scientific reports

Check for updates

OPEN

A lightweight transformer-based hybrid encoder-decoder model for chest X-ray medical report generation

Murat Ucan[1]✉, Buket Kaya[2], Mehmet Kaya[3]✉ & Reda Alhajj[4,5,6]✉

Diagnosing diseases from medical images and reporting them at the paragraph level is a significant challenge for deep learning-based autonomous systems. Existing work primarily focuses on achieving high accuracy, often paying less attention to the computational cost of training and testing. The goal of this work is to build a low computational cost and high-performance hybrid encoder-decoder architecture capable of producing autonomous medical reports. On the encoder side of our architecture, called FAST-MRG, features are extracted from images with a transformer-based encoder enriched with distillation techniques, while on the decoder side, a generative pre-training transformer generates paragraph-level text using the extracted features. Numerical analysis with word matching evaluation metrics, temporal analysis and observational analysis were performed to measure the success of the architecture. Our hybrid encoder-decoder architecture was trained and tested using chest X-ray images and reports from the Indiana University Chest X-ray collection dataset. The FAST-MRG architecture achieved scores of 0.373, 0.226 and 0.332 on the Bleu-1, Meteor and Rouge evaluation metrics, respectively. It also has an average time efficiency of 66% compared to previous work using similar GPU environments. The study demonstrates through experiments that meaningful reports are produced that can support doctors in diagnosis and treatment processes. In the study, the results are presented not only with measurable average values but also with a density distribution graph and the test results are analyzed in depth. With its low runtime and high performance, the proposed architecture can serve as a basis for future work.

Keywords Chest X-ray, Deep learning, Distillation, GPT, Medical report generation, Transformer

Medical imaging techniques are now widely used in almost all healthcare organizations thanks to technological advancements. However, medical professionals must be on hand to examine the radiographs produced by imaging devices. Smaller healthcare facilities and rural areas typically lack access to specialist physicians. Furthermore, doctors may make mistakes when analyzing medical imaging because of their hectic schedules and elevated stress levels. In order to help doctors make decisions, this condition has made it necessary to create a decision support system that provides results quickly and accurately.

Specialist doctors in many medical disciplines compose reports, which are subsequently archived in patients' files for future examination. When physicians review previous reports, they find documents composed in non-standard terminology. The absence of a standardized linguistic framework for written medical reports constitutes a significant issue in the field of medical documentation.

Another focus of studies focusing on the interpretation of medical images and disease diagnosis is the disease detection rate. It takes an average of ten minutes for the doctor to examine and report a chest X-ray image. In addition to this time, the time it takes for the specialist doctor to write the report and prescribe medication to the patient are also large periods of time before the start of treatment. Considering crowded hospitals, this waiting time is an important factor that affects reaching more patients and therefore public health. However,

[1]Department of Computer Technologies, Dicle University, 21200 Diyarbakir, Turkey. [2]Department of Electronics and Automation, Firat University, 23119 Elazig, Turkey. [3]Department of Computer Engineering, Firat University, 23119 Elazig, Turkey. [4]Department of Computer Science, University of Calgary, Calgary, AB, Canada. [5]Department of Computer Engineering, İstanbul Medipol University, İstanbul, Turkey. [6]Department of Health Informatics, University of Southern Denmark, Odense, Denmark. ✉email: murat.ucan@dicle.edu.tr; kaya@firat.edu.tr; alhajj@cpsc.ucalgary.ca

recent Covid-19 and other epidemics have shown that starting treatment as soon as possible is important for public health[1].

Recently, artificial intelligence technologies have been used in many areas and produce successful results. Different features of artificial intelligence in many fields such as autonomous vehicles, decision-support systems, defense industry and voice analysis make daily life easier in relevant areas. Researchers have conducted many studies in the field of interpretation of medical images with deep learning applications such as classification[2] and segmentation[3]. Classification in the field of medicine is simply defined as determining the disease class to which a medical image belongs[4]. Segmentation can be expressed, for example, as painting the cancerous surface in a different color in a medical image[5]. Although the studies carried out in these fields are valuable, the field of medicine is a critical field as it contains very complex dynamics and directly affects human life. A one-word classification result of the disease class or simply painting the diseased surface is not a sufficient output. In this context, the motivation of our study is to present a deep learning model that can produce detailed outputs at the paragraph level that can support doctors who are experts in their field in making decisions in diagnosing diseases.

The fact that deep learning architecture produces successful results in many areas and that it can also create medical reports is the most important indicator that supports the stated motivation of our study. Studies that produce output at the word level and studies that involve marking the diseased surface do not, on their own, have the competence to support doctors in making decisions for the diagnosis of diseases. Medical reports created in real-life hospitals are texts that give output at the paragraph level and include not only the existence of the disease but also other important findings about the disease. For this reason, our study aims to autonomously write reports that can produce paragraph-level output, similar to medical reports created in hospitals in real life, with deep learning architectures.

The existing studies reviewed in the literature have high computational costs in image-to-report generation processes and are not sufficiently successful in producing meaningful, clinically useful text. Furthermore, they are not integrated with distillation architectures using the attention mechanism as a simple encoder. This study proposes a new architecture called FAST-MRG, consisting of a distillation-enhanced visual encoder and a GPT-based decoder, to address these shortcomings. Our model shows significant improvements over previous approaches when evaluated both quantitatively (BLEU-n, METEOR, ROUGE) and qualitatively (density plot, observational analysis, statistical analysis). Based on the identified gaps in the literature and the study's objectives, this research addresses the following three main research questions (RQs):

- RQ1: Can a hybrid encoder-decoder architecture, combining a distillation-enhanced visual transformer (DeiT) and a GPT-based decoder, generate clinically accurate and meaningful paragraph-level medical reports from chest X-ray images?
- RQ2: Is it possible to significantly reduce computational costs and training time (runtime efficiency) compared to existing heavy deep learning models without compromising report generation performance?
- RQ3: How does the proposed FAST-MRG architecture compare to state-of-the-art methods in terms of standard linguistic evaluation metrics (BLEU, METEOR, ROUGE) and qualitative clinical consistency?

Within the scope of this study, studies were carried out on converting medical images into text. FAST-MRG architecture was developed within the scope of the study based on the encoder–decoder architecture. Indiana University Chest X-Ray Collection, an open access dataset, was used in training, validation and testing phases. The motivations behind our study and the main topics that will be focused on in the literature are listed below.

- In our study, a medical report will be developed that produces paragraph-level output, similar to the reports written by doctors in real hospitals;
- There is no standard medical language and reporting order among the reports written by doctors. The results produced in the study will contribute to a standard language context and report layout;
- The effectiveness of Encoder-Decoder architectures will be analyzed with the architecture developed in the study. Transformer-based Distilled Data-efficient Image Transformer architecture was used on the encoder side, and Generative Pre-training Transformer architecture was used on the decoder side. Comparing the impact of these architectures on reporting performance with other architectures is another focus of our study;
- The studies conducted will be able to support doctors in making a diagnosis during first level emergency intervention in hospitals where there are no specialist doctors in their field;
- Because hospitals are crowded, doctors cannot spare enough time to write reports. This situation not only makes it difficult to examine in sufficient detail, but also makes it common to make mistakes due to busy working hours and target pressure. The proposed architecture will support doctors in diagnosis, reporting, treatment and follow-up processes;
- Deep learning supported reporting process is a process that can be completed in seconds. By using the proposed system in hospitals, the reporting process will be significantly accelerated.

Literature review

Researchers have contributed to the literature by conducting numerous successful studies using deep learning architectures and medical images. Researchers in many fields such as gastroenterological images, ultrasound images, MRI images, microscopic images and x-ray images have been able to produce results on disease diagnosis by analyzing medical images. The common point of the studies in this category in the literature is that they only take input data from the image and produce results for the disease name at the word level.

Medical images are used for disease diagnosis in many areas. One of these is gastroenterological images. The presence of some diseases can be determined by processing the images taken through a camera sent from the

mouth to the stomach. Ucan et al.[6] used gastroenterological images in their study and were able to diagnose diseases in the 8-class data set with a success rate of 0.935. Classification and segmentation studies have also been carried out on ultrasound images. In their study, Cinar et al.[7] tried to diagnose images of brain tumor and normal classes using brain MRI images. Researchers also work with images taken at the microscopic level. Ucan et al.[8] classified images containing benign and malignant cells using cytological images. Many different studies are carried out to classify chest X-ray images. Some of these[9] focus on the presence of a single disease such as pneumonia. In their study, Karaddi and Sharma[10] sought a solution to the multi-class classification problem that can be diagnosed from chest X-ray images. The study, which classifies five different diseases that can be detected from chest X-ray images, can produce word-level output at the diagnosis stage.

The studies examined in this category are valuable and can support doctors who are experts in their field in diagnosing diseases. However, the results produced are classification-based studies based solely on the presence or absence of the relevant finding. Although it is valuable that this information can only be given to doctors who are experts in the field, it is very incomplete. Doctors need more information and interpretation at the time of diagnosis.

Research has also shown that there are studies that produce paragraph-level output from medical images. Singh et al.[11] focused on producing paragraph-level output on chest X-ray images. In the study using Encoder-Decoder architecture, CNN architecture was used in the encoder phase and LSTM architecture was used in the decoder phase. Wang et al.[12] proposed a text-image based deep learning architecture called TieNet. The researchers trained the CNN-RNN based architecture they developed with chest X-ray images and aimed to obtain textual outputs. BLEU-1, Meteor and Rouge-L were able to produce reports with success rates of 0.2860, 0.1076 and 0.2263, respectively, in evaluation metrics. Liu et al.[13] also used reinforcement learning in their method, which was developed using encoder and decoder architecture. Researchers who managed to obtain paragraph-level output from chest X-ray images used two separate data sets in their studies and presented their results separately. With the data set used in our study, they were able to produce reports with success rates of 0.369 and 0.359 in BLEU-1 and Rouge evaluation metrics, respectively.

Harzig et al.[14] conducted another study that enables getting a radiology report from medical chest radiographs. In their study, researchers proposed a system using ResNet-152 and LSTM architectures. In the HLSTM + att + Dual architectures proposed in the study, they achieved a success of 0.357 in the BLEU-1 metric. Alfarghaly et al.[15] used the pre-trained Chexnet[16] deep learning architecture, whose weights are publicly shared, as an encoder. The next stage of their proposed architecture is the semantic feature ex-traction stage. In the Decoder stage, they used attention mechanism and GPT-2. The researchers achieved 0.347 BLEU-1 success with the VSGRU architecture they proposed. In their study, Park et al.[17] preferred ResNet-152 architecture on the encoder side and LSTM architecture on the decoder side. They strengthened their proposed architecture with a multi-head attention layer. They were able to produce paragraph-level medical reports with a BLEU-1 success of 0.3731 in their most successful model, called mDiNAP-transformer-ewp, which can be compared with the method we proposed in our article.

Studies are carried out on visual data in many fields such as security, agriculture, military and health. The usage areas of visual data are expanding day by day and databases consisting of images are becoming important information bases day by day. However, the problem of analytical processing of this data brings with it the necessity of converting images into texts and numbers. In this context, researchers in the literature have carried out various studies on converting images into text.

You et al.[18] proposed a method based on CNN and RNN architecture using Flickr30k and MS-COCO datasets. The texts corresponding to the images in the data sets are mostly single-sentence texts. The researchers managed to achieve a BLEU-1 success rate of 0.647 with the method they proposed. Vinyals et al.[19] proposed a convolutional neural network and LSTM based architecture on Pascal, Flickr30k and COCO datasets. Researchers aim to produce sentence-level output from images. There are also studies that try to generate text using satellite images as input. Li et al.[20] tried to produce texts using Sydney-Captions, UCM-Captions and RSICD datasets containing satellite images and texts. The study focused on converting satellite images into text and identifying objects. Efforts are also being made to convert images into text via video images. In their study, Luo et al.[21] try to describe the video content with an accurate and short sentence. The researchers used publicly available MSVD and MSR-VTT datasets in their studies.

These studies are directly related to obtaining paragraph-level text from medical images. In both areas of study, images and texts are provided as an introduction to deep learning architecture. As output, sentence or paragraph level outputs are produced in both work areas. The most obvious difference between the fields of study is the semantic features of the texts produced. The texts associated with the images used in image captioning have normal word sequences and clear sentences. For this reason, it is more common to predict it with machine translations. However, in the paragraph-level texts that are tried to be obtained using medical images, the texts related to the images are written directly by real doctors who are experts in their field. Sentences, words and word sequences also include human factors. For this reason, it is relatively more difficult to predict the exact medical reporting of texts.

Materials and methods

Encoder-decoder architecture was used within the scope of the study to create paragraph-level medical reports from chest X-ray images. In the encoder section where chest x-ray images are given as input, the DeiT model, known as Distilled Data-efficient Image Transformer, was used. On the decoder side, GPT architecture is used. In Fig. 1, the architectural form of the study is given in general terms.

The suggested FAST-MRG architecture is a mix of an encoder and a decoder that makes medical reports from chest X-ray images at the paragraph level. The main new idea is to use a Distilled Data-efficient Image Transformer (DeiT) as the encoder and a Generative Pre-training Transformer (GPT) as the decoder. The encoder takes the

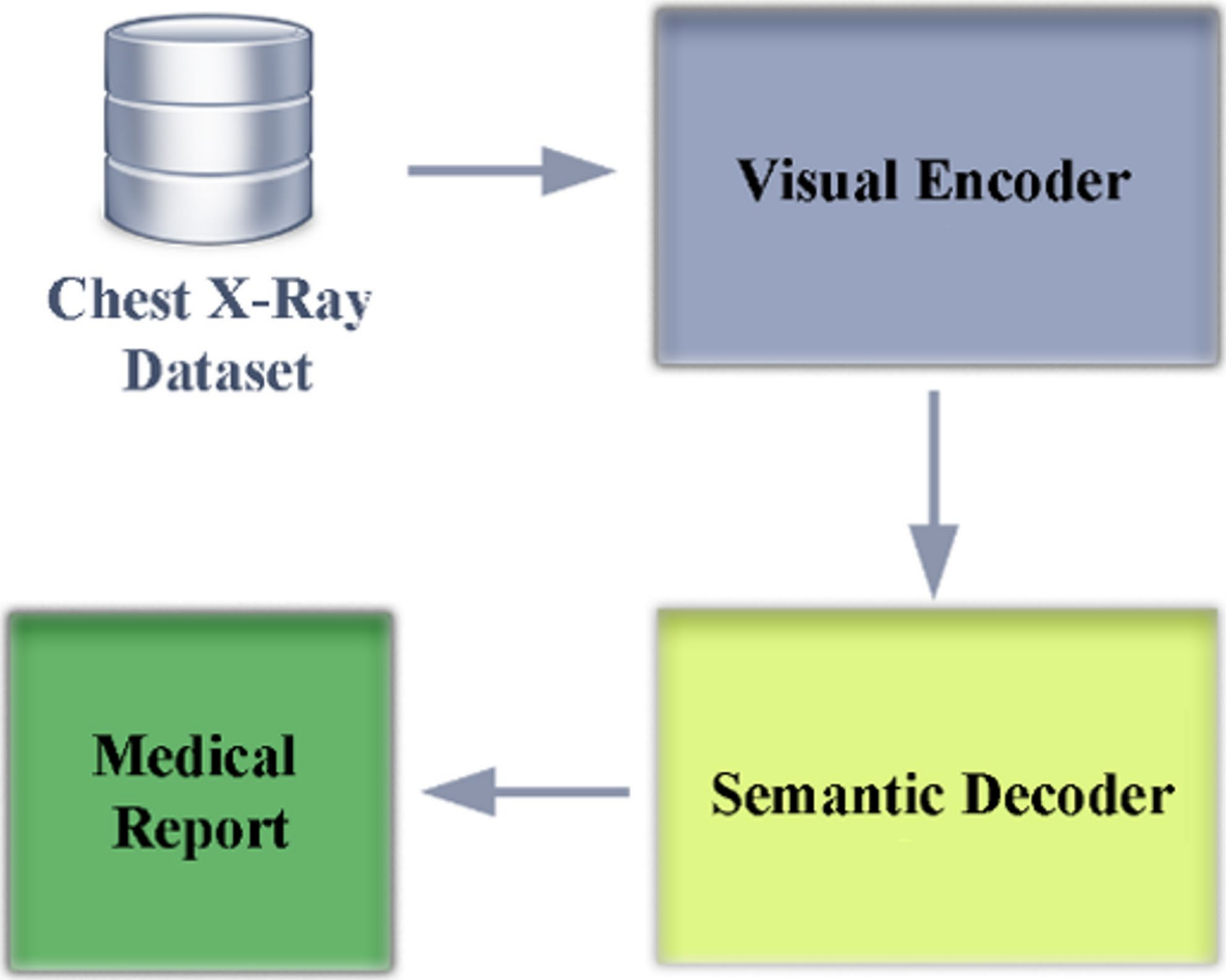


Fig. 1. General representation of the recommended medical report generation architecture.

input chest X-ray images (which have been resized to 224 × 224) and breaks them up into 16 × 16-pixel patches of the same size. The DeiT backbone uses a multi-head self-attention mechanism (with 12 heads and a head size of 64) to capture global dependencies across these patches. This is different from traditional CNN-based encoders. One important part of this architecture is hard distillation, which adds a specific distillation token to the network. This token lets the model learn from the hard labels of a teacher model, which makes it possible to learn representations quickly even with small medical datasets. The encoder's visual features and distillation tokens are what the GPT-based decoder uses as context. The decoder uses a masked self-attention mechanism to guess what the next word in the sequence will be. This makes sure that the text that is generated stays coherent. The FAST-MRG architecture autonomously synthesizes clinical findings into structured reports by combining the distilled encoder's ability to understand visuals with the GPT decoder's ability to speak fluently. This balances high performance with low computational cost.

Dataset and implementation details

Indiana University Chest X-Ray Collection dataset was used in the training, validation and testing stages of the studies[22]. The data set consists of images and associated textual reports written by medical doctors who are experts in the field. The Indiana University Chest X-ray collection consists of 6,469 chest X-ray images and pairs of paragraph-level report text associated with the images. The texts in the dataset are realistic reports written by real doctors, which is a positive advantage of the dataset. However, this situation caused the problem that the number of sentences in the reports and the number of words in the sentences were naturally not equal. This special situation of the data set was also taken into account within the scope of the proposed architecture. Image-report pairs was partitioned into three disjoint subsets using a randomized programmatic approach. First, the test set was isolated to evaluate the final performance on unseen data. Subsequently, the remaining data was split into training and validation sets. In the final configuration, the model was trained on 5,239 image-report pairs, hyperparameters were tuned using a validation set of 647 pairs, and the final performance metrics were reported

on a test set consisting of 583 pairs. This distribution ensures that the model is evaluated on a representative sample of the data while maintaining a sufficiently large corpus for the deep learning optimization process.

Both images and textual reports in the dataset were taken to some necessary pre-processing steps. The first of these is to resize the images in the data set to 224×224 and make them ready for use in the deep learning architecture. Figure 2 shows some random examples of images in the data set. As can be seen from the images in the figure, the chest X-ray images in the data set do not have a single human position. It includes frontal and side X-ray images. Additionally, the data set may include more than one image of a patient. Figure 3 shows some random examples from the medical reports in the data set.

Since the data set contains actual medical reports written by doctors, as shown in Fig. 3, they do not have a particular textual format. There are various differences between the number of sentences in the texts, the number of words in the sentence and the total number of words in the report. This situation even depends on human factors such as doctors' workload and emotional changes at the time of writing the report. This situation, which should be considered normal in terms of medical reports, is the most important factor that distinguishes the field of creating medical reports from medical images in the field of image captioning. Creating medical reports from medical images aims to create reports similar to those written by doctors, and human factors in medical reports should also be evaluated.

The Indiana University Chest X-ray dataset, which is widely used in the literature and available in open access, consists of image-report pairs. The images are stored in a folder in a format that hides patient names. The medical reports associated with the images are stored in .csv files. The relevant file contains data in columns such as image index, findings, impression, patient ID, and view position. This structure allows model training and evaluation to be performed in a consistent and repeatable manner.

The image-report pairs in the dataset used in the study are real reports written by expert doctors. This posed the problem of human elements being present in the dataset. As different doctors created the reports, there are common terminology, abbreviations or case differences in the medical reports in the dataset, similar to the real world. Therefore, some preprocessing was also performed on the texts of the medical reports in the dataset. In the first stage, common abbreviations in the medical texts were converted into longer or more formal terms, making the report language more consistent. In the dataset, "can't" was updated to "can not", "won't" to "will not" and "ll" to "will". This step was taken to make the text more similar to the way people usually speak and write in official languages. In the last step, all capitalized words in the text were converted to lower case, thus making the text case-insensitive. After that, all non-alphabetic special characters were removed from the report texts. Finally, all unnecessary spaces at the beginning and end of the report texts were removed. Only basic grammatical corrections, simplification of punctuation and capitalization corrections were applied to the report texts. However, we did not convert the medical expressions in the reports into a single standardized form. In the study, the original report structure of the dataset was preserved. This approach is important for comparing the results of the developed model with other studies in the literature.

In this study, only the Indiana University Chest X-ray dataset was used. The dataset is considered a benchmark in the literature for automatic medical report generation from chest X-rays and stands out as one of the largest and most balanced labeled datasets in this field. Additionally, since the number of open datasets for text generation from medical images is quite limited, the vast majority of researchers conduct their studies using this dataset. The Indiana University dataset, for the reasons stated, allows for scientific comparisons to be made and provides a suitable starting point for evaluating the basic performance of the developed models.

Model studies were conducted using an open-access dataset focusing solely on radiological findings obtained from chest X-ray images. The dataset does not include the patient's medical history, laboratory results, or physical examination findings. Therefore, the model analyzes only elements that can be understood from the image, such

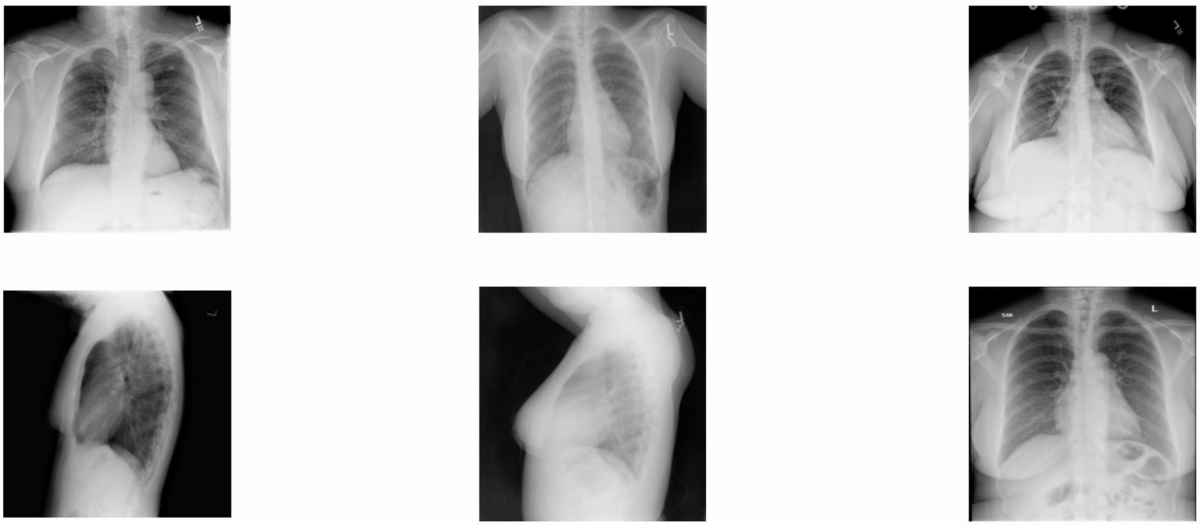


Fig. 2. Randomly selected images from the indiana university chest X-ray collection dataset.

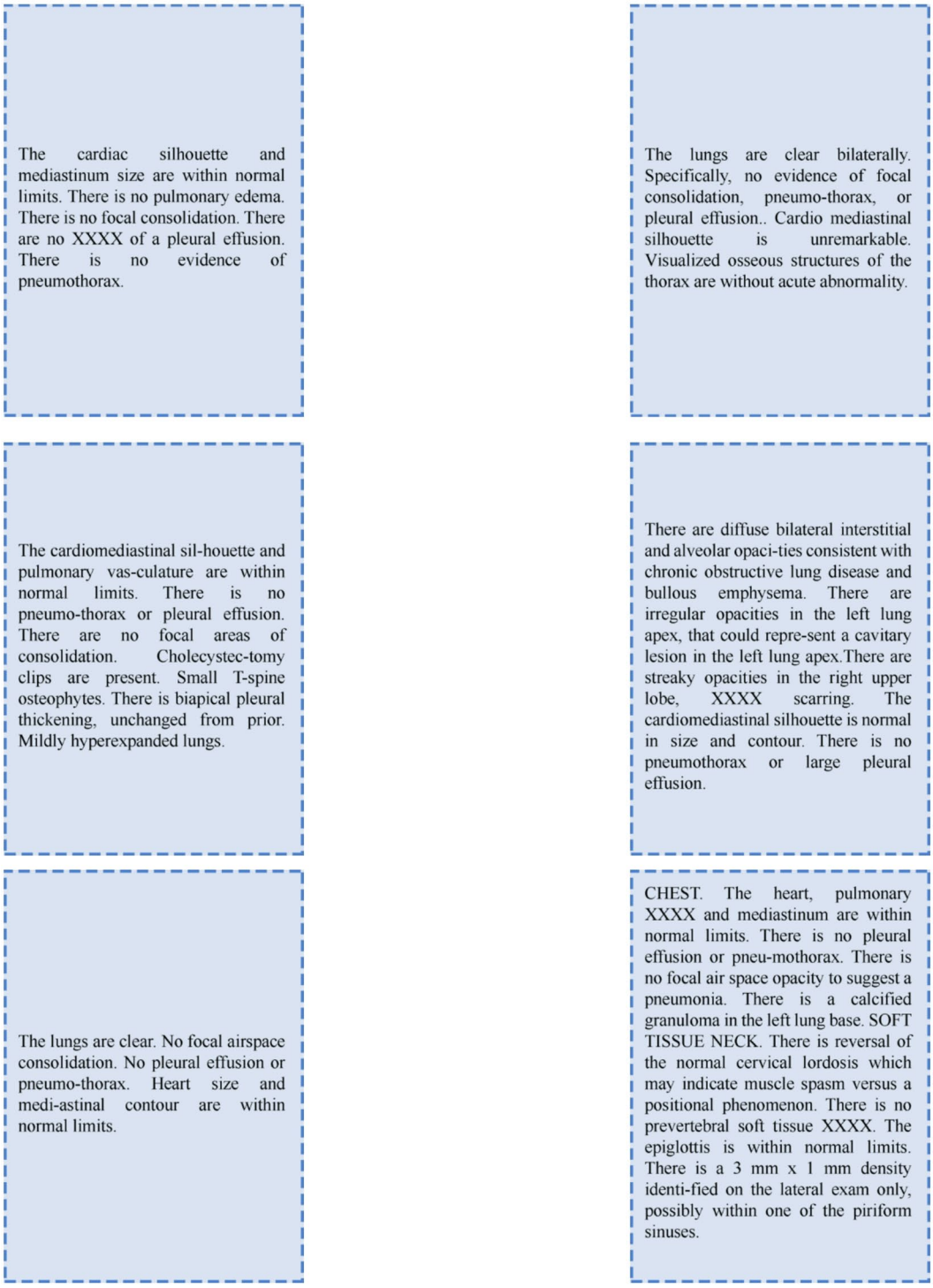


Fig. 3. Indiana university chest X-ray collection random medical reports dataset.

as congestion, increased density, heart silhouette, diaphragm, lungs, and costal structures. The model outputs aim to automate radiological interpretations within this framework.

Encoder: distilled data-efficient image transformer

Distilled Data-efficient Image Transformer (DeiT) deep learning architecture is a transformer-based architecture that aims to achieve higher success with less data[23]. It was generally developed to solve image classification and feature extraction tasks. In this respect, the architecture used is directly related to the field of study. Writing medical reports with medical images is a new field of scientific study.

Considering the publications in the literature and open source data sets, the field of creating medical reports is an area where limited data is used. The data sets currently used in the field of generating reports from medical images do not have very large sizes. The necessity of collecting both images and reports related to images during the data collection phase from the field prevents the creation of a high number of image report pairs. In this context, the DeiT deep learning model is a suitable deep learning model for the field of creating reports from medical images, thanks to its ability to achieve higher success by working with less data.

DeiT architecture has been successfully used in the literature for tasks such as image classification[24], object detection[25] and segmentation[26]. The architecture has the advantages of high success and fast operation when trained with little data. Its main disadvantage is that the architecture is more complex and has more parameters than traditional deep learning architectures. DeiT architecture can be examined in three basic subsections: body, transformer and classifier. The body uses a Convolutional Neural Network (CNN) to convert chest X-ray images into tokens when considered for the proposed study. CNN processes the image in a hierarchical manner and produces feature maps at different levels. The feature maps produced are then converted into tokens.

The second subsection of the DeiT architecture can be expressed as Transformer. In its most basic form, this section is used to learn the relationships between tokens. Transformer takes into account the relationship of each token with all other tokens by using a self-attention mechanism. Thanks to the self-attention mechanism used, the architecture is enabled to learn long-term dependencies in the visual. The third part of the DeiT architecture is called the classifier section. This stage is of critical importance, especially in studies aiming to produce word-level outputs. Softmax layer is used in the classifier section of the DeiT architecture. The Softmax function calculates the probability score for each class. However, our study focuses on extracting features from images and writing medical reports related to these features. For this reason, subsections of the DeiT architecture containing classifiers were not used. The representative architecture of the DeiT architecture is shown in Fig. 4.

As it is known, transformer-based architectures are successfully used in natural language processing applications. Researchers have applied the power that transformer architectures derive from the attention mechanism to the field of computer vision, and many sucessful new architectures have been developed. One of the important cornerstones of applications in the field of transformers and computer vision is the Vision Transformer (VIT) architecture[27]. The VIT architecture showed much better performance than traditional CNN-based architectures on the ImageNet[28] dataset, which is widely accepted in the field of computer vision. However, it owes this success to the VIT architecture, days of training with very powerful computers, and working on very large data sets. This context also expresses the necessity of the emergence of the DeiT architecture. DeiT is an efficient transformer-based architecture that can have higher success than traditional architectures. While the VIT architecture is not efficient to use in areas with a limited number of data sets, such as creating medical

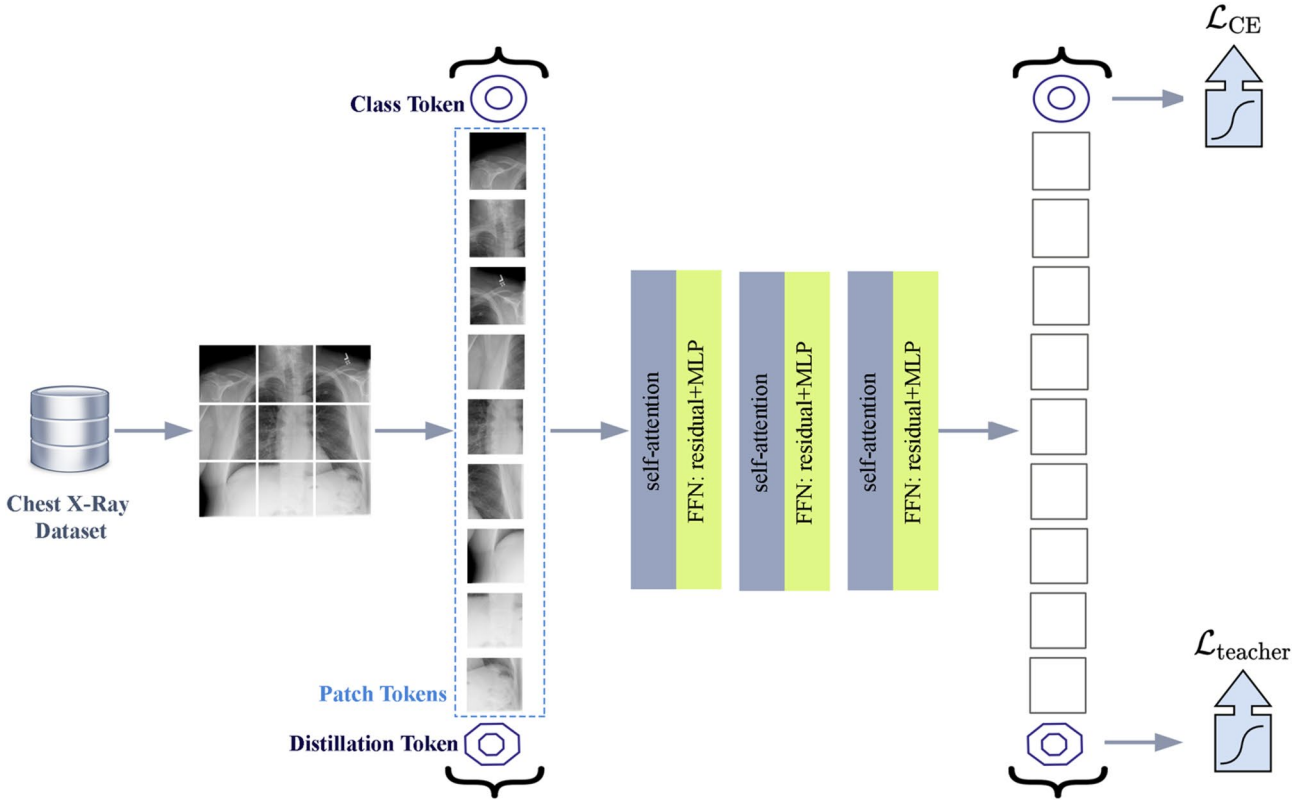


Fig. 4. The distilled data-efficient image transformer model, which forms the encoder part of our hybrid paragraph-level medical reporting architecture.

reports, the DeiT architecture can be used in the encoder part of the problem of not reporting medical images due to its existence.

The important contribution of the DeiT architecture is that it can use the transformer architecture and the concept of information distillation together. Hinton et al.[29] developed the distillation process for information in neural networks. Information distillation is a method of model compression in which a small model, also referred to as a student or teacher, is trained by imitating a previously trained larger model[30]. There are two different distillation methods called Soft and Hard. The Hard information distillation method was used in the DeiT model. Below is the general formula representation of Hard information distillation in Formula 1.

$$\mathcal{L}_{\text{global}}^{\text{HardDistill}} = \frac{1}{2}\mathcal{L}_{\text{CE}}\left(\psi\left(Z_s\right), y\right) + \frac{1}{2}\mathcal{L}_{\text{CE}}\left(\psi\left(Z_s\right), y_{\text{t}}\right) \tag{1}$$

In the above formula respectively, it expresses the concepts of ψ = Softmax, $\mathcal{L}_{\text{CE}}$ = cross-entropy loss, Z_s = logits of the student, y = ground truth, y_{t} = teacher's predicted labels. The main difference between soft distillation and hard distillation is the difference in the technique of transferring information to the student model. In soft distillation, the output of the teacher model consists of a probability distribution, while the output of the hard distillation model is considered a single correct label. Since the DeiT model is an architecture that focuses on higher success with relatively little data, hard distillation was preferred within the architecture. In addition, attention mechanism was used along with hard distillation in DeiT architecture. This has facilitated the generalization of the model, enabling higher success rates to be achieved[31].

As can be seen from Fig. 4, DeiT architecture has another important difference from ViT architecture. Distillation token has been added to the DeiT architecture separately from the ViT architecture. The distillation token is a deep learning technique that helps the student model imitate the output of the teacher model. Using distillation tokens enabled the architecture to train with less data and show higher performance. Similar to the ViT architecture, images are divided into 16×16 pieces and passed through an embedding layer to obtain fixed-size patch embeddings. The class token is passed through an encoder stack by adding image patches and distillation token location tags. The encoder stack consists of self-attention and FNN layers, which repeat three times, respectively.

To solve the problem of not having enough labeled data in medical reporting, the encoder architecture was not trained from scratch. Instead, it was given weights that had already been trained on the ImageNet dataset. The pre-training task used a hard-distillation method, in which the student model (DeiT) learns to reproduce the hard labels that a teacher convolutional neural network predicts. This means that the teacher's decisions are treated as the truth. When used with the distillation token mechanism, this method lets the model use data very efficiently. After this initialization, the whole architecture was fine-tuned on the Indiana University Chest X-ray collection to make the learned visual representations work better for making radiological reports.

Achieve the encoder's built-in lightweight feature while keeping a high level of representation, certain architectural dimensions were chosen. We cut the input images into 16×16 pixel patches and then linearly project them into an embedding dimension of 768. There are 12 heads in the self-attention mechanism, and each head has a size of 64. This setup lets the model handle long-range dependencies well in a small space, unlike traditional deep convolutional networks that need more depth to capture global features. By using these specific dimensions and the hard-distillation token instead of deeper convolutional layers, the architecture cuts down on the Floating Point Operations (FLOPs) and parameter redundancy that are common in medical image analysis models.

The encoder part of the proposed hybrid architecture is based on the Distilled Data-efficient Image Transformer (DeiT) architecture and is enhanced by the distillation technique. The multi-head self-attention blocks in the encoder process patches extracted from the input images. Similar to the standard ViT architecture, DeiT uses a multi-head self-attention mechanism to split the image into fixed-size patches. Each attention head is of size 64 and a total of 12 heads are used. Scaled dot-product attention is computed based on query (Q), key (K) and value (V) vectors obtained from learnable projection matrices, and layer normalization and residual connections are applied after each block. The unique aspect of the DeiT architecture is that it also adds a distillation token to the model to transfer information from the training model, which allows for more efficient representation learning on smaller datasets.

The hard distillation method was chosen for the proposed FAST-MRG architecture to improve the encoder's ability to extract features. When you use hard distillation, the teacher's prediction is treated as a true label (hard label), while soft distillation uses the teacher models probability distribution (soft targets) to pass on knowledge. This method is especially useful for making medical reports when there aren't many labeled data points compared to large, general datasets. The student model (DeiT) learns to copy the teacher networks sharp decisions instead of its uncertainty by using hard distillation. To make this happen, a special distillation token is added to the transformer layers next to the regular class token. The self-attention mechanism lets this token work with the other embeddings. This lets the model efficiently distill and encode complex visual patterns from the teacher model, which leads to better generalization performance at a lower cost.

Decoder: generative pre-training transformer

Another important stage in the process of creating medical reports is the decoding stage. At the encoding stage, the features obtained from the images must be interpreted meaningfully. After producing words in accordance with medical terminology, sentences must be formed by arranging the words in a logical order, and then paragraphs must be obtained by arranging the sentences in a logical order. Every stage of this process is important for medical reports to be more accurate, fluent and consistent. This stage requires the use of a strong language model. Generative Pre-training Transformer model is a frequently used language model in the

literature trained on large data sets[32]. GPT, a transformer-based architecture, can be used in many tasks such as text creation, translation, summarization and question answering. The basic architectural representation of GPT[33] architectures containing the attention layer is presented in Fig. 5.

In the first layer of the GPT architecture, words are converted into vector space. A large number of attention blocks are used within the architecture to avoid understanding the long-term dependencies of words in the language. With these attention layers, the interaction of each word with future words is modeled.

The generative pre-training component of the decoder is constructed using the GPT architecture, which is specifically designed to handle the complexity of natural language generation tasks. In this framework, the encoder visual feature maps and distillation tokens are turned into a series of input embeddings for the language model. The decoder uses a stack of masked self-attention layers, which lets the model pay attention to words that have already been generated while it predicts the next token in the report. This mechanism is very important for keeping the text's logical flow and medical accuracy. The GPT component makes sure that the final output is not just a bunch of random labels, but a structured, paragraph-level clinical narrative that helps people make decisions. It does this by modeling how each word interacts with future tokens in a vector space.

Evaluation metrics

As a result of the proposed architecture, a paragraph containing a large number of words is produced. For this reason, it is not possible to evaluate the results of the proposed architecture with metrics such as accuracy, precision and F1-score, as in classical deep learning architectures. Comparing machine-generated word groups with reports written by doctors is much more difficult and an important subject of study for researchers working in the field of natural language processing[34]. The Bleu metric put forth by Papineni et al.[35] is one of the evaluation metrics advised and most frequently used in this field. Bleu calculates with a mathematical formula how similar the translation processes made by machines are to the original text. There are 4 different metrics, Bleu-1, Bleu-2, Bleu-3 and Bleu-4, with n-gram values of (1, 0, 0, 0, 0), (0.5, 0.5, 0, 0, 0), (0.33, 0.33, 0.33, 0.33, 0) and (0.25, 0.25, 0.25, 0.25) respectively. Bleu metric was used in our study with an n-gram value of 4 to reveal the similarities of paragraph-level texts obtained from medical images with the original texts written by doctors. BLEU is a numerical evaluation metric that measures how similar the prediction reports generated by the deep learning model are to the reference text. Scores range from 0 to 1. As the BLEU score approaches 1, the similarity between the texts increases, indicating better reporting performance. Scores closer to 0 indicate lower similarity and signify poorer medical reporting performance.

The Meteor metric put forth by Banerjee et al.[36] is another significant evaluation metric that assesses text similarity. Meteor measures the similarity between reference text and machine-generated text using components such as term similarity, specificity and prevalence rather than n-gram matches. Another important metric in the literature that can evaluate machine-generated texts is the Rouge evaluation metric proposed by Lin[37]. Rouge is a recall-focused evaluation metric. It aims to determine how much of the crucial information in the reference text the automatically generated text summary covers. Thanks to this feature, the Rouge metric is frequently used in text summarization tasks.

Results

Indiana University Chest X-Ray Collection dataset was used with the proposed encoder-decoder based deep learning architecture within the scope of the research. In the original distribution of the data set, there are two separate CSV extension files called indiana_projections and indiana_reports. There are one or more frontal and side images of the patients. Some studies in the literature have tried to obtain a single medical report by using

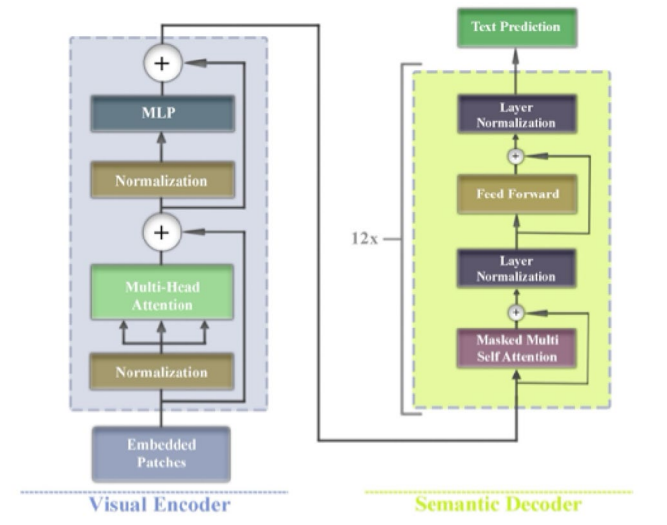


Fig. 5. Generative pretraining transformer model, which forms the decoder part of our hybrid paragraph-level medical reporting architecture.

Hyperparameter	Value
Learning rate	5×10^{-5}
Regularization strength for AdamW-based optimization	1×10^{-2}
Epoch	10
Train batch size	8
Test and validation bath size	8

Table 1. Overview of hyperparameters utilized in deep learning architectures for optimizing model performance and training efficiency.

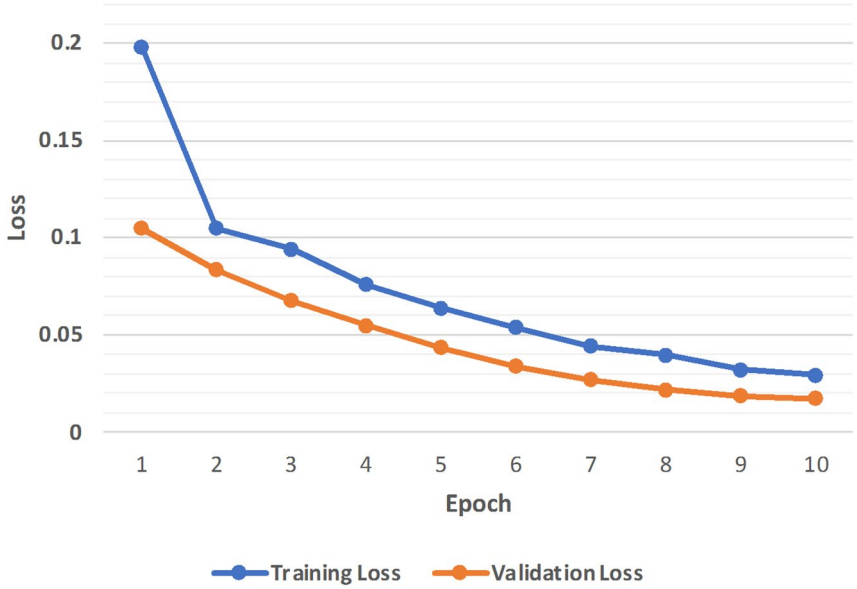


Fig. 6. Model training and validation loss graph of the proposed hybrid encoder-decoder architecture.

front and side x-ray images simultaneously. In studies with this approach, two images taken from the front and the side were used simultaneously in the training phase, and similarly, a single report writing method was used based on two images in the testing phase. This situation makes the current problem seem easier than it is and is far from the real-life problem approach. In our study, the approach of matching a single image with one report was taken as basis. Similarly, during the testing phase, tests and analyses were performed on writing a medical report using one image.

The dataset used in the research has been divided into subgroups for the training, validation, and test phases. As mentioned above, the concept of obtaining a medical report from a single image has been adopted. Thanks to this method, the dataset could be used more efficiently and comprehensively. The dataset contains a total of 6,469 pairs of chest X-ray images and reports. The dataset has been primarily divided into 10% test and 90% training-validation. Then, the training-validation subset is divided into 10% validation and 90% training. In the final state, there are 5239, 583, and 647 image-report pairs in the training, test, and validation subsets, respectively. This partitioning process was done randomly using Python codes without any human intervention. While determining the percentage split ratios of the training, validation, and test subsets, previous studies in the literature were also used. In order to make the most accurate comparisons and determine the current position of the study in the literature, a similar percentage division to previous studies has been used.

Reproducibility of the proposed FAST-MRG architecture, strictly defined hyperparameters were utilized throughout the experiments conducted on a P100 GPU. The model optimization was performed using the AdamW optimizer, chosen for its effective weight decay handling, with a regularization strength set to 0.01. The learning rate was fixed at 5×10^{-5} to ensure stable convergence without the use of a dynamic scheduler. Due to memory constraints and to maintain gradient stability, a batch size of 8 was consistently applied across training, validation, and testing phases. The entire training process was completed in 10 epochs, which was sufficient for the loss values to converge without overfitting, as evidenced by the loss graphs. Important parameters used in the training and testing processes are given in Table 1 below.

Training and validation loss graphs were drawn to understand how well the developed encoder-decoder-based deep learning architecture learned and its generalization performance. The training and validation loss graphs of our study are given in Fig. 6. In deep learning architectures, the error of the model during training is observed by examining the training loss graph. If the training loss is too high, it means that the model does not learn well, and if it is too low, it means that there may be an overfitting problem. In the verification loss graph,

the error of the model in the verification phase is examined. Validation loss, which is a much more important parameter than training loss, is evaluated based on the training loss graph. The fact that the validation loss graph is close to the training loss graph indicates that the deep learning model is progressing successfully. The fact that the validation loss graph is higher than the training loss graph indicates that the model has an overfitting problem. For these reasons, training by selecting appropriate hyperparameters is important in terms of model performance and adaptation to real-world problems. It can be seen that the training loss and validation loss data are decreasing in line with the expected results. The fact that the validation loss data does not exceed the training loss data shows that the model does not have an overfitting problem.

Quantitative analysis

The performance of the proposed architecture was also analyzed by comparing it with other studies in the literature using the same data set. Table 2 includes nine studies in the literature that use similar data sets and the metric measurements of the FAST-MRG architecture we proposed within the scope of the study. Bleu-1, Bleu-2, Bleu-3, Bleu-4, Meteor and Rouge evaluation metrics are important metrics frequently used in text matching studies. In other words, these are evaluation metrics that reveal, with different mathematical formulas, how well the chest x-ray reports obtained autonomously match the reports written by real doctors who are experts in the field during the creation of the data set.

The results stated in Table 2 clearly show that the proposed study achieved higher report writing success than previous studies in the literature. When the Bleu-1 evaluation metric is considered primarily, it is observed that only one study out of the nine other studies examined in the literature achieved a higher success. It can be seen that the CDGPT2 architecture outperforms the proposed FAST-MRG architecture by approximately 3 percent, achieving a success rate of 0.387. The Blue-1 metric shows how many of the words generated by the model appear in texts written by doctors. In other words, the metric makes a comparison at the word level. It is important to obtain high results at the word level, but it is also appropriate for a medical report to create similar sentences by choosing other synonymous words without breaking away from the context. For this reason, a performance decrease of approximately 3 percent can be ignored.

Blue-2 metric is an evaluation metric that determines how well groups of two words in the autonomously produced report match the word phrases in the reports written by doctors. The FAST-MRG architecture we proposed achieved the highest level of success in the Blue-2 evaluation metric. The proposed architecture achieved approximately 25 percent higher success than its closest competitors, NLG and HLSTM + att + Dual architectures. The Blue-3 and Blue-4 evaluation metrics compare autonomous reports and reports written by doctors using 3 and 4 word groups, respectively.

In this context, as the index of the blue metric increases, it can be said that it is a metric that aims to match at the sentence level from the word level. The FAST-MRG architecture proposed in this study achieved the highest level of success in the Blue-3 metric. It achieved approximately 48 percent higher success compared to its closest competitor. In the Blue-4 metric, our proposed FAST-MRG architecture achieved the highest success, scoring approximately 80 percent better than its closest competitor.

The Meteor metric is an evaluation metric that takes word order into account, unlike the Blue metrics. The Meteor metric is a natural language processing evaluation metric used to determine the fluency and naturalness of text. The proposed FAST-MRG architecture achieves a meteor score 32 percent higher than its closest competitor, ATT-RK. This achievement is also a measure of how much more fluent and natural the reports produced as a result of our study are than the closest competitor. The Rouge Metric is another evaluation metric that helps us understand how well the autonomously generated text and the actual reports in the dataset match. There are different variations of the Rouge metric such as Rouge-L, Rouge-N, Rouge-S. The FAST-MRG

	Word-overlap metrics					
Method	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**Meteor**	**Rouge-L**
CNN-RNN[19]	0.316	0.211	0.140	0.095	0.159	0.267
LRCN[38]	0.369	0.229	0.149	0.099	0.155	0.278
ATT-RK[18]	0.369	0.226	0.151	0.108	0.171	0.323
RTMIC[39]	0.350	0.234	0.143	0.096	–	–
VSGRU[15]	0.347	0.221	0.156	0.116	0.150	0.251
CDGPT2[15]	0.387	0.245	0.166	0.111	0.164	0.289
HLSTM + att + Dual[14]	0.373	0.246	0.175	0.126	0.163	0.315
NLG[13]	0.369	0.246	0.171	0.115	–	0.359
TieNet[12]	0.286	0.159	0.103	0.073	0.107	0.226
Gamma Enhancement[40]	0.363	0.371	0.388	0.412	–	–
Vi-Ba[41]	–	–	–	0.150	–	0.274
MERGIS[42]	0.296	–	–	–	0.128	0.335
META-CXR[43]	0.408	0.221	0.198	0.121	0.154	0.287
(Our) FAST-MRG	0.373	0.307	0.260	0.227	0.226	0.332

Table 2. Quantitative comparison of FAST-MRG architecture with previous studies using word-overlap evaluation metrics.

architecture proposed in this study achieved scores of 0.360, 0.195, 0.310, 0.343 in Rouge1, Rouge2, RougeL and RougeLsum evaluation metrics, respectively. Unfortunately, the studies in the literature are not very clear about which Rouge metric is used. However, it is stated that Rouge-L metric is used in some of the studies. For this reason, we have added the score data for the RougeL metric to the comparison table of our study. For example, in the related study where NLG architecture is proposed, it is not clearly stated which Rouge metric is used. Assuming that the Rouge-L metric is used, the NLG architecture scores about 8 percent higher than our proposed FAST-MRG architecture. However, in other metrics, our proposed FAST-MRG architecture achieved higher results than the NLG architecture.

In order to understand the particular impact of the hard distillation strategy utilized in our encoder, it is crucial to juxtapose the FAST-MRG architecture with conventional transformer-based methodologies that do not incorporate this mechanism. In our comparative analysis (Table 2), the Vi-Ba model exemplifies a baseline for a no-distillation scenario, employing a standard Vision Transformer (ViT) on the encoder side. It's clear that FAST-MRG is better than Vi-Ba on key metrics, with a 51% improvement in BLEU-4 and a 21% improvement in ROUGE-L. This difference shows that standard ViTs have a hard time learning strong features from the small Indiana University dataset because there isn't enough data. However, the distillation token in our architecture makes up for this by moving inductive biases from the teacher network. The fact that FAST-MRG works better than Vi-Ba shows that the distillation technique is needed for making medical reports when there isn't enough training data.

BLEU-1 evaluates the model's overall text generation success from a limited perspective because it only considers single-word matches. In contrast, BLEU-2 and higher metrics provide a more meaningful evaluation because they take into account the matching of multiple word groups with the reference text. In this context, the high performance of the FAST-MRG architecture we proposed within the scope of our study in BLEU -2, BLEU -3, and BLEU -4 scores demonstrates the model's capacity to generate word groups accurately and consistently. Therefore, despite the limited differences in BLEU -1, the overall results indicate that the model successfully generates text.

Time performance analysis

As explained in the previous sections of the paper, the DeiT deep learning architecture is used on the encoder side of the proposed architecture. DeiT architecture is a time-efficient architecture that can achieve higher success rates with less data. Again, as explained in the previous sections, P100 GPU is used for training and testing of the proposed architecture. One of the main reasons for choosing this GPU is to make the most accurate temporal comparison with other studies in the literature that use the same GPU. Alfarghaly et al.[15] used P100 GPU in their proposed VSGRU and CDGPT2 architectures. They used CheXNet on the encoder side of their architecture[16]. The training time using the original version of CheXNet and a single P100 GPU is approximately 22 h[44]. The training time using CheXNet with reduced data and dual P100 GPUs is given as approximately 6 h[45]. On the decoder side of the VSGRU architecture, the researchers reported that 4400 s of time was spent. Similarly, on the decoder side of the CDGPT2 architecture, 3200 s were reported. The FAST-MRG architecture proposed in our study has a total runtime of 28,391 s on the encoder-decoder side. Figure 7 shows the time comparison of our architecture graphically. As can be seen from the graph, our proposed FAST-MRG architecture is about 66 percent more time efficient than other architectures in the literature trained using the same GPU.

The lightweight nature of the FAST-MRG architecture is not merely theoretical but is demonstrated through measurable runtime metrics compared to baseline models. By replacing computationally intensive CNN-based encoders (such as CheXNet used in VSGRU and CDGPT2) with the Distilled Data-efficient Image Transformer (DeiT), the model significantly reduces computational overhead. Experimental results obtained under identical hardware conditions (Tesla P100 GPU) reveal that FAST-MRG requires a total runtime of only 28,391 s. This

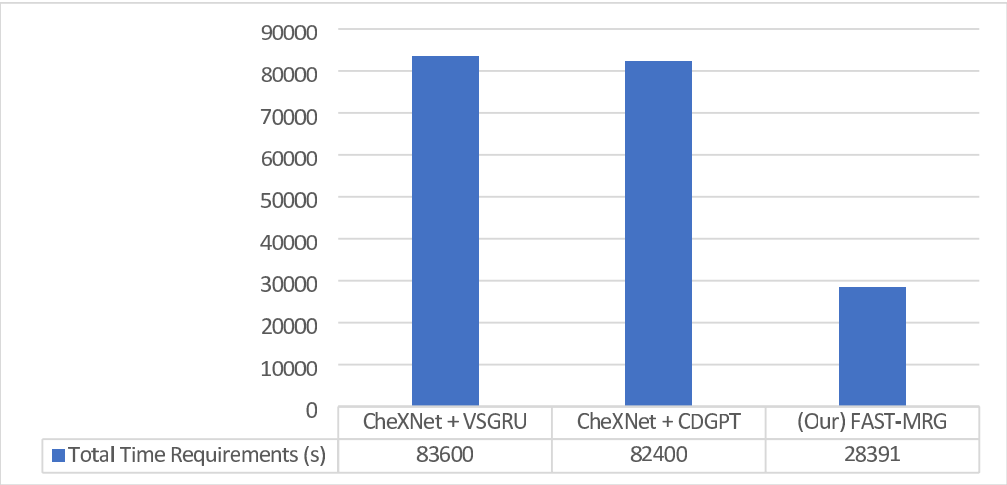


Fig. 7. Total time requirement analysis graph of the proposed architecture and other architectures in the literature trained using the same GPU.

stands in sharp contrast to the extended processing times of baseline architectures, resulting in a measurable 66% increase in time efficiency. These figures confirm that the hybrid design successfully balances high-performance reporting with a lightweight computational footprint, making it suitable for resource-constrained clinical environments.

It is important to note that a direct and standardized temporal comparison in this domain is inherently difficult, as the vast majority of existing studies focus exclusively on accuracy metrics and do not report training durations or computational costs. Due to this scarcity of data in the literature, the comparative baseline presented in Fig. 7 was derived by aggregating the partial component times (separate encoder and decoder durations) reported in related works using similar hardware. Therefore, this analysis should be interpreted as an estimated benchmarking effort intended to highlight the relative computational efficiency and lightweight nature of the proposed DeiT-based architecture, rather than a precise chronometric competition against models that do not document their temporal footprint.

Discussion

Thanks to advancing technology and advancing medical investments, medical imaging techniques are successfully used in almost all hospitals, including first level local medical centers. Unfortunately, there are not enough specialized doctors available to interpret medical images. Considering the time off for humanitarian reasons and the busy working hours of specialists, it is essential to have autonomous medical reporting systems to be used in emergencies. In addition, an autonomous system that can support specialist doctors in decision-making can minimize the error rate in the diagnosis of diseases and make significant contributions to save more lives. Classification-based studies in the literature are far from providing details by only indicating the presence or absence of the disease. In this context, our study aims to autonomously generate paragraph-level medical reports from chest X-ray images.

This study successfully fulfilled the research questions posed in the introduction through the proposed FAST-MRG methodology and the obtained experimental results. Addressing RQ1, the study demonstrated that the hybrid architecture, utilizing DeiT for feature extraction and GPT for text generation, effectively produces paragraph-level reports similar to those written by doctors. As shown in the observational analysis (Fig. 8), the model successfully generated accurate findings for "Normal" and "Accurate" cases, capturing the semantic context of the medical images. The density plots further confirm that the generated texts maintain a consistent quality distribution across the test set, avoiding zero-density failures often seen in weaker models. Addressing RQ2, the proposed lightweight architecture directly addressed the efficiency question. By employing the distilled data-efficient image transformer (DeiT), which is designed to perform well with less data, and a streamlined GPT decoder, the FAST-MRG model achieved a 66% improvement in time efficiency compared to previous architectures (VSGRU, CDGPT2) trained on similar P100 GPU environments. This confirms that high performance can be achieved with significantly lower computational costs. Addressing RQ3, the quantitative analysis provided a clear answer regarding the model's comparative performance. FAST-MRG outperformed several state-of-the-art methods, particularly in higher-order n-gram metrics. It achieved the highest scores in BLEU-2 (0.307), BLEU-3 (0.260), and BLEU-4 (0.227), surpassing the closest competitors by margins of approximately 25% to 80%. Additionally, the METEOR score of 0.226 indicates superior fluency and naturalness in the generated reports compared to studies like ATT-RK. These results validate the proposed method's superiority in generating linguistically and clinically coherent text.

The study is based on an encoder-decoder architecture. On the encoder side, the transformer-based DeiT architecture is used, which is a new approach known for achieving higher performance with less data. On the decoder side, GPT infrastructure, a popular architecture for natural language processing applications, is used. The performance metrics analyzed clearly show that the proposed architecture achieves high performance. However, the human eye checks are as valuable as the numerical results produced by the evaluation metrics. Figure 8 shows the four chest X-ray images selected from the test subset of the dataset and the ground truth reports of the images written by real doctors. Additionally, there is sharing of the prediction texts produced by the proposed FAST-MRG architecture.

The FAST-MRG architecture proposed in this study is a hybrid encoder–decoder approach. The hybrid concept is an approach that aims to achieve superior performance at both the visual and linguistic levels by combining different powerful model structures. In the encoder section of the model, a transformer architecture enriched with distillation is used to extract meaningful features from images. On the decoder side, a GPT-based language model is used to generate natural and meaningful reports at the paragraph level. Unlike traditional methods, which only produce limited labels from visual features, this structure provides much more flexible and contextually rich text generation. Unlike traditional systems, the proposed hybrid approach combines information extraction power and text generation capabilities under one roof. While the encoder section handles the task of extracting information from images, the decoder section takes on the task of writing more consistent and meaningful reports.

The results of medical reports produced autonomously by the deep learning architecture can actually be analyzed in four different subcategories. The first one is the image labeled as normal, that is, without any disease. It can be seen that the proposed FAST-MRG architecture produces a close report for a chest X-ray image of an individual without any disease. The change in word order does not cause any shift in meaning. The expert doctors who supported the study noted that both reports were equal in the human observation technique. The example shared in the Accurate section of Fig. 8 is actually a great opportunity to re-explain a problem in the dataset within the study. Some words in the dataset that were thought to be subject to ethics committee restrictions were labeled as XXXX. As in other studies in the literature, we used medical reports for training and testing without compromising the originality of the dataset. Therefore, making sense of these words poses a particular challenge for our autonomous decision-support mechanism. Nevertheless, our FAST-MRG

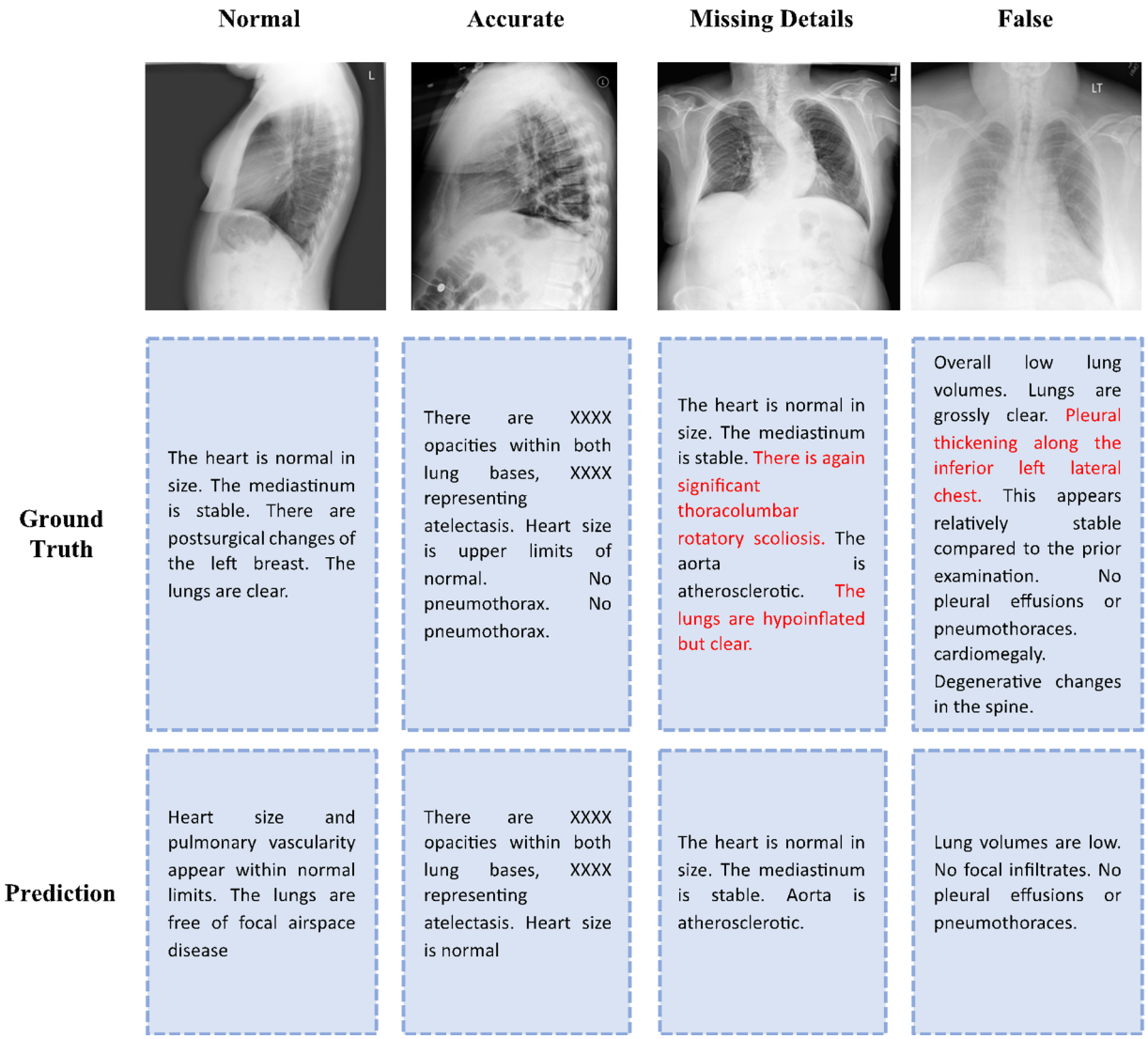


Fig. 8. Example medical reports autonomously generated by our FAST-MRG deep learning architecture from test images.

architecture was able to successfully predict even the report associated with a chest X-ray image containing the hidden words. In the same section, the sentence "No pneumothorax" repeated in the Ground Truth text clearly shows another problem with the dataset. The dataset contains reports written by real doctors. This also shows that the texts in the dataset contain human spelling and grammatical errors. Nevertheless, the report produced by our FAST-MRG architecture is clearly understandable and does not contain any erroneous details that could lead to incorrect treatment.

In the example shared in the Missing Details section in Fig. 8, we observe that the architecture still generates an incomplete report even though it has the correct parts. Unfortunately, the findings of the scoliosis disease could not be obtained from the image and as a result, the report could not be written for the relevant disease in the report. When the dataset was examined in detail specifically for this disease, it was seen that there were very few images of the disease. In fact, the dataset is a dataset focusing on chest X-ray diseases. Scoliosis is a disease that includes spinal curvature findings in the skeletal muscle system. If the owners of the dataset do not evaluate the findings of scoliosis at all or increase the number of images in the dataset to an acceptable sample level, such problems can be avoided. In the example given in the part of Fig. 8 labeled False, the presence of pleural thickening could not be detected and was not included in the report content.

The quantitative evaluation in this study uses standard n-gram metrics (BLEU, METEOR, ROUGE) to make sure it can be compared to the baseline studies in Table 2. However, we know that these metrics mostly measure lexical overlap and not clinical accuracy. To address this limitation and guarantee that the lightweight design of the FAST-MRG model does not hinder the identification of significant findings, we conducted a thorough qualitative analysis with the involvement of expert physicians. The analysis of Fig. 8 shows that this human evaluation was based on the clinical validity of the generated reports, putting them into three groups: Normal,

Accurate, Missing Details, and False. This was done based on medical correctness, not just word matching. This expert review serves as a proxy for clinical efficacy.

Another important representation in which the success of the proposed architecture can be analyzed is density graphs. In the literature, the results of many studies are given only as an average with evaluation metrics. Although it is correct to share the results as an average, it is not sufficient alone to understand the success of the architecture. Sharing only the average data makes it impossible to understand the series that make up the average. The density map shows the distribution of the data as individual results. Figure 9 shows the density maps obtained with the test data of our FAST-MRG architecture.

In this study, some of the model's failed outputs have been presented with examples. However, the systematic classification of error types is of great importance, especially in terms of clinical applications. With the contributions of the expert doctors involved in the project, the model outputs have been examined in four main categories: Normal, Accurate, missing details and false. This classification method has been useful in determining the types of errors the model is more prone to. However, this detailed analysis has been summarized to avoid exceeding the scope of the current study. In future studies, more in-depth analyses for each category are planned to be conducted.

An error analysis reveals that the model demonstrates substandard performance in the scoliosis class. The primary rationale for this phenomenon is the paucity of reports of scoliosis in the dataset. Furthermore, scoliosis is fundamentally a spinal disorder, rather than a lung or chest disease. Consequently, it serves as a remote exemplar of chest X-ray-based disease classification and reporting, a focal point of this study. Such classes should be treated with caution in model evaluations due to both their low representation and limited

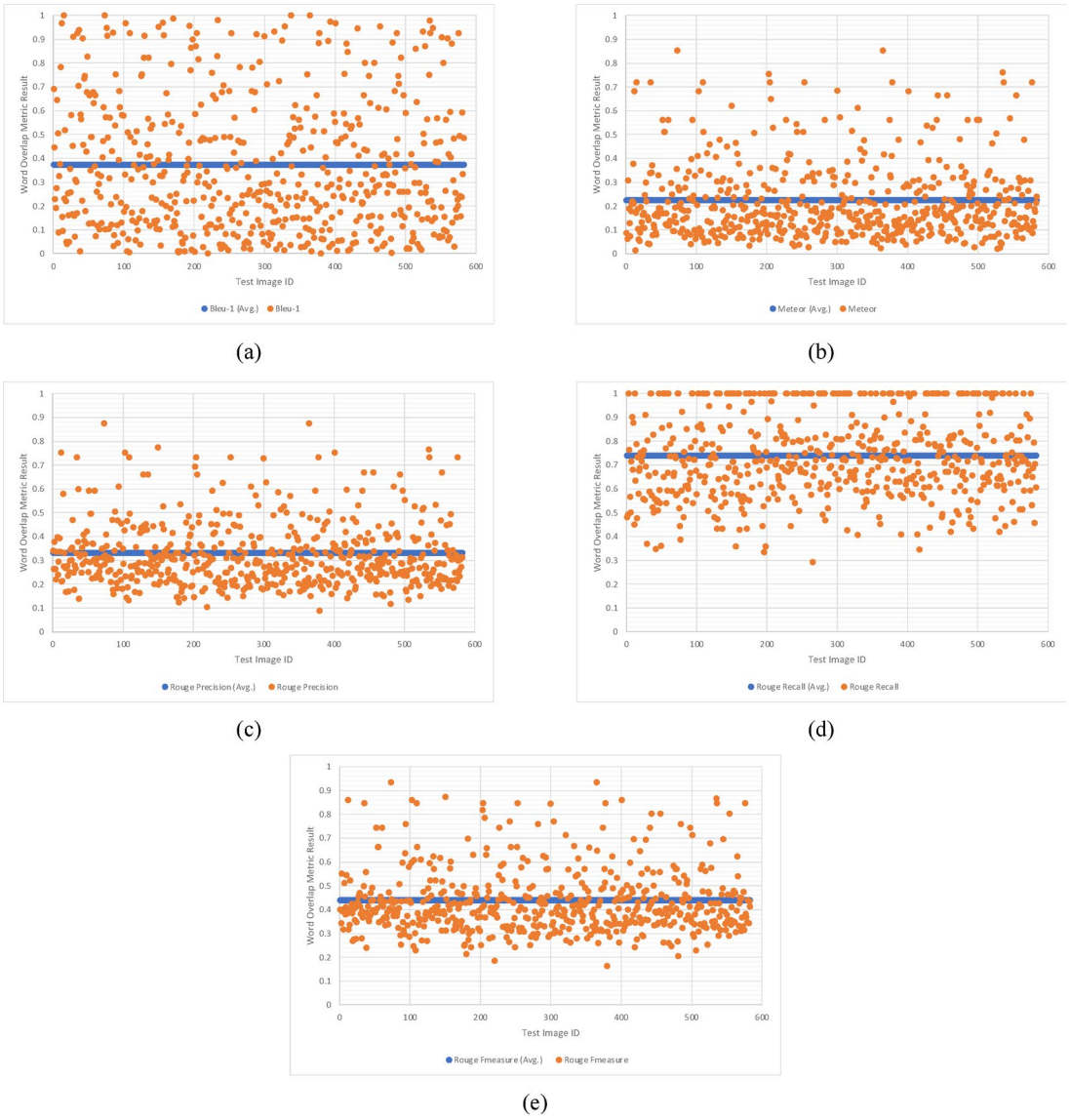


Fig. 9. Density Scatter Plot of Proposed FAST-MRG Architecture test results for different evaluation metrics **a** Bleu-1, **b** Meteor, **c** Rouge precision, **d** Rouge recall, **e** Rouge fmeasure.

Word-overlap evaluation metrics	Sample standart deviation	Confidence level (%)	Confidence interval
BLEU-1	0.4243	95	0.2663–0.6136
BLEU-2	0.3728	95	0.2183–0.5335
BLEU-3	0.3379	95	0.1901–0.4862
BLEU-4	0.3145	95	0.1733–0.4576
Meteor	0.1569	95	0.1433–0.1701
Rouge	0.1339	95	0.1223–0.1447

Table 3. Word-overlap evaluation metrics with bootstrapped confidence intervals based on test results.

relevance. Therefore, instead of directly evaluating performance for low-frequency and irrelevant classes such as scoliosis, we focused on measuring overall model performance over more meaningful classes. There is also an important reason why we did not eliminate image-report pairs for this disease during training and testing. If we had not used patients with scoliosis findings in the training and testing process, we would have lost the ability to compare with similar studies in the literature and the reliability of our comparison tables would have decreased. For this reason, we thought it appropriate to carry out a method that we can compare with the literature in the current situation.

The test subset of the dataset consists of 583 images. The x-axes of the 5 graphs in Fig. 9 are the id values of these images in the test dataset. The orange circles show the score obtained in the relevant evaluation metric as a result of the comparison of autonomously generated medical reports and ground truth texts. The BLEU lines show the average score on the respective evaluation metric.

From all five of the density plots in Fig. 9, it can be observed that the density of the performance metric shows a significant distribution. It is clearly observed that there is no density at or near zero. In addition, the dataset is completely randomly and program-matically divided into training, validation and test subsets. When the intensity distri-bution of the test results is examined, it is observed that there is no similar score density in successive images, which is another proof of the successful distribution. Bleu-1 is a score type that is calculated by evaluating single word matches with reference and source texts. For this reason, it is natural that the density in this metric is distributed along the lower and upper boundaries. When the graph of the Meteor metric is analyzed, it is seen that there is a distribution close to the average and the scores deviating from the average do not show continuity. The Rouge evaluation metric has three different subgroups. Precision is calculated based on the rate at which the phrases found in the autonomously written report are also found in the reference text. Recall is calculated based on the rate at which the phrases found in the reference text are found in the autonomously written reports. F-Measure is a combination of Precision and Recall metrics. A high F-Measure value indicates that the report is both comprehensive and faithful to the reference text. When the graphs of all three rouge metrics are analyzed, the absence of values close to the zero boundary shows that the architecture achieves successful results on the entire test dataset. Similarly, it can be observed that the results of the performance metrics are close to the average. This is valuable as research is being conducted on a deep learning architecture that can directly support doctors in decision making.

When the test results are close to the average, this is ideal for medical report generation. The least desirable situation is when some results are predicted very well and some are predicted very poorly. Because even if the average of the developed architecture is good, there is a risk of incorrect or incomplete treatment for patients with very poor predictions.

In order to establish a more solid foundation for the significance of the evaluation metrics, a 95% confidence interval was calculated using the bootstrap method on the test results. Thus, it has been more clearly demonstrated whether the improvements offered by the model compared to previous studies are random or statistically significant. In this context, the standard deviation value and the corresponding confidence interval for each word-overlap-based evaluation metric are provided in Table 3. This analysis allows for a better understanding of the variation in text similarity measurements and provides a clearer determination of whether the results are random.

Within the proposed FAST-MRG architecture, a transparent and comprehensive evaluation beyond simple average metrics was provided, enabling a multidimensional analysis of the model's limitations and reliability. As shown in Fig. 8, error cases categorized as Missing Details and Incorrect Predictions have been clearly analyzed. These examples highlight that, although the model excels at detecting primary lung opacities, it may struggle with rare skeletal findings or subtle tissue anomalies due to dataset imbalances. To visualize how these errors are distributed across the entire population rather than being isolated examples, we created intensity distribution graphs for the full test set of 583 images (Fig. 9). These graphs confirm that the errors are not systematic across all data, revealing that the density clusters around the high-performance average, despite outliers with lower scores (representing the challenging cases shown in Fig. 8). Finally, to ensure the statistical reliability of these results, we calculated 95% confidence intervals using the bootstrap method (Table 3). This statistical rigor confirms that the improvements provided by the model are significant and stable within the [0.1733–0.4576] bounds calculated for BLEU-4, thereby robustly validating the consistency of the architecture.

The test set size of 583 images is a strong statistical sample, but the confidence intervals in Table 3 (for example, BLEU-1: 0.2663–0.6136) show a wide range. The main reason for this width is the high standard deviation in the test results (0.4243 for BLEU-1). The reason for this difference is that the ground truth data is not structured. As mentioned in Section "Dataset and implementation details", the reports rely heavily on human factors and do not have a standardized linguistic template. Also, the density plot analysis in Fig. 9 shows that BLEU-1 scores tend

to be more spread out at the lower and upper ends than in the middle, which naturally makes the confidence interval wider. Our study, give a clear benchmark that takes into account the natural linguistic diversity of real-world medical reporting by reporting these intervals without artificially narrowing them.

While the proposed FAST-MRG architecture demonstrates high performance in generating medical reports, its implementation in clinical practice entails significant ethical responsibilities and risks. As detailed in the density analysis (Fig. 9), variability in model predictions exists, and poor predictions pose a direct risk of leading to incorrect or incomplete treatment. Furthermore, as observed in the error analysis, the model may occasionally omit rare findings (e.g., scoliosis) or fail to detect subtle pathologies due to dataset imbalances. Therefore, this system is intended solely as a Clinical Decision Support System (CDSS). It must be operated under a human-in-the-loop framework where a specialist physician reviews and validates every generated report before it affects patient care. Since this study was conducted in a laboratory setting without real-time clinical trials, direct autonomous deployment without expert supervision is currently not recommended. Future work will focus on integrating uncertainty estimation mechanisms to alert physicians when the model's confidence is low, thereby enhancing patient safety.

Conclusions

The main problem addressed in our study is the autonomous generation of fast and reliable paragraph-level medical reports from chest X-ray images using deep learning techniques. A hybrid encoder-decoder model was developed in this study. Our deep learning model uses the transformer-based DeiT architecture on the encoder side and the GPT architecture on the decoder side. Compared to similar studies in the literature, it was concluded that the developed architecture achieved higher success in a shorter training time. The results obtained in the study were analyzed in detail using different methods such as quantitative analysis, time analysis, statistical confidence analysis, and qualitative analysis. The proposed FAST-MRG architecture achieved word alignment evaluation metric scores of 0.373, 0.226, and 0.332 for BLEU-1, Meteor, and Rouge evaluation metrics, respectively.

The autonomous generation of medical reports is a critical area of work that can directly impact human life by preventing incorrect and incomplete diagnoses. The high success rate achieved by this study could serve as a guide for the use of this architecture in real-life problems. The FAST-MRG architecture has a much shorter processing time compared to similar studies in the literature. Additionally, the architecture is efficient in training and can achieve higher success rates with less data. FAST-MRG achieved higher performance despite only a single test image being input. In this context, the proposed architecture could guide new studies that could be used as a decision support system in hospitals in real-life settings. The limitations of the study are defined as diseases that can be detected from chest X-ray images and related images. Another important limitation of the study is that the results were obtained by conducting experiments in a laboratory setting; real-time tests were not performed. The developed model has the potential to be applied to other medical imaging fields, such as gastroenterology, MRI, CT, and all other diseases related to images, in future studies.

Data availability

The datasets used and analysed during the current study are available from the corresponding author upon reasonable request. The source code related to the proposed method will be made publicly available upon publication and is available from the corresponding author upon reasonable request during the review process.

Received: 6 August 2025; Accepted: 16 February 2026
Published online: 11 March 2026

References

1. Sobahi, N., Sengur, A., Tan, R.-S. & Acharya, U. R. Attention-based 3D CNN with residual connections for efficient ECG-based COVID-19 detection. *Comput. Biol. Med.* **143**, 105335. https://doi.org/10.1016/j.compbiomed.2022.105335 (2022).
2. Bütün, E., Uçan, M. & Kaya, M. Automatic detection of cancer metastasis in lymph node using deep learning. *Biomed. Signal Process. Control* **82**, 104564. https://doi.org/10.1016/j.bspc.2022.104564 (2023).
3. Cinar, N., Ozcan, A. & Kaya, M. A hybrid DenseNet121-UNet model for brain tumor segmentation from MR images. *Biomed. Signal Process. Control* **76**, 103647. https://doi.org/10.1016/j.bspc.2022.103647 (2022).
4. Goutham, V., Sameerunnisa, A., Babu, S., Prakash, T. B. Brain tumor classification using EfficientNet-B0 model, in *Proceedings of the 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Vol. 3, 2503–2509 (IEEE, 2022)
5. Chen, B., Zhang, Z., Lin, J., Chen, Y. & Lu, G. Two-stream collaborative network for multi-label chest X-ray image classification with lung segmentation. *Pattern Recognit. Lett.* **135**, 221–227. https://doi.org/10.1016/j.patrec.2020.04.016 (2020).
6. UCan, M., Kaya, B., Kaya, M. Multi-class gastrointestinal images classification using EfficientNet-B0 CNN model, in *Proceedings of the 2022 International Conference on Data Analytics for Business and Industry (ICDABI)* 1–5 (2022).
7. Cinar, N., Kaya, M. & Kaya, B. A novel convolutional neural network-based approach for brain tumor classification using magnetic resonance images. *Int. J. Imaging Syst. Technol.* **33**, 895–908. https://doi.org/10.1002/ima.22839 (2023).
8. UCAn, M., Kaya, B., Kaya, M. Comparison of deep learning models for body cavity fluid cytology images classification, in *Proceedings of the 2022 International Conference on Data Analytics for Business and Industry (ICDABI)* 151–155 (IEEE, 2022).
9. Ibrahim, A. U., Ozsoz, M., Serte, S., Al-Turjman, F. & Yakoi, P. S. Pneumonia classification using deep learning from chest X-ray images during COVID-19. *Cognit. Comput.* https://doi.org/10.1007/s12559-020-09787-5 (2021).
10. Karaddi, S. H. & Sharma, L. D. Automated multi-class classification of lung diseases from CXR-images using pre-trained convolutional neural networks. *Expert Syst. Appl.* **211**, 118650. https://doi.org/10.1016/j.eswa.2022.118650 (2023).
11. Singh, S., Karimi, S., Ho-Shon, K., Hamey, L. From chest X-rays to radiology reports: A multimodal machine learning approach, in *Proceedings of the 2019 Digital Image Computing: Techniques and Applications (DICTA)* 1–8 (IEEE, 2019).
12. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R. M. TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. https://doi.org/10.48550/arXiv.1801.04334 (2018).

13. Liu, G., Hsu, T.-M.H., Mcdermott, M., Boag, W., Weng, W.-H., Szolovits, P., Ghassemi, M. *Clinically Accurate Chest X-ray Report Generation*.
14. Harzig, P., Chen, Y.-Y., Chen, F. & Lienhart, R. Addressing data bias problems for chest X-ray image report generation. *arXiv preprint* https://doi.org/10.48550/arXiv.1908.02123 (2019).
15. Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M. & Fahmy, A. Automated radiology report generation using conditioned transformers. *Inform. Med. Unlocked* **24**, 100557. https://doi.org/10.1016/j.imu.2021.100557 (2021).
16. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. 2017.
17. Park, H., Kim, K., Park, S. & Choi, J. Medical image captioning model to convey more details: Methodological comparison of feature difference generation. *IEEE Access* **9**, 150560–150568. https://doi.org/10.1109/ACCESS.2021.3124564 (2021).
18. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J. Image captioning with semantic attention, in *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4651–4659 (2016).
19. Vinyals, O., Toshev, A., Bengio, S., Erhan, D. Show and tell: A neural image caption generator, in *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3156–3164 (2015).
20. Li, Y., Zhang, X., Cheng, X., Tang, X. & Jiao, L. Learning consensus-aware semantic knowledge for remote sensing image captioning. *Pattern Recognit.* **145**, 109893. https://doi.org/10.1016/j.patcog.2023.109893 (2024).
21. Luo, X. et al. Global semantic enhancement network for video captioning. *Pattern Recognit.* **145**, 109906. https://doi.org/10.1016/j.patcog.2023.109906 (2024).
22. Demner-Fushman, D. et al. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **23**, 304–310. https://doi.org/10.1093/jamia/ocv080 (2016).
23. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H. Training data-efficient image transformers & distillation through attention, in *Proceedings of the International Conference on Machine Learning*, 10347–10357 (PMLR, 2021).
24. Ferdous, G. J., Sathi, K. A., Hossain, M. A., Hoque, M. M. & Dewan, M. A. A. LCDEiT: A linear complexity data-efficient image transformer for MRI brain tumor classification. *IEEE Access* **11**, 20337–20350. https://doi.org/10.1109/ACCESS.2023.3244228 (2023).
25. Song, H., Sun, D., Chun, S., Jampani, V., Han, D., Heo, B., Kim, W., Yang, M.-H. An extendable, efficient and effective transformer-based object detector (2022).
26. He, A. et al. H2Former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Trans. Med. Imaging* **42**, 2763–2775. https://doi.org/10.1109/TMI.2023.3264513 (2023).
27. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. https://doi.org/10.48550/arXiv.2010.11929 (2020)
28. Deng, J., Dong, W., Socher, R., Li, L.-J., Kai L., Li F.-F. ImageNet: A large-scale hierarchical image database, in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
29. Hinton, G., Vinyals, O., Dean, J. Distilling the knowledge in a neural network (2015).
30. Alotaibi, A., Alafif, T., Alkhilaiwi, F., Alatawi, Y., Althobaiti, H., Alrefaei, A., Hawsawi, Y., Nguyen, T. ViT-DeiT: An ensemble model for breast cancer histopathological images classification, in *Proceedings of the 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)* 1–6 (IEEE, 2023).
31. Gou, J., Yu, B., Maybank, S. J. & Tao, D. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **129**, 1789–1819. https://doi.org/10.1007/s11263-021-01453-z (2021).
32. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* (2019).
33. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. Improving language understanding by generative pre-training (2018).
34. Sharma, S., Asri, L. El, Schulz, H., Zumer, J. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation (2017).
35. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation, in *Proceedings of the Proceedings of the 40th annual meeting of the Association for Computational Linguistics* 311–318 (2002).
36. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in *Proceedings of the Proceedings of the acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* 65–72a (2005).
37. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries, in *Proceedings of the Text Summarization Branches Out* 74–81 (2004).
38. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T. Long-term recurrent convolutional networks for visual recognition and description, in *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2625–2634 (2015).
39. Xiong, Y., Du, B., Yan, P. Reinforced transformer for medical image captioning, in *Proceedings of the Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10* 673–680 (Springer, 2019).
40. Tsaniya, H., Fatichah, C. & Suciati, N. Automatic radiology report generator using transformer with contrast-based image enhancement. *IEEE Access* **12**, 25429–25442. https://doi.org/10.1109/ACCESS.2024.3364373 (2024).
41. Ucan, M., Kaya, B., Kaya, M. & Alhajj, R. Medical report generation from medical images using vision transformer and Bart deep learning architectures. In *Proceedings of the Social Networks Analysis and Mining* (eds Aiello, L. M. et al.) 257–267 (Springer, 2025).
42. Nimalsiri, W., Hennayake, M., Rathnayake, K., Ambegoda, T. D., Meedeniya, D. Automated radiology report generation using transformers, in *Proceedings of the 2023 3rd International Conference on Advanced Research in Computing (ICARC)* 90–95 (IEEE, 2023).
43. Edirisinghe, D., Nimalsiri, W., Hennayake, M., Meedeniya, D. & Lim, G. Chest X-ray report generation using abnormality guided vision language model. *IEEE Access* **13**, 157651–157673. https://doi.org/10.1109/ACCESS.2025.3606961 (2025).
44. zoogzog CheXNet Implementation in PyTorch. Available online: https://github.com/zoogzog/chexnet (accessed on 01 March 2025).
45. evakli11 CheXNet Implementation in PyTorch. Available online: https://github.com/evakli11/cs541dlfinalproject_chexnet (accessed on 01 March 2025).

Acknowledgements

The authors would like to express their sincere gratitude to Mehtap Çiçekçi for her valuable contributions to the clinical evaluation of the generated reports and for her support in the validation of the study results.

Author contributions

Conceptualization, M.U. and M.K; methodology, B.K. and M.K.; software, M.U. and R.A; valida-tion, B.K., M.K., and R.A.; formal analysis, M.U. and B.K.; investigation, M.U. and R.A.; resources, B.K. and R.A.; writing—original draft preparation, M.U. and M.K.; writing—review and editing, B.K. and R.A.; visualization, M.U. and B.K.;

supervision, M.K. and R.A.; project administration, M.K.

Funding
This research was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant No 123E171 and in part by the Firat University Scientific Research Projects Unit (FUBAP) under Grant MF.25.150.

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval and consent to participate
This study did not involve direct human participation but used publicly datasets. The data supporting the findings of this study are based on a publicly available dataset.

Additional information
Correspondence and requests for materials should be addressed to M.U., M.K. or R.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026