IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 January 2024, accepted 11 February 2024, date of publication 19 February 2024, date of current version 23 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3367360

RESEARCH ARTICLE

CNX-B2: A Novel CNN-Transformer Approach For Chest X-Ray Medical Report Generation

FAWAZ F. ALQAHTANI[1,3], MASHOOD MOHAMMAD MOHSAN[2], KHALAF ALSHAMRANI[1], JAHAN ZEB[2], SALIHAH ALHAMAMI[1], AND DAREEN ALQARNI[1]

[1]Department of Radiological Sciences, College of Applied Medical Sciences, Najran University, Najran 61441, Saudi Arabia
[2]Department of Computer and Software Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan
[3]Health Research Centre, Najran University, Najran 61441, Saudi Arabia

Corresponding author: Fawaz F. Alqahtani (ffalqahtani@nu.edu.sa)

This work was supported by the Deputy for Research and Innovation Ministry of Education, Kingdom of Saudi Arabia through the Institutional Funding Committee at Najran University, Kingdom of Saudi Arabia, under Grant NU/IFC/2/MRC/-/6.

ABSTRACT Medical imaging techniques are the most popular non-invasive methods to diagnose chest diseases. Chest X-ray scans are employed commonly to detect chronic obstructive pulmonary diseases and other respiratory diseases. Despite the significance of these diagnostic methods, the process of disease detection and the subsequent task of CXR report writing is tedious for radiologists. Therefore, Automated radiological report generation is a highly desirable task for radiologists. Previous studies were focused on the automated generation of medical reports to achieve greater quantitative scores rather than focusing on the quality of reports. Such approaches suffer from the problem of generating normal reports for CXR with diseases. Additionally, the absence of clear segregation between normal and abnormal samples in publicly available datasets its impossible to evaluate the performance of models in generating rare abnormal reports. To address these issues, we propose CNX-B2 which is a Convolutional Neural Network (CNN) combined with a Transformer approach to generate medical reports. The proposed encoder is designed to be both hybrid and efficient, capturing meaningful spatial features through inherent convolution biases. This enables the transformer-based decoder to robustly convert these features into coherent medical reports. Secondly, we also introduce a new radiological report dataset to evaluate model performances on abnormal reports separately. Our proposed model is further evaluated on the IU-Xray dataset, achieving competitive scores of 0.479 BLEU-1, 0.188 METEOR, and 0.586 CIDER.

INDEX TERMS Convolution neural networks, transformers, medical report generation, natural language processing.

I. INTRODUCTION

Diseases that target the chest region are specifically dangerous as they can lead to fatality. Lungs play a vital role in the respiratory system and any damage to them can cause inefficiency of the human body or could have life-threatening implications too. Chronic respiratory diseases are among the most common non-communicable diseases worldwide, largely due to the ubiquity of noxious environmental, occupational, and behavioural inhalation exposures. In addition to chronic obstructive pulmonary disease (COPD) and asthma, chronic respiratory diseases include interstitial lung disease, pulmonary sarcoidosis, and pneumoconiosis, such as silicosis and asbestosis. On average, 58000 deaths per year occur only due to pneumonia [1]. The primary practice to detect such diseases is the use of various medical imaging techniques such as Chest X-ray (CXR), Ultrasound and Magnetic Resonance imaging (MRI) [2]. CXR is the most popular non-invasive imaging technique to detect chest diseases. Understanding and interpreting CXR and then reporting findings is a time-consuming task, labouring and error-prone task for radiologists. Automated radiological report generators can lighten the workload of radiologists and can also assist them.

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

Recent attempts to generate reports have been motivated by image captioners. Both tasks share the same generic problem, an image to sequence one. A radiological report consists of patient data, history, radiologist's findings, and impressions along with multiple CXR scans. It takes approximately 15-20 minutes to complete this labour procedure. This lengthy and hectic procedure can lead to error if done by inexperienced radiologists also it is a time-consuming procedure for experienced radiologists. The problem of it being a tedious task was greatly highlighted during the 2019 pandemic when a large number of CXR images needed to be analyzed [3]. The objective of automated report generation is to take CXR image pixels as input and generate findings and impressions as part of a radiology report.

Due to the clinical importance of the problem, researchers have proposed many architectures [4], [5], and [6]. Almost all research has utilized the Indiana University dataset [7] it contains abnormal and normal cases. The frequency of normal sentences in this dataset is very high and was highlighted in [8] which makes the previous proposed architectures report generation capabilities biased. Another major issue of IU dataset is the lack of distribution between normal and abnormal reports which makes it difficult to measure architecture performances on rare frequent sentences and diseases. In this work, we propose a new dataset to overcome these limitations.

A comprehensive review of the literature concluded that the initial attempts to generate radiological reports employed a CNN image encoder to infer latent features and a Recurrent Neural Network (RNN) to decode those features to generate medical findings. The recent advancements in language modelling have brought out a novel architecture named Transformers. It is the only architecture that is recurrence and convolution free. From 2017, Transformers [9] has not only dominated language models such as RNN and RNN with attention but also has dominated CNN in the Computer Vision field. To improve the current accuracy of the report generator this article has proposed a novel configuration in the Encoder-Decoder arrangement.

The contribution of this article is summarized as follows:

1) A novel CNX-B2 model is proposed which is based on CNN-Transformer approach. ConvNeXT is employed as image encoder which has the ability to capture visual features due to spatial biasness but its design is modified according to Vision Transformers making it more efficent. BioBERT is a medical language based Transformer which is employed to generate radiological reports as the decoder.
2) A new dataset was proposed to overcome the limitation of IU dataset along with a distribution of normal and abnormal reports. This dataset is public and can be used for evaluating model performance on normal and abnormal reports.
3) The proposed architecture is evaluated on IU dataset and as well as on our proposed dataset. CNX-B2 has outperformed CDGPT2 [10], a previous attempt of combining CNN and Transformer and achieved remarkable BLEU scores against earlier attempts.

The remaining part of the article is structured as follows. Section II examines related literature. The methodology is explained in Section III. The experimental details and results are discussed in Section IV.

II. REVIEW LITERATURE

Automated Radiological Report Generation is a derivative technique to describe clinical details of Chest X-ray images. It is a combination of computer vision and Natural Language Processing which have a strong societal impact. Description retrieval, template filling, and hand-crafted NLP techniques were some of the earlier methods of report writing.

There were many advancements in automated medical report generation later, but the base arrangement of each method was to utilize an image encoder for converting CXR images into a latent space and then bring a decoder into play to generate medical reports. The problem was generically identified as an image-to-sequence problem. We have divided the review literature based on the encoder-decoder architectures used in Automated radiological report generation.

CNN-RNN is the most employed encoder-decoder configuration. RNN [11] are employed in time series forecasting, classification, signal processing, and NLP due to their ability to learn over multiple time steps. Long Short-Term Memory (LSTM) and Gate Recurrent Unit (GRU) [12] are popular architectures maintaining context over extended periods. Later Attention mechanisms enabled context learning without time step constraints, allowing the RNN's to encode or decode relevant information without fixed-length context vectors [13].

The CNN encoder captures visual dominant features from CXR using its inherent spatial biases and Recurrent Neural Network (RNN) decodes these spatial features to text. To further increase the accuracy many techniques were introduced. Li et al. [4] utilized a unique hierarchical decision-making procedure known as Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent). A high-level retrieval policy module decided whether to use a low-level generation module to create a new sentence or to obtain a template sentence from an existing template database for each sentence. This technique assisted the generator to detect abnormal findings at a higher rate. Moreover, Xue et al. [14] achieved multimodality and employed the attention mechanism,

The idea of having a separate generator or decoder for normal and abnormal cases was further explored using LSTM networks. The encoded vector was passed to a topic generator, consisting of a single layer LSTM, to generate a sequence of high-level topics and was assisted by a gate module to produce distribution over normal and abnormal cases. The Attention-based Abnormal-Aware Fusion Network (A3FN) generated structured, topic-coherent and abnormality-aware radiological reports [5]. To furthermore increase the ability of an RNN to generate rare diseases, techniques such as

TABLE 1. Literature review summary table.

Sr#	Method	Year	Approach
1	HRGR-Agent [4]	2018	CNN, RNN and Retrieval Policy Module
2	CNN + 3 RNN [14]	2018	CNN, Sentence and Recurrent Paragraph Generative model
3	A3FN [5]	2019	CNN and separate LSTM abnormal report generator
4	ResBlock + Multi-attention [15]	2019	Multi-attention, and Fusion module, Sentence and Word RNN
5	Raregen [8]	2020	Few shot GAN, Garph Network and hierarchical LSTM
6	SentSAT + KG [16]	2020	Densenet, Graph Convolution and LSTM
7	R2Gen [18]	2020	Relational Memory driven Transformers
8	CDGPT2 [10]	2021	Chexnet combined with GPT2
9	VFE + GKE + SKE + RG [6]	2022	A unique combination of CNN, Knowledge graph and Transformer
10	JPG [19]	2022	Traditional CNN combined with a complete Transformer
11	AERMNET [17]	2023	CNN combined with LSTM and Relational Memory Module
12	AdaMatch-based LLM [20]	2023	Two Vision Transformers, BioClinicalBERT and a Large Language Model (LLM)
13	DCL [21]	2023	Two unimodal encoders, a multimodal encoder, a report decoder and three dynamic graph modules

few shot, Graph CNN, relational memory networks, and Generative Adversarial Networks were also explored [8], [15], [16], [17].

Transformers [9] was a revolutionary approach for sequence-to-sequence tasks such as Machine Translation. It was recurrence and convolution free network. Li et al. [22] first employed Transformers for text recognition. It leveraged the Transformer architecture for both image understanding and wordpiece-level text generation. Memory-Driven Transformer was proposed for automated medical report generation. The decoder layer incorporates Memory-Conditioned Layer Normalization, enhancing report generation capability [18].

Recently, Large Langauge Model (LLM) was employed using Cyclic techniques for report generation [20]. Contrastive Learning is a technique to improve representation learning. Li et al. [21] proposed a Dynamic graph combined with Contrastive Learning in Transformers. This improved visual and text representation in medical report generation task. Furthermore, 3D shared subspace was also explored for representation improvement [19].

CNN-Transformers approach employs a CNN to capture spatial features from CXR and a Transformer, dominant in language generation, decodes these features to generate reports. Alfarghaly et al. [10] was the first to employ this approach. It used visual and semantic features to condition the decoder for faster training. Furthermore, various knowledge graphs and word embeddings were also combined with spatial features to assist Transformers-based decoders [6]. Although these approaches show improvements but a pure CNN-Transformer approach for report generation is still unexplored. Table 1 summarizes the review literature.

III. PROPOSED METHOD

Figure 2 illustrates the proposed methodology. ConvNext [23], a CNN based encoder, is employed to encode image features and BioBERT [24], a transformer based, language model processes the image features (Hidden values) to generate radiological medical reports.The output of ConvNext is used to calculate Q query, and K key vectors in cross attention layers of decoder. The output of encoder is fed to BioBERT being a variant of BERT models lack the capability of text generation so additional layers of cross attention were added in it to be used as decoder.

A. ENCODER

ConvNext models are constructed from Residual Networks (ResNets) but having similar architectural design improvements of transformers excluding attention layers. In 2021, Hierarchical transformers such as SWIN transformer [25] proposed a novel "sliding window" mechanism employing attention within them indicating the use of convolution was still relevant. ConvNext being a pure convolution based model adapts the advanced architectural feature of transformer network visualizing the importance of CNN as a generic back bone or as encoders. It contains a stack of stages each consisting of different number of layers with distinct kernel sizes of convolution in it. The design changes in ResNet to produce ConvNeXt model are stated below.

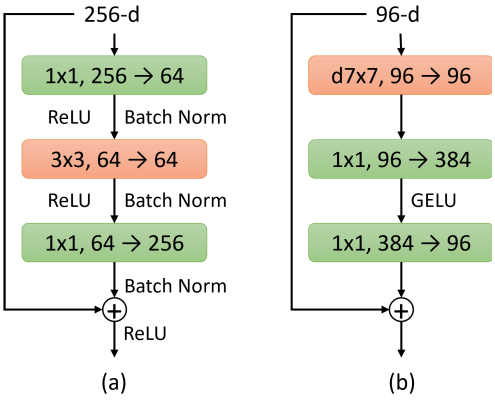
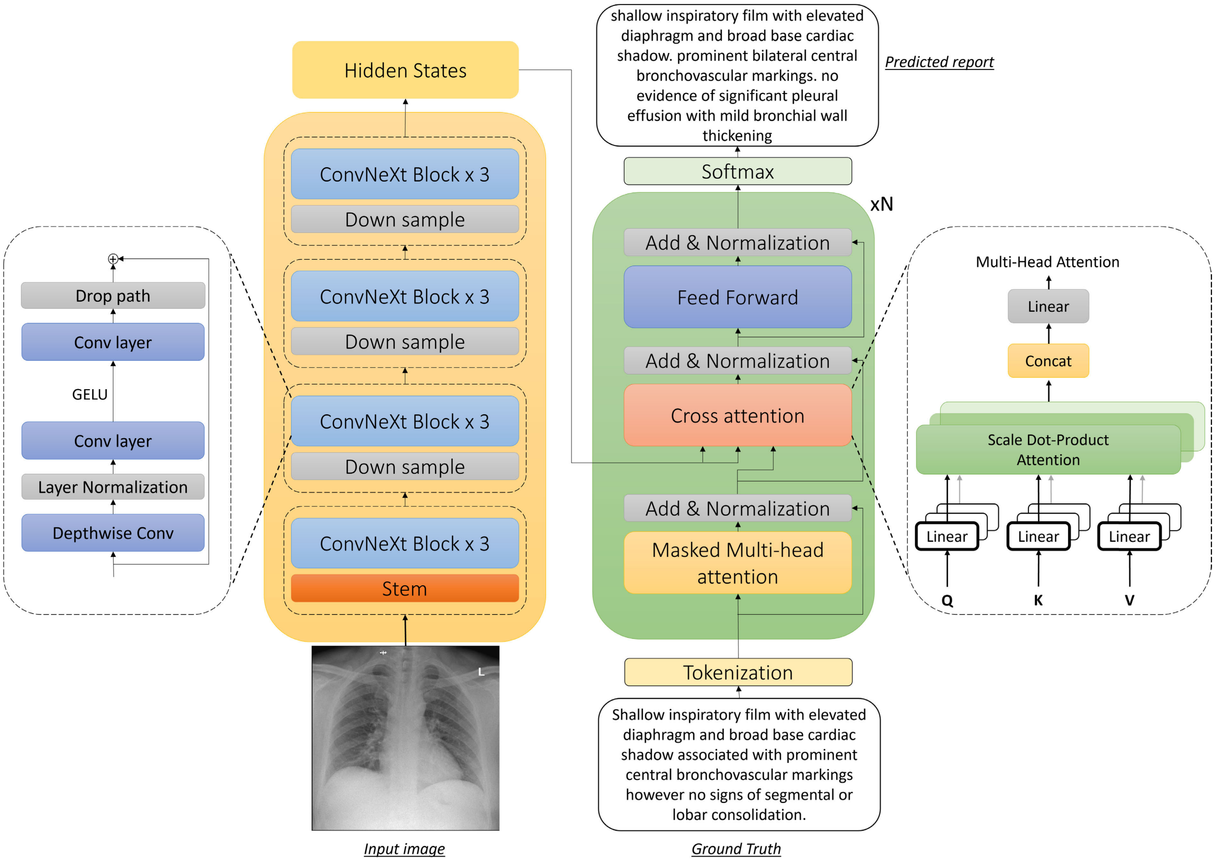


FIGURE 1. (a) A simple ResNet block. (b) A ConvNeXt block after macro designing, ResNext-ify and other alteration of a ResNet block.

1) MACRO DESIGNING

CNN networks has multi-stage block like designs and each of them has different feature map resolution. SWIN models adopted them in transformers with a stage ratio of 1:1:3:1 and for larger variants of it had 1:1:9:1 and whereas standard ResNet has 3:4:6:3 stages so in ConvNext it was converted to 1:1:3:1 stage ratio. Secondly, "stem cell"



shallow inspiratory film with elevated diaphragm and broad base cardiac shadow. prominent bilateral central bronchovascular markings. no evidence of significant pleural effusion with mild bronchial wall thickening

Shallow inspiratory film with elevated diaphragm and broad base cardiac shadow associated with prominent central bronchovascular markings however no signs of segmental or lobar consolidation.

FIGURE 2. Figure represents the proposed architecture. The left side illustrates the Image Encoder consisting of ConvNeXt block and right side is the BioBERT language decoder along with cross attention for radiologist report generation.

structure also known as the input layer of Resnet employs a 7 × 7 convolution layer with a stride of 2, as natural images are inherent with redundancy, the stem cell generally down sample the input images. SWIN transformer uses ''Patchify'' strategy with a more aggressive approach to downscale input image with non-overlapping convolution in order to accommodate the multi-stage design. ConvNext ''Stem cell'' layer was replaced by SWIN transformer ''Patchify'' with a convolution layer of 4 × 4 with a stride of 4.

2) RESNEXT-IFY

The ResNeXt [26] principle is to separate convolution kernels into groups in bottleneck layer. This in short expands the width of model and significantly reduce FLOPs. For ConvNext we have employed depth-wise convolution as it is similar to the weighted sum operation in attention mechanism. The use of 1 × 1 convolution and depth wise convolution combines the spatial and channel features separately. This reduced FLOPS of the model but also reduced the accuracy so following the strategy of RexNeXt we also increased the width of ConvNext from 64 to 96.

3) INVERTED BOTTLE NECK AND KERNEL SIZE

A major design in Transformers is the inverted bottle neck and it layer is 4x times wider than the input feature map. A similar scheme was applied in ConvNeXt model. This step although increases the FLOPs of the block but for overall model, it was decreased due to lack operation for down sampling in the architecture. Another design parameter of SWIN transformer is the use 7 × 7 kernel in the inverted block where as ConvNext used 3 × 3. To increase the size of kernel we first had to move the depth wise layer up and then increase the kernel size to 7 × 7. Figure 1 illustrates the design changes.

4) MICRO DESIGN

Some micro scale design features of Resnet was also replaced by SWIN transformer features. The Gaussian Error Linear Unit (GELU) a smoother version of ReLU is employed as the activation function of the model. Another major difference between a transformer block and Resnet block is the use of less activation function. Therefore in Resnet block all activation function were removed expect for the middle layer, replicating it with SWIN model. Similarly, the count of normalization layer was also reduced and Batch Normalization was replaced by Layer Normalization.

B. DECODER

The architecture of BioBERT was to serve as an Encoder [24]. It was a stack of identical layers and each layer consisted of two sub-layers: The multi-headed Self Attention sub-layer and Position wise feed-forward sub-layer. In our approach, we employed BioBERT as a decoder. We added cross-attention between the two sublayers so it can process the hidden states or context vector from the image-based encoder, ConvNeXt. To add a third sub-layer to the two sub-layers that are already present in each Decoder, the decoder executes multi-head attention over the hidden states from the encoder, which are then turned into queries and keys. Each sub-layer employs residual connections in a way that is comparable to the encoder's behaviour before layer normalization. The ''Masked Multi-head attention'' layer is used by the decoder to stop focusing on succeeding places. Due to the masking technique and the moving of the output embedding offset by one place as a result of Casual Language Modelling (CLM), the prediction of location i can only rely on the known outputs at positions lower than i. The hyperparameters of BioBERT we used are listed in table 2.

1) TEXT INPUT REPRESENTATION

The decoder receives a set of tokens as input. The feature vectors of each token's d dimension are retrieved using the Embedding look-up table of V vocabulary length, and positional encoding of the sequence is then appended. The entire report, token by token in a causality manner, is provided as decoder input during training, but only the initial start token is provided during inference in order to construct subsequent words using CLM.

2) SELF ATTENTION

The key element of a Transformer model is the self attention mechanism which extracts most important or attends relevant features from input text. After the combination of embedding vector along with positional encoding, resultant passed though various linear layers to generate Q Query, K Key and V Value vectors for calculation of Self attention. The following equation represents the computation of attention mechanism:

$$\text{Self Attention}\,(Q, K, V) = \text{softmax}((Q \cdot K^{\mathrm{T}})/\sqrt{d_k}) \cdot V \tag{1}$$

Dot product multiplication of Q and K vectors makes computation efficient in Transformers and makes it possible to compile multi headed self attention (MSA) layers and large language models. Gradient explosion is avoided by applying the scaling factor of $\sqrt{d_k}$ dimension of key vector. Later, softmax function is implemented to assign probabilities to most attended values from V vector. The multi headed self attention mechanism helps gather information from several sub spaces, h times, so any relevant features are never left behind. The projected results from MSA are concatenated altogether to maintain the same dimension as the input feature.

$$\text{Multihead}(Q, K, V) = \text{concat}(h_1 \ldots h_n) \cdot W^o \tag{2}$$

where

$$\text{head}_i = \text{Self Attention}(W_i^q q_j, W_i^k K, W_i^v V)$$

Equation 2 depicts calculation of a MSA layer in which n represents total number of attentions in one MSA layer whereas $\mathrm{W_i}^{\mathrm{q}}$, $\mathrm{W_i}^{\mathrm{k}}$ and $\mathrm{W_i}^{\mathrm{v}}$ represents weight matrices of Query, Key and Value respectively of each single attention head.

3) CROSS ATTENTION

In sequence to sequence tasks such as machine translation, question answering and text summarizing, the transformer architecture has a special sub layer known cross attention layer in decoder which attends different parts of encoded hidden vector and decoder vector to generate text. The hidden feature vector generated by the encoder is also known as latent vector, which has concealed all information of one domain or language.

$$\text{Cross Attention}\,(Q_{dec}, K_{enc}, V_{enc}) = \frac{\text{softmax}(Q_{dec} \cdot K_{enc}^{\mathrm{T}})}{\sqrt{d_{k_{enc}}} \cdot V_{enc}} \tag{3}$$

In our case, BioBERT is an encoder based architecture and we are translating images to text in form of medical reports so cross attention, also known as encoder-decoder attention, is introduced with random initialization of weights. Equation 3 demonstrates the computation of a single cross attention head where Q_{dec}, K_{enc} and V_{enc} are Query from previous Decoder layer, Key and Value are vectors generated from Encoder's hidden states respectively. Based on the context of the output sequence created so far, cross attention enables the decoder to choose to pay attention to different sections of the input sequence. Due to its significance in many sequence-to-sequence tasks, Transformer models have excelled at these tasks, achieving state-of-the-art outcomes.

IV. EXPERIMENTATION

The proposed novel methodology for medical report generation was experimented on two datasets and all implementation details are mentioned below.

A. IMPLEMENTATION DETAILS

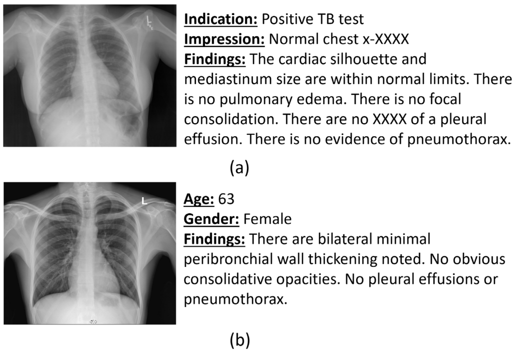
The proposed architecture for radiological report generation was implemented on PyTorch utilizing Hugging Face library. The model was trained on dual Nvidia RTX 2070 8GB system for 20 epochs with a batch size of 4 on each GPU. The other hyper parameters are mention in table 2. The training and testing phase for each dataset was done separately and no weights were shared between them.

B. EVALUATION METRICS

To asses the radiological report generation capability of our proposed method. We employed BLEU [27], METEOR [28], ROUGE L [29] and CIDER [30]. BLEU even being used for language translation tasks is also employed in report generation tasks in literature. CIDER metric was proposed to capture human judgment of consensus in image captioning tasks, but is not employed much in medical report generation.

TABLE 2. List of parameter and hyper parameters of proposed architecture.

Sr#	Parameters	Encoder	Decoder
1	Model	ConvNeXt	BioBERT
2	Hidden layers	-	12
3	Number of stages	4	-
4	Depth	3, 3, 27, 3	-
5	Attention heads	-	12
6	Intermediate size	-	3072
7	Hidden size	128, 256, 512, 1024	768
8	Patch size	4	-
9	Image size	224	-
10	Vocabulary size	-	28996
11	Max length	-	50
12	Beams	-	4
13	Layer normalization	$1e^{-12}$	$1e^{-12}$



Indication: Positive TB test
Impression: Normal chest x-XXXX
Findings: The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.

(a)

Age: 63
Gender: Female
Findings: There are bilateral minimal peribronchial wall thickening noted. No obvious consolidative opacities. No pleural effusions or pneumothorax.

(b)

FIGURE 3. (a) A sample IU dataset [7] containing X-ray image, Indication, Impression and Findings. (b) A sample of our dataset containing Xray image, Age, Gender, and Findings.

Original implementation of COCO image captioning.[1] was utilized to measure scores to maintain standardization.

C. DATASETS

The proposed method was evaluated on two datsets: IU dataset and Our dataset. Training and Testing of both datasets were conducted separately.

1) IU DATASET

The Indiana University (IU) chest X-ray dataset [7] is one the most widely used dataset which contains 3995 reports along with 7470 multiple CXR images. It is a public dataset created for training and testing of X-ray images to test algorithm for classification, summary generation or report generation tasks. Each report contains patient details, impressions, reports and tags for Indiana database. Personal information of patients revealing identity is hidden by XXXX. Unfortunately, train and test split were never released hence making it difficult for bench marking purposes along with that, many information is missing and pairs are also incomplete.

2) OUR DATASET

We present a novel dataset that is consistent with distribution of normal and abnormal reports. The Picture Archiving and Communication System of King Khalid Hospital in Najran, Saudi Arabia was searched for all chest x-ray performed between November 2019 and November 2022. A total of 1250 image and report pairs were collected. The analysed data included 528 abnormal cases and 722 normal: 471 female and 779 male cases. The mean age of these subjects was approximately 53 years. Our dataset is a small and consistent dataset elaborating variety in medical reports along with clear distribution of abnormal and normal pairs for quicker harmonious benchmarking.

Figure 3 provides a sample CXR images along with the Impression, Findings, Indication, Age, and Gender of patients from both datasets. Table 3 contains the details of the Train, Validation, and Test split of both datasets.

TABLE 3. Train, validation, and test split details of Our dataset and IU dataset [7].

Sr#	Split	IU dataset [7]	Ours dataset
1	Train	2403	873
2	Validation	500	253
3	Test	300	124

V. RESULTS AND ANALYSIS

In this section, we present a comprehensive analysis of the results obtained from our experimentation, shedding light on key findings and their implications. We performed quantitative and qualitative analyses separately. Finally, we compared the results of our proposed architecture with existing methods.

A. QUANTITATIVE RESULTS

The proposed model employs a pre-trained CNN image model, trained on natural images, and a language model without pre-trained weights. In our experimental investigation, tests with three language models were performed on IU and Our dataset. The effectiveness of the models was evaluated using automatic assessment metrics like the n-gram similarity between the produced sentences and the ground truth reports. GPT2 [31], MiniLM [32] and BioBERT [24] transformer-based language models are employed for experimental study. Table 4 displays the experimental results, displaying the effectiveness of each model.

TABLE 4. Results of experimental investigation on IU [7] and our dataset.

Sr#	Encoder	Decoder	BLEU 1	BLEU 2	BLEU 3	BLEU 4
		Indiana University Dataset [7]				
1	ConvNeXt	GPT2	0.356	0.221	0.152	0.098
2		BioBERT	**0.479**	**0.363**	**0.261**	**0.173**
3		MiniLM	0.371	0.257	0.183	0.109
		Our Dataset				
5	ConvNeXt	GPT2	0.341	0.318	0.302	0.284
6		BioBERT	**0.445**	**0.409**	**0.389**	**0.375**
7		MiniLM	0.364	0.307	0.281	0.263

The IU dataset does not have a distribution between normal and abnormal cases. This makes it difficult to evaluate models performance on abnormal cases. Our dataset on the

[1] https://github.com/yikang-li/coco-caption

INPUT	GROUND TRUTH	CNX-B2 (ours)	AERMNet	CDGPT2
	No evidence of active disease. Low lung volumes are present. The heart size and pulmonary vascularity appear within normal limits. Lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen. Surgical clips are present in the abdomen.	normal chest heart size normal. lungs are clear. xxxx are normal. no pneumonia, effusions, edema, pneumothorax, adenopathy, nodules or masses.	cardiomegaly without heart failure , low lung volumes left retrocardiac opacities , subsegmental atelectasis . low lung volumes . opacity at the left base consistent with overlying subsegmental atelectasis . no definite pleural effusion seen , no typical findings of pulmonary edema .	low lung volumes. cardiomediastinal silhouette is within normal limits. lungs are clear. no pneumothorax or large pelural effusion.
	No acute cardiopulmonary disease. . The cardiomediastinal silhouette is normal size and configuration. Pulmonary vasculature within normal limits. The lungs are well-aerated. There is no pneumothorax, pleural effusion, or focal consolidation.	1. no acute radiographic cardiopulmonary process. the cardiomediastinal silhouette and vasculature are within normal limits for size and contour. the lungs are normally inflated and clear. osse	no acute cardiopulmonary abnormality . no airspace disease , effusion or noncalcified nodule . normal heart size and mediastinum . visualized of the chest are within normal limits	no acute cardiopulmonary findings. . the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. the lungs are clear of focal airspace disease, pneumothorax, or pleural effusion. there are no acute bony findings.
	No acute cardiopulmonary abnormality. The heart is normal size. The mediastinum is unremarkable. There is no pleural effusion, pneumothorax, or focal airspace disease. The XXXX are unremarkable.	no acute cardiopulmonary abnormality. the lungs are clear bilaterally. specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion. cardiomediastinal silhouette is un	no acute cardiopulmonary abnormality . the lungs are clear , and without focal airspace opacity . the cardiomediastinal silhouette is normal in size and contour , and stable . there is no pneumothorax or large pleural effusion	1. no acute radiographic cardiopulmonary process. the cardiomediastinal silhouette is within normal limits for size and contour. the lungs are normally inflated without evidence of focal airspace disease, pleuraleffusion, or pneumothorax. no acute asseus abnormality

FIGURE 4. A comparison of generated reports by various models on multiple samples of IU dataset [7]. The colour green means similar and red indicates wrong findings.

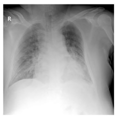
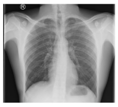
INPUT	GROUND TRUTH	PREDICTED
	ilateral scattered areas of alveolar shadows for clinical correlation and follow up.	ilateral scattered areas of alveolar shadows for clinical correlation and follow up.
	Shallow inspiratory film with elevated diaphragm and broad base cardiac shadow associated with prominent central bronchovascular markings however no signs of segmental or lobar consolidation.	shallow inspiratory film with elevated diaphragm and broad base cardiac shadow. prominent bilateral central bronchovascular markings. no evidence of significant pleural effusion with mild bronchial wall thickening
	Lung fields are clear, CP angles are clear , Cardiac size is not enlarged	lung fields are clear, cp angles are clear, cardiac size is not enlarged

FIGURE 5. A comparison of generated reports by proposed model on multiple samples of our dataset.

other hand evaluates model performance on both cases and specially highlighting the false positive reports.

B. QUALITATIVE RESULTS

Relying merely on numeric numbers, as shown in table 4, may not give an exhaustive review due to the imbalance and bias towards normal reports in the IU dataset. To evaluate the overall quality of the reports that were generated, we thus carried out a comparison between various test samples on both datasets. The results provided by our proposed model closely match the actual data in IU dataset, which corresponds to the most typical occurrences. Figure 4 displays multiple test samples of IU. These reports are clear and offer information in-depth enough to compare to the reference report.

We compared CNX-BR2 report generation with recent SOTA's for qualitative comparison. We used AERMNet and CDGPT2 models which are public. The generated reports from unseen CXR of our and SOTA's model can viewed at figure 4. In the first example (topmost) CNX-B2 generated the most similar report to ground truth. The AERMNet model indicated diseases such as "*Cardiomegaly*" and indicated "*left retrocardiac opacities*" whereas the ground truth suggests that CXR is normal. Another important finding is that most of the reports generated by our CNX-B2 and CDGPT2 are similar except for the first case. It is to be noted that AERMNet and CDGPT2 are trained on a different IU dataset split and if they were trained on similar splits then reports may vary. Unfortunately, due to a lack of abnormal reports, a proper evaluation of all models cannot be performed.

On the other hand, Figure 5 displays some abnormal and normal samples from test split of our dataset. The first sample is abnormal and the proposed model predicted exactly similarly to the reference report indicating a complete True negative. An interesting finding can be highlighted in the second sample where the proposed model utilized a synonym maintaining the context of the report. "no signs" was replaced by "no evidence" and this type of accurate prediction is missed in quantitative analysis. Lastly, a normal report sample is generated as it is.

C. COMPARISON WITH LITERATURE

After the experimental research had completed the model, we moved forward with a comparison to current state-of-the-art methods. The findings of the automated metrics for the generated reports in comparison to the literature are shown in

TABLE 5. Comparison with review literature.

Sr#	Model	Dataset	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR	ROUGE	CIDER
1	HRGR-Agent [4]	IU [7]	0.438	0.298	0.208	0.151	-	0.322	0.343
2	CNN + 3 RNN [14]		0.464	0.358	**0.270**	**0.195**	0.274	0.366	-
3	A3FN [5]		0.443	0.337	0.236	0.181	-	0.347	0.374
4	ResBlock + Multi-attention [15]		0.476	0.340	0.238	0.169	-	0.347	0.297
5	Raregen [8]		0.448	0.340	0.255	0.178	-	0.371	0.378
6	SentSAT + KG [16]		0.441	0.291	0.203	0.147	-	0.367	0.304
7	R2Gen [18]		0.470	0.304	0.219	0.165	0.187	0.371	-
8	CDGPT2 [10]		0.387	0.245	0.166	0.111	0.164	0.289	0.257
9	VFE + GKE + SKE + RG [6]		**0.496**	0.327	0.238	0.178	-	0.381	0.382
10	JPG [19]		0.479	0.319	0.222	0.174	**0.193**	0.377	-
11	AERMNET [17]		0.486	0.321	0.236	0.183	0.219	**0.398**	0.560
12	AdaMatch-based LLM [20]		0.416	0.300	0.207	0.144	0.162	0.365	-
13	DCL [21]		-	-	-	0.163	**0.193**	0.383	**0.586**
14	Ours		0.479	**0.363**	0.261	0.173	0.188	0.354	0.408
15	Ours	Ours	0.445	0.409	0.389	0.375	0.279	0.60	0.603

Table 5. While using a Traditional CNN pre-trained on CXR pictures for the Encoder, CDGPT2 used a Transformer with pre-trained weights for the Decoder. In spite of this strategy, CDGPT2 was able to obtain a BLEU-1 score of 0.38.

Our proposed model has the greatest BLEU-2 score when comparing all of these methods on the IU dataset, demonstrating its overall competence in producing reports that are very similar to those produced by radiologists. It has outperformed CDGPT2. Other studies such as HRGR-Agent, ResBlock Multi-attention and SentSAT + KG have employed RNN for language decoder and a traditional CNN for image encoder. Comparing all these techniques on the IU dataset, our proposed model demonstrates remarkable scores, indicating its overall effectiveness in generating reports that closely resemble those written by radiologists.

In comparison to previous CNN-Transformer approaches [10], [18], the CNX-B2 models achieve the best quantitative scores in BLEU-2, BLEU-3, METEOR, and CIDER metrics. This demonstrates the effectiveness of having a hybrid CNN encoder for a transformer decoder in medical report generation.

D. DISCUSSION

Even though Transformers are dominating CNN in vision-based tasks, this does not mean that CNN is cornered. The potential of Convolutional layers still plays a critical role in capturing spatial features. Our CNX-B2 highlights that having an image encoder of CNN generates useful reports. The ConvNeXt adopts efficient methods of a vision transformer without using attention layers. The proposed approach, CNN-Transformer, competes with previous CNN-RNN or complete Transformers-based research as highlighted in table 5.

In IU dataset, many reports exhibit repetitive and identical descriptive sentences. The doctor's report frequency plot reveals a skewed distribution, with abnormal sentences frequently occurring (frequency = 1) across the entire dataset. Specifically, sentences with a frequency of less than 3 account for 6,290 out of 8,022 unique sentences [33]. To evaluate CNX-B2 on abnormal we evaluate it on our dataset. The robustness and understanding of CNX-B2 are shown in figure 5 where it uses a synonym with the correct context. The training of CNX-B2 involved the utilization of pre-trained weights, facilitating its easy integration into real-time applications by various institutions. This adaptability allows it to quickly grasp language understanding from a small corpus, rather than undergoing training from scratch.

The computational complexity of the CNX-B2 is 222 GFLOPS with 224.31 M parameter. It is one of the limitations of our research and we hope future research can explore developing a lightweight approach for medical report generation.

VI. CONCLUSION

In this research, we studied the effect of hybrid CNN as an image encoder for radiology medical report generation. We aim to address two issues: 1) The lack of distribution of abnormal and normal reports in datasets, and 2) How to improve model report generation capability. To address them, We first introduced a novel dataset collected from King Khalid Hospital in Najran, Saudi Arabia. This dataset contains a segregation of reports which assisted in the evaluation of our approach. Secondly, We propose a CNX-B2, a CNN-Transformer, approach which employs ConvNeXt (hybrid CNN) capable of capturing spatial features and BioBERT as a decoder for report generation. Our experimentation demonstrates that CNX-B2 competes with previous approaches in Natural Language Generation (NLG) metrics. The qualitative analysis demonstrates CNX-B2 has a good understanding of language which makes it justified to be employed as an automated radiological medical report generator.

VII. FUTURE WORK AND LIMITATIONS

Our CNX-B2 is not the perfect approach. One of the major limitations is that it has a computational complexity of 222 GFLOPS and contains 224.31 million parameters. This makes it impossible to be inferred on mobile devices. We intend to propose lightweight approaches in future for inference purposes. Secondly, some of the generated reports by CNX-B2 are not similar to ground truths even by attaining good language understanding. However, this problem also happens in previous works. Therefore, our future plan is to develop automated medical report generators using vision and language-based mobile networks and achieve better quantitative and qualitative results. Furthermore, We also aim to evaluate automated generated reports from radiologists.

REFERENCES

[1] T. Gupte, A. Knack, and J. D. Cramer, "Mortality from aspiration pneumonia: Incidence, trends, and risk factors," *Dysphagia*, vol. 37, no. 6, pp. 1493–1500, Dec. 2022.

[2] D. Sutton, *Textbook of Radiology and Imaging Set IND Reprint*, vol. 2. Amsterdam, The Netherlands: Elsevier, 2014. [Online]. Available: https://books.google.ae/books?id=8NrboAEACAAJ

[3] N. L. Demirjian, B. K. K. Fields, C. Song, S. Reddy, B. Desai, S. Y. Cen, S. Salehi, and A. Gholamrezanezhad, "Impacts of the coronavirus disease 2019 (COVID-19) pandemic on healthcare workers: A nationwide survey of United States radiologists," *Clin. Imag.*, vol. 68, pp. 218–225, Dec. 2020.

[4] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1537–1547.

[5] X. Xie, Y. Xiong, P. S. Yu, K. Li, S. Zhang, and Y. Zhu, "Attention-based abnormal-aware fusion network for radiology report generation," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, Chiang Mai, Thailand. Cham, Switzerland: Springer, Apr. 2019, pp. 448–452.

[6] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102510. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522001578

[7] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016.

[8] X. Jia, Y. Xiong, J. Zhang, Y. Zhang, and Y. Zhu, "Few-shot radiology report generation for rare diseases," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 601–608.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017.

[10] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Informat. Med. Unlocked*, vol. 24, Jan. 2021, Art. no. 100557. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914821000472

[11] J. Elman, "Finding structure in time," *Cognit. Sci.*, vol. 14, no. 2, pp. 179–211, Jun. 1990.

[12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[14] Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Granada, Spain. Cham, Switzerland: Springer, Sep. 2018, pp. 457–466.

[15] X. Huang, F. Yan, W. Xu, and M. Li, "Multi-attention and incorporating background information model for chest X-ray image report generation," *IEEE Access*, vol. 7, pp. 154808–154817, 2019.

[16] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12910–12917.

[17] X. Zeng, T. Liao, L. Xu, and Z. Wang, "AERMNet: Attention-enhanced relational memory network for medical image report generation," *Comput. Methods Programs Biomed.*, vol. 244, Feb. 2024, Art. no. 107979.

[18] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1439–1449.

[19] J. You, D. Li, M. Okumura, and K. Suzuki, "JPG-jointly learn to align: Automated disease prediction and radiology report generation," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 5989–6001.

[20] W. Chen, L. Shen, X. Li, and Y. Yuan, "Fine-grained image-text alignment in medical imaging enables cyclic image-report generation," 2023, *arXiv:2312.08078*.

[21] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest X-ray report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3334–3343.

[22] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "TrOCR: Transformer-based optical character recognition with pre-trained models," in *Proc. AAAI*, Feb. 2023. [Online]. Available: https://www.microsoft.com/en-us/research/publication/trocr-transformer-based-optical-character-recognition-with-pre-trained-models/

[23] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.

[24] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[26] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.

[28] A. Lavie and A. Agarwal, "Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl. (StatMT)*, 2007.

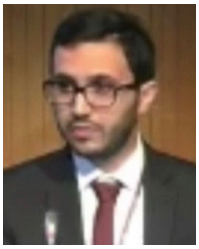
[29] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[30] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[32] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Proc. 34th Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 5776–5788.

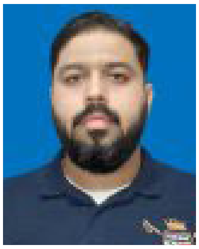
[33] P. Harzig, Y.-Y. Chen, F. Chen, and R. Lienhart, "Addressing data bias problems for chest X-ray image report generation," 2019, *arXiv:1908.02123*.



FAWAZ F. ALQAHTANI received the Ph.D. degree from Sheffield Children's Hospital, The University of Sheffield, U.K. He has more than 13 years of academic experience (since 2010) in teaching and research. He is currently an Associate Professor with the Radiological Sciences Department, Najran University, Saudi Arabia. He has a strong scientific background in spine imaging and artificial intelligence methods in medical imaging field. Furthermore, he is actively exploring and conducting various experimental and computational research work, including the potential of using glass technologies in medical imaging and radiological sciences applications.

MASHOOD MOHAMMAD MOHSAN received the B.S. degree in computer science from the University of Agriculture (UAF), Faisalabad, in 2020, and the M.S. degree in computer software engineering from the College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Rawalpindi. He is currently doing collaborative research in medical imaging.

KHALAF ALSHAMRANI received the M.Sc. degree in diagnostic radiography from Cardiff University, in 2014, and the Ph.D. degree in paediatric radiology from the Medical School, The University of Sheffield, in 2019. His research interests include skeletal imaging in paediatric, artificial intelligence in radiology, and bone density research in children.



JAHAN ZEB received the Master's degree in computer software engineering from the National University of Science and Technology, Pakistan. He is currently an Assistant Professor with CEME, NUST.

SALIHAH ALHAMAMI received the B.Sc. degree in diagnostic radiology from the Applied Medical Sciences, Najran University, Saudi Arabia, in 2022. She is currently with Najran University Hospital, as a Radiographer, and the Quality Officer with the Radiology Department. Her current research interests include artificial intelligence, vascular ultrasound, and radiation protection.

DAREEN ALQARNI received the bachelor's degree in diagnostic radiology from Najran University, Najran, Saudi Arabia, in 2022. She is currently a Diagnostic Radiologist with Najran University Hospital, and an Assistant with the Quality Department of Diagnostic Radiology. Her current research interests include diagnostic radiology, especially ultrasound and artificial intelligence.