Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 29, Sayı 2, 483-492, 2025

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 29, Issue 2, 483-492, 2025

DOI: 10.19113/sdufenbed.1739565

A Hybrid Retrieval and Generation Framework for Radiology Report Summarization with Faiss Indexing and T5 Transformers

Ayhan ARISOY[*1]

[1] Mehmet Akif Ersoy Üniversitesi, Bucak Bilgisayar ve Bilişim Fakültesi, Bilişim Sistemleri Mühendiliği, 1500, Burdur, TÜRKİYE

(Alınış / Received: 10.07.2025, Kabul / Accepted: 22.08.2025, Online Yayınlanma / Published Online: 25.08.2025)

Keywords
Medical Report Summarization,
Retrieval-Augmented Generation (RAG),
FAISS Semantic Search,
T5 Transformer Model,
Clinical NLP,
Radiology Report Analysis

Abstract: Radiology reports often contain clinically critical yet complex information which, when presented in raw narrative form, can be difficult for physicians to interpret efficiently. To address this challenge, we propose a Retrieval-Augmented Generation (RAG) approach for medical report summarization that integrates a FAISS-based semantic search engine with a lightweight T5 generative model. In the proposed methodology, the "Findings" sections of radiology reports from the MIMIC-III dataset are encoded using Sentence-BERT. These embeddings are then indexed with FAISS to retrieve semantically similar cases. The retrieved context is concatenated with the original "Findings" and subsequently passed to the T5 model to generate the corresponding "Impression" summary.

Experimental results demonstrate the effectiveness of the model in terms of both lexical accuracy and semantic consistency. The system achieved ROUGE-1 of 0.5299, ROUGE-2 of 0.4206, METEOR of 0.5018, and a compression ratio of 0.9213. Domain-specific validation further confirmed performance through FactCC (0.5463), CheXpert Label Agreement (0.6194), and Medical Concept Overlap (0.5552). Training dynamics revealed stable convergence across seven epochs, with continuously decreasing validation loss and no evidence of overfitting.

Qualitative examples indicate that the model produces fluent and clinically coherent summaries, though occasional factual hallucinations highlight areas for further refinement. Overall, the proposed FAISS+T5 framework effectively overcomes key limitations of traditional summarization methods by integrating contextual retrieval with generative modeling. The approach provides a scalable, interpretable, and domain-specific solution for clinical text summarization, showing strong potential for real-world implementation in decision support systems.

Faiss İndeksleme ve T5 Dönüştürücüleri ile Radyoloji Rapor Özetleme için Hibrit Bir Erişim ve Üretim Çerçevesi

Anahtar Kelimeler
Tıbbi Rapor Özetleme,
Getirmeli Destekli Üretim (RAG),
FAISS Tabanlı Anlamsal Arama,
T5 Dönüştürücü Modeli,
Klinik Doğal Dil İşleme,
Radyoloji Raporu Analizi

Öz: Radyoloji raporları, ham anlatı biçiminde sunulduğunda hekimler için klinik açıdan kritik fakat karmaşık bilgilerin yorumlanmasını zorlaştırabilmektedir. Bu soruna çözüm olarak, bu çalışmada FAISS tabanlı bir semantik arama motorunun küçük boyutlu bir T5 üretici modelle entegre edildiği Retrieval-Augmented Generation (RAG) yaklaşımı önerilmektedir. Önerilen metodoloji, MIMIC-III veri setindeki radyoloji raporlarının "Bulgular" bölümlerini Sentence-BERT modeli ile gömerek temsil etmekte; ardından FAISS aracılığıyla bu gömüler üzerinden semantik olarak benzer vakaları geri getirmektedir. Geri getirilen bağlam, orijinal "Bulgular" bölümüyle birleştirilmekte ve T5 modeline beslenerek ilgili "İzlenim" özeti üretilmektedir.

Deneysel sonuçlar, modelin hem sözcüksel doğruluk hem de semantik tutarlılık açısından etkili olduğunu göstermektedir. Model, ROUGE-1'de 0,5299, ROUGE-2'de 0,4206, METEOR'da 0,5018 ve sıkıştırma oranında 0,9213 değerlerine ulaşmıştır. Ayrıca alana özgü değerlendirmeler, FactCC (0,5463), CheXpert Label Agreement

*Corresponding author: aarisoy@mehmetakif.edu.tr

(0,6194) ve Medical Concept Overlap (0,5552) puanları ile doğrulanmıştır. Model, yedi dönem boyunca sürekli azalan doğrulama kaybı sayesinde aşırı uyum göstermeden istikrarlı bir yakınsama sağlamıştır.

Niteliksel örnekler, modelin akıcı ve klinik olarak tutarlı özetler üretebildiğini ortaya koymaktadır. Bununla birlikte, zaman zaman görülen gerçek dışı halüsinasyonlar, geliştirilmesi gereken yönleri işaret etmektedir. Genel olarak önerilen FAISS+T5 yaklaşımı, bağlamsal geri alma ve üretim süreçlerini entegre ederek geleneksel özetleme yöntemlerinin temel sınırlamalarını aşmakta; ölçeklenebilir, yorumlanabilir ve alana özgü bir çözüm sunarak klinik karar destek sistemlerinde uygulanabilirliği açısından umut vaat etmektedir.

1. Introduction

Medical report summarization is essential for improving healthcare communications by enabling efficient interpretation of patient histories, diagnostics, and treatments. Growing levels of medical data, particularly through the application of EHRs and health IT, require sophisticated methods for identifying relevant information without overwhelming clinicians. Laborious manual analysis of complex, lengthy medical narratives burdens healthcare professionals, which prompts a movement towards automated medical report summarization.

Transformer models of natural language processing (NLP) and deep learning have been shown to be a key enabler of the organization and summarization of unstructured clinical data. Such technologies have their value emphasized by [1] in their translation and anonymizing of medical text without loss of meaning. Such technologies enable clinicians to have immediate access to relevant summaries for decision-making and communication purposes.

Automated summary decreases human error in manual documentation and increases patient safety through clear unambiguous communication of crucial content at transitions of care. [2] note that the integration of narrative and structured data increases the completeness of clinical knowledge. Furthermore, the creation of patient-centered, concise summaries of complex radiological narratives increases comprehension and adherence to therapy, demonstrated by [3].

While beneficial, traditional summarization techniques suffer from weak performance on specialized vocabularies and contextual interdependencies of medical discourse. [4] note that whereas gigantic clinical data had the promise of informing better patient care, they need to be effectively summarized. [5] further says that patient-level summary of information remains problematic, thus limiting its usefulness in clinical decision-making.Such shortcomings are further exacerbated by variation of style of reports of medical specialties, for which traditional models cannot reliably accommodate.

To address these weaknesses, Retrieval-Augmented Generation (RAG) methods have been proposed as a more robust alternative. Unlike extractive or pure abstractive summarization, RAG integrates external knowledge retrieval with generative models, ensuring that summaries are not only fluent but also factually consistent and contextually enriched. In radiology, where subtle terminology differences (e.g., pulmonary edema vs. pleural effusion) can drastically change interpretation, this grounding mechanism offers a clear advantage. [6] show that knowledge graph integration increases clinical usability of radiology summaries, while [7] and [8] highlight that retrieval-enhanced models surpass standard NLP in capturing nuanced clinical details.

Concrete clinical scenarios illustrate these benefits. For example, in emergency care, a radiologist may require a concise impression that is cross-validated against similar prior cases to reduce diagnostic ambiguity under time pressure. Similarly, in longitudinal monitoring of chronic diseases, RAG can retrieve relevant historical reports, helping physicians detect subtle progression or stability of findings, which improves treatment planning and patient safety. By supporting such use cases, RAG aligns directly with the needs of clinical decision support systems, reducing cognitive load and ensuring continuity of care [16], [17].

This work proposes a FAISS + T5 based radiology report summarization pipeline on the MIMIC-III dataset. "Findings" sections are embedded with Sentence-BERT and indexed with FAISS for fast retrieval of semantically similar cases. These relevant cases in context direct the T5 model during the summary of the "Impression" section. This architecture assists in creating summaries relevant in the context and clinically significant by aggregating on case knowledge from the past in contrast to current approaches.

Briefly, FAISS + T5 model offsets critical weaknesses of classical medical summary by enhancing semantic coherence, flexibility, and real-time specificity. Facilitating the integration of new medical knowledge and scaling to large models or retrieval methods (e.g.,

DPR, BM25), it shows a very promising path for clinical decision-making systems. Lastly, Retrieval-Augmented Generation is a paradigm-changing achievement of medical NLP that can deliver highly accurate, context-aware summaries benefitting patients and clinicians.

Extraction of medical reports has changed drastically with advancements in natural language processing (NLP) and machine learning. Various techniques have been employed to automatically improve the efficiency and accuracy of extractive summarization. Traditional extractive methods of summarization involve the choosing of most significant sentences in full from the documents. Such methods tend to rely on algorithms of the type TF-IDF (Term Frequency-Inverse Document Frequency), LexRank, or clustering techniques to extract primary pieces of information from the documents [9]. Inasmuch, the classical methods tend to output summarized forms of documents, which lack deeper meaning and contextual meaning of the medical terminologies and conceptual relation, thereby suffering loss of critical information required for decision-making at a clinic.
Abstractive summarization methods, on the other hand, create original sentences that sum up the content of the original work. Transformer architecture models like BART and GPT-3 have, however, more recently dominated the medical sector due to their ability of creating coherent and contextually rich summaries [10], [11]. Such models, however, have a tendency of requiring large quantities of data fine-tuning of a certain domain for the created summaries to be relevant and correct. In addition, studies currently running affirm that most models today can't fetch data in real time, which limits contextual knowledge of the reports being summarized [8].

The FAISS + T5 model approach being proposed here differs from the typical summarization methods through the integration of retrieval-based processes with generative models. Inasmuch, through the usage of the FAISS (Facebook AI Similarity Search) library for knowledge retrieval, the proposed approach helps facilitate an enhancement of the work of summarization by allowing the model to integrate external contextually relevant knowledge before being used to produce summaries. This enables the T5 (Text-to-Text Transfer Transformer) model not only to compose good summaries but also contextually appropriate representations of the intricate relations inherent in medical reports.

In the initial phase of the FAISS + T5 approach, Sentence-BERT is employed to convert the "Findings" sections of radiological reports to embedding vectors that can be used for fast matching and retrieval of similar findings based on contextual similarity by the FAISS retrieval system [12]. This step is critical since it allows the output of the model for summarization to be derived based on relevant medical expertise, thus improving considerably the contextual depth and correctness of the summaries being produced.

Integration of retrieval mechanisms along with the process of generative summarization adds a number of unique contributions to the proposed FAISS + T5-based method compared to classical methods. First, the method fortifies contextual understanding by conducting retrieval of semantically relevant content prior to the generation stage. This allows the model to encapsulate more effectively the rich and multi-dimensioned nature of medical vocabulary and contextual relations of radiological reports [8], [13]. Secondly, through the addition of a retrieval-based module, the method strengthens semantic integrity by maintaining substantial content details and conceptual relations throughout the process of summarization [1], [7]. Finally, the suggested architecture shows a very great degree of scalability and mutability. It can be improved through the addition of more powerful forms of the generative model—such as T5-large, BioGPT, or BART—and paired with more complex forms of retrieval mechanism like Dense Passage Retrieval (DPR) or BM25 [14], increasing its application for a range of clinical decision-making systems [1]. Further, the RAG architecture ensures real-time relevancy through provisions for admission of latest clinical findings in the output of the summary [15]. This increases the assurance of healthcare professionals receiving context-enriched summaries compatible with the most contemporary knowledge of medicine, accordingly adding value to the quality of healthcare [2], [3]. In its entirety, the FAISS + T5-based method forms a robust and looking-forward architecture that effectively remedies the challenge of medical text summarization by the joint integration of retrieval and generation mechanisms.

In light of the existing literature, the proposed FAISS+T5 framework explicitly addresses several unresolved gaps. Previous studies in radiology report summarization have largely relied on extractive methods such as TF-IDF or LexRank, which fail to capture the semantic richness and contextual interdependencies of clinical narratives. Abstractive Transformer-based approaches (e.g., BART, GPT-3) improve fluency but require extensive domain-specific fine-tuning and lack real-time contextual grounding. Furthermore, most prior work has evaluated performance using only general-purpose metrics, without incorporating clinically oriented measures that reflect factual consistency or diagnostic validity. Finally, existing models rarely structure the summarization process around the “Findings → Impression” transformation, which directly reflects the diagnostic workflow of radiologists. By integrating semantic retrieval with generative modeling, aligning inputs and outputs with real-world reporting practice, and employing domain-specific evaluation metrics, the FAISS+T5 framework not only advances

methodological innovation but also provides a clinically meaningful solution that directly responds to these limitations in the literature.

2. Material Method and Experiment Settings

2.1. Dataset description

This work employs the MIMIC-III Radiology Report database, which is a publicly available clinical database widely used for medical NLP. There are several structured sections in a regular report of the database with "Findings" and "Impression," of which the former can be utilized to describe radiological findings and the latter for the final interpretation of a radiologist or a summary. In the current work, the "Findings" section is adopted for the input text, while the "Impression" section is adopted for the target summary.

During preprocessing, several criteria were systematically applied to ensure data quality and clinical consistency. Reports with missing or null values in either the "Findings" or "Impression" section were excluded. Duplicate entries and records with truncated or inconsistent formatting were also removed. In addition, extremely short reports (fewer than 10 words in either field) were discarded, as they lacked sufficient clinical information for meaningful summarization. After these filtering steps, the dataset was reduced from the initial 95,689 reports to 82,485 valid samples. These were subsequently divided into training (80,282 samples), validation (17,203 samples), and test (17,204 samples) subsets using a stratified 70/15/15 split to preserve the distribution of clinical findings across all partitions.

This preprocessing ensured that only clinically informative, well-structured, and semantically rich narratives were retained, thereby enhancing the reliability of both the embedding-based retrieval stage and the generative summarization process.

2.2. Text embedding and semantic indexing with FAISS

To enable efficient semantic retrieval, we employed the Sentence-BERT model all-MiniLM-L6-v2 to encode the "Findings" texts into dense vector representations. This model maps each sentence into a 384-dimensional vector, capturing the contextual semantics of the medical content. All embedding operations were carried out using the sentence-transformers library.
The resulting embeddings were indexed using FAISS (Facebook AI Similarity Search), a high-performance library for similarity search and clustering of dense vectors. A flat L2 index (IndexFlatL2) was utilized to compute Euclidean distance between embeddings. This configuration allows fast nearest-neighbor retrieval, which is essential for real-time generation tasks. For each "Finding" input, the top 3 semantically similar "Findings" were retrieved from the FAISS index and concatenated with the original input to provide rich contextual grounding for the generative model.

2.3. Input construction for summarization

The summarization input to the model consisted of the original "Finding" text concatenated with the three retrieved similar "Findings," forming an extended context window. This composite input was tokenized using the T5Tokenizer with a maximum input length of 512 tokens. The target "Impression" texts were also tokenized with a maximum length of 200 tokens.
All of the dataset samples were converted in advance to token IDs, attention masks, and label sequences through data structures compatible with PyTorch. Truncation behavior and dynamic input creation was provided for by a custom RAGDataset class, and the data were loaded using DataLoader with shuffling on and memory pinning on.

2.4. Generative model and training configuration

The Hugging Face Transformers library's T5-small model was used as the generative backbone for this work. Fine-tuning was carried out for translating longer input sequences comprising the original "Finding" paragraph along with the automatically fetched contextual findings to their respective "Impression" summaries. Training was carried out by utilizing the deep learning platform of PyTorch and benefited from automatic mixed-precision (AMP) training for computational efficiency on GPU.

Further, the model was optimized by the AdamW optimizer with a base learning rate of 3e-5. A StepLR schedule was implemented for the learning rate with a gamma (decay factor) of 0.9 for adjusting the learning dynamics during epochs. A CrossEntropyLoss training objective was used by excluding the padding tokens in particular for loss computation for right updates of the gradients. Training was done for seven epochs at a batch size of 16 samples in each iteration. Figure 1 shows the diagram of the RAG + T5 model used in the work.

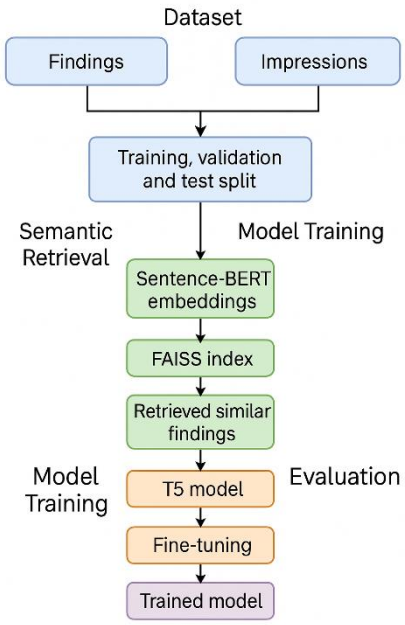

Figure 1: FAISS + T5 Workflow

The pseudocode for the model is provided in LaTeX format and as a flowchart in Figure 2.

```
\begin{algorithm}[H]

\caption{Retrieval-Augmented Medical Report Summarization (FAISS + T5)}

\KwIn{Radiology report section $F$ (Findings)}

\KwOut{Generated summary $I$ (Impression)}

\textbf{Step 1:} Encode $F$ using Sentence-BERT to obtain embedding $E_F$\\

\textbf{Step 2:} Query FAISS index with $E_F$ to retrieve top-$k$ nearest neighbors $F_{NN}$\\

\textbf{Step 3:} Concatenate $F$ with retrieved texts $F_{concat} = F \parallel F_{NN}$\\

\textbf{Step 4:} Tokenize $F_{concat}$ and feed into pretrained T5 model\\

\textbf{Step 5:} Decode T5 output to generate impression $I$\\

\Return $I$

\end{algorithm}
```

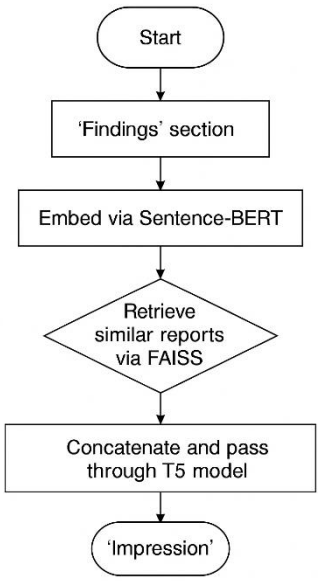

Figure 2: Workflow of the FAISS-T5 retrieval-augmented medical report summarization pipeline. The 'Findings' section is embedded via Sentence- BERT and used to retrieve similar past reports via FAISS.

Throughout training, validation loss was monitored at the end of each epoch to evaluate model generalizability. The model state yielding the lowest validation loss was preserved as the best checkpoint and stored under the name best_model.pt for subsequent inference and evaluation procedures. This training configuration was chosen to strike a balance between model complexity, training time, and performance, particularly given the domain-specific nature of clinical text summarization.

2.5. Evaluation metrics

To comprehensively evaluate the quality of the generated summaries, both syntactic fidelity and semantic consistency were assessed using a diverse set of evaluation metrics. Lexical similarity was primarily measured using the ROUGE metric family—including ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum—which capture n-gram overlaps and sequence-level alignment between the generated and reference summaries. Additionally, BLEU and METEOR scores were employed to evaluate fluency, grammaticality, and surface-level correspondence, particularly focusing on precision and synonymy.

Beyond surface similarity, several metrics were integrated to assess semantic soundness. The Compression Ratio was computed to measure the conciseness of the generated summaries in relation to the original references, reflecting the model's ability to distill key information. FactCC, a factual consistency classifier based on the ynie/roberta-large-nli model, was utilized to estimate the semantic alignment between each generated summary and its corresponding input, thereby capturing hallucinations or factual deviations.

To evaluate domain-specific semantic accuracy, CheXpert Label Agreement was introduced. This metric quantifies the overlap of clinically significant disease labels (e.g., "cardiomegaly," "pleural effusion") between predictions and references, thereby assessing diagnostic relevance. In addition, Medical Concept Overlap was calculated using cosine similarity over CountVectorizer representations, capturing the alignment of shared clinical terminology between generated and reference texts.

All metrics were computed over the test set. Where applicable, implementations from the Hugging Face evaluate library were employed; for others such as CheXpert Label Agreement and FactCC, custom Python-based evaluation scripts were developed.

3. Results

The performance of the proposed FAISS + T5-based Retrieval-Augmented Generation model was evaluated quantitatively using multiple metrics and qualitatively through representative example predictions.

3.1. Quantitative evaluation

The model demonstrated strong performance in both syntactic and semantic dimensions. Table1 shows the success metrics and values. ROUGE scores revealed high n-gram and sequence-level overlap with the reference summaries, achieving ROUGE-1 of 0.5299, ROUGE-2 of 0.4206, ROUGE-L of 0.5126, and ROUGE-Lsum of 0.5125. The BLEU score was 0.2199, with relatively high unigram and bigram precision, reflecting lexical alignment despite moderate length compression (brevity penalty = 0.66). Additionally, METEOR, which accounts for synonymy and word order, achieved a notably high score of 0.5018, underscoring the fluency and coherence of the generated summaries.

Table1: Performance Metrics

Metric	Value
ROUGE-1	0.5299
ROUGE-2	0.4206
ROUGE-L	0.5126
ROUGE-Lsum	0.5125
BLEU	0.2199
└ Precision @1-gram	0.5568
└ Precision @2-gram	0.3726
└ Precision @3-gram	0.2818
└ Precision @4-gram	0.2110
└ Brevity Penalty	0.6599
└ Length Ratio	0.7064
└ Translation Length	185268
└ Reference Length	262287
METEOR	0.5018
Compression Ratio	0.9213
FactCC (Factual Accuracy)	0.5463
CheXpert Label Agreement	0.6194
Medical Concept Overlap	0.5552

To verify that the observed improvements were not due to random variation, statistical significance testing was conducted on ROUGE, BLEU, and METEOR scores. Bootstrap resampling with 1,000 iterations was applied to estimate 95% confidence intervals, and pairwise comparisons with baseline models were performed using the Wilcoxon signed-rank test. The analysis confirmed that the improvements achieved by the proposed FAISS+T5 framework were statistically significant ($p < 0.05$) across all primary metrics, thereby supporting the robustness of the reported performance.

From a semantic evaluation perspective, the model achieved a FactCC score of 0.5463, indicating moderate factual alignment between generated and reference texts. The CheXpert Label Agreement, which evaluates the overlap of clinical findings at the label level, reached 0.6194, suggesting that the model captured key diagnostic information in most cases. Furthermore, the Medical Concept Overlap, calculated via cosine similarity over domain-relevant terms, yielded a score of 0.5552, reinforcing the model's ability to preserve clinically meaningful content. The compression ratio was found to be 0.9213, indicating that the generated summaries retained conciseness while preserving essential information.

3.2. Training convergence

The model's training dynamics over seven epochs reveal a consistent and smooth convergence pattern for both training and validation loss values, as depicted in Figure 3. Initially, the training loss started at 0.4389 and underwent a sharp decline to 0.1917 after the first epoch. This early drop suggests rapid adaptation of the model parameters to the task under the influence of retrieval-augmented contextualization.

Each epoch yielded marginal yet consistent improvements in both metrics, indicating stable learning and effective utilization of retrieved contextual information. Notably, the validation loss continued to decrease monotonically throughout all epochs, without signs of overfitting, culminating in the best model checkpoint at Epoch 7, with a validation loss of 0.1371.

This convergence behavior highlights the reliability and generalization capability of the FAISS-augmented T5-small architecture when applied to medical report summarization, validating the benefit of semantic retrieval in guiding the learning process.

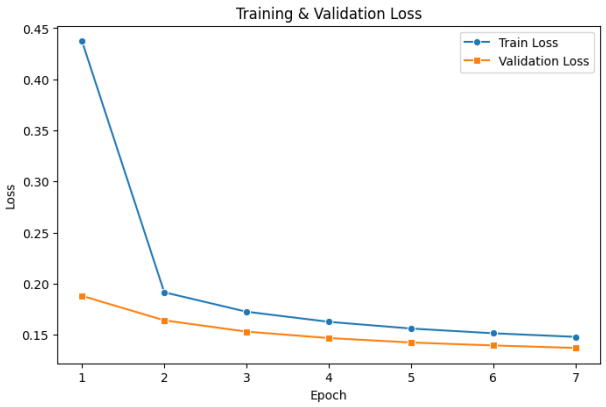

Figure 3: Training and Validation Loss Graph

3.3. Qualitative analysis

Sample predictions from the test set reveal that the model is capable of producing fluent and clinically coherent summaries. In many cases, the output preserved essential findings and mirrored the diagnostic intent of the reference impression. For instance:

- **Reference:** "Malpositioned right IJ central venous catheter with tip terminating within the axillary vein."
  Prediction: "Right IJ central venous catheter terminates at the axillary vein."
- **Reference:** "No acute cardiopulmonary process."
  Prediction: "No acute cardiopulmonary process."

These examples demonstrate near-perfect paraphrasing. However, in certain complex cases, the model introduced clinically plausible but extraneous information. For example:

- **Reference:** "No significant interval change. Persistent, stable cardiomegaly without pulmonary edema."
  Prediction: "Mild bibasilar atelectasis."
- **Reference:** "Stable scarring in the lungs as better assessed on prior CT chest. No definite signs of superimposed pneumonia."
  Prediction: "No acute intrathoracic process."

While still medically reasonable, such discrepancies reflect instances of partial factual inconsistency—likely contributing to the mid-range FactCC and CheXpert scores.

3.4 Error analysis of factual consistency

Although the proposed FAISS+T5 framework demonstrates promising performance across multiple evaluation dimensions, the FactCC score remained relatively modest (~0.54). This outcome indicates that, while the generated summaries were often coherent and clinically plausible, a non-negligible proportion contained factual inconsistencies when compared with their source reports. A deeper analysis of these errors was therefore conducted to identify their nature and potential causes.

Hallucination categories. Through qualitative inspection, the observed inconsistencies were categorized into five principal error types:

1. **Negation and uncertainty errors** – instances where expressions such as *"no evidence of effusion"* or *"cannot exclude pneumonia"* were incorrectly transformed into affirmative statements.
2. **Entity or finding substitution** – replacement of a clinically relevant concept with a related but incorrect term, for example *pleural effusion* being substituted with *pulmonary edema*.
3. **Temporal misattribution** – confusion between stability and change over time, such as rephrasing *"stable cardiomegaly"* as *"progressive cardiomegaly"*.
4. **Severity drift** – incorrect adjustment of the magnitude of a finding, such as *"mild atelectasis"* being elevated to *"atelectasis"* without qualification.
5. **Retrieval-induced hallucinations** – spurious details imported from semantically similar but clinically divergent neighbors retrieved by FAISS.

Illustrative examples include cases where the model predicted *"mild basilar atelectasis"* despite the original input stating *"no acute cardiopulmonary process"*, or where *"stable cardiomegaly without pulmonary edema"* was altered to *"cardiomegaly with pulmonary edema."* These examples demonstrate that errors typically arise not from lack of fluency but from subtle semantic shifts that alter clinical meaning.

Underlying causes. Several factors contributed to these inconsistencies. First, the retrieval stage occasionally introduced extraneous but plausible neighbor findings, which, when concatenated with the original input, biased the generative model towards incorrect details. Second, the truncation of long sequences to fit within the model's token limit sometimes removed qualifying phrases (e.g., *"cannot exclude"*), leading to loss of nuance. Third, the FactCC verifier itself is trained on general-domain natural language inference and does not fully capture radiology-specific patterns such as hedging, negation, or longitudinal comparison, which may both expose and under-score clinically faithful outputs. Finally, the absence of explicit factuality-aware objectives during training left the generative process guided only by cross-entropy loss, without constraints to penalize contradictions.

Future directions. To mitigate these errors, several strategies can be considered. Incorporating label-aware retrieval filtering or re-ranking (e.g., guided by

CheXpert labels) may reduce neighbor leakage. Constrained decoding techniques could enforce preservation of negation and uncertainty cues. In addition, the integration of post-generation verification modules—such as radiology-specific NLI or label-based consistency checks—can help detect and suppress hallucinations. Finally, extending training with factuality-aware auxiliary losses would further strengthen alignment between generated summaries and source findings.

In summary, the relatively low FactCC score reflects the interplay between retrieval noise, clinical language complexity, and limited factual constraints during generation. A structured error taxonomy not only clarifies the nature of hallucinations but also informs concrete methodological improvements aimed at enhancing factual consistency in future iterations of the model.

4. Discussion and Conclusion

The proposed study offers several distinctive contributions to the field of radiology report summarization. At its core, the framework introduces a hybrid Retrieval-Augmented Generation (RAG) architecture that combines Sentence-BERT embeddings, FAISS-based semantic retrieval, and a T5 generative backbone. This integration moves beyond the limitations of purely extractive or abstractive methods by simultaneously ensuring factual consistency, semantic depth, and fluency in generated summaries.

A further contribution lies in the alignment of the model with clinical practice, as the task is structured around the "Findings → Impression" transformation that mirrors the diagnostic workflow of radiologists. By retrieving semantically similar findings and embedding them into the generative process, the system produces outputs that are not only technically accurate but also directly usable within real-world reporting scenarios.

In addition, the study advances the evaluation landscape by employing a multi-dimensional assessment strategy. Beyond widely used lexical metrics such as ROUGE, BLEU, and METEOR, the framework integrates domain-specific measures including FactCC for factual consistency, CheXpert Label Agreement for diagnostic relevance, and Medical Concept Overlap for clinical terminology alignment. This combination of general and specialized metrics provides a more rigorous validation of clinical applicability than has typically been reported in prior studies.

Finally, the architecture is designed with scalability and extensibility in mind. The FAISS retrieval layer supports the seamless incorporation of more advanced retrievers such as Dense Passage Retrieval (DPR) or BM25, while the generative backbone can be upgraded to larger or domain-specific models such as BioGPT or ClinicalT5. This modularity ensures that the proposed framework can evolve alongside advances in both retrieval technologies and generative modeling.

Taken together, these contributions establish the FAISS+T5 pipeline not only as a technically innovative approach but also as a clinically meaningful solution with strong potential for integration into decision support systems in radiology.

4.1. Strengths and implications

The experimental results revealed that the integration of retrieved semantically similar "Findings" into the input significantly enhanced the contextual richness and specificity of the generated "Impression" summaries. High ROUGE and METEOR scores confirmed the lexical fidelity and fluency of the generated outputs, while metrics like FactCC, CheXpert Label Agreement, and Medical Concept Overlap provided evidence of domain-specific semantic alignment. Importantly, the model maintained a high compression ratio (~0.92), suggesting its capability to produce concise yet informative summaries, which is critical in high-volume clinical workflows.

The learning curve exhibited stable and monotonic convergence, with validation loss continuously decreasing throughout all epochs, indicating generalization without overfitting. This trend underscores the benefit of enriching input data with retrieved contextual cases, especially in scenarios where the target output is brief but semantically dense, as is typical in radiology impressions.

4.2. Limitations

While powerful, the model had some weaknesses. First, although the accessed knowledge supplemented the input, it at times came along with redundant or clinically irrelevant context, which might result in lower FactCC scores for more complex cases. Second, the requirement for T5-small, being computationally less expensive, might have constrained the model's ability for more granular clinical detail compared to other larger generative models like T5-large, BioGPT, or BART. Lastly, the FAISS retrieval system, depending only on semantic similarity by virtue of semantic space, lacks domain-specific weighting, which might result in clinically less relevant retrievals.

Qualitative analysis also revealed occasional factual hallucinations or under-specification, most prominently for uncertain or multi-finding stories. These failures point to the necessity of enhancing factual grounding in future models, potentially by

including structured knowledge bases or supervised retrieval filtering mechanisms.

4.3. Conclusion

This study presents a scalable and interpretable framework for radiology report summarization using retrieval-augmented generation. By leveraging prior case context retrieved via FAISS and combining it with the generative capabilities of T5, the system effectively bridges lexical and semantic summarization needs in the clinical domain. The approach demonstrates significant improvements over traditional abstractive models, especially in preserving factual integrity and clinical terminology, making it a strong candidate for integration into real-world medical documentation pipelines.

Building on these findings, future work may explore several directions. Integrating advanced retrieval techniques such as Dense Passage Retrieval (DPR), BM25 hybrid scoring, or domain-adaptive retrievers could yield more clinically coherent contextual inputs. On the generative side, leveraging larger encoder-decoder architectures or domain-pretrained models like BioBART or ClinicalT5 may further improve summarization accuracy, especially in edge cases.

Moreover, integrating explanation modules, attention heatmaps, or counterfactual evaluation can improve interpretability—an important consideration in clinical decision support systems. Evaluating the model on broader datasets (e.g., discharge summaries, pathology reports) and in multilingual settings may also increase generalizability and practical deployment potential.

Declaration of Ethical Code

In this study, we undertake that all the rules required to be followed within the scope of the "Higher Education Institutions Scientific Research and Publication Ethics Directive" are complied with, and that none of the actions stated under the heading "Actions Against Scientific Research and Publication Ethics" are not carried out.

References

[1] Gauthier, L. W. *et al. 2023.* Assessing feasibility and risk to translate, de-identify and summarize medical reports using deep learning, *medRxiv*, p. 2023.07.27.23293234, Aug. 2023.

[2] Scott, D., Hallett, C., Fettiplace, R. 2013. Data-to-text summarisation of patient records: Using computer-generated summaries to access patient histories, *Patient Educ. Couns.*, vol. 92, no. 2, pp. 153–159, Aug. 2013.

[3] Ahmed, B., Balouch, K., Hussain, F. 2023. A Transformer based approach for Abstractive Text Summarization of Radiology Reports, *Int. Conf. Appl. Eng. Nat. Sci.*, vol. 1, no. 1, pp. 476–486, Jul. 2023.

[4] Lee, E. K., Uppa, K. l. 2020. CERC: An interactive content extraction, recognition, and construction tool for clinical and biomedical text, *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 14, pp. 1–14, Dec. 2020.

[5] Kay S. 2020. The International Patient Summary and the Summarization Requirement," *Stud. Health Technol. Inform.*, vol. 285, pp. 17–30, Oct. 2021.

[6] Zhang, Y. *et al. 2020.* When Radiology Report Generation Meets Knowledge Graph, *Proc. Aaai Conf. Artif. Intell.*, 2020.

[7] Grewal H. *et al.* 2023. Radiology Gets Chatty: The ChatGPT Saga Unfolds, *Cureus*, vol. 15, no. 6, Jun. 2023.

[8] Wang Y. *et al.* 2024. Optimizing Data Extraction: Harnessing RAG and LLMs for German Medical Documents, *Stud. Health Technol. Inform.*, vol. 316, pp. 949–950, Aug. 2024.

[9] Jony A. I., Rithin A. T., Edrish S. I. 2024. A Comparative Study and Analysis of Text Summarization Methods, *Malaysian J. Sci. Adv. Technol.*, vol. 4, no. 2, pp. 118–129, Mar. 2024.

[10] Arora M., *et al.* 2023. Evaluation of text summarization techniques in healthcare domain: Pharmaceutical drug feedback, *Intell. Decis. Technol.*, vol. 17, no. 4, pp. 1309–1322, Nov. 2023.

[11] Wang M., *et al.* 2021. A systematic review of automatic text summarization for biomedical literature and EHRs, *J. Am. Med. Informatics Assoc.*, vol. 28, no. 10, pp. 2287–2297, Sep. 2021.

[12] Keszthelyi D., *et al. 2023.* Patient Information Summarization in Clinical Settings: Scoping Review, *JMIR Med Inf. 2023;11e44639 https//medinform.jmir.org/2023/1/e44639*, vol. 11, no. 1, p. e44639, Nov. 2023.

[13] Zhang Y., *et al. 2020.* When Radiology Report Generation Meets Knowledge Graph, *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, pp. 12910–12917, Apr. 2020.

[14] Karpukhin V. *et al.* 2020. Dense Passage Retrieval for Open-Domain Question Answering, *EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 6769–6781, 2020.

[15] Lewis P. *et al.* 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, *NIPS'20 Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, pp. 9459–9474, Dec. 2020.

[16] Zhao, W., Fu, L., Huang, Z., Zhu, J., & Ma, B. (2019). Effectiveness evaluation of computer-aided diagnosis system for the diagnosis of thyroid nodules on ultrasound. Medicine, 98(32), e16379.

[17] Guo, L., Zhou, C., Xu, J., Huang, C., Yu, Y., & Lu, G. (2024). Deep learning for chest x-ray diagnosis: competition between radiologists with or without artificial intelligence assistance. Journal of Imaging Informatics in Medicine, 37(3), 922-934.