Enhancing Image Insight By Transforming Images Into Descriptive Captions And Audio

Ms. Sangeetha Tony[1*] and Ms. Jisha Mary Jose[1†]

[1]Computer Science and Engineering, Rajagiri School of Engineering and Technology, Kochi, 682039, Kerala, India.

[†]These authors contributed equally to this work.

Abstract

The automation of interpreting medical imaging, especially Chest X-rays, has increasingly been a focus in AI research. This work uses multiple methods of AI to facilitate extraction of features, generate descriptions and provide interactive insights. The dataset used is collected by the National Institute of Health (NIH), and contains the Chest X-ray images and medical reports.ResNet-50 is used in this system for feature extraction, identifying important features from the Chest X-ray images. These extracted features, along with the medical reports, are provided to an LSTM model for training and generating descriptions for new images. Furthermore, the generated captions are converted into audio using the Google Text-to-Speech (gTTS) converter and saved in .mp3 format. The system also integrates the Gemini AI model to support interactive question-answering based on the generated descriptions, providing users with helpful insights in conversational form. These insights are later saved into a .txt file for future use. Evaluation is carried out using the BERT score.

Keywords: Chest X-ray Description, Medical Imaging, Deep Learning, ResNet-50, LSTM, Google Text-to-Speech (gTTS), Gemini AI, Multimodal AI, BERT Score.

1 Introduction

Medical imaging is important to healthcare, on the whole for the analysis, monitoring and management of a huge range of ailments. Among the maximum common imaging strategies in chest radiographs are frequently used to come across and compare conditions along with pneumonia, tuberculosis, lung most cancers, and different breathing illnesses. The interpretation of those photographs requires some knowledge, which often demands skilled radiologists who can identify and slender abnormalities. This manner is time-consuming and calls for get right of entry to to experienced specialists. This poses a sizeable mission in remote healthcare settings wherein such information might not be effectively to be had. By addressing those situations, this project proposes the improvement of an AI-powered system that automates chest X-ray interpretation, combining deep learning technologies for feature extraction, caption generation for clarity, converting caption to audio and getting further insights about the caption by including Gemini chatbot.

Many existing systems have looked at captioning images and audio output in order to improve accessibility. A common approach follows a CNN-based model framework such as ResNet50[16],VGG16, DenseNet[15] for feature extraction and for creating captions, GAN[21], LSTM[17][18][19] Bidirectional Long Short-Term Memory (BiLSTM)[20] are used, then uses some tool, usually Google Text-to-Speech (gTTS)[23][24][26] to synthesize the caption into audio. This approach specifically benefits those who are visually impaired. Some systems also include some type of chatbot (e.g. Gemini[25], Google Assistant) to allow for interactive question and answering based on the image content. While existing multimodal approaches have been successful with general object captioning and accessibility, their implementation in the medical realm, specifically for using chest x-ray images, remains novel. An effort has been made to develop a multimodal approach in the context of healthcare by creating meaningful descriptions of chest X-ray images. The descriptions emphasize the existence of possible medical conditions and are accompanied by a simple audio output to help people who are blind or have low vision. An interactive chatbot is part of the output as well, allowing a user to ask contextual questions based on the image. This increases usability and clinical relevance. The model personalized and fine-tuned to emphasize medically relevant vocabulary, and this adds value over a more generic model.

The rest of this paper is organized as follows. Section 2 reviews existing works related to image caption, speech synthesis, and chatbot integration in multimodal AI systems. Section 3 presents the proposed system architecture for chest X-ray description generation, audio conversion, and chatbot interaction. Section 4 discusses the experimental setup, results, and evaluation metrics. Finally, Section 5 concludes the paper and outlines possible future enhancements.

2 Related Works

Dr. K. Alice et al.[1] introduced an image-to-text-to-speech system to serve visually impaired or blind users. It employs a CNN-based object detection model, audio output via Google Text-to-Speech (gTTS), and includes a Gemini chatbot for follow-up contextual questions. The model generates precise image descriptions from structured data splits after training on custom datasets (vegetables/fruits and birds).It improves learning and access, and can be further extended to improve real-time processing and support for multiple languages.

Satti et al.[2] suggested encoder-decoder model for image captioning. ResNet50 has been used as an encoder and LSTM as a decoder. The model is trained using Flickr8k dataset and produces grammatical captions through the employment of pre-trained features and sequential context. The model achieved BLEU score of 0.96 and accuracy of 96.21%, which signifies excellent captioning performance.

A captioning framework using real and GAN-generated images was proposed by Md. Zakir Hossain et al.[3] to enhance the diversity of the dataset and the quality of the captions. Attention mechanisms were used to focus on important regions in images, resulting in better BLEU scores and more informative captions. This approach makes the captions more accessible to visually impaired individuals.

A model of image captioning using CNNs (VGG16 and ResNet50) for feature extraction and RNNs for caption generation was proposed by Sudhakar J. et al.[4] The captions would then be synthesized into audio using an application programming interface (API) from Google for text to speech (TTS). ResNet50 performed better than VGG16 on the Flickr8k dataset with 73% accuracy, offering a beautiful combination of vision, language, and speech for accessibility.

K. Bhargav Ram et al.[5] suggested a deep learning-based image captioning and voice synthesis system for enhancing accessibility and user experience. The model uses VGG16 for feature extraction of visuals and LSTM for captioning. It captures significant visual features like shape, texture, and object features. The captions are converted with gTTS for audio output. The model has a BLEU score of 52.7%, indicating its ability to create descriptive and coherent subtitles. The architecture combines computer vision, NLP, and speech synthesis to provide usage scenarios like multilingual speech generation, live subtitle generation, and social media automation.

Liu et al.[6] introduced a sentence-level image-language contrastive learning-based multi-grained radiology report generation model. The approach enhances the coherence and relevance of the generated medical reports. In contrast to CNN-LSTM models, the presented framework considerably improves consistency between image features and text descriptions, enhancing diagnostic precision and clinical utility.

Kumari et al.[7] have put forward Vision 360, an encoder-decoder model that utilizes CNNs for feature extraction and LSTMs for generating text. Though general-purpose in architecture, the model emphasizes the necessity of domain-specific training if the model is to be deployed on medical imaging, highlighting the role of radiological language modeling.

Sailaja et al.[8] examined CNN-RNN-based medical image captioning. The work emphasized the necessity of high-quality datasets and proper clinical vocabulary. It recommends that pre-trained medical language models can be used to improve captioning accuracy for diagnostics.

Ueda et al.[9] proposed a dynamic encoder-switching approach to improve caption generation through the use of adaptation across various types of medical images. This approach enhances contextual understanding and linguistic diversity, yielding improvements in radiology report generation through encoder customization.

Reddy et al.[10] discussed how TTS, STT, and deep learning could be applied to speech processing in a healthcare setting. The system allows for hands-free use, greater accessibility, and automated documentation by ASR. All these methods facilitate greater efficiency and accessibility by using AI-augmented diagnosis.

Arif et al.[11] proposed an attention-augmented CNN-LSTM model for image captioning. The attention mechanism enables the model to concentrate on the appropriate image patches, enhancing the description of important features like lesions or abnormalities. This method is especially useful in medicine for producing clinically significant captions.

Mohsan et al.[12] explored a transformer-based model incorporating Vision Transformers (ViTs) and language transformers for automated medical diagnosis via radiology report generation. Unlike CNNs, ViTs eschew spatial bias and more comprehensively encode global image context. Their work improved upon standard practices on the Indiana University Chest X-Ray dataset, which supports the potential of transformer-based models for automated medical diagnostics.

El Medhoune et al.[13] introduced a combined CNN-Transformer model for medical image captioning. The CNN captures local features and the Transformer uses self-attention to capture long-range dependencies. The model demonstrated enhanced accuracy, coherence, and clinical significance as a remedy for automated medical report generation compared to conventional CNN-RNN models.

Ravinder et al.[14] proposed a two-phase model, which first did feature extraction via YOLOv4 and an LSTM-based soft attention mechanism to generate the caption. The model was tested on the PEIR dataset, where it achieved a benchmark BLEU score of 81, offering a human-like performance in visual understanding through effective medical captioning.

2.1 Motivation

The works reviewed have demonstrated the efficacy in combining CNNs for feature extraction, LSTM for generating captions, and gTTS for audio output to improve accessibility. Based on these investigations, ResNet-50 to extract features from within the chest X-rays and LSTM to generate a health-related summary as text. In conclusion, gTTS is used to change every specific textual explanation. These specific descriptions were carefully converted directly into a definite audio format.Google Gemini is used in order to add understandings that strengthen, furthermore, the system's ability for assisting health experts to make elaborate as well as engaging reports.

3 Proposed System

The objective of this work is to generate descriptive captions for chest X-ray images with the potential to provide medical interpretations. In this work RESNET-50 is used to obtain the features from the X-ray images which are then used to generate descriptions using LSTM. The text generated from LSTM was then converted to speech using Google Text to Speech API (gTTS) to enhance accessibility. Google Gemini also uses the generated descriptions to help provide further interpretation of the description that is created. In summary this process enhances the medical diagnostic process with a detailed source of information and a chance to interact with the information.

3.1 Architecture

The architecture in Figure1 starts with chest X-ray images and reports as input, which are fed into ResNet-50 to extract features. ResNet-50 extracts vital visual information like anatomical detail and abnormalities. The features are then fed into an LSTM network, which produces a descriptive text from the processed data. The generated description is provided as output and further transformed into speech using gTTS for accessibility purposes. In addition, the system incorporates the Google Gemini API, which gives more insights through processing of the image content and adding relevant medical information to the generated description. This holistic method ensures better interpretation and accessibility of chest X-ray data.

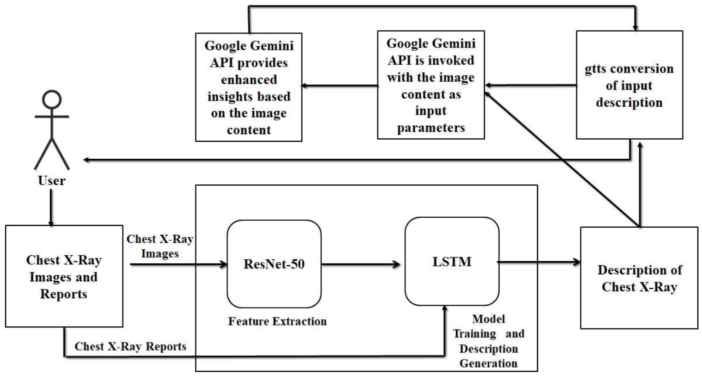

Google Gemini API provides enhanced insights based on the image content
Google Gemini API is invoked with the image content as input parameters
gtts conversion of input description
User
Chest X-Ray Images and Reports
Chest X-Ray Images
ResNet-50
LSTM
Description of Chest X-Ray
Feature Extraction
Model Training and Description Generation
Chest X-Ray Reports

Fig. 1 System architecture of image description model

3.2 Dataset

The dataset employed in this system is acquired from the National Institute of Health and consists of 7410 Chest X-ray images and respective Chest X-ray reports obtained from Indiana University. In the process of preparing the dataset, the reports are imported, and the important information from the findings and impressions sections is extracted and merged to form a complete caption for every image. These captions are then saved in a CSV file where every row is an individual image along with its corresponding caption to maintain organized data for model training.

3.3 Implementation and Training

The procedure starts with feature extraction of the visual features from the Chest X-ray images by ResNet-50, a deep convolutional neural network capable of extracting intricate image patterns. By passing the images to the ResNet-50 model, it detects the significant features like the primary anatomical structures and possible abnormalities, which are critical for the diagnosis of the conditions in the X-rays.These features are given in the form of high-dimensional vectors and are stored in .npy format in order to provide quick and easy access in the subsequent training process as they contain compressed and concise information regarding the image.

Subsequently, the respective captions describing the content of the images are read from a CSV file. These captions are used as target output when training the model. The features from the extracted image and their captions are employed to train an LSTM (Long Short-Term Memory) network, a form of recurrent neural network (RNN) to process and produce sequences, such as text. The LSTM learns the mapping from the visual features (from the X-ray) to their respective textual descriptions, producing captions from the features extracted through ResNet-50. In order to make sure that the model learns well without overfitting the training data, it is trained for 20 epochs, during which it gets a chance to improve its predictions progressively. Every epoch consists of feeding the features and captions into the network and optimizing the weights in order to reduce the prediction error.

After training is finished, the model is saved in.keras format and is thus ready for use in the future, for example, generating captions for new X-ray images. The training history is also saved, which contains important statistics like accuracy and loss. Accuracy reflects how good the model is at producing captions that are equal to the ground truth, whereas loss measures the difference between predicted and actual captions. These metrics are useful for evaluating the model’s performance and for visualizing how well the model learned during training.

3.4 Description and Audio Generation

The process starts with loading the trained LSTM model to produce coherent text descriptions of Chest X-ray conditions. The trained LSTM model, which was previously trained on the extracted image features and related captions, is utilized to make inferences based on the features of new, unseen Chest X-ray images. When a new image is uploaded, the model generates a text description that notes significant abnormalities or findings on the X-ray and purports to offer a diagnosis or diagnosis-related findings.

Lastly, the generated text description is converted into speech using gTTS (Google Text-to-Speech), which is a Google service. With this integration, generated captions are able to be spoken in a human-sounding voice, so information becomes accessible for those who prefer the information be provided as an audio output, as opposed to viewing the text. The text is transformed by gTTS into an audio file (commonly MP3 format), and that file is then playable, permitting hands-free.

3.5 Gemini Insights

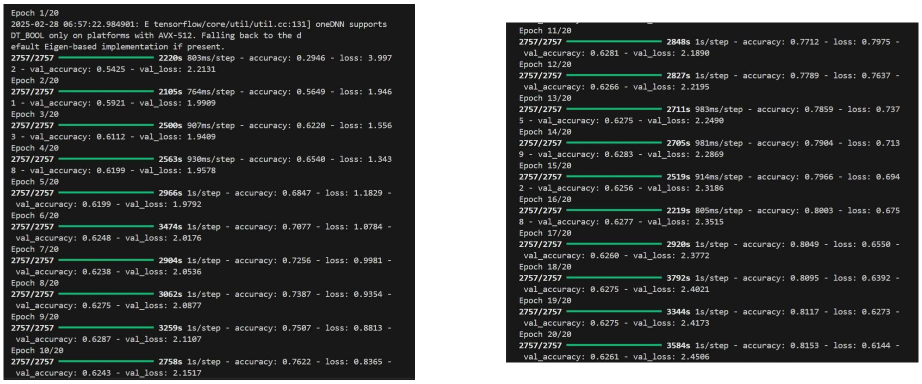
The system incorporates the Gemini API to improve the quality of description generated and offer additional information. Once the LSTM model creates a written description of the Chest X-ray, the description is passed on to the Gemini AI chatbot for improvement. The chatbot improves the content, clarifying, correctness, and

medical relevance. In addition to completing the description, the Gemini AI chatbot offers the ability to generate ideas based on the text as well, such as possible diagnosis, medical conclusions or recommendations.
The chatbot is also capable of answering user questions, returning descriptive explanations or supporting information to the generated description. For convenience so that such insights can be accessed in the future, the system stores the improved description and concise insights in the form of a.txt file. The integration ensures that the output is informative as well as well-presented, adding value to the overall comprehension of the Chest X-ray results.

4 Experimental Results

4.1 Training using LSTM Model

The training was conducted using a structure of an LSTM (Long Short-Term Memory) model, which is ideally suited for tasks involving sequential data, like text. The training included a tokenization of the image captions, creation of a vocabulary and preparation of the data was completed by converting the words into sequences of numbers.For the LSTM model, an embedding layer for our word representation, an LSTM layer of 256 units, and a dropout layer to avoid overfitting were employed. The model was trained for 20 epochs with a batch size of 64 with a categorical cross-entropy loss function and the Adam optimizer. The training of images is shown in Figure 2. Upon training, the model obtained a training accuracy of 0.8153, indicating effective learning in text generation and informative text description generation from the Chest X-ray images.



```
Epoch 1/20
2025-02-28 06:57:22.984901: E tensorflow/core/util/util.cc:131] oneDNN supports
DT_BOOL only on platforms with AVX-512. Falling back to the d
efault Eigen-based implementation if present.
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2220s 803ms/step - accuracy: 0.2946 - loss: 3.997
2 - val_accuracy: 0.5425 - val_loss: 2.2131
Epoch 2/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2105s 764ms/step - accuracy: 0.5649 - loss: 1.946
1 - val_accuracy: 0.5921 - val_loss: 1.9909
Epoch 3/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2500s 907ms/step - accuracy: 0.6220 - loss: 1.556
3 - val_accuracy: 0.6112 - val_loss: 1.9409
Epoch 4/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2563s 930ms/step - accuracy: 0.6540 - loss: 1.343
8 - val_accuracy: 0.6199 - val_loss: 1.9578
Epoch 5/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2966s 1s/step - accuracy: 0.6847 - loss: 1.1829 -
 val_accuracy: 0.6199 - val_loss: 1.9792
Epoch 6/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 3474s 1s/step - accuracy: 0.7077 - loss: 1.0784 -
 val_accuracy: 0.6248 - val_loss: 2.0176
Epoch 7/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2904s 1s/step - accuracy: 0.7256 - loss: 0.9981 -
 val_accuracy: 0.6238 - val_loss: 2.0536
Epoch 8/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 3062s 1s/step - accuracy: 0.7387 - loss: 0.9354 -
 val_accuracy: 0.6275 - val_loss: 2.0877
Epoch 9/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 3259s 1s/step - accuracy: 0.7507 - loss: 0.8813 -
 val_accuracy: 0.6287 - val_loss: 2.1107
Epoch 10/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2758s 1s/step - accuracy: 0.7622 - loss: 0.8365 -
 val_accuracy: 0.6243 - val_loss: 2.1517
```

```
Epoch 11/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2848s 1s/step - accuracy: 0.7712 - loss: 0.7975 -
 val_accuracy: 0.6281 - val_loss: 2.1890
Epoch 12/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2827s 1s/step - accuracy: 0.7789 - loss: 0.7637 -
 val_accuracy: 0.6266 - val_loss: 2.2195
Epoch 13/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2711s 983ms/step - accuracy: 0.7859 - loss: 0.737
5 - val_accuracy: 0.6275 - val_loss: 2.2490
Epoch 14/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2705s 981ms/step - accuracy: 0.7904 - loss: 0.713
9 - val_accuracy: 0.6283 - val_loss: 2.2869
Epoch 15/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2519s 914ms/step - accuracy: 0.7966 - loss: 0.694
2 - val_accuracy: 0.6256 - val_loss: 2.3186
Epoch 16/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2219s 805ms/step - accuracy: 0.8003 - loss: 0.675
8 - val_accuracy: 0.6277 - val_loss: 2.3515
Epoch 17/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 2920s 1s/step - accuracy: 0.8049 - loss: 0.6550 -
 val_accuracy: 0.6260 - val_loss: 2.3772
Epoch 18/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 3792s 1s/step - accuracy: 0.8095 - loss: 0.6392 -
 val_accuracy: 0.6275 - val_loss: 2.4021
Epoch 19/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 3344s 1s/step - accuracy: 0.8117 - loss: 0.6273 -
 val_accuracy: 0.6275 - val_loss: 2.4173
Epoch 20/20
2757/2757 ━━━━━━━━━━━━━━━━━━━━ 3584s 1s/step - accuracy: 0.8153 - loss: 0.6144 -
 val_accuracy: 0.6261 - val_loss: 2.4506
```

Fig. 2 Training using LSTM

Figure 3 shows the accuracy of the testing and validation of the model over 20 epochs. The blue line shows the accuracy of the testing which goes up at a constant rate, suggesting the model is learning from the testing data. The orange line shows the accuracy of the validation which plateaued early during the training process, and remained relatively unchanged over the epochs. Figure 4 shows the training and validation loss curves over 20 epochs. The blue line indicates training loss, which improves

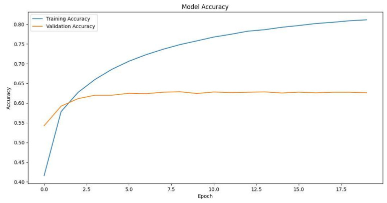


Fig. 3 Model Accuracy using LSTM

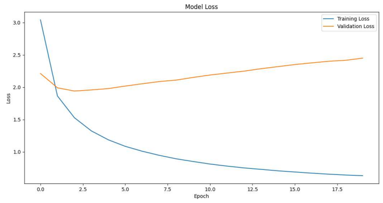


Fig. 4 Model Loss using LSTM

consistently across the epochs, reflecting that the model is progressively enhancing its performance on the training set. Conversely, orange line marking validation loss starts to decrease but begins to rise again after some number of epochs.

Figure 5 depicts the histograms for the loss and accuracy values between training and validation across epochs. To the left, the histogram of loss values indicates that training loss is primarily skewed towards low values, showing how the model is capable of reducing error within the training set. On the contrary, the validation loss values are scattered across a wider range, showing that the model finds it difficult to perform at the same level on unknown data. On the right-hand side, the histogram of accuracy values shows that accuracy during training is spread towards upper accuracy ranges, while validation accuracy values are concentrated around a lower, narrower range.

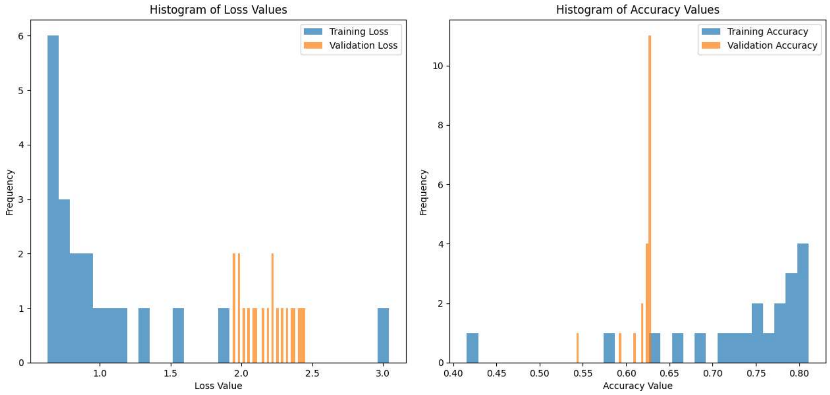


Fig. 5 Histogram plots of training and validation loss (left) and accuracy (right) values across epochs, illustrating model performance distribution and generalization capability.

4.2 Word Length Progression in Reference Vs. Candidate

The graph in Figure 6 depicts the development of the lengths of words across various word positions for a particular text. The Reference (blue solid line) and the Candidate (orange dashed line) as separate representations of the text, indicating a divergence in lengths of words for each position. The x-axis indicates word position while the y-axis indicates word length. The Reference has a moderate level of variation, with the Candidate showing more variation, with moments that had large spikes in word length. These variations offer an analysis for structural difference between layers of text composition and word distribution.

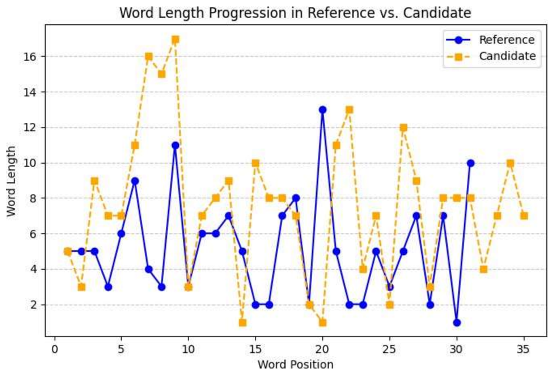


Fig. 6 Word length progression comparison between reference and candidate sentences across word positions, highlighting variation in linguistic complexity.

4.3 BERT Score

BERT Score is a measure that judges the levels of similarity between two texts based on model embeddings obtained from BERT (Bidirectional Encoder Representations from Transformers) [22]. Classic measures like BLEU or ROUGE simply verify exact matches at the word level. BERT Score goes a step further and adds an expected semantic meaning through comparing word embeddings rather than counting word overlaps. BERT Score accomplishes this by first converting each word in the reference text and in the candidate text into high dimensional vectors using a pre-trained BERT model. Then for each word in the candidate text, it measures the similarity of that word with its most similar words in the reference text using cosine similarity.

The Table 1 shows the BERTScore for the generated description as well as the candidate key.Based on the measures of similarity for each word, BERT Score calculates a Precision score (the ratio of the candidate text that overlaps with the reference text), Recall score (how closely the representation of the reference text overlaps with the candidate text), and an F1 score (the harmonic mean combining recall and precision). For example, consider the following:

- **Reference:** ”Chest X-ray shows the cardiomediastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no focal areas of consolidation. Cholecystectomy clips are present. Small

T-spine osteophytes. There is biapical pleural thickening, unchanged from prior. Mildly hyperexpanded lungs."

- **Candidate:** "Chest X-ray reveals COPD, lingular scarring, and atherosclerosis. Possible pneumonia or other infection. Significant cardiac history (CABG, valvuloplasty). Osteopenia and spinal changes noted."

For this pair of texts, a Precision of 0.8451, Recall of 0.8224, and F1 score of 0.8367 were obtained. These values show that the candidate text is quite closely aligned with the reference text.Table 1 shows the detailed BERTScore values for this reference-candidate pair.

Table 1
BERTScore values

Metric	Score
Precision	0.8451
Recall	0.8224
F1	0.8336

5 Conclusion

Utilizing deep learning models, such as ResNet-50 for image classification and LSTMs for text generation, in combination with speech synthesis interfaces such as Google Text-to-Speech (gTTS), is a remarkable advancement in creating accessible, easy to use applications capable of being able to describe images. These applications essentially transform visual information into descriptive text to audio information, providing the means for users affected by visual impairment, to experience and interact with their environment more cognitively. The ability to audibly discern accurate, coherent descriptions of images helps create a sense of independence for visually impaired individuals and improves the quality of their lives by allowing them to participate more easily in many contexts that may otherwise rely on viewing visual content, such as reading books, social media interactions, or even diagnostics in a clinical setting.

The future of image description systems is full of potential for further development and improvement. A clear direction is to advance accuracy and coherence of descriptions through exploring contemporary models, like Vision Transformers (ViT), or the CLIP model, which can draw similar conclusions based on both visual and textual data. These models have the potential to create descriptions that are more detailed and contextually aware. We should also consider developing these systems to be multilingual, to respond to audiences that may not speak English, or who may come from different traditions and customs. Further development of systems that can provide image-to-speech description systems in real time may be appropriate for different types of visual content, whether in live video streams, in an educational space, or even in the daily life experience.

References

[1] Dr. K. Alice, Ayush Srivastava, and Vikram Saurav. (2024) ’Image Insight - Transforming Images into Narratives for Seamless Learning’, *International Conference on Intelligent Systems for Cybersecurity (ISCS)*

[2] Satti, Satish Kumar, Goluguri N. V. Rajareddy, Prasad Maddula, and N. V. Vishnumurthy Ravipati. (2023). ’Image Caption Generation using ResNET-50 and LSTM’, *International Conference on Intelligent Systems for Cybersecurity (ISCS)*.

[3] Hossain, Md. Zakir, Ferdoous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. (2021). ’Text to Image Synthesis for Improved Image Captioning’, *IEEE Access*.

[4] Sudhakar, J., Viswesh Iyer, V., and Sharmila, S. T. (2022). ’Image Caption Generation Using Deep Neural Networks’, *International Conference for Advancement in Technology (ICONAT)*, Goa, India.

[5] Bhargav Ram, K., Venkatesh, B., Sala Pooja Sai Sree, and Chunduru Anilkumar. (2023). ’Image Caption and Speech Generation Using LSTM and GTTS API’, *ICAISS 2023*.

[6] Liu, A., Guo, Y., Yong, J.-H., and Xu, F. (2024). ’Multi-Grained Radiology Report Generation With Sentence-Level Image-Language Contrastive Learning’, *IEEE Transactions on Medical Imaging*, 43(7).

[7] Kumari, A., Chauhan, A., and Singhal, A. (2022). ’Vision 360: Image Caption Generation Using Encoder-Decoder Model’, *International Conference on Cloud Computing, Data Science & Engineering (Confluence)*.

[8] Sailaja, M., Harika, K., Sridhar, B., Singh, R., Charitha, V., and Rao, K. S. (2022). ’Image Caption Generator using Deep Learning’, *ASSIC*.

[9] Ueda, A., Yang, W., and Sugiura, K. (2023). ’Switching Text-Based Image Encoders for Captioning Images With Text’, *IEEE Access*.

[10] Madhusudhana Reddy, V., Vaishnavi, T., and Pavan Kumar, K. (2023). ’Speech-to-Text and Text-to-Speech Recognition using Deep Learning’, *ICECAA*.

[11] Arif, M. H. (2024). ’Image to Text Description Approach based on Deep Learning Models’, *Bilad Alrafidain Journal for Engineering Science and Technology*.

[12] Mohsan, M. M., Akram, M. U., Rasool, G., Alghamdi, N. S., Baqai, M. A. A., and Abbas, M. (2024). ’Vision Transformer and Language Model-Based Radiology Report Generation’, *IEEE Access*.

[13] El Medhoune, H., My Abdelouahed, S., Zouitni, C., and Aarab, A. (2024). ’Combining CNN and Transformer for Enhancing Medical Image Captioning’, *ICDS*.

[14] Ravinder, P., and Srinivasan, S. (2024). ’Automated Medical Image Captioning with Soft Attention-Based LSTM Model Utilizing YOLOv4 Algorithm’, *Journal of Computer Science*, 20(1), pp. 52–68.

[15] Boulesnane, A., Mokhtari, B., Segueni, O. R., and Segueni, S. (2024). ’Uterine Ultrasound Image Captioning Using Deep Learning Techniques’, *arXiv preprint arXiv:2411.14039*.

[16] Kanimozhiselvi, C. S., V., K., P., K. S., and S., K. (2022). ’Image Captioning Using Deep Learning’, *ICCCI*, pp. 1–7.

[17] Azhar, I., Afyouni, I., and Elnagar, A. (2021). ’Facilitated Deep Learning Models for Image Captioning’, *CISS*, pp. 1–6.

[18] Li, S., and Huang, L. (2021). ’Context-based Image Caption using Deep Learning’, *International Conference on Signal Processing (ICSP)*, pp. 820–823.

[19] Satyanarayana Reddy, S. S., Kumar, A., and Jyaram, M. (2021). ’An Image Sentence Generation Based on Deep Neural Network Using RCNN LSTM Model’, *ICIIP*, pp. 364–368.

[20] Mahalakshmi, P., and Fatima, N. S. (2022). ’Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Techniques’, *IEEE Access*, 10, pp. 18289–18297.

[21] Yang, M., Li, Y., Zhang, Z., Zhu, J., and Sun, M. (2020). ’An Ensemble of Generation- and Retrieval-Based Image Captioning With Dual Generator Generative Adversarial Network’, *IEEE Transactions on Image Processing*, 29, pp. 9627–9640.

[22] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). ’BERTScore: Evaluating Text Generation with BERT’, *arXiv preprint arXiv:1904.09675*.

[23] Ahmed, S., Khan, M. A., and Saleem, Y. (2023). ’A Novel Framework for Automatic Caption and Audio Generation’, *Procedia Computer Science*, 218, pp. 1443–1450. https://doi.org/10.1016/j.procs.2022.12.180

[24] Khan, A., Shaikh, Z., Shaikh, A., and Shaikh, N. (2022). ’Image to Speech Using CNN, LSTM and gTTS’, *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 10(6), pp. 1235–1239.

[25] Imran, M., and Almusharraf, N. (2024). ’Google Gemini as a Next Generation AI Educational Tool: A Review of Emerging Educational Technology’, *Smart Learning Environments*, 11(22).

[26] Sheenu, K. R., and Santhosh, M. (2024). ’Image Captioning with Audio Generation Using Vision Transformer and gTTS’, *International Journal of Creative Research Thoughts (IJCRT)*, 12(4), pp. e149–e156.