in many areas causes delays in diagnosis. Recent developments in deep learning have shown promise in automating image interpretation through caption generation; however, the majority of current systems lack disease-specific contextual understanding and are trained on English-language datasets, which limits their clinical utility and multilingual applicability. The disease-aware medical image captioning model developed for the Kurdish (Sorani) language is presented in this research. It produces textual descriptions for chest X-ray pictures that are both semantically rich and clinically correct. The suggested model uses a Bahdanau attention-equipped LSTM decoder to guide caption generation and incorporates disease-specific class embeddings to extract spatial features using CheXNet (DenseNet-121), which was pretrained on ChestX-ray14. To train and assess the system, a new Kurdish dataset was created, which included 7,775 chest X-ray pictures from nine different diagnostic categories together with carefully chosen Kurdish descriptions. A BLEU-4 score is 0.4015, a ROUGE-L score is 0.8722, and a BERTScore F1 is 0.9151. The

model beat the current benchmarks when performance was assessed using BLEU, ROUGE-L, BERTScore, and CHRF++. These findings show how well the system produces linguistically coherent, disease-relevant captions in a language environment that has limited resources. This work shows a high-quality Kurdish dataset to help future research in multilingual clinical AI applications, in addition to developing the state of medical image captioning for presented languages that are rare.

1. Introduction

The diagnosis and management of thoracic disorders like atelectasis, effusion, cardiomegaly, and pneumothorax, and medical imaging are essential. Because of their accessibility and affordability, chest X-rays are the most often utilized imaging modality (Rajpurkar et al., 2017). However, it takes specific knowledge to interpret chest X-rays, and there is a lack of qualified radiologists in many areas. When diagnosis and treatment are delayed from the result of this problem, that may affect patient outcomes. Deep learning-based automatic image captioning systems are a means of solving this issue by creating textual captions of medical images, helping the doctors in their decisions, and developing accessibility in settings with limited resources. The study used recurrent neural networks (RNNs) or transformer models for generating sequences and convolutional neural networks (CNNs) for understanding images; current developments in medical image captioning have illustrated and characterized radiological findings automatically (Liu et al., 2021). These models are useful, but their usefulness is limited in medical and multilingual situations because they were trained on English language datasets and are used for general image captioning purposes.

Moreover, a lot of current systems provide captions without taking into account the underlying disease class, which leads to imprecise or excessively generic descriptions that might not be clinically relevant (Boag et al., 2019). Localized

and disease-aware captioning systems (Li et al., 2019), able to produce medically accurate descriptions in underrepresented languages like Kurdish (Sorani), are becoming more and more necessary. Specificity and precision are crucial in the medical field. The dependability of such systems may be jeopardized if the caption generation is not conditional on the disease type. For example, a caption that appropriately represents a pulmonary effusion differs greatly from one that describes a collapsed lung. Here, we present a disease-aware image captioning system for chest X-rays that produces Kurdish medical descriptions based on the type of sickness that has been diagnosed. As the image encoder, the design extracts rich feature maps using CheXNet, a DenseNet121-based convolutional model pretrained on extensive radiography data (ChestX-ray14). To direct the creation of captions, these characteristics are combined with a disease-specific one-hot vector and a worldwide attention mechanism. An LSTM network that learns to generate clinically meaningful Kurdish descriptions that match the image and diagnosis is used to construct the decoder. To assess the model's linguistic and clinical correctness, it uses measures such as BLEU, ROUGE-L, METEOR, CHRF++, and BERTScore in addition to disease-specific keyword matching. The following are principal contributions:

- (1) A comprehensive pipeline for Kurdish text generation and disease-aware attention captioning of medical images.
- (2) A local dataset includes a collection of chest X-ray pictures with Kurdish subtitles and labels for nine different disease categories.
- (3) Enhances caption quality using conventional greedy or beam search generation; we demonstrate that the inclusion of illness conditioning enhances caption quality.

The remainder of the document is organized as follows: Section 2: The literature on deep learning techniques and medical image captioning is reviewed. In Section 3 the dataset and preprocessing techniques are shown. In Section 4, the design of our suggested model is described in detail. In section 5 the training and evaluation processes are described. Results and constraints

are discussed in Section 6. The paper is finally concluded, and new directions are suggested in Section 7.

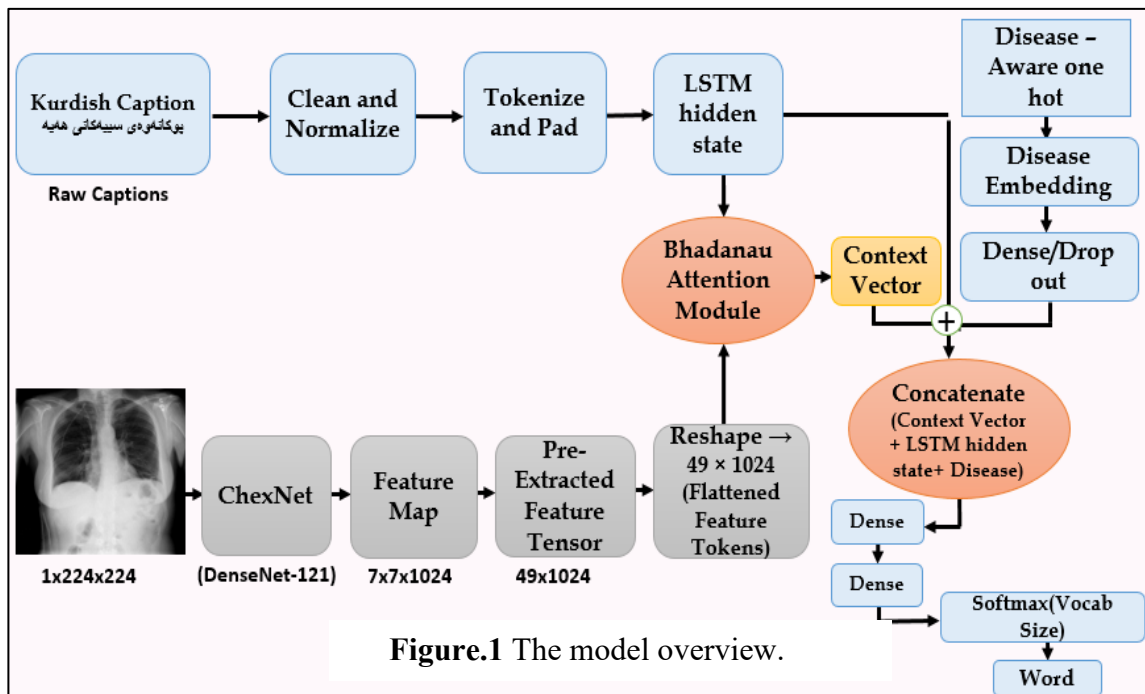
2-Literature Review

Visual characteristics and medical ideas are extracted from radiological images using a multi-label model based on VGG-16. For improved contextual accuracy, these are integrated with semantic data to create captions using an LSTM decoder with beam search (Beddiar et al., 2022). The framework proposed by Beddiar et al. (2022) Lack of disease-specific training results in general captions that could overlook important clinical information, like the difference between pneumothorax and pleural effusion. Furthermore, clinical relevance is not adequately reflected by traditional NLP metrics like BLEU and METEOR. Early diagnosis and treatment planning are made possible by OCT's ability to provide comprehensive retinal pictures. More recently, methods have used CNN-LSTM architecture to produce educational captions. With the use of a sequence processor, a decoder, and an image feature extractor (like VGG-16), the system may be easily adjusted to other feature representations (Vellakani & Pushbam, 2020). The model with (Vellakani & Pushbam, 2020). While a hierarchical RNN (with layers at the sentence and word levels) produces comprehensive medical captions, a global label pooling (GLP) technique uses max pooling to forecast abnormality heatmaps from feature maps. The phrase RNN generates topic vectors based on CNN-extracted features and anomalies, which the word-level RNN then translates into comprehensible medical descriptions (Yin et al., 2019). The assessment of the medical accuracy and practical usefulness of the captions is limited because it only uses basic NLP metrics without clinical validation (Yin et al., 2019). Frequently, baseline medical image captioning models produce descriptions that are too general and exclude information unique to a given ailment. This is addressed by the DiTag and mDiTag models, which improve the detection and description of abnormalities by using visual contrasts between normal and patient images that are retrieved by ResNet-152 at several scales (Park et al., 2020). The model's inaccurate tag predictions are frequently made

by the multi-label classification (MLC) component (Park et al., 2020), which lowers caption quality and adds noise to the semantic feature vector. The model's capacity to generalize to a variety of pathological instances is restricted by the dataset's imbalance and limited size, which includes an excessive number of normal cases and a dearth of samples for specific abnormalities. They were using bias between findings in abnormality and normal using a Cooperative Multi-Agent System (CMAS) that they used reinforcement learning to improve the creation of medical reports. The system, which is incorporated into a CNN–RNN architecture, improves clinical accuracy and language quality by using different agents and a dual incentive method (Pang et al., 2023). Although the suggested reinforcement learning method increases readability and therapeutic relevance, it mostly relies on well-thought-out reward systems, which can be intricate and arbitrary. Furthermore, the multi-agent structure makes training more challenging and the model more complex, which may restrict scalability and practical implementation (Pang et al., 2023). To find important local areas and encode them as semantic elements that will direct the creation of captions, AMAnet employs a multi-label classification network. Compared to baseline models, it generates more accurate and clinically relevant reports thanks to a sentinel gating mechanism that balances attention between visual inputs and LSTM memory (Yang et al., 2021). The effectiveness of the model is largely dependent on the level of visual attention, which can be affected by noise or changes in image quality. Furthermore, SPEA-II may limit real-time application and increase computing complexity even when it is utilized for parameter optimization (A. Singh et al., 2022).

3. Materials and Methods

The model produces disease-aware image captions for chest X-rays and combines textual inputs and diagnostic labels in a single architecture. **Figure 1**



3.1. Dataset

A custom dataset was developed and employed in this study to ensure relevance and specificity to the research objectives. The dataset includes chest X-ray. Chest X-ray images from a variety of patient demographics, including men, women, children, and the elderly, are included in the dataset. These image data were collected over a period of approximately five months from two medical institutions: Rizgary Hospital and Mihrabani Hospital. There was a risk of data loss over time because medical imaging data in some hospitals had not been digitalized. Consequently, the information had to be collected right away in order to be preserved. In several instances, professional-grade smartphones were used to take pictures because of the lack of digital storage infrastructure

or limitations on data transfer via USB drives or CD media. Hospitals with digital systems, on the other hand, made it simpler and more organized to retrieve imaging data. The custom dataset includes 7775 grayscale chest X-ray images classified into nine diagnostic classes of cases: atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, normal, pneumothorax, and consolidation. The category distribution of the dataset utilized in this investigation is shown in

Table 1. Atelectasis, cardiomegaly, consolidation, effusion, infiltration, normal, mass, nodule, and pneumothorax are among its nine diagnostic categories. The table lists the number of photos, age range, gender representation of the patients, standard X-ray view positions (PA/AP), and whether or not appropriate Kurdish captions are available for each class.

Attribute s Class	Atelectas is	Cardiomeg aly	Consolidati on	Effusio n	Infiltrati on	Norm al	Mass	Nodul e	Pneumotho rax
Image Count	868	860	860	860	860	887	860	860	860
Gender	Female/ Male	Female/Ma le	Female/Ma le	Female /Male	Female/ Male	Femal e/Mal e	Femal e/Mal e	Femal e/Mal e	Female/Mal e
Age Range	18-80	20-90	18-78	28-85	18-20	18-85	20-85	20-90	20-92
View Position	PA / AP	PA / AP	PA / AP	PA / AP	PA / AP	PA / AP	PA / AP	PA / AP	PA / AP
Caption Availabili ty	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table.1 show a category of chest X-ray' dataset

Three carefully annotated Kurdish captions detailing the radiological findings were gathered for each X-ray image. Direct hospital visits were used to create these captions, and knowledgeable annotators gave thorough explanations based on their clinical analysis of the medical photos. The Kurdish language captions were reviewed to make sure of the linguistic quality and correctness of the captions. To ensure that the medical information was presented in Kurdish in an understandable and accurate manner, these specialists checked each caption's grammar, vocabulary, and sentence construction. Saving linguistic integrity was interested in ensuring that the dataset was suitable for clinical interpretation as well as future studies on multilingual medical picture captioning. The flexibility, structure, and compliance with deep learning workflows of JSON made it the ideal format for storing the image captions in this investigation. In contrast to flat file formats like CSV or TXT, JSON maintains the hierarchical structure required for rich annotation while enabling the effective grouping of many captions per image. It is easy to use common Python modules because it is machine-friendly and human-readable. JSON is so good for long-term portability and system combination because it is also widely supported across platforms and is independent of the language used. It also makes it possible to add extra annotation layers or patient metadata in the future without changing the schema as a whole. JSON is the optimal choice for handling caption data.

3.2. Feature Extraction

CheXNet has been used for encoding the content of chest X-ray images. Beside this, the model weights are provided by brucechou1983 (CheXNet_Keras_0.3.0_weights.h5). A convolutional neural network with DenseNet121 that was pretrained on the **ChestX-ray14** dataset. This model has proven to be effective in classifying thoracic diseases (Rajpurkar et al., 2017). It is accurate for extracting features from grayscale images that are pertinent to medicine (Ouis & Akhloufi, 2024). For encoding information of radiological from chest X-ray images, CheXNet, a good feature extractor based on DenseNet-121

and pretrained on the ChestX-ray14 dataset, was used in this paper. By setting (include_top=False), the model was loaded without the classification head, removing the fully linked layers for disease classification and the final global average pooling. Instead, each image was given a spatial feature map of size $7 \times 7 \times 1024$ after feature representations were taken out of the final convolutional block. For applications that need fine-grained spatial context, such as medical image interpretation, the method maintains the geographical localization of diseased regions. The image attributes can be efficiently used as input to attention-based models such as Bahdanau Attention by restructuring this into a sequence of 49 (7×7) vectors, each with 1024 dimensions. This enables the decoder to concentrate on particular areas while creating a caption. By avoiding retraining the entire network and ensuring transfer learning from a medically relevant domain, this approach to CheXNet greatly lowers computational costs and overfitting concerns on tiny medical datasets. The spatial shape of the resultant feature vector is either (49, 1024) (i.e., $7 \times 7 \times 1024$ flattened), obtaining rich semantic and spatial data. The features have been extracted for every image and saved as a dictionary for all three split datasets: train, validation, and test. A normalized picture ID is used to index each image feature, and all files are cleaned by removing their extensions and capitalizing the first character to guarantee constant caption key matching. For quick loading during model training and inference, the retrieved features were serialized and saved.

3.3. Caption Preprocessing

The Kurdish captions were subjected to a number of preprocessing procedures in order to allow the model to learn structured textual representations. including cleaning, tokenization, and normalization. Three unique captions, each carefully chosen, were preprocessed on the Kurdish captions in order to make it possible for the model to learn structured textual outputs. Three captions are added to each chest X-ray image, which are selected, cleaned, and mapped regularly to the whole image features.

3.3.1 Caption Cleaning

All raw captions have been separated into three parts—train, validation, and test—then saved. Each caption was

- **Flattened** when nested within the lists.
- **Cleaned** by eliminating punctuation, additional whitespace, and non-Kurdish characters using Unicode-aware regular expressions.
- Using and tokens to denote the beginning and end of a sentence.
- To prevent training on repetitive sentences, reduplicated sentences were used.

The following preparation workflow was finished. The cleaned captions were stored and arranged methodically into sorted files, capitalizing only the first character to make sure the image IDs align with the format of extracted features.

3.3.2. Normalization

After cleaning captions, the mapping dictionary was harmonized across all splits. For uniformity with feature keys, the first character of each image ID was capitalized, and file extensions were removed, utilizing a set seed to ensure reproducibility.

3.3.3. Maximum Caption Length

The model determined the maximum sequence length throughout the corpus after combining all of the captions. The fixed input length for training the decoder was defined by the longest caption, which had 18 tokens.

3.3.4. Tokenization

The cleaned Kurdish text was transformed into integer sequences using Keras' Tokenizer. In order to handle uncommon or unseen words during training and inference, (<unk>) was used as the initial out-of-vocabulary (OOV) token in the Tokenizer. Due to dialectal variances, uncommon terminology, and morphological complexity, this token acts as a stand-in for any word that is

absent from the taught lexicon. This is particularly crucial in Kurdish. Incorporating (<unk>) guarantees that the model maintains its resilience to linguistic fluctuations and improves its generalization when faced with unknown terms. The tokenizer was trained on all captions across all splits after being initialized with an out-of-vocabulary (<unk>) token. Including every word found in the training captions, the final vocabulary size was 407 tokens. In the subsequent model pipeline, a sequence decoder was trained using the tokenizer and cleaned caption data.

3.4. Model Architecture

The model architecture incorporates a disease-aware picture captioning framework made up of two primary modules: a decoder and an encoder. The encoder uses visual features from chest X-ray pictures that are obtained by CheXNet, and one-hot encoded class vectors are used to add disease-specific information. An attention-based decoder incorporates these inputs to provide Kurdish captions with therapeutic significance. The model may generate pertinent and contextually aware descriptions that are suited to radiological information by incorporating both spatial image representations and illness markers. The system creates precise and clinically relevant Kurdish descriptions for chest X-rays by combining visual elements, textual context, and disease-specific cues.

3.4.1. Image Encoder

A. Image Feature Extraction (CheXNet):

As mentioned in the previous section, extracting high-level semantic representations from grayscale images of chest X-rays, the CheXNet has been employed, a neural model based on DenseNet121 that was pretrained on the ChestX-ray14 dataset. A (49×1024) feature map is created by resizing and processing each input image, where the number of spatial regions (7×7 grid) is 49, and the feature vector's depth for each region is 1024 (Rajpurkar et al., 2017). To guarantee precise alignment

with caption keys during training and evaluation, the collected features are sorted and saved.

B. Class diseases conditioning:

One-hot encoded illness vector that represents one of the nine present thoracic disorders (atelectasis, cardiomegaly, consolidation, effusion, infiltration, mass, nodule, normal, and pneumothorax) in accordance with a semantic conditioning approach. This method, which was influenced by earlier research and is now common in radiology captioning pipelines, uses explicit disease labels to direct the language model as it is decoding (Zheng & Yu, 2023). The implementation concatenates the picture characteristics and intermediate caption representation with the illness label that has been taken from the image filename and encoded using a fixed mapping. This technique guarantees that the Kurdish captions produced are contextually aware and consistent with the clinical results.

3.4.2. Decoder

A Keras Tokenizer that has only been trained on the training captions is used to clean, normalize, and tokenize the target captions. The maximum caption length is set at 18 tokens, while the vocabulary size is 407 characters. Every sequence is run through a 256-unit LSTM after being integrated into a 256-dimensional vector. The decoder's internal context for word creation and attention alignment is the final LSTM hidden state. **The Bahdanau attention mechanism has been used** in order to allow the decoder to dynamically concentrate on the most pertinent spatial characteristics while predicting each word, integrating Bahdanau attention. This attentional system calculates a score of significance between each spatial feature and the hidden state of the LSTM at that moment. These spatial features' weighted sum (context vector). The output layer and fusion are concatenated by the decoder:

1. Context vector using attention-weighted
2. Hidden state with the final LSTM
3. The class vector with densely projected diseases

To predict the next word in the caption, this combined vector is run through a dense ReLU-activated layer and a final Dense (vocab_size, activation='softmax') layer.

3.5. Training Pipeline and Supervision Strategy

To generate tuples with the format (image_feature, input_sequence, disease_vector) → target_word, utilizing a custom data generator, enabling dynamic caption learning and supervision at the word level. The following is the structure of the training process:

I. Supervision of Sequence: A sliding window is used to break down each caption into several training pairings. The caption "پوكانه وهی سییه کان" for instance, is changed to:

- (" ,) → 'پوكانه وهی'
- (" , 'پوكانه وهی') → 'سییه کان'
- and so on. All input sequences are padded to a maximum length of 18 tokens.

II. Captions—Features Alignment: To avoid training without matching, all captions and features are aligned beforehand using key intersections. The result of datasets consists of:

- 4664-aligned for features and captions in training
- 1554-aligned for features and captions in validation pairs
- 1557-aligned for features and captions in test These are saved as sorted JSON and .pkl files after verification.

III. Wrapping Dataset: TensorFlow's tf.data is used to wrap the generator. Dataset API using the pipeline listed below: .shuffle(1000), .repeat(), .batch(64), and .prefetch(tf.data.AUTOTUNE).

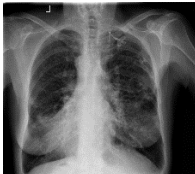
IV. Parameters of Training: Loss Function: SparseCategoricalCrossentropy, Optimizer: Adam (default parameters), Batch Size: 64, Epochs: 20, and Steps per

Epoch: computed dynamically using complete token sequences; for instance, 123,180 sequences → 1925 steps per epoch

V. Callback: Simple Logger, in which loss and accuracy are recorded per epoch by a custom Keras callback. The model is automatically saved with the best validation accuracy.

4. Results

To evaluate the produced Kurdish medical captions' quality and clinical applicability, a wide range of evaluation metrics for natural language creation, both traditional and modern, were used. **Table 2** illustrates the result of the image with the generated caption, while words in red color indicate significant or pertinent phrases that were left out of the model's prediction, and words highlighted in blue color show accurately anticipated concepts that match the ground truth.

Chest X-Ray Images	Ground Truth	Predicted
	ئاوی کۆکراوهی ده‌رکه‌وتوو له نێوان په‌رده‌ی سییه‌کان، ئه‌نجامی تیشکی وێنه‌که ئاوی کۆبووه نیشان دهدات له نێو په‌رده‌ی سییه‌کان، له‌نێوان په‌رده‌ی سییه‌کان ئاوی کۆبووه نیشاندراوه	ئاوی کۆکراوه‌ی ده‌رکه‌وتوو له‌ نێوان په‌رده‌ی سییه‌کان
	هیچ گۆرانکاریه‌کی نه‌خۆشی دیار نه‌یه له وێنه‌ی تیشکی سینگه‌که، دروستی و نه‌بوونی کێشه له سی و دڵ و به‌شه‌کانی پرونکراوه‌ته‌وه له وێنه‌ی تیشک، له‌م وێنه هه‌موو به‌شێکی سینگ و سی ته‌ندروسته نه‌خۆشی نییه	ئه‌نجامی تیشکی وێنه‌که نیشاندهری روونی ته‌ندروستیه
	له‌م وێنه‌دا به‌دیارکه‌وتنی گرتی ورد له سییه‌کانی نه‌خۆش پرونکراوه‌ته‌وه، گرتی بچکۆله دیارخراوه له وێنه‌ی تیشکی سینگ، گرتی بچکۆله‌ی دیاره له ناو تیشکی سییه‌کانی نه‌خۆش	هه‌بوونی گرتی له‌نێو سی له تیشکی نه‌خۆشه‌که دیارخراوه

cause of this is the lexical overlap in descriptions because the visual characteristics of the two circumstances are comparable. In particular, nodule-related captions frequently use phrases that set them apart, like "بچکۆله", "ورد", or "بچکۆلهی". They are not as commonly linked to mass. Higher-probability sequences are typically preferred by beam search, which can occasionally cause misunderstanding between the two classes. On the other hand, greedy decoding showed better accuracy in producing captions that accurately distinguish nodules from masses, even though it had lower total metric scores. With beam search The nodule will be a problem because it is the same mass; the only difference is the nodule has (بچکۆله, بچکۆلهی), but the mass doesn't have it, but by greed this problem is fixed. Although all evaluation metrics in beam are better than greedy, except for predicting captions in mass and nodule, which will be fixed by greedy. Table 3 illustrates the corrected generation of the caption of the Nodule class with a greedy search.

Chest X-Ray	Ground Truth	Predicted
	لەم وێنەدا بە دیارکەوتنی گرتی ورد لە سییه کانی نەخۆش روونکراوەتەوه، گرتی ورد لە سی نەخۆش روون و ئاشکرا دیاره، لە وێنەى تیشكى ئەم نەخۆشه گری ی بچکۆلهی ههیه لهناو سییه کان	لە وێنەى تیشكى ئەم نەخۆشه گری ی بچکۆلهی ههیه لهناو سییه کان

Table.3 Result of image with generated caption Nodule class with greedy search

captions. The following metrics were **BLEU scores for beam-generating captions with BLEU-1, BLEU-2, BLEU-3, and BLEU-4**: Employed BLEU's NLTK (Bilingual Evaluation Understudy) use of smoothing to measure the n-gram accuracy of the reference and predicted captions: (BLEU-1: 0.6182), (BLEU-2: 0.4805), (BLEU-3: 0.4323), and (BLEU-4: 0.4015). These results show that the model does a moderate job of capturing patterns at the unigram and larger phrase levels. The

longest common subsequence (LCS) between the reference and predicted captions is assessed using ROUGE-L. The evaluated ROUGE-L for the beam-generating caption is 0.8722. The model evaluated **BERTScore (multilingual)** in order to determine the semantic similarity between the embeddings of the reference and predicted captions. Employed BERTScore with XLM-Roberta. This works especially well for low-resource, morphologically rich languages like Kurdish: (Precision: 0.9131), (Recall: 0.9172), and (F1-Score: 0.9151). With **CHRF++**, character-level n-gram F-score is evaluated using CHRF++, which is helpful for morphologically complicated languages. CHRF++ F-score: 46.3099. Finally computed accuracy that is Nearly 90% of the predictions made during the evaluation phase were accurate, according to the suggested model's high accuracy of 0.8984. This impressive performance demonstrates how well the model matches generated captions to the relevant visual characteristics and diagnostic context in chest X-ray pictures. **Table 4** and **Figure 2** show the result of the evaluating metrics for our model with beam-generated captions.

BLUE-1	BLUE-2	BLUE-3	BLUE-4	ROUGE-L	CHRF++	BERT-P	BERT-R	BERT-F1	Accuracy
0.6182	0.4805	0.4323	0.4015	0.8722	41.6092	0.9131	0.9172	0.9151	0.8984

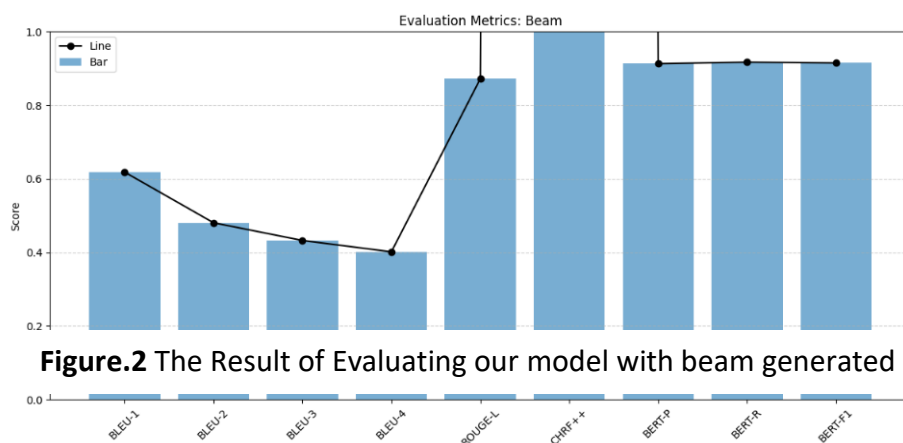


Figure.2 The Result of Evaluating our model with beam generated caption.

5. Discussion and Comparison

The continuous attempts to enhance automated radiology reporting's interpretability, semantic coherence, and clinical accuracy are highlighted by this comparison. The features are an attention-based LSTM decoder and a disease-aware captioning pipeline based on CheXNet (DenseNet-121) that is especially suited for the Kurdish medical context.

I. Model Effectiveness

With a BLEU-4 score of 0.4015, the model's predicted and reference captions show a moderate n-gram overlap. Though there is little lexical overlap, the generated captions appear to be semantically accurate, which is an important characteristic in medical applications where paraphrasing is frequently used, according to the BERTScore F1 of 0.9151. This comparison between BERTScore and BLEU shows that our multilingual, morphologically rich target language (Kurdish) is better suited for embedding-based evaluation. And traditional lexical metrics (like ROUGE-L) yield 0.8722 which is strong due to the same matching reference strings.

II. Impact of Disease Awareness

The model uses diagnosis-specific priors to direct generation by incorporating disease class vectors into the decoder. More targeted and pertinent captions are made possible by this class conditioning process, particularly for illnesses with visually overlapping symptoms (e.g., effusion vs. atelectasis). In contrast to baseline models without class awareness, our qualitative assessment demonstrates that the decoder regularly incorporates disease-relevant terms in its outputs.

III. Comparison with Non-Attention Baselines

The capacity of traditional CNN-RNN captioning algorithms to localize diseases is limited since they treat visual information as global vectors. On the other hand, our model allows for fine-grained focus by utilizing spatial attention across

49×1024 feature maps. This design improves precision in phrase selection (e.g., “collapse of lung” vs. generic terms like “abnormality”), especially in longer or multi-clause captions.

IV. Error Analysis

Despite strong performance, some generated captions suffer from repetitive phrasing or lack of syntactic variety. This is partially due to limited caption diversity per image and the absence of beam search or temperature sampling during generation. Additionally, rare diseases may be underrepresented in the training set, limiting class-specific fluency.

A direct comparison within the same linguistic context is not feasible at this time due to the lack of Kurdish-language medical picture captioning systems. However, a comparison study with well-known state-of-the-art models in other languages, mainly English, has been supplied in order to put the contribution of this research into context. In this research a state-of-the-art design has been implemented to caption Kurdish text in medical images. The performance and novelty of the proposed model compared with other research are provided in **Table 5**. The primary discoveries or contributions presented in this study, in **Table 6**, the comparison provides their contribution, the dataset utilized, and the proposed architectural elements are highlighted in each section. The previous research covers a wide range of approaches, such as multi-label classification techniques, transformer-based language models, encoder-decoder frameworks, and disease-aware mechanisms.

Authors	Findings	Architectures	Dataset
(Beddiar et al., 2022)	presented a technique for creating a vocabulary based on medical terms and descriptions; ablation research was carried out to evaluate the model's constituent parts.	VGG-16 , LSTM	ImageCLEF 2021

(A. Singh et al., 2022)	verified on benchmark datasets after proposing an encoder-decoder Show-Attend-Tell model using SPEA-II for parameter adjustment.	CNN , LSTM	VQA-Med
(Selivanov et al., 2023)	suggested an image captioning model that achieved better NLG metrics on MIMIC-CXR by integrating SAT and GPT-3 with a unique preprocessing methodology.	DenseNet-12140, 512 for VGG-1641, 2048 for InceptionV342 and ResNet-10143 ,LSTM and GPT-3	MIMIC-CXR (IU X-ray)
(Zheng & Yu, 2023)	suggested an interactive captioning system that optimizes keyword prediction and report production by utilizing evidential uncertainty and user feedback.	CNN + LSTM , evidential learning and interactive feedback	IU X-Ray Dataset
Our Model	Developed a disease-aware Kurdish captioning model using CheXNet and LSTM with Bahdanau attention; outperforms prior models across multiple metrics.	CheXNet(121DENSE)+LSTM+ Disease Aware	Our Dataset

Table.5 The Compare for the model and other models

Thoroughly compared the suggested Kurdish medical image captioning model's performance with current state-of-the-art methods documented in the literature in order to evaluate it objectively. Both conventional n-gram overlap metrics and contemporary semantic similarity metrics, such as BERTScore (Precision, Recall, F1), ROUGE-L, CHRF++, and BLEU (1–4), are taken into account in the evaluation. The findings of these investigations across various datasets and architectures are compiled **Table.6**.

Some techniques investigated uncertainty modeling and domain-specific optimizations, while others concentrated on enhancing linguistic coherence or visual feature encoding. Our model performs exceptionally well, especially when it comes to BLEU and BERTScore measures, which show improved semantic alignment and linguistic correctness with reference captions. The robustness of our method is further reinforced by the inclusion of CHRF++, which is particularly pertinent for morphologically rich languages.

Reference	Architectures	Blue-1	Blue-2	Blue-3	Blue-4	Rouge-L	CHRF++	BERTScore P	BERTScore R	BERTScore F1
(Yin et al., 2019)	CNN-LSTM, R2Gen+GOVE	0.445	0.292	0.201	0.154	0.344	-	-	-	-
(Chen et al., 2020)	Memory-driven Transformer (MDT)	0.353	0.218	0.145	0.103	0.277	-	0.333	0.273	0.276
(Xu et al., 2020) ^v	Reinforce CNN-LSTM	0.412	0.279	0.206	0.157	0.342	-	-	-	-
(Zeng et al., 2020) [*]	Double CNN-LSTM	0.47 ±0.02	0.40 ±0.02	0.30 ±0.01	-	-	-	-	-	-
(S. Singh et al., 2021)	CNN-LSTM	23.07	11.86	7.05	4.75	23.15	-	-	-	-
(Alfarghaly et al., 2021)	Condition GPT2	0.387	0.245	0.166	0.111	0.289	-	-	-	-
(Lee et al., 2022)	R2Gen+GOVE	0.4566	0.2910	0.1967	0.1365	-	-	-	-	-
(Pino et al., 2022)	CNN-LSTM ATT	0.361	0.226	0.152	0.106	0.314	-	-	-	-
Ours	CNN-LSTM-Disease Aware	0.6182	0.4805	0.4323	0.4015	0.8722	41.6092	0.9131	0.9172	0.9151

Table.6 the Compare Metrics for our model and other models

Strong semantic alignment with ground-truth medical descriptions is demonstrated and suggested by the class-conditioned captioning approach, especially when assessed using contemporary embedding-based metrics.

6. Conclusion

This paper created the disease-aware medical image captioning system that is designed for the Kurdish language. The study integrated an LSTM decoder, Bahdanau attention, and a pretrained CheXNet model to make effective, accurate, and semantically rich Kurdish captions for chest X-ray images that used high-quality visual feature extraction. The conditioning vectors of disease-specific use eliminate the generic and clinically ambiguous outputs that are common in older models and guarantee that the generated captions are in line with diagnostic significance. Multilingual medical AI research is further supported by the development of a fresh dataset that includes thousands of chest X-rays with manually annotated Kurdish captions across nine illness categories. The suggested model outperformed current state-of-the-art systems in thorough evaluations using BLEU, ROUGE-L, CHRF++, and BERTScore, especially in semantic alignment and linguistic coherence. These findings demonstrate how well class-aware attention mechanisms and localized linguistic modeling work together in the medical field. However, there are still certain restrictions. Furthermore, by combining larger-scale datasets, syntactic augmentation, or transformer-based decoders with reinforcement learning techniques, future research can solve the model's occasional production of repeated or syntactically stiff sentences. To sum up, our work lays the groundwork for automated medical reporting in Kurdish and other low-resource languages that is sensitive to disease.

7. Future Work

Using the suggested model as a web-based tool to enable real-time medical image captioning in clinical settings—especially in areas with limited resources—is one of the study's future approaches. To enhance generalization and linguistic variety, additional fine-tuning will be carried out utilizing bigger and more varied datasets. Extending the framework into a multimodal system that integrates image classification and caption creation is another crucial avenue for enabling concurrent disease identification and descriptive reporting. The goal of this

integrated strategy is to improve medical practitioners' interpretability and diagnostic support.

Acknowledgement

The medical imaging data (X-ray) images utilized in this investigation was made available by Mihrabani Hospital and Rizgary Hospital, for which we are especially grateful. Additionally, thanks are given to the medical experts and annotators at Rizgary Hospital, who shared their knowledge in creating the Kurdish medical captions with a linguistic specialist of Kurdish medical captions. Without their assistance and cooperation, this job would not have been feasible.

Reference

- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. 3–9. <http://arxiv.org/abs/1711.05225>
- Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 6666–6673. <https://doi.org/10.1609/aaai.v33i01.33016666>
- Boag, W., Harry Hsu, T. M., McDermott, M., Berner, G., Alsentzer, E., & Szolovits, P. (2019). Baselines for Chest X-Ray Report Generation. *Proceedings of Machine Learning Research*, 116, 126–140.
- 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. *Proceedings - IEEE International Conference on Data Mining, ICDM, 2019-Novem(Icdm)*, 728–737. <https://doi.org/10.1109/ICDM.2019.00083>
- Chen, Z., Song, Y., Chang, T., & Wan, X. (2020). *Generating Radiology Reports via Memory-driven Transformer*. 1439–1449.
- Park, H., Kim, K., Yoon, J., Park, S., & Choi, J. (2020). Feature difference makes sense: A

- medical image captioning model exploiting feature difference and tag information. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 95–102. <https://doi.org/10.18653/v1/2020.acl-srw.14>
- Vellakani, S., & Pushbam, I. (2020). An enhanced OCT image captioning system to assist ophthalmologists in detecting and classifying eye diseases. *Journal of X-Ray Science and Technology*, 28(5), 975–988. <https://doi.org/10.3233/XST-200697>
- Xu, W., Qi, C., Xu, Z., & Lukasiewicz, T. (2020). *Reinforced Medical Report Generation with X-Linear Attention and Repetition Penalty*. <http://arxiv.org/abs/2011.07680>
- Liu, F., Wu, X., Ge, S., Fan, W., & Zou, Y. (2021). Exploring and distilling posterior and prior knowledge for radiology report generation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 13748–13757. <https://doi.org/10.1109/CVPR46437.2021.01354>
- Singh, S., Karimi, S., Ho-Shon, K., & Hamey, L. (2021). Show, tell and summarise: learning to generate and summarise radiology findings from medical images. *Neural Computing and Applications*, 33(13), 7441–7465. <https://doi.org/10.1007/s00521-021-05943-6>
- 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>
- Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., & Fahmy, A. (2021). Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24, 100557. <https://doi.org/10.1016/j.imu.2021.100557>
- Beddiar, D. R., Oussalah, M., Seppänen, T., & Jennane, R. (2022). ACapMed: Automatic Captioning for Medical Imaging. *Applied Sciences (Switzerland)*, 12(21), 1–24. <https://doi.org/10.3390/app122111092>
- Lee, H., Cho, H., Park, J., Chae, J., & Kim, J. (2022). Cross Encoder-Decoder Transformer with Global-Local Visual Extractor for Medical Image Captioning. *Sensors*, 22(4), 1–13. <https://doi.org/10.3390/s22041429>
- Pino, P., Parra, D., Besa, C., & Uribe, S. (2022). *Inspecting state of the art performance and NLP metrics in image-based medical report generation*. *NeurIPS*, 15–17. <https://doi.org/10.52591/ixai202012128>
- 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.

<https://doi.org/10.1155/2022/9638438>

- Zheng, E., & Yu, Q. (2023). Evidential Interactive Learning for Medical Image Captioning. *Proceedings of Machine Learning Research*, 202, 42478–42491.
- Pang, T., Li, P., & Zhao, L. (2023). A survey on automatic generation of medical imaging reports based on deep learning. *BioMedical Engineering Online*, 22(1), 1–16.
<https://doi.org/10.1186/s12938-023-01113-y>
- Selivanov, A., Rogov, O. Y., Chesakov, D., Shelmanov, A., Fedulova, I., & Dylov, D. V. (2023). Medical image captioning via generative pretrained transformers. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-31223-5>
- Ouis, M. Y., & Akhloufi, M. A. (2024). ChestBioX-Gen: contextual biomedical report generation from chest X-ray images using BioGPT and co-attention mechanism. *Frontiers in Imaging*, 3. <https://doi.org/10.3389/fimag.2024.1373420>

ژێرنووسی وێنەی پزیشکی بە زمانی کوردی بە بەکارهێنانی زیرەکی دەستکرد

پوخته

دەستنیشانکردنی تێکچوونەکانی ناوچەی سینگ و سییهکان وەک گەورەبوونی دڵ، کۆبوونەوهی ئاو لە نێوان پەردە سییهکان، بوونی هەوا لە نێوان پەردەکانی سییهکان زیاتر بەنده بە وێنەگرتنی تایبەتی پزیشکی، بەتایبەتی تیشکی ئێکسی سینگ. بەلام بۆ لێکدانەوهی ئەم وێنە پێویستیان بە پزیشکی لێهاتووی تیشک هەیە و کەمیان لە هەندێ ناوچەکاندا دەبیته هۆی دواکەوتنی دەستنیشانکردن. پێشکەوتنەکانی ئەم دواپێیە لە (deep learning) بۆ ئۆتوماتیکیکردنی لێکدانەوهی وێنە لە پرۆژە دروستکردنی کاپشنەوه نیشانداوه، هەرچەندە زۆربەی سیستەمەکانی ئێستا تێگەیشتنێکی کەمیان تایبەت بە دیاریکردنی نەخۆشییان هەیە وە لەسەر کۆمەڵە داتاکانی زمانی ئینگلیزی راپێنراون، ئەمەش سوودی کلینیکی و جێبەجێکردنی زمانی تر سنووردار دەکات. مۆدیلی نووسینی وێنەی پزیشکی (رەچاوکردنی - نەخۆشی) کە بۆ زمانی کوردی (سۆرانی) پەڕە پێدراوه لەم لێکۆڵینەوهیەدا خراوەتەر وو. لەم لێکۆڵینەوهیە پێشنیاری (Bahdanau attention-equipped LSTM) کراوه بۆ رێنماییکردنی دروستکردنی سەردێر و جێگیرکردنی جوړی تایبەتی نەخۆشی لەخۆدەگریت بۆ

دهرهینانی تایبه تمه ندیه له به شه کانی وینه به به کارهینانی (CheXNet (DenseNet-121, که پیشوه خته له سه ر ChestX-ray14 راهینرابوو. بۆ راهینان و هه لسه نگاندنی سیسته مه که، داتا سیتهیکی نویی کوردی دروستکرا، که بریتی بوو له ۷۷۵ وینهی تیشکی ئیکس سینگ له نو جۆری جیاوازی دهستنی شانکردن له گه ل پیناسه کوردیه به وردی هه ل بژێردراوه کان. نمره ی BLEU-4 0.4015, نمره ی ROUGE-L 0.8722, و نمره ی BERTScore F1 0.9151, مۆدیله که له پیوه ره کانی ئیستای برده وه کاتیک پته وی کارکردن به به کارهینانی BLEU, ROUGE-L, BERTScore, و CHRF++ هه لسه نگیندرا. ئه م دۆزینه وانه نیشان دده ن که سیسته مه که تا چه ند به باشی له ژینگه یه کی زمانیدا که سه رچاوه ی سنوورداره, نووسینی زمانه وانی یه گگرتوو و په یوه ندیدار به نه خۆشی به ره ه م ده هینیت. ئه م کاره کۆمه له داتایه کی کوردی کوالیته ی به رز نیشان دده ات بۆ یارمه تیدانی توێژینه وه کانی داها توو له به رنامه کانی AI کلینیکی فره زمانیدا، جگه له په ره پیدانی دۆخی ژیرنووسی وینه ی پزیشکی بۆ زمانه پێشکه شکراره کان که ده گمه نیان پێشکه ش کردوو. وشه ی سه ره کی: Deep Learning; LSTM; Bahdanau Attention Natural Language Generation; مۆدیلی په چاو کردنی نه خۆشی ; زمانه سه رچاوه که مه کان .

وصف الصور الطبية باللغة الكردية باستخدام الذكاء الاصطناعي

ملخص

يعتمد تشخيص اضطرابات الصدر، مثل تضخم القلب، وانصباب، استرواح الصدر، بشكل كبير على التصوير الطبي، وخاصةً تصوير الصدر بالأشعة السينية. إلا أن تفسير هذه الصور يتطلب أخصائي أشعة ماهرين. نقص أطباء التشخيص في المناطق النائية يؤدي إلى تأخير التشخيص. وقد أظهرت البحوث الحديثة في (deep learning) أتمتة تفسير الصور من خلال إنشاء التعليقات التوضيحية؛ إلا أن معظم الأنظمة الحالية تفنقر إلى فهم تشخيصي محدد، وتُدرَّب على مجموعات بيانات باللغة الإنجليزية، مما يحد من فائدتها السريرية وتطبيقاتها اللغوية الأخرى. تُقدِّم هذه الدراسة نموذج وصف الصور الطبية المُطوَّر للغة الكردية (السورانية). في هذه الدراسة، يُقترح استخدام وحدة معالجة الصور الطبية طويلة المدى (Bahdanau attention-equipped LSTM)، ويتضمن تحديد موقع خاص بنوع المريض لاستخراج سمات منطقة الصورة باستخدام شبكة (CheXNet (DenseNet-121، والتي تم تدريبها مسبقاً على ChestX-ray14 لتدريب النظام وتقييمه، تم إنشاء مجموعة بيانات كردية جديدة، تتكون من 7775 صورة أشعة سينية لتسعة أنواع تشخيصية مختلفة مع أوصاف كردية مختارة بعناية. مع درجة BLEU-4 البالغة 0.4015، ودرجة ROUGE-L البالغة 0.8722، ودرجة BERTScore F1 البالغة 0.9151، تفوق النموذج على المعايير الحالية عند تقييم المتانة باستخدام BLEU و ROUGE-L و BERTScore و CHRF++. توضح هذه النتائج مدى جودة إنتاج النظام لكتابة لغوية متماسكة وذات صلة مرضية في بيئة لغوية محدودة الموارد. يوضح هذا العمل مجموعة بيانات كردية عالية الجودة لمساعدة الأبحاث المستقبلية في تطبيقات الذكاء الاصطناعي السريري متعدد اللغات، بالإضافة إلى تطوير حالة ترجمة الصور الطبية للغات المعروضة بشكل نادر. الكلمات المفتاحية: Deep Learning; LSTM; Bahdanau Attention Natural؛ نموذج وعي المريض؛ لغات الموارد النادرة.