TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

ADVANCED COLLEGE OF ENGINEERING AND MANAGEMENT

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

KALANKI, KATHMANDU

[CT755]

A MAJOR PROJECT REPORT ON

"ATTENTION BASED AUTOMATED RADIOLOGICAL REPORT GENERATION USING MULTIMODEL ARCHITECTURE"

Submitted By:

Ankit Chhetri	**ACE077BCT012**
Bhaskar Subedi	**ACE077BCT023**
Biplov Belbase	**ACE077BCT031**

Project Supervisor:

Er. Utsav Pokhrel

A Major Project Final report submitted to the Department of Electronics and Computer Engineering in the partial fulfillment of the requirements for degree of Bachelor of Engineering in Computer Engineering

Kathmandu, Nepal

2025/03/07

“ATTENTION BASED AUTOMATED RADIOLOGICAL REPORT GENERATION USING MULTIMODEL ARCHITECTURE”

Submitted by:

Ankit Chhetri	ACE077BCT012
Bhaskar Subedi	ACE077BCT023
Bilov Belbase	ACE077BCT031

Supervised by:

Er. Utsav Pokhrel

A MAJOR PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER OR ELECTRONICS & COMMUNICATION ENGINEERING

Submitted to:

"DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING"

ADVANCED COLLEGE OF ENGINEERING AND MANAGEMENT

Balkhu, Kathmandu

2025/03/07

LETTER OF APPROVAL

The undersigned certify that they have read and recommended to the Institute of Engineering for acceptance, a project report entitled **"ATTENTION BASED AUTOMATED RADIOLOGICAL REPORT GENERATION USING MULTIMODEL ARCHITECTURE"** submitted by:

Ankit Chhetri	ACE077BCT012
Bhaskar Subedi	ACE077BCT023
Biplov Belbase	ACE077BCT031

In the partial fulfillment of the requirements for the degree of Bachelor's Degree in Computer Engineering.

……..

Project Supervisor

Er. Utsav Pokhrel

Department of Electronics and Computer Engineering

……..

Head of Department

Senior Asst. Prof. Prem Chandra Roy

Department of Electronics and Computer Engineering

2025/03/07

COPYRIGHT

ACKNOWLEDGMENT

We would like to express our profound gratitude and deep regards to our respected supervisor, Associate Professor **Er, Utsav Pokhrel**, for his insightful advice, motivating suggestions, invaluable guidance, help, and support in the successful completion of this project and also for his constant encouragement and advice throughout our Bachelor's program.

We express our deep gratitude to **Er. Prem Chandra Roy**, Head of the Department of Electronics and Computer Engineering, **Er. Dhiraj Pyakurel,** Deputy Head, Department of Electronics and Computer Engineering, **Er. Laxmi Prasad Bhatt,** Academic Project Coordinator, Department of Electronics and Computer Engineering, for their support, co-operation, and coordination.

The in-time facilities provided by the department throughout the Bachelor's program are also equally acknowledgeable.

We would like to convey our thanks to the teaching and non-teaching staff of the Department of Electronics & Communication and Computer Engineering, ACEM for their invaluable help and support throughout Bachelor's Degree. We are also grateful to all our classmates for their help, encouragement, and invaluable suggestions.

Finally, yet more importantly, we would like to express our deep appreciation to our grandparents, parents, and siblings for their perpetual support and encouragement throughout the Bachelor's degree period.

Project Members:

Ankit Chhetri	ACE077BCT012
Bhaskar Subedi	ACE077BCT023
Biplov Belbase	ACE077BCT031

ABSTRACT

In Nepal, where there are a limited number of radiologists and healthcare resources especially in rural areas where the doctor-patient ratio is very low, AI comes in handy. The use of AI in medical diagnosis is transformable and has great potential. This study presents Attention-Based Automated Radiological Report Generation Using Multimodel Architecture, which proposes or addresses the need to generate accurate radiological reports from chest X-ray images along with the shortage of radiologists in Nepal. Multimodel framework,(CheXNet) is pre-trained model for image feature extraction with gated recurrent units (GRU) decoder which is enhanced by Bahdanau attention mechanism to synthesize context-aware clinical reports. The preprocessing technique involves employing XML parsing and data cleaning, image normalization, and tokenization. Indiana University X-ray dataset is trained where the system processes input images through CheXNet for the extraction of hierarchical features, while the GRU in addition to the attention decoder aligns both visual and textual data for generating context-aware descriptions. A beam search optimization strategy is used which ensures coherent and clinically accurate reports. BLEU score metrics are validated for text similarity. From the study, we found that it can reduce diagnosis delays, reduce workload, and improve healthcare access. This study uncovers the transformation potential of using multimodel architectures to revolutionize diagnostic healthcare, particularly in low-resource environments or rural areas.

Keywords: *Attention Mechanisms, Automated Radiological Report Generation, CheXNet, Healthcare Accessibility, Multimodel Deep Learning*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF EQUATIONS

Title **Page**

LIST OF ABBREVIATIONS

BLEU	:	Bilingual Evaluation Understudy
CNN	:	Convolutional Neural Network
CVT	:	Convolutional Vision Transformer
CXR	:	Chest X-Ray
GRU	:	Gated Recurrent Unit
LLM	:	Large Language Model
LSTM	:	Long Short-Term Memory
NLP	:	Natural Language Processing
PNG	:	Portable Network Graphics
RNN	:	Recurrent Neural Network
ViT	:	Vision Transformer

CHAPTER 1: INTRODUCTION

1.1 Background

Good health is the first requisite of happiness and success in the life of people. People with sound physical and mental health have better productivity. Nepal's constitution has recognized right to health as the fundamental right of the citizens. It stipulates that the people have rights to free basic health services, emergency health services and access to information about health. Still, many Nepalese have no access to affordable and basic health facilities because state-run health centers lack sufficient infrastructure and human resources. This is a reason why the private hospitals operating in many parts of country cater to around 80 per cent health services, with the government hospitals providing only 20 per cent. This is really a matter of serious concern. The situation in the far-flung areas is far worse. Doctors often hesitate to work in the remote areas owing to low pay and geographical difficulties. The doctor-patient ratio in Nepal is 1:850 in the Kathmandu Valley and 1:150,000 in the rural areas. The World Health Organization recommends a doctor-patient ratio of 1:1,000. There are 32,218 registered doctors and 72,550 registered nurses in the country. Of them, only around 15,000 have been working in government health facilities [1]. Although public health service is state responsibility, its privatization has increased its cost beyond the capacity of majority of Nepalese. They have to spend a huge amount of their savings on the medical treatment and purchase of expensive medicines. The added financial burden is painful for the poor.

1.2 Motivation

Radiological report generation is crucial for the diagnosis of different diseases and injuries, with the help of X-ray images, different health problems can be analyzed. Manual report generation can be tedious and prone to variability among radiologists. In the context of Nepal, where there is an extremely low ratio of doctors to patients and the access to healthcare is not easily accessible in rural areas, the use of automated intelligent report generation can be beneficial in many ways as it saves time and effort and improves patient outcomes. The success of attention-based models in NLP and computer vision makes them promising for generating accurate and context-aware radiology reports. This project targets the enhancement of early and accurate diagnosis of chest-related diseases. Aside from all these pros, there lie different challenges as well, sometimes, AI-based approaches fail to capture fine-grained details in medical images. By deploying portable X-ray units in rural areas and conducting various

training programs for healthcare workers, we can significantly improve healthcare accessibility and reduce sole dependency on medical specialists.

1.3 Problem Statement

Manual report generation is a crucial but time-consuming task that requires expert interpretation of medical images. Accurate chest X-rays are very vital to predicting different chest-related diseases. In rural areas of Nepal, where there are a limited number of specialists and a limited number of resources, fixing these problems is super important and crucial. Traditional automated methods such as CNN often struggle to capture complex dependencies and generate accurate reports. Attention based multimodel architecture and attention mechanisms can be used to accurately generate reports and improve efficiency, reduce workload, and improve patient care.

1.4 Objective

The general and specific objectives of the project are listed below:

1.4.1 General Objectives

The general objective of the project is:

- Enhance the process of chest X-ray report generation and diagnosis from medical images.
- Improving Efficiency and Consistency in the reporting process.

1.4.2 Specific Objectives

The specific objectives of the project are:

- Integration of Mutimodel Architecture.
- Using Attention Mechanism to improve report accuracy.
- Reducing radiologist workload.
- Ensuring clinically relevant reports.

1.5 Significance of the study

Our project is important for many reasons; the major impact it will have is in helping to improve healthcare access and quality within Nepal. We can do this by improving the process of generating chest X-ray reports and diagnoses, especially in rural and under-resourced settings, ensuring timely access to diagnostic tools to individuals that aid in the earlier detection and treatment of chest-related problems. It can also reduce the cost associated with manual report

generation. It helps not only in rural areas but also can be implemented at any place, under any conditions, if the problem arises or if the arrival of specialists is delayed. This improves the overall healthcare system, increases accuracy, speeds up the diagnosis, and reduces human error. It supports specialists, resulting in better outcomes. It also reduces the burden. With the advancement in AI, the healthcare field is being transformed.

CHAPTER 2: LITERATURE REVIEW

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 [2].

Medical report generation process proposed a CNN-RNN architecture to generate captions for images [3]. These results were however too simple and lacked details. As more work was done in the field, attention was introduced with model's attention with RNN and CNN [4].

CNNs have proven to be extremely effective in image classification, object detection, and segmentation tasks, and have been utilized in a variety of applications including as self-driving cars, medical diagnostics, and facial recognition. CNNs were shown to be capable of classifying view orientations of chest radiographs with excellent accuracy [5].

Chest radiographs were used to construct a CNN-based model for the automated classification of pulmonary tuberculosis, obtaining high performance and indicating the promise for deep learning in the disease detection [6]. CNN was used to detect and classify abnormalities on chest radiographs, with good sensitivity and specificity [7].

The attention mechanism has revolutionized neural network architectures by enabling models to focus on the most relevant features of input data. In vision and language tasks, such as radiological report generation, attention enhances the synergy between convolutional layers for spatial understanding and recurrent networks for sequence generation. This integration improves both interpretability and accuracy in medical imaging applications. [8]

For report generation, Jing et al. built a multi-task learning framework, which consists of co-attention and a hierarchical LSTM that predicts the tags, localizes the regions with abnormalities, and uses these for the radiological image annotation and report paragraph

generation. They performed their experiments on two publicly available datasets: IU CXR and PEIR Gross [9].

Cho et al.'s invention of the Gated Recurrent Unit (GRU) makes a remarkable stamp in recurrent neural networks (RNNs), by providing a better substitute for Long Short-Term Memory (LSTM) networks without its performance being disturbed. In order to improve the learning of long-term dependencies, the GRU was created to solve the vanishing gradient issue in conventional RNNs. GRUs are capable of process sequential data such as text, speech and time series data which are similar to LSTM. Also GRU is performing well in natural language processing task such as speech recognition, text to voice, language translation, etc.[10]

Bahdanau et al. proposed the additive attention mechanism that dynamically enables the decoder to focus on relevant parts of the input sequence. This approach improves handling of long sequences by computing alignment scores and generating context vectors for better predictions [11].

CHAPTER 3: REQUIREMENT ANALYSIS

3.1 Functional Requirements

Image Input: Our system should accept chest X-ray images in standard PNG format.

Image Preprocessing: Preprocess input images to normalize for analysis.

Feature Extraction: Utilize CheXNET to extract relevant features from X-ray images.

Text Generation: Employ GRU with attention to generate descriptive medical reports based on extracted features.

Report Formatting: Ensure generated reports are formatted professionally and include necessary medical terminology.

3.2 Non-functional Requirements

Performance: Our system should be capable of processing X-ray images and generating reports within a reasonable timeframe.

Security: Implement robust security measures to protect patient data and ensure compliance with healthcare privacy regulations.

Usability: A user-friendly interface that is intuitive for both radiologists and healthcare professionals to interact with.

Training and Support: Offer comprehensive training and support resources for end-users to effectively utilize the system.

3.3 System Requirements

3.3.1 Hardware Requirements

1. **GPUs**: High-end GPUs for training deep learning models.
2. **Memory**: Minimum 8 GB RAM to manage datasets and support GPU operation.
3. **Storage**: At least 256 GB of SSD storage to store datasets and model checkpoints efficiently.

3.3.2 Software Requirements

1. **Deep Learning Frameworks**:
    - **Keras**: Deep learning framework.
2. **Libraries and Dependencies**:
    - **NumPy**: For numerical operations.
    - **Pandas**: For data manipulation and analysis.

- **NLTK**: For natural language processing tasks.

3. **Pre-trained Models and Tokenizers**:
    - **CheXNet** is based on DenseNet-121 but customized for medical image analysis
4. **CUDA**: For GPU acceleration.
5. **Jupyter Notebooks**: For interactive development and testing.

CHAPTER 4: METHODOLOGY

The methodology for our project consists of five main phases: data collection, preprocessing, feature extraction, report generation using GRU with attention, and evaluation. Here is a detailed breakdown of each step:

4.1 Data Collection

To develop a robust system for automated medical report generation, the data of Indiana University of X-ray and their report was used, ensuring diversity and comprehensiveness:

- **Data Source:** Indiana University(X-ray images and Radiology reports).
    - ***Chest X-ray*** –There are 7,471 images in .png file format which contain front view and lateral view of each patient's chest.
    - ***Radiology Report*** –There are about 3955 patient's text reports available in .XML format.

The data set contains chest X-ray images and radiology text reports. Each image has been paired with four captions such as Comparison, Indication, Findings and Impressions that provide clear descriptions of the salient entities and events. The finding caption gives maximum information present in images. The goal of this case study is to predict the findings of the medical report attached to the images.

4.2 Data Preprocessing

4.2.1 XML Parsing Creating Data Points

In this section we passed the raw XML data then parsed and structured as data points. Then the data points were stored in csv files for future model requirements.

This XML file has a lot of information related to patients such as: image_id, text captions like — comparison, indication, findings, impression etc. The findings feature from these files were extracted and considered them as reports because they are more useful for the medical report. The image_id from these files were also extracted to get the x-rays corresponding to each report.

4.2.2 EDA and Data Preprocessing

In this phase the text data was preprocessed to remove unwanted tags, texts, punctuation and numbers. Performed basic decontractions i.e. words like won't, can't and so on

converted to will not, can't and so on respectively. Following process was performed to handle cells with empty value and NaN value.

- Empty cells in the image name column were dropped.
- Text data having empty value and NaN were replaced with "No Impression".

After the data preprocessing step, we got total of 3851 rows present in the final data points.

4.3 Structure Data

There are only two image types: Front and Lateral, but each patient has multiple x-rays associated with them. The maximum number of images associated with a report were 5 while the minimum is 0. The highest frequency of being associated with a report were 2 images.
The data points that have 1,3,4,5 images were handled using the following approach.

- The data points were limited to two images per data point. When there were five images, they were split as 4+1 (all images + last image) to create four data points. The last image was ensured to be Lateral if the remaining images were Frontal. Multiple images were converted into two images by implementing the following steps.

1. If 5 images then total 4 data points were created
   - 1^{st} image + 5^{th} image
   - 2^{nd} image + 5^{th} image
   - 3^{rd} image + 5^{th} image
   - 4^{th} image + 5^{th} image
2. If 4 images then total 3 data points were created
   - 1^{st} image + 4^{th} image
   - 2^{nd} image + 4^{th} image
   - 3^{rd} image + 4^{th} image
3. If 3 images then total 2 data points were created
   - 1^{st} image + 3^{rd} image
   - 2^{nd} image + 3^{rd} image
4. If 2 images then it is according to the requirement
5. At last, if 1 image, it was replicated to make it 2.

4.4 Baseline Model [Encoder-Decoder]

A sequence-to-sequence model is a deep learning model that takes a sequence of items (in this case, features of an image) and outputs another sequence of items (reports).

The encoder processes each item in the input sequence, it compiles the information it captures into a vector called the context. After processing the entire input sequence, the encoder sends the context over to the decoder, which begins producing the output sequence item by item. The steps followed are following:

4.4.1 Add Token in text data

After creating new data points from existing data points, <start> and <end> token were added into text data and the decoder input and output were prepared. The **start** and **end** tokens were special tokens added at the beginning of the sentence and end of the sentence respectively to help the model learn the sentences structure.

4.4.2 Tokenization

Since machines only understand numerical values, text data cannot be directly fed into deep learning and machine learning models. Therefore, the text data was converted into numerical data using a Tokenizer. Tools provided by the TensorFlow deep learning library were used to perform this operation.

4.4.3 Image Feature

The transfer learning was used for converting the image to a feature vector, and the pre-trained CheXnet competition model weights were utilized.

The CheXnet model, a DenseNet121-layered architecture was trained on 112,120 chest X-ray images for the classification of 14 diseases. The weights of that model were loaded, and the image was passed through that model while ignoring the top layer.

As two X-rays correspond to each patient, each image was preprocessed according to the input of the DenseNet121 model, and the model's output for both images were concatenated at the end.

4.4.4 Encoder-Decoder Architecture

Then the concatenated image tensors were passed to encoder, image features were fed into a dense layer having 512 neurons and then a dropout was added layer for tuning. In the decoder part, an embedding layer, a dropout layer, and an LSTM layer are included. The input sequence, i.e., the encoder output is passed to the embedding layer. Then, the output of this layer is passed to the dropout layer and finally fed into the LSTM

LSTM layer is Long Short-Term Memory networks — are a special kind of RNN, capable of learning long-term dependencies.

The outputs of the encoder and decoder were then added using the Add layer of Keras. This output was passed to a Time Distributed Dense layer. The Time Distributed Dense layer was applied at the end because the output is sequential, and it needed to be applied to every temporal slice of the output.

Note: The dropout layers are added only for fine tuning the model.

4.4.5 Model Inference

In the inference stage, the argmax based Greedy search was used to find the output sentence. Greedy search is a vanilla implementation for generation of output which is selecting a single word with maximum probability from the entire vocabulary.

4.5 CheXNet Architecture (DenseNet-121)

The CheXNet model is an adaptation of the DenseNet-121 architecture, which is known for its efficient feature reuse and compact design. Here is the breakdown of the DenseNet-121 architecture.

1. Input layer
    - Input shape: 224×224×3 (Resized chest X-ray image).
    - Preprocessing: Normalization and resizing.
2. Convolution and Pooling:
    - Initial convolution: 7×7 kernel with stride 2 and padding.
    - Max pooling: 3×3 kernel with stride 2.

Convolution operation:

$$Y_{(i,j)} = \sum_{i=1}^{K_h-1} . \sum_{j=1}^{K_w-1} X_{(i+m,j+n)} . W_{(m,n)} + B \quad (1)$$

- $Y_{(i,j)}$: The value at position (i,j) in the output feature map.
- $X_{(i+m,j+n)}$: The value at position (i+m,j+n) in the input feature map (i.e., the region covered by the kernel at position (i,j).
- W(i,j): The weight of the kernel at position (m,n).
- B: The bias term added to the result of the convolution.

- K_h K_w: The height and width of the kernel.
- m, n: Indices iterating over the kernel dimensions.

The output size of a convolution layer can be calculated as:

$$\text{Output Size (H, W)} = [\frac{(\text{Input Size (H, W)} - \text{Kernel Size} + 2 \times \text{Padding})}{Stride} + 1] \quad 2$$

- Input Size (H, W): Height and width of the input.
- Kernel Size: Height and width of the convolution kernel.
- Padding: Number of pixels added around the input (to preserve spatial dimensions).
- Stride: Step size of the convolution operation

The output size of a pooling layer can be calculated using the same formula:

$$\text{Output Size (H, W)} = [\frac{(\text{Input Size (H, W)} - \text{Kernel Size})}{Stride} + 1] \quad 3$$

3. Dense Blocks:
    - Dense Block 1: Contains multiple layers (each having batch norm, ReLU, 1×1 convolution, and 3×3 convolution).
    - Transition Layer 1: 1×1 convolution followed by 2×2 average pooling.
    - Repeated for Dense Blocks 2, 3, and 4.
4. Global Average Pooling: Converts the final feature map into a single feature vector.

$$\text{GAP Output} = \frac{1}{H \times \text{W}} \sum_{i=1}^{H} . \sum_{j=1}^{W} X_{i,j} \quad 4$$

H: Height of the feature map.

W: Width of the feature map.

Xi,j: Value at position (i,j) in the feature map.

5. Fully Connected Layer: Custom final layer for 14-class classification (for 14 thoracic diseases in CheXNet).

4.6 Gated Recurrent Unit

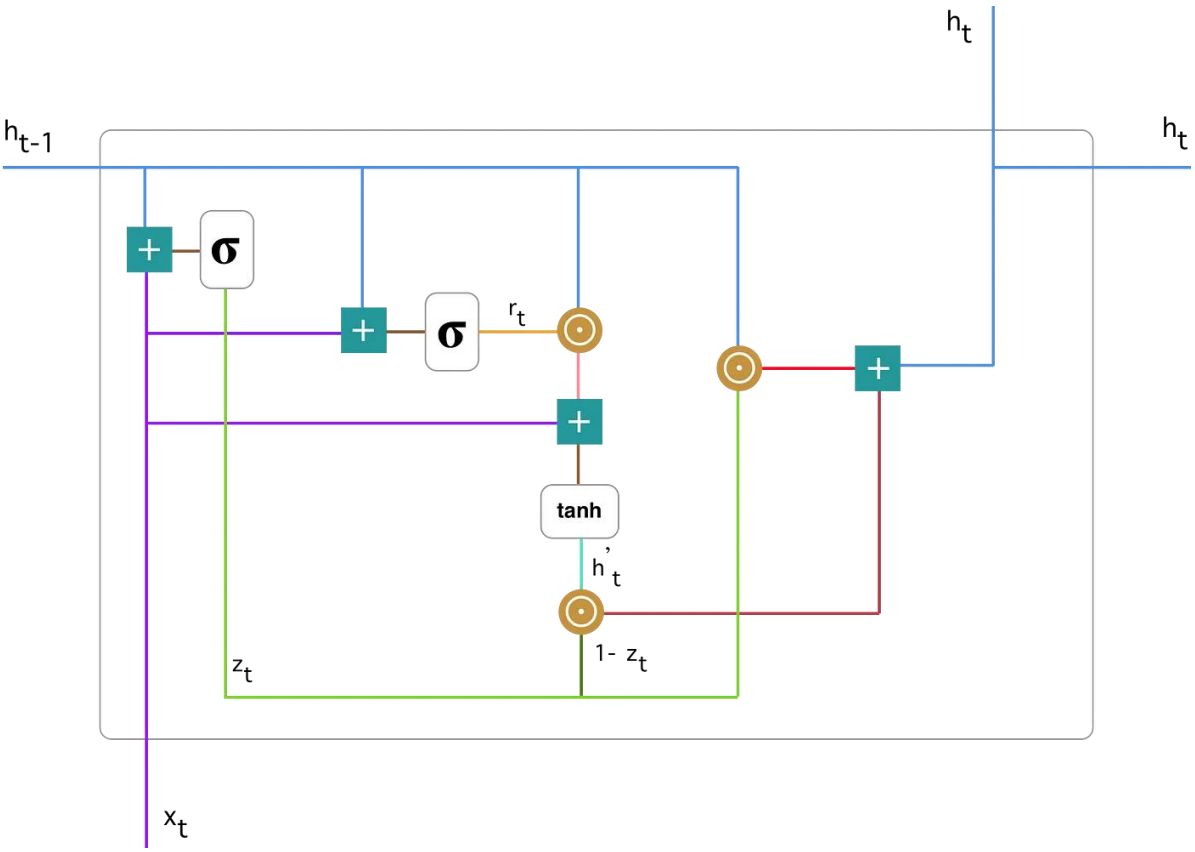

Figure 4.6 Gated Recurrent Unit (GRU)

*Source:https://miro.medium.com/v2/resize:fit:828/format:webp/1*6eNTqLzQ08AABo-STFNiBw.png*

GRU stands for Gated Recurrent Unit, which is a type of recurrent neural network (RNN) architecture that is similar to LSTM (Long Short-Term Memory). Like LSTM, GRU is designed to model sequential data by allowing information to be selectively remembered or forgotten over time. However, GRU has a simpler architecture than LSTM, with fewer parameters, which can make it easier to train and more computationally efficient. The main difference between GRU and LSTM is the way they handle the memory cell state. In LSTM, the memory cell state is maintained separately from the hidden state and is updated using three gates: the input gate, output gate, and forget gate. In GRU, the memory cell state is replaced with a "candidate activation vector," which is updated using two gates: the reset gate and update gate.

Update Gate:

The update gate decides how much of the past information should be carried forward and how much new information should be added.

$$Z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \quad (5)$$

Reset Gate:

The reset gate decides how much of the previous state should be ignored. Essentially, this gate is used from the model to decide how much of the past information to forget.

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \quad (6)$$

Candidate State:

The candidate state represents the new state based on the reset gate's influence.

$$h'_t = \tanh(Wx_t + r_t \odot Uh_{t-1}) \quad (7)$$

Final Hidden State:

The final hidden state is a combination of the candidate state and the previous state, modulated by the update gate.

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \quad (8)$$

4.7 Main Model [with Attention]

Attention in deep learning can be broadly interpreted as a vector of importance weights: in order to predict or infer one element, such as a pixel in an image or a word in a sentence, we estimate using the attention vector how strongly it is correlated with other elements and take the sum of their values weighted by the attention vector as the approximation of the target.

For this model, the Bahdanau additive attention mechanism was used. Here is a brief overview of the architecture of the encoder-decoder model with the additive attention mechanism:

Input- The model is fed with both image vector and report text with embedding dimension in which both inputs are added and sent as the context vector to decoder.

Decoder stage- GRU is used to extract high-level features from the input, providing a deeper understanding of the input features.

Additive Attention- It provides weight vectors (alpha) to every sequence of words and gets added up with word level features from each time stamp into sentence level features vector. This is a simplified form of Bahdanau's Attention.

The encoder is the same as the baseline model. For the decoder, a one_step_decoder layer was created, which takes in the decoder_input, the encoder_output, and the state value. The decoder_input, which is any character token number, was passed through the embedding layer, and then the embedding output and the encoder_output were passed through the attention layer, which produced the context vector. The context vector was then passed through the RNN (with GRU used here), with the initial state being that of the previous decoder. Dropout layers were used for tuning and regularization of the model.

The decoder calls the one-step attention layer for each of the decoder time-steps and computes the scores and attention-weights. All the outputs of each time-steps are stored in the 'all-outputs' variable. The outputs from each decoder step are the next word in the sequence. 'All-outputs' will be our final output.

4.7.1 Loss Function

Sparse Categorical Cross-entropy loss is used to measure the accuracy of the generated reports compared to the actual reports, guiding the optimization process.

4.7.2 Model Inference

For the final model, In the inference stage we have used the Beam search to find the output sentence. In Beam search at each time step at the decoder part,we select the top K-words(K=Beam length) with maximum likelihood of occurrence and generate words. We instantiate K independent versions of the model and use them to generate words but as Beam width increases the inference time and memory consumption increases due to which making predictions will be slow.

The reason behind choosing beam search over greedy search is that the objective function of Beam search is maximizes the conditional probability of all candidate words and choose the max probable outcome.

4.8 Model Evaluation

4.8.1 Metrics

The performance of the integrated model is evaluated using the BLEU score, which is used to measure the precision of the generated text.

4.8.2 Validation

Our model will be validate using a separate validation set to ensure that it generalizes well to unseen data.

CHAPTER 5: SYSTEM DESIGN AND ARCHITECTURE

5.1 System Architecture

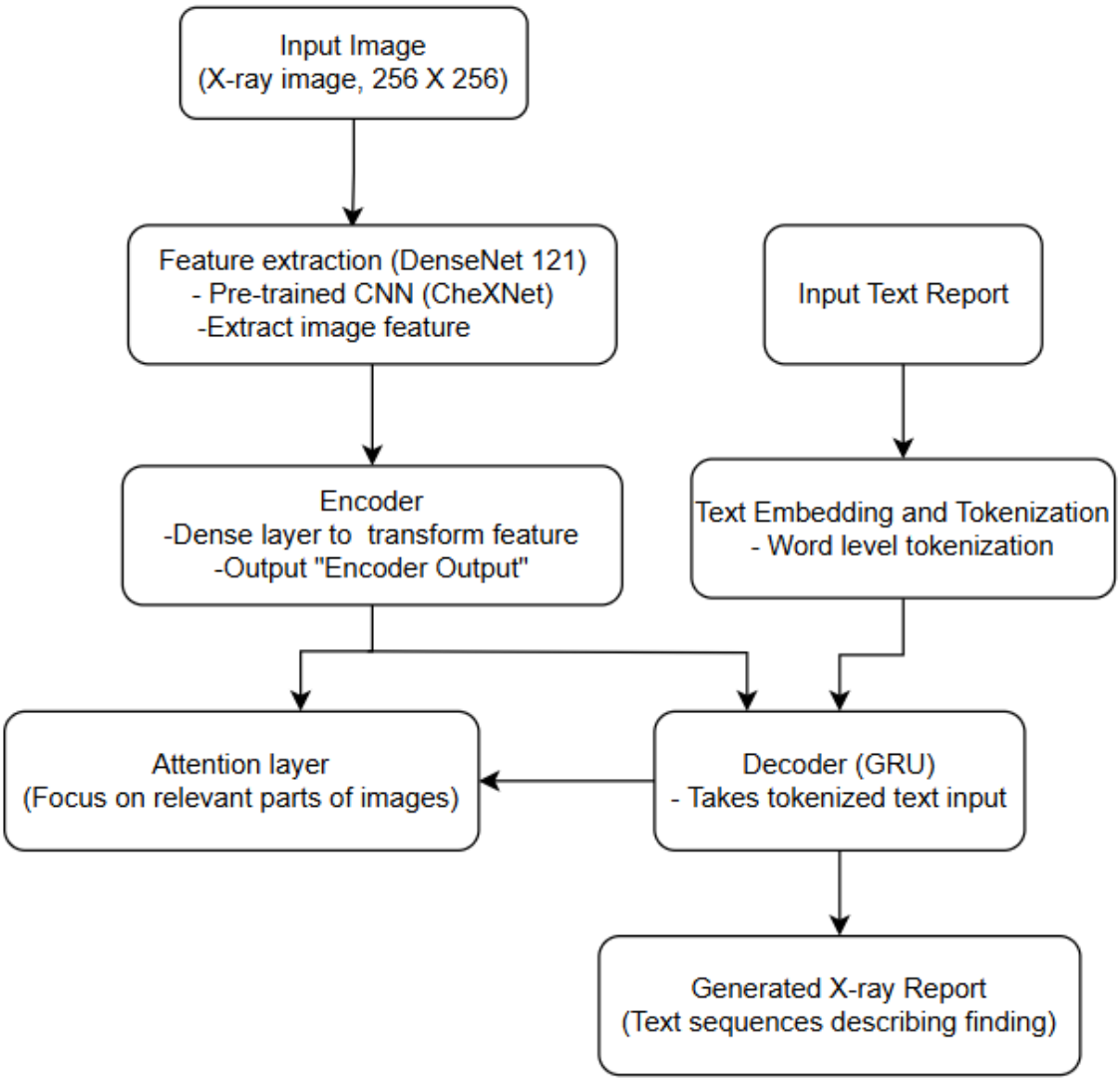

Figure 5.1 System Architecture

5.2 Flowchart

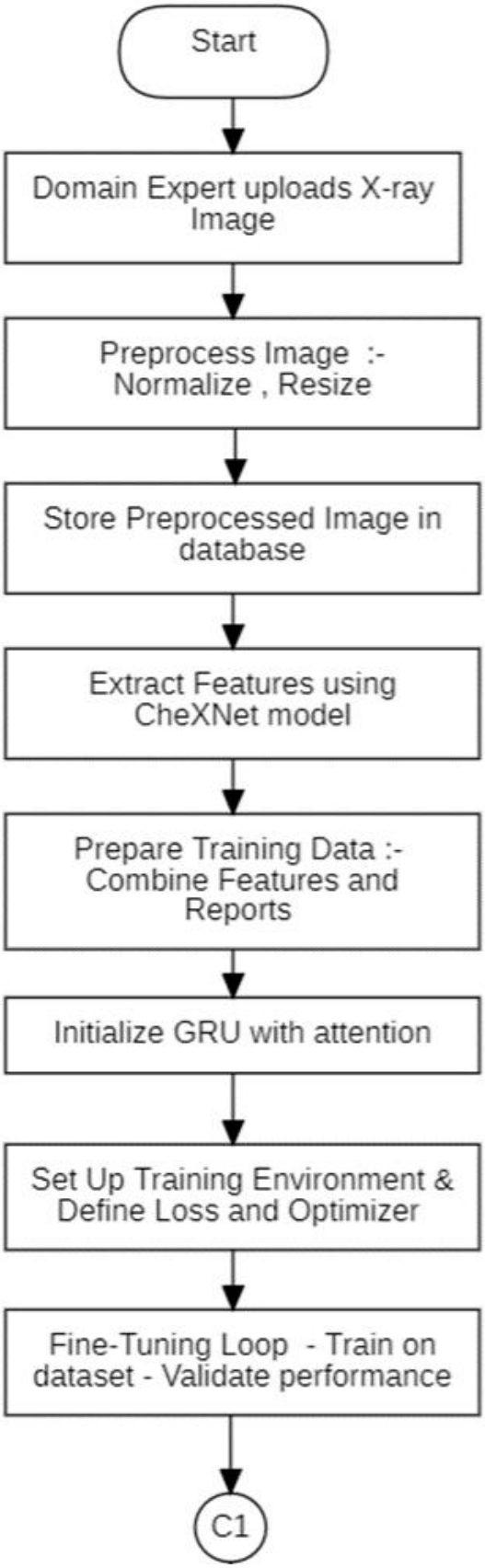

Figure 5.2 Flowchart

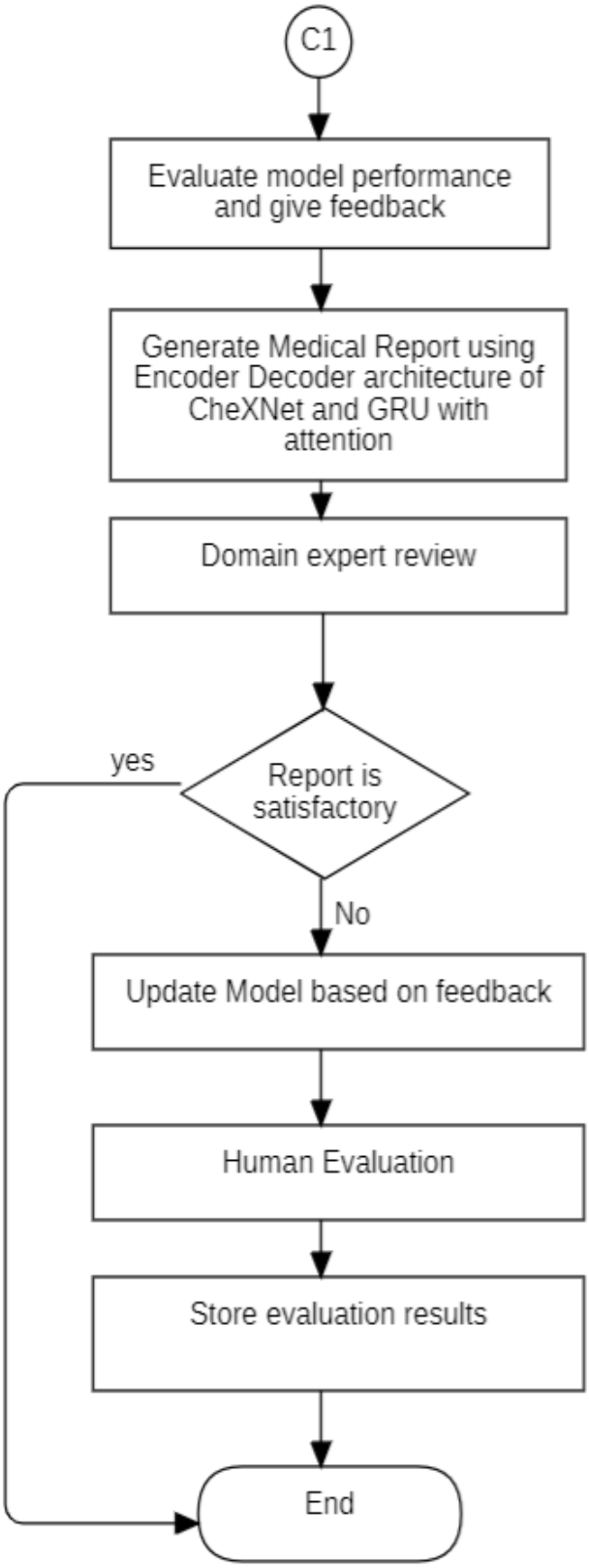

Figure 5.2 Flowchart

5.3 Use case Diagram

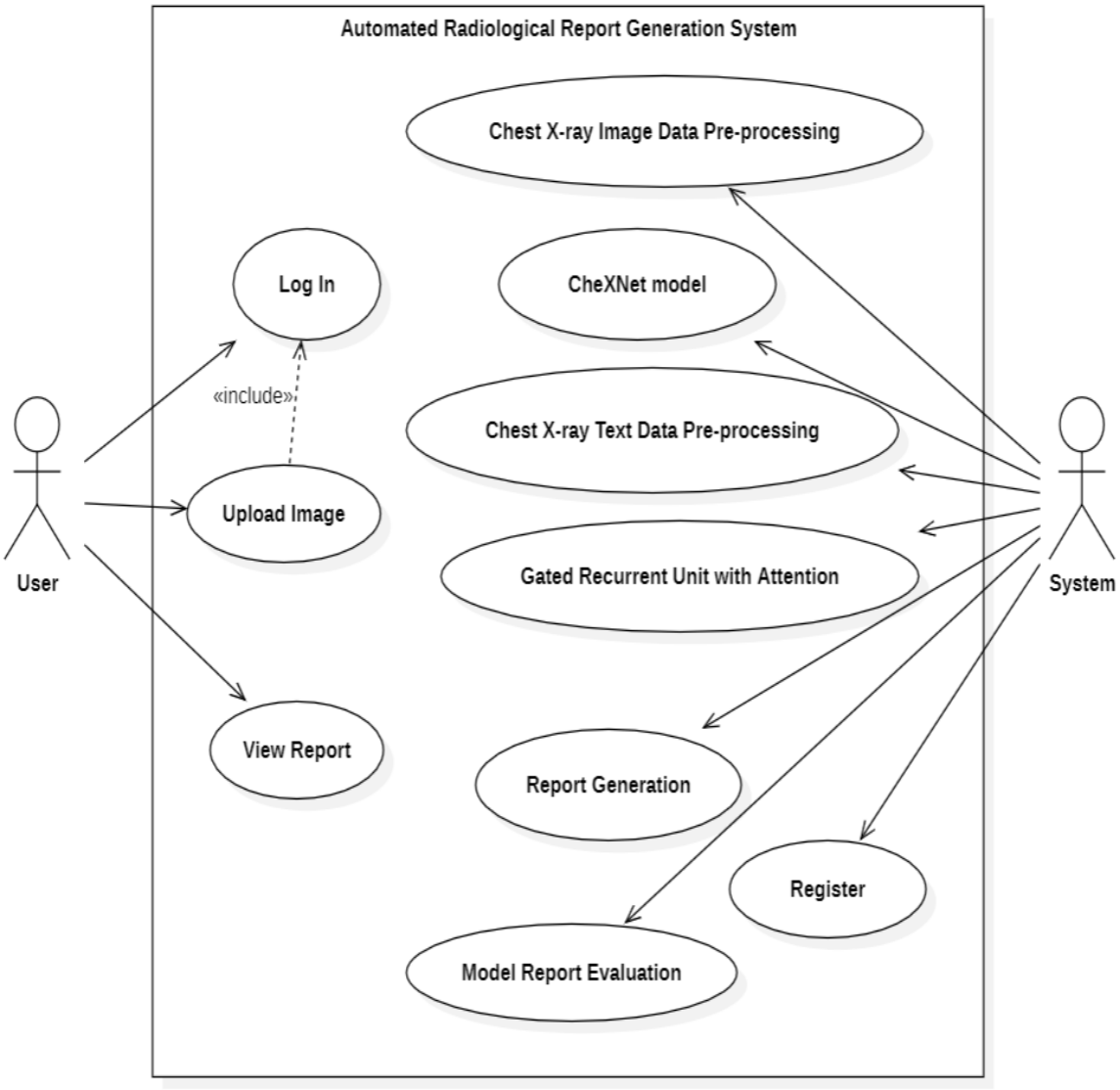

Figure 5.3 Use-case diagram

5.4 DFD Level 0

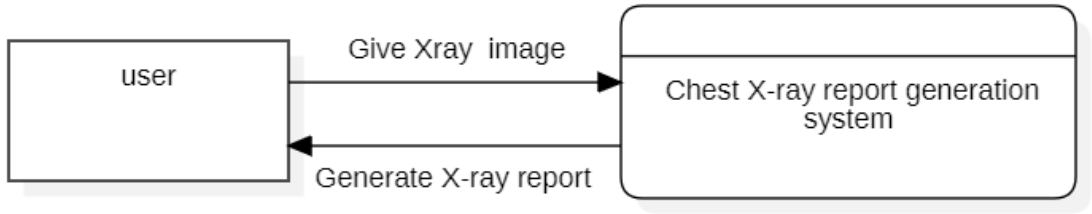

Figure 6.4 DFD level 0

5.5 DFD Level 0

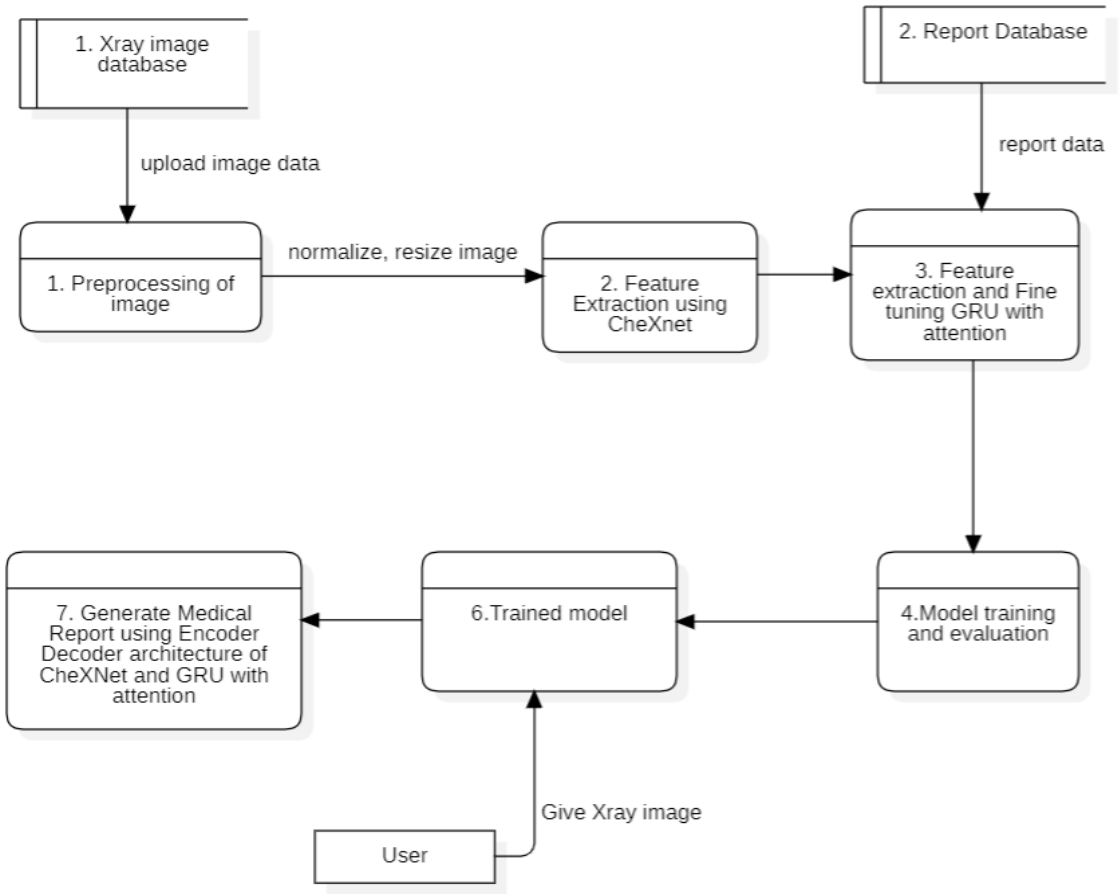

Figure 5.5 DFD level 1

CHAPTER 6: RESULTS AND ANALYSIS

6.1 Results and Analysis

In order to produce medical reports from photos, we began this research with a simple encoder-decoder paradigm. At first, the BLEU-1 score was only 0.106. This BLEU score shows that the model's ability to generate accurate and fluid reports from the image input was lacking.

The attention mechanism is incorporated to address this problem. Attention is crucial for the model to understand complex medical images because it allows to focus on relevant areas of the input image while generating the output text. Furthermore, the original decoder was replaced with a GRU. Gathering data from past and future time steps facilitates the model's ability to produce reasonable text.

Indeed, these changes have yielded favorable results. After these modifications were made, the BLEU-1 score increased from 0.106 to 0.3787, which indicates an improvement in the generated reports. In the final model, we applied beam search to generate three outputs and choose the best one. This modification was helpful for text generation because the model is able to produce more relevant and accurate medical reports.

The BLEU scores of main model are:

BLEU-1: 0.3787

BLEU-2: 0. 2418

BLEU-3: 0.1735

BLEU-4: 0.1026

The addition of attention mechanism, using a GRU, and implementing beam search all played significant roles in improving the model’s ability to generate more accurate and meaningful medical reports.

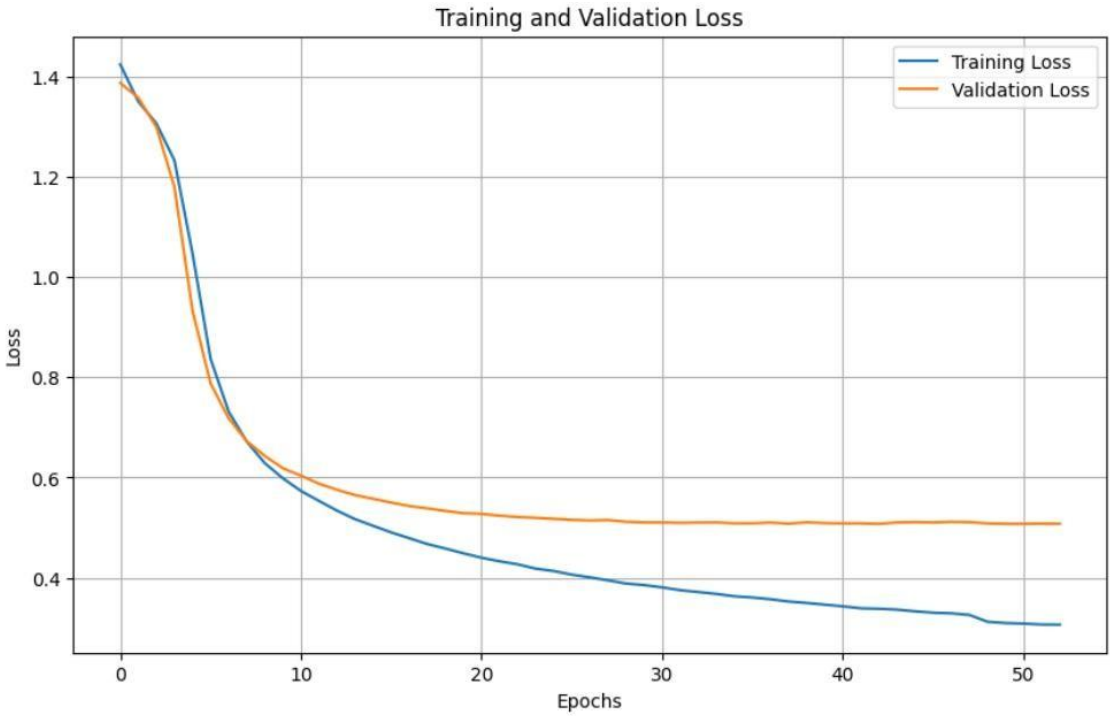

Figure 6.1 Training and Validation Loss

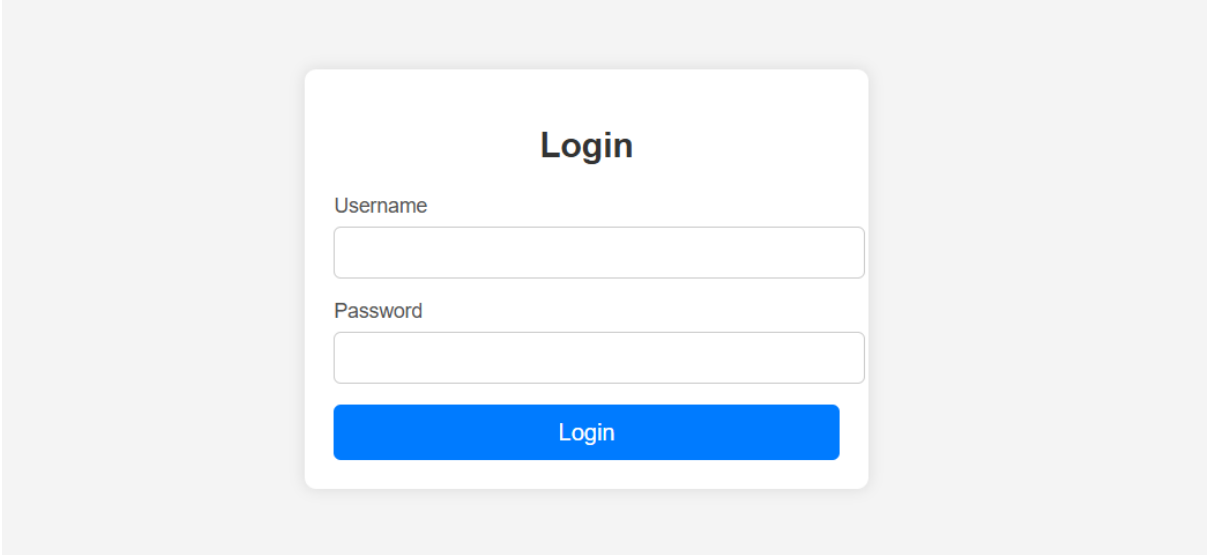

Figure 6.2 Login Page

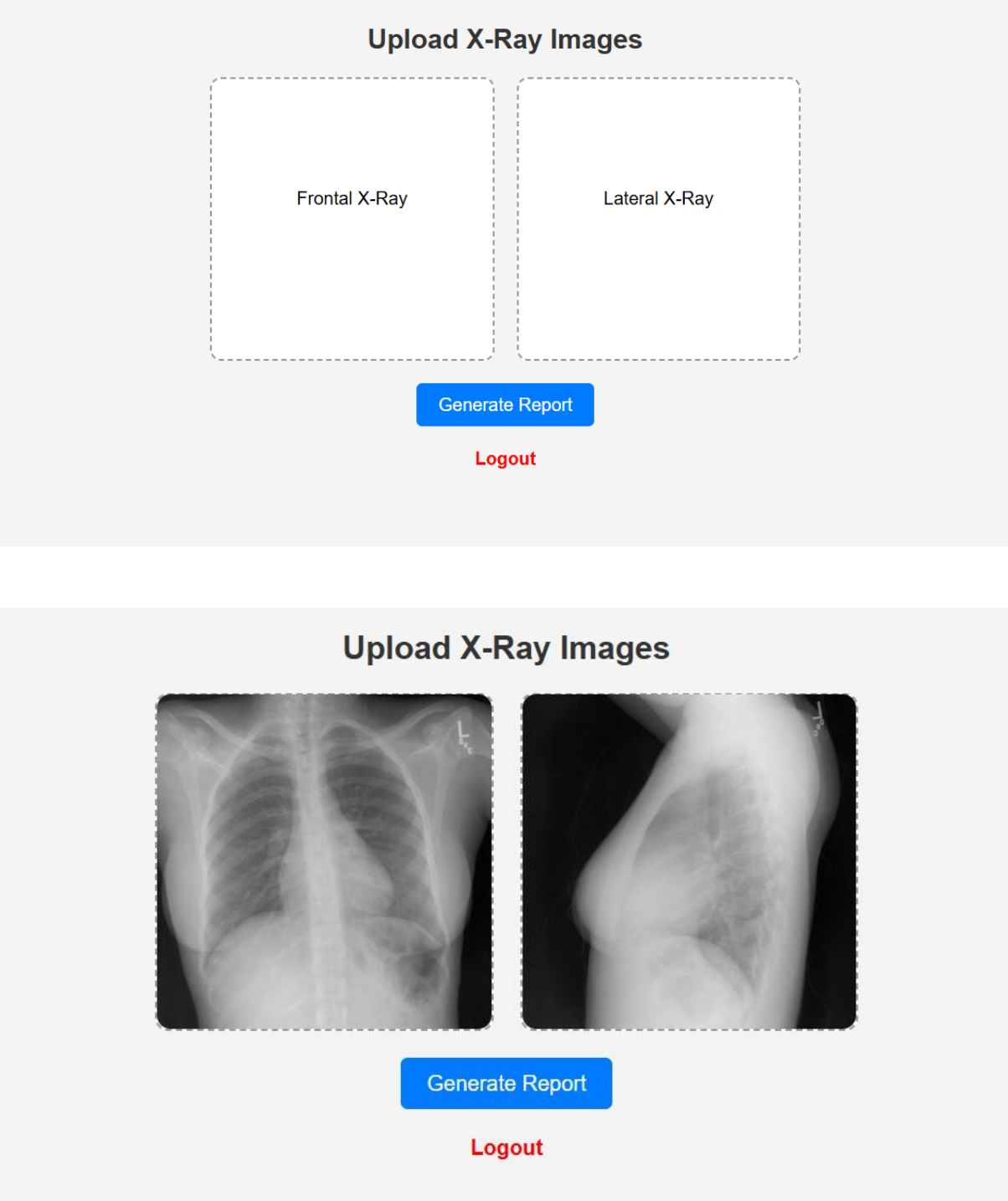

Figure 6.3 Home Page

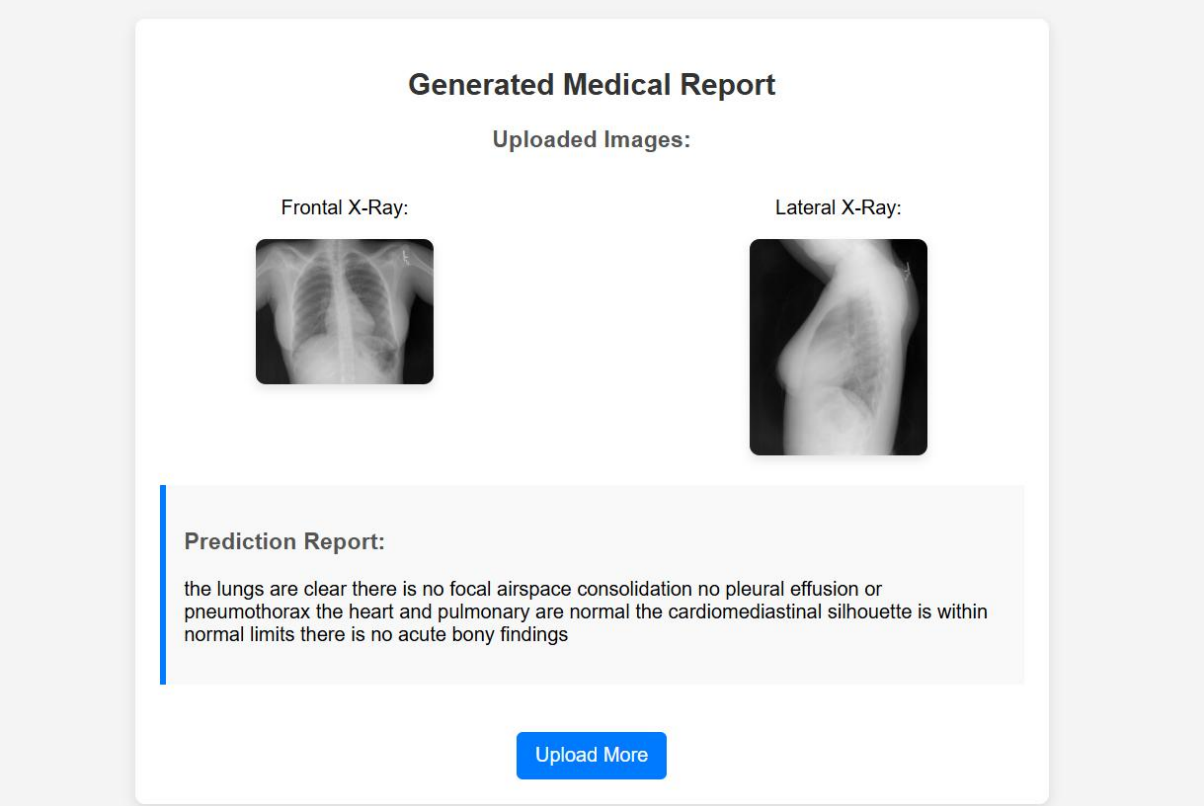

Figure 6.4 Output

CHAPTER 7: CONCLUSION, LIMITATION AND FUTURE ENHANCEMENT

7.1 Conclusion

This Project aims to generate a clinically accurate chest X-ray report based on the frontal and lateral X-ray images. We collected data from Indiana University publicly available chest x-ray dataset, and significant work went into EDA and Pre-Processing. For the baseline model, we created simple encoder decoder model which did not give us decent result. Improved the baseline results by building Attention model, the results of attention model are promising in comparison with baseline model and the impression statements are meaningful according to the X-ray image.

7.2 Limitations

- The model cannot generate accurate chest x-ray due to the limited number of training samples.
- The model is more biased towards generating no disease (normal case) because dataset contains majority data points of normal case.

7.3 Future Enhancement

Some possible future enhancements of the project are listed below:

- Training the model on larger dataset like MIMI-CCXR to improve the model prediction.
- Developing complex architectures by using ViT, CVT, and LLM to better capture information of images while training on larger datasets.

REFERENCES

[1] "Hurdles to hospitals," The Rising Nepal, Mar. 13, 2024. [Online]. Available: https://risingnepaldaily.com/news/23880#:~:text=The%20doctor-patient%20ratio%20in,patient%20ratio%20of%201%3A1%2C000.

[2] F. F. Alqahtani, M. M. Mohsan, K. Alshamrani, J. Zeb, S. Alhamami, and D. Alqarni, "CNX-B2: A novel CNN-transformer approach for chest X-ray medical report generation," IEEE Access, vol. 12, 2024, pp. 26626–26635.

[3] I. Allaouzi, M. B. Ahmed, B. Benamrou, and M. Ouardouz, "Automatic caption generation for medical images," in Proc. 3rd Int. Conf. Smart City Appl., New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 1–6.

[4] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (Lecture Notes in Computer Science), vol. 11769. Midtown Manhattan, NY, USA: Springer, 2019, pp. 721–729.

[5] A. Rajkomar, S. Lingam, A. G. Taylor, M. Blum, and J. Mongan, "High-throughput classification of radiographs using deep convolutional neural networks," J. Digital Imaging, vol. 30, no. 1, Feb. 2017, pp. 95–101.

[6] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," Radiology, vol. 284, no. 2, Aug. 2017, pp. 574–582.

[7] M. Cicero, A. Bilbily, E. Colak, D. Kontos, and J. Mermelstein, "Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs," *Investigative Radiology*, vol. 52, no. 5, May 2017, pp. 281–287.

[8] D. Soydaner, "Attention mechanism in neural networks: where it comes and where it goes," *Neural Computing and Applications*, vol. 34, no. 16, pp. 13371–13385, May 2022.

[9] J. B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.

[10] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proc. 2014 Conf. Empirical Methods Natural Lang. Process. (EMNLP), Doha, Qatar, pp. 1724–1734, Oct. 2014.

[11] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.