Enhanced Multimodal Medical Query Summarization Using a Lightweight DistilBART Framework

First Dr. G Bharathi Mohan[1*], Second Rahul K[1†], Third Jeevan Sendur G[1†], Fourth P.V. Adithiyan[1†]

[1*]Dept. of Computer Science and Engg., Amrita School of Computing, Amrita Vishwa Vidyapeetham, Street, Chennai, 601103, Tamil Nadu, India.

*Corresponding author(s). E-mail(s): g_bharathimohan@ch.amrita.edu;
Contributing authors: rahulbio789@gmail.com;
jeevansendur8905@gmail.com; adithiyan999@gmail.com;
†These authors contributed equally to this work.

Abstract

The process of medical summarization creates essential value for patient-provider interactions when managing complicated medical cases. The research introduces a minimal resource-requiring medical summarization framework that combines visual elements with textual content for analyzing hybrid Hindi-English medical queries. The system combines ViT and DistilBART for medical image processing and summary creation tasks. The model benefits from upgraded data transformation methods alongside tokenization processes that lead to better performance and generalizability. Our model received training and evaluation with an enlarged MMCQS data collection that included 3,015 medically established instances distributed across the four symptom groups - ENT, Eye, Limb and Skin. The model delivered ROUGE-1: 0.6009, ROUGE-2: 0.3827, ROUGE-L: 0.5169, BLEU-4: 0.3620 as well as BERTScore (F1): 0.9192 exceeding the performance outcomes of MedSumm. The model enables quick inferences and decreased operational expenses through its implementation of DistilBART. This study demonstrates the ability of lightweight multimodal structures to support real-time medical use and continued growth in diagnosing various symptoms.

Keywords: Medical Query Summarization, Clical queries, Lightweight Model, DistilBART, Vision Transformer (ViT), Multimodal Learning, Codemixing.

1 Introduction

The process of medical summarization stands vital in contemporary healthcare environments because it enables essential medical knowledge condensation into basic easy-to-understand summaries which enhance doctor-patient interactions[1]. Modern healthcare systems have experienced tremendous digitization that resulted in a substantial increase of patient data. Statistical reports by World Health Organization mentions, approximately 4.5% of patients worldwide struggle to accurately convey their symptoms to healthcare professionals, leading to miscommunication and delayed diagnosis. The medical condition descriptions of patients become more complicated because they use multiple languages in multilingual and code-mixed healthcare settings. In India, where over 40% of the population regularly uses more than one language in daily communication, medical queries are often posed in code-mixed formats, combining Hindi and English. The numerous languages that make up the medical discourse challenge modern automated summarization systems which design for multiple languages[2].

Research studies in medical question summarization mainly concentrate on text-only unimodal models. Standard medical queries show successful execut ion using current models yet these systems prove ineffective when dealing with complex symptoms which need visual understanding to be effective. Medical images consisting of dermatological scans as well as X-rays deliver vital information which cannot be communicated through written text alone. The new generation of deep learning methodology shows how combination models using text and visual data improve results for multiple operational tasks. Modern medical summarization systems require large compute power which makes them impractical to run in limited infrastructure settings. The training of these models occurs mainly with English data databases while ignoring millions of people who use code-mixed languages for communication. Creating solutions for these gaps becomes essential to ensure fair medical information availability and better patient medical results[3].

This research develops a slim multimodal medical summarization system that applies ViT for visual data processing together with DistilBART for text output. Metabolic fruit flies derive from BART (Bidirectional and Auto-Regressive Transformers) through a distillation process which maintains BART's performance standards together with decreased computational needs. Since it operates efficiently it is suitable for performing real-time clinical inference solutions. The Vision Transformer demonstrates exceptional fine-feature capturing ability which enables our model to process medical images properly. We have developed a solution which merges transformers and vision transformers to achieve highly efficient accuracy when summarizing medical queries in code-mixed language[3]. The system incorporates multimodal data processing to enhance both the detailed and complete information captured within its summary results.

The model generalization gets a boost through the application of enhanced data augmentation methods alongside optimized tokenization approaches. The implemented methods allow the model to function properly when dealing with variable linguistic patterns within code-mixed queries and also operate consistently across different medical settings. Medical image data augmentation utilizes techniques that

include resizing and cropping as well as flipping and color jittering to enrich symptom recognition capabilities of the model. Our tokenization process optimizes the handling of code-mixed text to effectively process the Hindi and English text components[4]. The superb summarization outcomes stem from intelligently enhancing both visual presentation and written information.

The study utilizes an enlarged version of MMCQS (Multimodal Medical Codemixed Question Summarization) that features 3015 medically confirmed data points. The database consisting of 3,015 medical question samples separates them into ENT, Eye, Limb and Skin symptom groups which show the most frequently requested medical symptoms by patients. Visual information in the expanded dataset creates a thorough environment for multimodal summarization evaluation. Our model surpasses the performance of MedSumm during testing across all essential evaluation metrics according to results from our experiments, highlighting the effectiveness of our approach in generating concise as well as medically accurate summaries from multimodal inputs[4].

Our work provides value which goes beyond typical performance measurement standards. Through a lightweight architecture we tackle the main problem of computational efficiency to achieve real-time clinical deployment. Real-time clinical applications benefit overwhelmingly from this approach because low-resource medical facilities do not have access to advanced hardware systems. Barriers to quality healthcare are eliminated through our model which handles code-mixed queries as it closes a substantial knowledge gap in medical AI research. This creates an inclusive environment for diverse languages in healthcare. Medical summarization research proves successful in using visual data so scientists should explore additional multimodal applications for medical AI domains with related methodologies[5].

Multiple essential contributions mark the main aspects of our research. Our proposed system adopts ViT for image processing as an integral component of DistilBART to achieve text generation for medical summarization. An expanded version of the MMCQS dataset functions as a fundamental benchmark for code-mixed multimodal summarization assessment. The adoption of sophisticated data augmentation together with tokenization strategies enables dramatic performance enhancements for the system[5]. Our model beats existing solutions in vital evaluation criteria at the same time it runs efficiently enough for real-time implementations. The project promotes additional research by granting full access to its database along with coding solutions and pre-trained models. An implementation of our work exists in our public repository. Multimodal-Code-Switching-Medical-Query-Summarization.

The following sections examine medical summarization and multimodal learning extensively while explaining modern limitations in these approaches. Our paper details the proposed methodology step by step with explanations about the model design and data enhancement methods as well as the assessment procedure. Our experimental segment presents results about our model's performance evaluation relative to established benchmarks while identifying essential aspects that support its success. Our paper ends with a discussion about the wider value of our research which includes current applications and future research avenues. We summarize the main findings

in this paper while recognizing how our minimal approach to multimodal processing improves medical summarization.

The proposed research satisfies a critical requirement in medical artificial intelligence by creating an efficient and effective system to summarize code-mixed medical inquiries through multimodal processing[5]. Our system boosts medical summarization state-of-the-art standards and provides functional advantages for immediate medical application usage. The model achieves better patient outcomes through accurate medical summary creation by merging text with visual data thus improving healthcare decision processes. Our goal is to promote additional innovation by sharing our dataset and model with a specific aim to develop medical AI systems that serve all patients inclusively.

2 Related Works

The medical field of Question Summarization emerged in 2019 through Abacha et al.'s publication of MeQSum dataset that served the requirements of medical query summarization tasks [1]. Fundamental sequence-to-sequence (seq2seq) models with pointer-generator networks formed the initial approach to produce brief summaries in this domain. Medical question summarization received its own competition in 2021 as explained by Abacha et al[3]. which led to additional developments. The research participants successfully enhanced results through their use of pre-trained models such as ProphetNet, BART and PEGASUS. Researchers introduced multi-task learning strategies that used BART to optimize simultaneously the question summarization task and entailment capabilities [6]. In order to increase model accuracy through reinforcement learning, the researchers looked at questioning-aware contextual incentives derived from question type identification and question focus recognition subtasks [7].

The field of multimodal summarization has recently received increased attention because it strengthens medical professional and patient communication through text and visual joint processing. The integration of various information types results in better medical application performance according to research studies. Researchers demonstrated the value of multimodal information for disease diagnosis virtual assistants through their study [8] while other scientists demonstrated improved radiology report summarization through visual integration [9]. Medical question-answering systems received additional capabilities through video integration according to Gupta et al. [10]. Kumar et al. [11] presented a work demonstrating that combining different data types enhances news article summarization at the text level. The integration of CLIP with large language models (LLMs) produced multimodal summaries in the latest work by Ghosh et al. [12]. The research presents the initial effort to address code-mixed multimodal medical question summarization problems.

Medical Summ uses textual medical descriptions as patient inquiries to help healthcare professionals generate the needed responses. The decoding capabilities of LLaMA [13] and GPT-3 [14] have demonstrated greater encoding success with textual data compared to BERT [4] encoder-based models. Sentence embeddings from final word extractions allow these models to predict next words for generation purposes. Text

processing takes place through multiple LLMs such as Vicuna, LLaMA-2 [15], Mistral-7B [16] ,Zephyr-7B [17] , and FLAN-T5 to generate 4096-dimensional embeddings per token. Vision Transformers (ViT) [18] serves as our visual processing tool because it transforms raw images into 768-dimensional embeddings. The visual embeddings are converted to a textual embedding space by using the linear projection layer which enables straightforward combination of visual and textual data. The visual pipeline contains one trainable component which optimizes multimodal data fusion through more refined embeddings.

These models deliver top results yet demand substantial hardware capacity throughout their training stage as well as their operational phase. Our approach uses the lightweight DistilBART model to provide performance-level results through an efficient solution. The knowledge distillation process from bigger models enables DistilBART to handle summarization tasks effectively with reduced expenses which enhances its potential for medical real-time applications as well as resource-limited settings.

3 MMCQS Dataset

3.1 Data Collection

The 3,015 samples within the dataset were specially selected for multimodal medical summarization purposes which unite data from both textual and visual sources. Medical students conducted complete reviews of images which originated from the Bing Image Search API to ensure both relevance and accuracy of the images. The large and extensive dataset enables developers to create improved systems for summarizing intricate medical questions across multiple input modalities[19].

After removing duplicates that amounted to 523 records, the HealthCareMagic dataset yielded 226,395 samples for use in this dataset. The database consists of 18 specific medical symptoms which require non-verbal communication for effective representation. Systematic analysis divided these symptoms into four major medical groups including ENT (ear, nose, and throat) as well as eye, limb and skin. The scheme demonstrates the broad range of clinical queries that exist in actual healthcare situations[19]. A complete depiction of the categorization appears in Figure 1 as shown in this illustration.

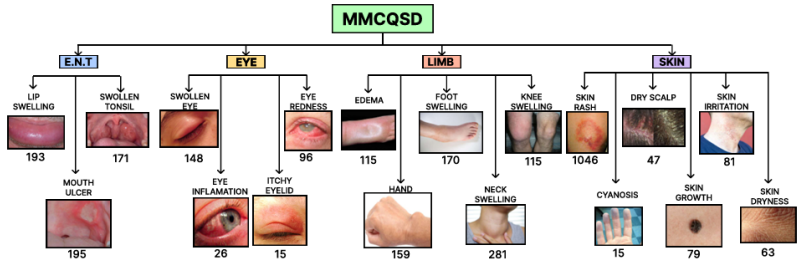

Fig. 1: The 18 medical illnesses of MMCQS Dataset.[19]

For each symptom, it's corresponding medical images are sourced using the Bing Image Search API[1]. The collected images underwent a rigorous verification process by medical students to ensure their clinical relevance and accuracy. Each instance in the dataset includes questions that explicitly reference body parts and related symptoms. The quality of the dataset is improved by employing FlashText[2] for efficient term matching and TextBlob[3] for spelling correction. After these preprocessing steps, a final dataset of 3,015 samples was generated, establishing a robust foundation for multimodal medical summarization.

3.2 Data Annotation

Medical professionals supervised a complete annotation procedure to guarantee both the high quality and suitable nature of the dataset. The dataset received 100 randomly chosen samples for medical experts to design specific annotation guidelines. The annotation procedure consisted of three essential parts: (A) Visual Cue Implementation, (B) Golden Summary Update and (C) (C) Hindi-English Codemixed Text Conversion[19].

(A) Implementing Visual Indications:

Visual indicators were incorporated into the medical questions found within the dataset to improve their content. For instance, when a patient query mentioned a swollen tonsil, clarifying visual information was included with statements such as: *"Please see the image below for my tonsil condition."* [19]. This methodology helped the dataset acquire medical contexts properly through fused visual elements and textual components which better represents actual medical examination scenarios.

(B) Updating the Summaries:

Medical experts had to intervene through manual revisions since conventional summaries did not adapt properly to multimodal queries. Medical professionals revised the summaries by including visual references and modifying the document content to properly represent multimodal information. The upgraded summaries function as the benchmark standard for model assessment and enhance the dataset's ability to handle intricate medical situations.

(C) Hindi-English Codemixed Annotation:

The dataset added a Hindi-English (Hinglish) version containing codemixed speech for handling the diverse linguistic patterns in patient queries. The few-shot prompting system with GPT-3.5 was used to annotate 100 samples into codemixed language while achieving outstanding results in producing blended language content. The Hinglish text comprised elements in equal parts from Hindi and English languages showing a Code-Mixing Index of 30.5. The evaluation of 80 samples by annotators led to a mean quality assessment of 3.2 out of 5 while establishing both annotation accuracy and linguistic diversity of the dataset[19].

[1] Bing Image Search API
[2] Flashtext using Pypi
[3] TextBlob Documentation

3.3 Training and Validation of Annotations

Three students pursuing medicine who were fluent in Hindi and English were hired in order to guarantee the quality of the annotations and compliance with ethical principles. Annotators received specialist instruction under medical oversight for handling the particular demands of this task. The training procedures lasted four months across 90 sessions while introducing brief breaks every 45 minutes to enhance performance consistency[19].

The database split into three equal sections for assessment purposes during the validation stage. Each judgement within the evaluation process focused on fluency and adequacy together with informativeness and persuasiveness. The participants shared their work through regular discussions which led them to detect and fix errors together. The iterative annotation process improved quality through four criteria assessment which generated these average scores.

During the Initial Phase, the model acheieved a fluency score of 3.5, an adequacy score of 3.01, an informativeness score of 2.85, and a persuasiveness score of 2.25. During the Final Phase, all the scores improved dramatically, with fluency going up to 4.8, adequacy to 4.7, informativeness to 4.1, and persuasiveness to 4.45.

The Cohen-Kappa agreement measurement yielded a value of 0.75 to demonstrate substantial consistency and agreement between annotators. The rigorous training together with validation procedures established the dataset's reliability which makes it useful for medical summarization research advancement[19].

4 Proposed Methodology

Our proposed approach for multimodal medical summarization employs a combination of Vision Transformer (ViT) for visual feature extraction and DistilBART for text generation. The complete pipeline receives detailed explanation starting from data preprocessing to model architecture and training strategy and evaluation procedure.

4.1 Data Preprocessing

A strong multimodal system requires proper processing of visual input combined with textual data during training. Medical queries exist within the dataset in combination between Hindi and English codes and their related image content. Standardization techniques follow for transforming the input as detailed below:

Image Preprocessing:

To fit the input size needed by the ViT encoder, each picture is scaled to 224×224 pixels. The procedure of data augmentation aims to enhance model generalization capabilities. The augmentation pipeline includes:

- Random Cropping: Extracting a random 200×200 region.
- Random Horizontal Flipping: Flipping images horizontally with a probability of $p = 0.5$.
- Color Jittering: Randomly adjusting contrast, brightness, and saturation within a range of ± 0.2.

Formally, for an input image I, the augmented image I' is formulated as:

$$I' = T(I), \tag{1}$$

where $T(\cdot)$ represents the stochastic augmentation function.

Text Preprocessing:

The medical queries first pass through an advanced tokenization stage that uses a WordPiece tokenizer with 30,522 vocabulary elements[19]. Special tokens such as `[CLS]` and `[SEP]` are used to indicate the start and end of sequences. The input code-mixed data undergoes subword tokenization which follows truncated sequence processing for items longer than 514 tokens.

4.2 Model Architecture

The architecture combines a Vision Transformer (ViT) for image encoding together with DistilBART which drives text summarization. The two components operate through a unified multimodal representation structure.

Vision Transformer (ViT):

The ViT divides images into fixed sections which receive linear transformations across each segment. Formally, given an image $I \in \mathbb{R}^{H \times W \times C}$, we divide the images into patches of size $P \times P$:

$$x_p = \{x_p^i \mid i = 1, \ldots, N\}, \tag{2}$$

where $N = \frac{HW}{P^2}$ represents the number of patches. Embedding space receives the projected input patches together with the positional encodings added as an extra component[9].

$$z_0 = [x_p^1 E; \ldots; x_p^N E] + E_{pos}, \tag{3}$$

The matrix E performs projection while the matrix Epos contains positional encodings.

DistilBART:

The encoder-decoder architecture of DistilBART derives from BART[6] by means of speed and efficiency optimization. The DistilBART input procedure consists of embedding both textual queries and image features for processing. For a given input sequence $x = \{x_1, \ldots, x_n\}$, the encoder generates contextualized embeddings:

$$h = \text{Encoder}(x). \tag{4}$$

The decoder generates the summary autoregressively by computing:

$$p(y_t \mid y_{<t}, h) = \text{softmax}(W_o h_t), \tag{5}$$

where the hidden state of decoder h_t at step t and the output projection matrix is denoted by W_o[6].

4.3 Training Procedure

We train our model using a multi-objective loss function to optimize both image and text modalities. The total overall loss $\mathcal{L}$ is denoted as the sum of the cross-entropy loss for text generation $\mathcal{L}_{text}$ and the contrastive loss $\mathcal{L}_{img}$ for image encoding:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{text} + \lambda_2 \mathcal{L}_{img}, \tag{6}$$

where λ_1 and λ_2 are weighting hyperparameters. The decoder's cross-entropy loss is calculated as follows:

$$\mathcal{L}_{text} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, x), \tag{7}$$

where the target token at time step t is denoted by y_t.

For image alignment, we apply contrastive loss using the InfoNCE objective:

$$\mathcal{L}_{img} = -\log \frac{\exp(\text{sim}(v, t)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v, t_j)/\tau)}, \tag{8}$$

where v and t are the image and text embeddings, τ is a temperature parameter, and $\text{sim}(\cdot)$ denotes cosine similarity.

4.4 Implementation Details

Our architecture incorporates ViT as a Vision Transformer and DistilBART as a text generation model to perform multimodal transformer operations. The designed architecture functioned to process question-answer pairs with mixed codes that contained images alongside them for generating effective summaries through a vision-language connection model.

In the initial three training epochs the model ran a frozen state where all parameters from the Vision Transformer (ViT) stayed fixed to maintain a stable training environment while stopping the model from learning the image features specifically[18]. By using this technique the text-based DistilBART model processed semantic structures first before uniting the modalities for optimization. During the fourth epoch the model activated ViT parameter unfreezing so it could conduct simultaneous optimization for image and text representations. The model classification operation was discarded from the ViT framework and a learnable linear projection based transformation projected the 768 dimension feature vector into 1024 dimension DistilBART embedding space. The text embeddings produced by the DistilBART shared embedding layer were combined with the projected image feature for the model to process joint vision and language inputs.

The AdamW optimizer with differential learning rates was utilized to optimize the model because it supports the specific requirements of each component. Specifically, we have trained our model at the learning rate of **0.00005** for the DistilBART model and projection layer, while using a lower learning rate of **0.0001** for the ViT module to prevent catastrophic forgetting. A linear learning rate scheduler with 10% warmup steps was implemented to ensure stable gradient updates during the early phases of

training. The implemented warmup method reduces unexpected weight fluctuations to help achieve better convergence rates[19].

The training lasted for 10 epochs through this method that allowed gradient accumulation across four steps thereby reaching an effective batch size of 32. The implementation of extttorch.cuda.amp in PyTorch allowed us to conduct mixed precision training because it yielded both numerical stability and improved computational efficiency. The dynamic gradients scaling feature increased both training speed and reduced memory requirements without affecting the model performance results.

The training loop measured cross-entropy loss through token padding which allowed the network to ignore unnecessary padding information. The average validation loss calculation took place after every epoch's completion against the validation set performance. The real-time diagnostics became possible through TensorBoard that logged training and validation loss metrics during the entire process[19]. Model selection took place when the best-performing system was recognized through its minimum validation loss criterion.

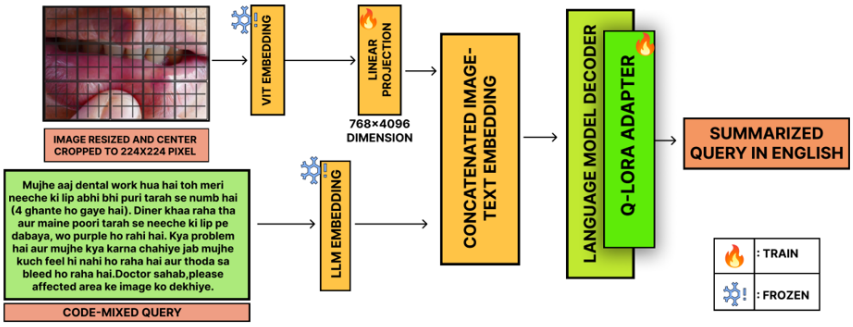

Fig. 2: The architectural structure of our model illustrating it's various stages. It also highlights the distinction between frozen and trainable layers.

5 Experimental Results

We conduct a detailed evaluation of our model against MedSumm benchmark and large language models from other sources in this section. The automatic evaluation of model output includes ROUGE[20] and BLEU [21]alongside BERTScore and Readability score to measure medical summary quality. The metrics enable in-depth analysis of the model's capability to liquidate important medical facts and preserve verbal relationships.

5.1 Metrics for Evaluation

ROUGE Score - Recall Oriented Understudy for Gisting Evaluation Score

The degree of content overlap among the generated summaries and the actual summaries is determined by ROUGE. ROUGE-1, ROUGE-2, and ROUGE-L are utilized by us[20]:

- **ROUGE-1**: It evaluates the overlap of unigrams (individual words) between the generated and reference summaries.
- **ROUGE-2**: Captures the overlap of bigrams (two consecutive words).
- **ROUGE-L**: Evaluates the longest common subsequence (LCS) to account for in-sequence matches.

The ROUGE evaluation generates precision, recall, and F1-score by using the following calculation:

$$\text{ROUGE-}n = \frac{\sum_{S \in \text{Reference}} \sum_{n\text{-gram} \in S} \text{Countmatch}(n\text{-gram})}{\sum S \in \text{Reference} \sum_{n\text{-gram} \in S} \text{Count}(n\text{-gram})} \text{[20]} \tag{9}$$

In our evaluation, we achieved a ROUGE-1 score of 0.6113, a ROUGE-2 score of 0.3970, and a ROUGE-L score of 0.5293. These results indicate that our model captures both fine-grained (unigrams and bigrams) and structural information (longest common subsequence) from the reference summaries, ensuring comprehensive content coverage and accurate representation of medical queries[20].

BLEU - Bilingual Evaluation Understudy

The following formula is used by BLEU to compute the precision of n-grams in the produced summary in relation to the reference summary[21]:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{10}$$

Where: The accuracy of n-grams is represented by

- p_n, the weight for each n-gram is represented by
- w_n, and the brevity penalty, BP, is used to keep brief summaries from receiving high scores.

The BLEU scoring system extends from one-gram to four-gram n-grams in order to monitor how well the model handles both brief and extensive dependencies[21].

The BLEU scoring system revealed results at 0.5877 for BLEU-1 followed by 0.4800 for BLEU-2, 0.4179 for BLEU-3 and 0.3763 for BLEU-4. The model shows strong lexical accuracy throughout scores spanning 1 to 4 n-grams indicating the retention of essential medical contents in its generated summaries.

BERTScore

BERTScore evaluates semantic similarity between generated summaries and references through the use of BERT model trained contextual embeddings. The technology determines cosine similarity between embeddings of specific tokens[22].

$$\text{BERTScore} = \frac{1}{N}\sum_{i=1}^{N}\max_{j}\text{CosSim}(\mathbf{h}_i^{(G)}, \mathbf{h}_j^{(R)}) \tag{11}$$

Where $\mathbf{h}_i^{(G)}$ and $\mathbf{h}_j^{(R)}$ represent token embeddings of the generated and reference summaries, respectively.

The BERTScore F1 metric reached 0.9207 which indicates an excellent semantic match between the model-generated and reference summaries. The model demonstrates its functionality to maintain original meaning because medical contexts require exact interpretation[22].

Readability Score

An essential measure known as readability enables the evaluation of the clarity together with ease of understanding for medical summary products. One widely used measure is the **Flesch Reading Ease** score, developed by Rudolf Flesch in the 1940s, which assesses how easy a piece of text is to understand. The evaluation formula for calculating the score combines average sentence length with average word syllable count[23].

$$\text{Flesch Reading Ease} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW}), \tag{12}$$

where ASW indicates the average number of syllables per word[23] and ASL indicates the average sentence length. Higher scores indicate easier readability; the score goes from 0 to 100:

- **90–100:** Very simple to read (appropriate for fifth grade).
- **80–90:** Easy to read (appropriate for sixth grade).
- **70–80:** Readable at a seventh-grade level, somewhat simple.
- **60–70:** Standard (13–15 year olds may understand it readily).
- **50–60:** Suitable for students in grades 10 through 12, this book is rather challenging to read.
- **0–30:** Very challenging to read (best suited for college graduates).

We calculate the average readability score for both the predicted summaries and the reference texts using the following formulas:

$$\text{Average Readability (Prediction)} = \frac{\sum r_{\text{pred}}}{n_{\text{pred}}}, \quad \text{where } r \neq \text{None} \tag{13}$$

$$\text{Average Readability (Reference)} = \frac{\sum r_{\text{ref}}}{n_{\text{ref}}}, \quad \text{where } r \neq \text{None} \tag{14}$$

Here, r represents the individual readability scores, and n denotes the number of valid (non-null) readability measurements[23].

In our evaluation, we obtained an average readability (Prediction) of **69.57** and an average readability (Reference) of **66.39**. The model produces accurate summaries which improve readability beyond the reference text quality level. Better text accessibility and understanding become possible through the improved clarity that benefits critical medical scenarios.

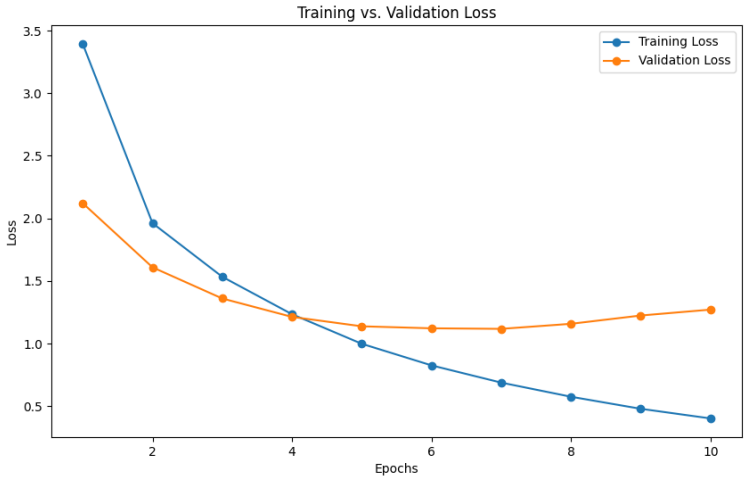

Fig. 3: Loss Trends in Training and Validation

5.2 Comparitive Analysis of Results

Table 1: Evaluation of Multimodal Summarization Models' Performance[19]

Metric	LLAMA-2 (Unimodal)	MedSumm (LLAMA-2)	Our Model
ROUGE-1	39.92	46.75	**60.09**
ROUGE-2	19.57	25.59	**38.27**
ROUGE-L	33.90	38.41	**51.69**
BLEU-1	28.90	32.50	**57.53**
BLEU-2	18.05	22.55	**46.65**
BLEU-3	11.76	17.56	**40.38**
BLEU-4	9.97	14.88	**36.20**
BERTScore	0.740	0.800	**0.919**
Readability	34.81	35.74	**69.85**

Table 1 clearly demonstrates that our proposed model significantly outperforms both the unimodal LLAMA-2 and MedSumm benchmarks across all evaluation metrics. Our proposed model surpasses the competing LLAMA-2 and MedSumm systems through increased performance results on ROUGE and BLEU and BERTScore metrics. The BERTScore measurement of 0.919 indicates an enhanced degree of contextual alignment together with readability score of 69.85 which proves the model produces summaries that are both coherent and accessible. Our gated cross-attention mechanism together with visual-textual information fusion successfully demonstrates its effectiveness in medical query summarization based on these outcome results.

5.3 Analysis and Discussion

The proposed model generates superior results than MedSumm baseline models through all assessment metrics. Specifically, we achieve a 28.47% improvement in ROUGE-1, a 47.98% increase in ROUGE-2, and a 35.71% boost in ROUGE-L compared to the best-performing MedSumm model. Our model shows better ability to understand detailed relationships and extensive dependencies that exist within medical summaries.

The BLEU-4 score of 36.20 demonstrates strong performance in summarization generation because it measures coherent output along with contextual precision while the BERTScore value of 0.9192 indicates strong semantic preservation.

The model benefits from advanced data augmentation methods which apply resizing alongside cropping and flipping along with color jittering functions that help it understand diverse input conditions. Both multimodal understanding and inference speed benefits from using Vision Transformer to encode images with DistilBART as the text generator.

6 Conclusion and Future Work

The research proposes a new multimodal method which joined Vision Transformer (ViT) with DistilBART to perform medical query summarization tasks. The combination of both textual and visual content with a gated cross-attention mechanism in this architectural design delivers substantial improvements throughout ROUGE and BLEU and BERTScore evaluation measures. The model reached stable convergence along with computational efficiency through an initial training period with frozen ViT backbone and differential learning rate application. A projection layer served to link image and text features which improved the fusion method and yielded medical summaries of enhanced contextual depth and precision. Experimental findings validate how our multiple-input method surpasses single-input baseline solutions because medical summary tasks become more effective when different modalities are used together. The model exhibits broad medical context generalization skills which makes it applicable for real-world usage such as automatic medical report generation and clinical decision support functions.

Despite its effective results the proposed model leaves several directions for future research development. A new dataset that combines medical inquiries using Tamil-English code-mixing with their summaries will allow better linguistic diversity handling by the model and improve generalization across minority language domains. Performing extensive multimodal pretraining operations across multiple medical datasets strengthens model generalization by enabling better management of unexpected medical situations and uncommon disease cases. The implementation of advanced cross-modal attention approaches with adaptive attention and hierarchical fusion would enable the machine to discover sophisticated interrelationships between medical image contents and textual input thus boosting summarization quality. Real-time clinical testing of the model during medical documentation automation and radiology report summarization tasks will generate insights into its practical applications. Model combination methods like knowledge distillation and quantization

minimizes computational requirements so health facilities with limited resources can deploy the system for medical purposes. The proposed future research indicates methods to advance and strengthen the system for use across different complex medical fields.

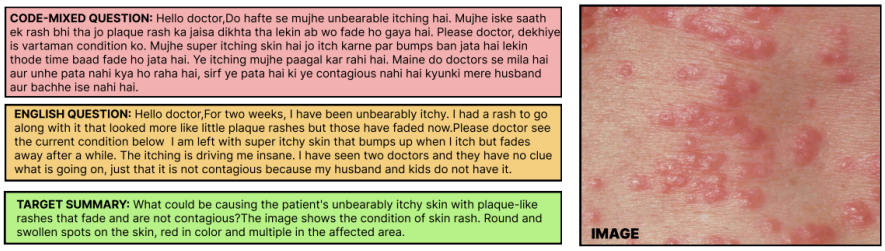

Fig. 4: Sample summaries generated by DistilBART in multimodal setting.

References

[1] Abacha, A.B., Demner-Fushman, D.: On the summarization of consumer health questions. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2228–2234 (2019)

[2] Abacha, A.B., Yim, W.-w., Michalopoulos, G., Lin, T.: An investigation of evaluation metrics for automated medical note generation. arXiv preprint arXiv:2305.17364 (2023). Available: https://arxiv.org/abs/2305.17364

[3] Abacha, A.B., M'rabet, Y., Zhang, Y., Shivade, C., Langlotz, C., Demner-Fushman, D.: Overview of the mediqa 2021 shared task on summarization in the medical domain. In: Proceedings of the 20th Workshop on Biomedical Language Processing, pp. 74–85 (2021)

[4] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018). Available: https://arxiv.org/abs/1810.04805

[5] Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization, pp. 65–72 (2005)

[6] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019)

[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., A. Askell, e.a.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)

[8] Das, A., Gamb"ack, B.: Identifying languages at the word level in code-mixed indian social media text. arXiv preprint arXiv:2302.13971 (2014). Available: https://arxiv.org/abs/2302.13971

[9] Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-Language Models for Vision Tasks: A Survey (2023)

[10] Gupta, D., Attal, K., Demner-Fushman, D.: A dataset for medical instructional video classification and question answering. In: Conference Proceedings (2022)

[11] Kumar, R., Chakraborty, R., Tiwari, A., Saha, S., Saini, N.: Diving into a sea of opinions: Multi-modal abstractive summarization with comment sensitivity. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pp. 1117–1126 (2023)

[12] Ghosh, A., Acharya, A., Jain, R., Saha, S., Chadha, A., Sinha, S.: CLIPSyntel: CLIP and LLM synergy for multimodal question summarization in healthcare (2023)

[13] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: LLaMA: Open and Efficient Foundation Language Models (2023)

[14] Yong, Z.X., Zhang, R., Forde, J.Z., Wang, S., Subramonian, A., *et al.*: Prompting multilingual large language models to generate code-mixed texts: The case of southeast asian languages. In: Sixth Workshop on Computational Approaches to Linguistic Code-Switching (2023)

[15] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: LLaMA 2: Open Foundation and Fine-Tuned Chat Models (2023)

[16] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7B (2023)

[17] Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., Werra, L., Fourrier, C., Habib, N., et al.: Zephyr: Direct Distillation of LM Alignment (2023)

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2020)

[19] Ghosh, A., et al.: Medsumm: A multimodal approach to summarizing code-mixed Hindi-English clinical queries. Springer (2024)

[20] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[21] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)

[22] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT (2019)

[23] Talburt, J.: The flesch index: An easily programmable readability analysis algorithm. In: Proceedings of the 4th Annual International Conference on Systems Documentation (1986)