Summarizing Medical Records Using the T5 Text-to-Text Framework

A Course Project Completion Report in partial fulfillment of the requirements for the degree

Bachelor of Technology

in

Computer Science & Artificial Intelligence

BY

Name	**Hall Ticket**
KUSHI RAJ KANCHU	2203A52030
DAYYALA PRANAY	2203A52013
ANANYA PAGADALA	2203A52046
SANDESH NAYAK	2203A52087

Submitted to

DR. SANDEEP KUMAR

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCESR
UNIVERSITY, ANANTHASAGAR, WARANGAL

April, 2025.

SR UNIVERSITY

SCHOOLOFCOMPUTERSCIENCE&ARTIFICIAL INTELLIGENCE

CERTIFICATE

This is to certify the Project Report entitled “**MEDICAL TEXT SUMMARIZATION**” is a record of Bonafide **KUSHI RAJ KANCHU(2203A52030), DAYYALA PRANAY(2203A52013), ANANYA PAGADALA(2203A52046)** and**SANDESH NAYAK(2203A52087)**, in partial fulfillment of the award of the degree of**BACHELOR OF TECHNOLOGY**in **COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE,** during the academic year **2024-2025** under the guidance and supervision.

Supervisor
Dr.Sandeep Kumar
Professor & Associate Dean (Data Science)
SR University

Head of the Department
Dr. M. Sheshikala
Prof & HOD (CS&AI)
SR University

Tableof Contents

1.1 INTRODUCATION

The healthcare sector creates huge volumes of text data in the modern digital world. Clinical notes, laboratory reports, radiological results, summary of patient release, electronic health records (EHRS) and medical research articles are included in it. The net volume and complexity of these documents can make doctors, nurses and researchers demanding to quickly access and understand the most suitable details, although this information is essential for understanding the patient's health and expanding medical knowledge. Summary of medical texts has become a practical way to deal with this problem. It helps users to quickly understand important information by automatically distilling lengthy medical documents into brief and bright summaries using natural language processing techniques (NLP).

There are mainly two types of medical text summary: mining and abstract. Selection and assembly of the most important sentences or phrases from the source material is how the summarization works. Although it is relatively easy to construct and often provide reliable results, the summary may not sound natural or flow well. On the contrary, abstract summary means understanding the whole text and coming up with fresh sentences that mediate the same concepts in a way that is more organic and human. This more sophisticated approach uses deep learning models such as T5, Biogpt and Bert. In general, abstract summary is easier to read, but if the model is not properly trained, they can sometimes introduce mistakes.

In the clinical environment where decision -making must be accepted quickly and time is the essence, summary is particularly important. Doctors often have a limited amount of time to review test results, treatment plans and patient history.

They can better decide and better understand the patient's condition using a well - written summary. Sumarization can also benefit from medical scientists who have to read and evaluate many scientific articles. In addition, patients can help to simplify difficult medical terminology into comprehensible explanations. The summary also reduces the paperwork that healthcare professionals have to complete, which can reduce stress and burnout.

However, the creation of useful tools for a summary for the medical area is not without its problems. One of the main obstacles is complex terminology used in medical texts, which includes specialized phrases, technical terms and abbreviations. Proper interpretation of these requires knowledge specific to the domain. Since patients' records are confidential and demanding for accessing the AI Pro model, there is also a lack of high quality, marked medical data. In addition, medical texts often contain time or contextual information that is difficult to understand for machines, such as the course of the disease or previous treatment. In addition, although abstract summarization can lead to summary of fluids, it sometimes creates inaccurate or irrelevant content that is unacceptable in medical applications where accuracy is necessary. Finally, common evaluation techniques, such as the Bleu score, are the goal to improve the accuracy, reliability and ease of integrating tools for summary into real health systems in the future. In order to provide real -time summary, when patients are treated, these tools could be directly associated with hospital EHR systems. To summarize important ideas from books and articles, they could also help in medical education.

These tools can help patients understand health care and make it easier to understand medical instructions. Future models will also have to explain how and why specific content was included in a summary to support confidence and adoption.

1.2 NEED OF PROJECT

The healthcare sector generates a huge number of unstructured text data every day from electronic health records (EHR), clinical notes and summons for discharge to diagnostic reports and medical research publications. Although this data is invaluable to provide patient care and research management, its mere volume and complexity often impress health workers. Manual reading, interpretation and analysis of such lengthy documents is time consuming and can lead to missing critical information, which potentially affects the quality and time of care.

In a rapidly developing clinical environment, there is an urgent need for a system that can automatically summarize medical texts into brief, clear and meaningful versions. Such a system can significantly reduce the cognitive burden on doctors, nurses and other healthcare professionals by allowing quick access to key patient information. This in turn can improve the speed and accuracy of diagnoses, streamline treatment planning and increase the patient's overall results.

In addition, with the rapid growth of scientific publications and growing demand for evidence -based practices, medical scientists often have to go through hundreds of articles to find relevant information. The intelligent summarization tool can extract and represent the most important findings of large volumes of literature, help scientists remain updated and make informed decisions more efficiently.

In this context, the development of an accurate and reliable system is not only a technical challenge summary of a medical text - it is a necessity. It has the potential to transform the way in which health care data is consumed and understood. The aim of this project is to use techniques using natural language processing (NLP) to bridge the gap between data overload and action insight, which eventually supports better provision of health care, reduce burnout and support improved communication across medical ecosystem.

1.3 LITERATURE WORK

Literature on medical text summarization highlights the shift from rule-based to deep learning models for improved accuracy and contextual understanding. Recent studies emphasize the effectiveness of transformer-based architectures like BioGPT and Pegasus in biomedical domains.

Table1: Medical text analysis of existing state of art methods

S.No	Author	Year	Methodology	Name of Dataset	Remarks/Results	Limitations
01	Balamurugan Palanisamy	2025	Pretrained Transformers(GPTs) BERT, ROBET	COVID-19 Public Media Dataset	Accuracy-89% Precision-80% F1-score-78% Recall-77%	Pretrained models like BERT may struggle with longer or complex texts, and risk overfitting.
02	Ayesha Khaliq	2024	BERT-Encoder LDA, HGNN, GAT	PubMed	F1-score:46.03 Rouge-1:21.42 Rouge-1:39.71	The integrated framework may require high computational resources and training time
03	Muhammad Hafizul	2023	Automatic text summarization	-PubMed -Medical sum	Rouge-2:0.348 MOABC-:0.342 LSSA-0.335	Optimization-based ATS methods are computationally expensive and slow.
04	Azzedine Aftiss	2024	K-Means clustering BERT, PageRank Longformer-Encoder-	Cochraane and MS^2	ROUGE-1:29.41% ROUGE-2:6.57% ROUGE-L:18.31%	The hybrid model lacks explicit discussion on scalability for real-time

			Decoder		BERTScore-85.95%	applications.
05	Ghandeer Althari	2022	BERT, ALBERT XLNet, ELECTRA	BioScope corpus	Accuracy-99% Precision-98.97% Recall-99.86% F1 score-98.22%	Transformer models have high-dimensional parameters and are time-consuming to fine-tune.
06	Jimin lee	2024	RoBERTa-large Pegausus RoBERTa-base DISTILLRoBERTa-base	MSLR Cochrane	Accuracy-89% Precision-84% F1-score-86% Recall-80%	Human evaluation data may include biases and limited real-world generalizability.
07	Seyyede Zahra	2024	ReQuEST BART	CQADup-Stack	Rouge-L:37.37	The paper does not clearly outline specific limitations of the ReQuEST model.
08	Yuan-chi-yang	2021	Rouge-1	COVID-19	Rouge-1:65% F1-score:45.43%	Limited metric focus (mainly ROUGE-1), and does not account for semantic coherence.
09	Nor Asilah Wati Abdul Hamid	2024	ATS, MOABC FbTS	BBC News	MoABC-0.312 ATS-0.34	Optimization techniques like MOABC can be computationally expensive and less interpretable.
10	Christoph Schommer	2024	LED(Longformer-Encoder-Decoder	MS^2	ROUGE-1:28.79% ROUGE-2:8.22% ROUGE-L:17.93%	The paper lacks discussion on the adaptability of the model across diverse biomedical domains

1.4RESEARCH GAPS

- **Lack of generalization across domains:**

  Many models Summary try to generalize well across different data sets (eg medical, legal, finance). Research with multiple domain training, adaptation of domains or portable models that can work well in different industries.

- **Ineffective calculation and scalability:**

  High computing costs of transformer -based models, especially when working with large data sets, represent an important challenge. Research is needed to develop light, efficient models or optimization techniques that reduce computing load while maintaining performance.

- **Prejudices in data sets and model predictions:**

  Prestressing data sets and lack of diversity of data training can lead to beveled results. It is necessary to research methods to identify and alleviate distortion in data sets, which ensures that sumarization models are fair, impartial and robust across different demographic or context groups.

- **Metrics of limited evaluation:**

  The metrics of the current evaluation (eg Rouge, Bertscore) often cannot capture the holistic assessment of the model, especially in terms of accuracy, accuracy, induction and usability in the real world. Future work should focus on the development of more complex metrics of evaluation specific to the task that reflect the quality of human summary.

1.5OBJECTIVES

- To improve text summary of medical report using transformer.
- To optimize the proposed models for better accuracy, robustness across domains and real -time application, while improving the rating metrics for more comprehensive evaluation.

1.6 PROPOSED WORK

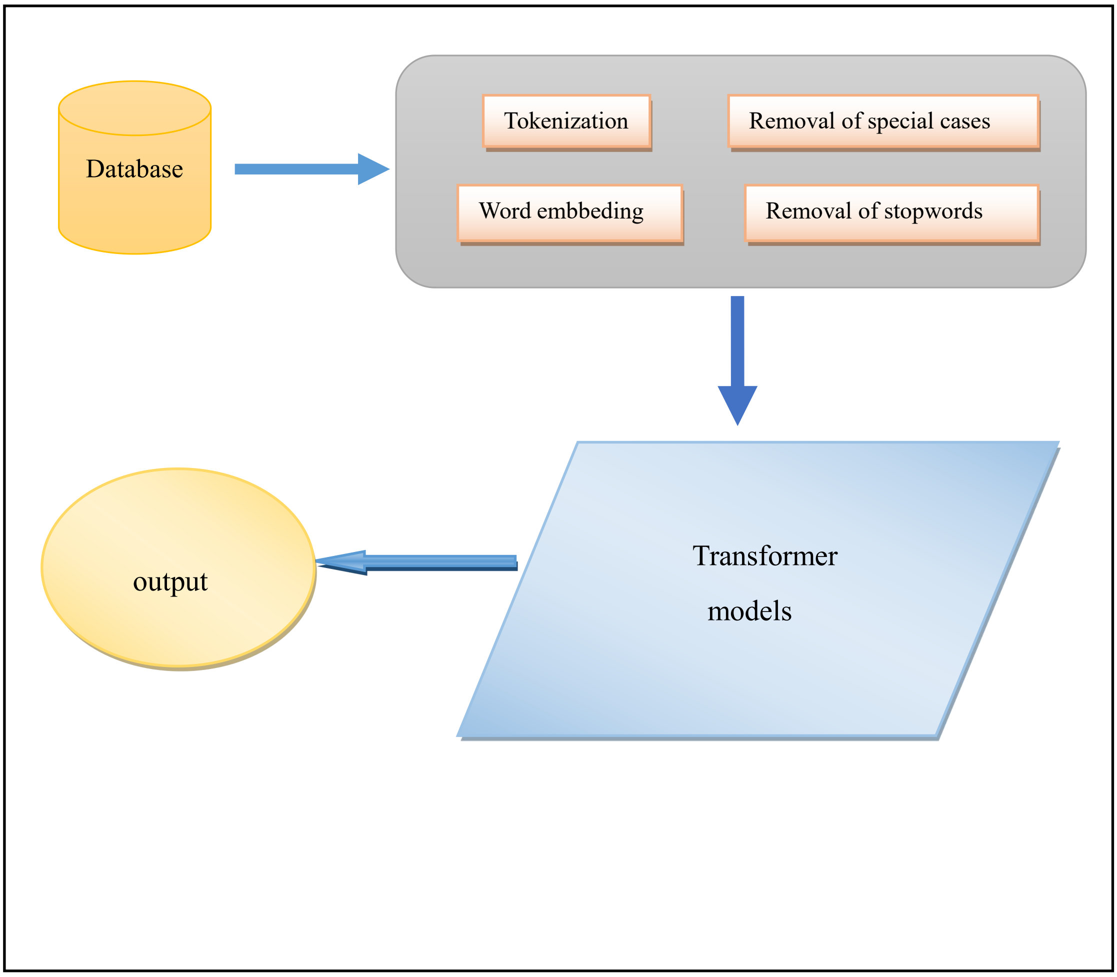
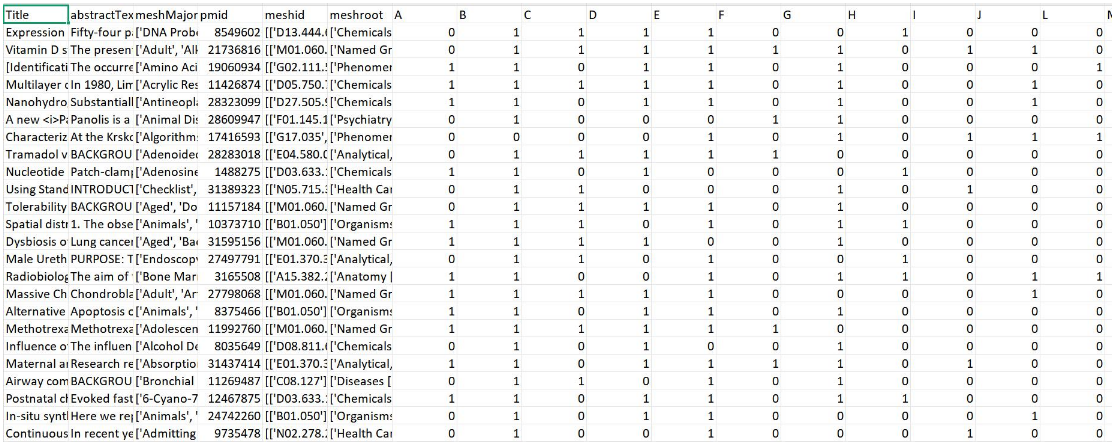


Figure 1: Flow chart of the proposed work

Step 1: Dataset Used

The data file is derived from PUBMED ABSTRACTS and is used for multiple branded classification tasks in biomedical domain. Each record usually contains abstract of biomedical research and associated labels with multiple eye (headings of medical subjects). It allows models to learn from scientific text and predict more relevant biomedical categories for each abstract. This processed version probably includes cleaned text, standardized labels and formats ready for use for model training. It is generally used in the tasks of the processing of natural language (NLP), such as the classification of medical documents, obtaining information and semantic indexing.



Title	abstractTex	meshMajor	pmid	meshid	meshroot	A	B	C	D	E	F	G	H	I	J	L	M
Expression	Fifty-four p	['DNA Prob	8549602	[['D13.444.6	['Chemicals	0	1	1	1	1	0	0	1	0	0	0	
Vitamin D s	The presen	['Adult', 'Alk	21736816	[['M01.060.	['Named Gr	0	1	1	1	1	1	1	0	1	1	0	
[Identificati	The occurre	['Amino Aci	19060934	[['G02.111.5	['Phenomer	1	1	0	1	1	0	1	0	0	0	1	
Multilayer c	In 1980, Lim	['Acrylic Res	11426874	[['D05.750.7	['Chemicals	1	1	1	1	1	0	1	0	0	1	0	
Nanohydro	Substantiall	['Antineopla	28323099	[['D27.505.9	['Chemicals	1	1	0	1	1	0	1	0	0	1	0	
A new <i>Pa	Panolis is a	['Animal Dis	28609947	[['F01.145.1	['Psychiatry	0	1	0	0	0	1	1	0	0	0	0	
Characteriz	At the Krsko	['Algorithms	17416593	[['G17.035',	['Phenomer	0	0	0	0	1	0	1	0	1	1	1	
Tramadol v	BACKGROU	['Adenoidec	28283018	[['E04.580.0	['Analytical,	0	1	1	1	1	1	0	0	0	0	0	
Nucleotide	Patch-clamp	['Adenosine	1488275	[['D03.633.1	['Chemicals	1	1	0	1	0	0	0	1	0	0	0	
Using Stand	INTRODUCT	['Checklist',	31389323	[['N05.715.3	['Health Car	0	1	1	0	0	0	1	0	1	0	0	
Tolerability	BACKGROU	['Aged', 'Do	11157184	[['M01.060.	['Named Gr	0	1	1	1	1	0	1	0	0	0	0	
Spatial distr	1. The obse	['Animals', '	10373710	[['B01.050']	['Organisms	1	1	1	0	1	0	1	1	0	0	0	
Dysbiosis o	Lung cancer	['Aged', 'Bac	31595156	[['M01.060.	['Named Gr	1	1	1	1	0	0	1	0	0	0	0	
Male Ureth	PURPOSE: T	['Endoscopy	27497791	[['E01.370.3	['Analytical,	0	1	1	0	1	0	0	1	0	0	0	
Radiobiolog	The aim of	['Bone Marr	3165508	[['A15.382.2	['Anatomy |	1	1	0	0	1	0	1	1	0	1	1	
Massive Ch	Chondrobla	['Adult', 'Ar	27798068	[['M01.060.	['Named Gr	1	1	1	1	1	0	0	0	0	1	0	
Alternative	Apoptosis c	['Animals', '	8375466	[['B01.050']	['Organisms	1	1	0	1	1	0	1	0	0	0	0	
Methotrexa	Methotrexa	['Adolescen	11992760	[['M01.060.	['Named Gr	1	1	1	1	1	1	0	0	0	0	0	
Influence o	The influen	['Alcohol De	8035649	[['D08.811.6	['Chemicals	0	1	0	1	0	0	1	0	0	0	0	
Maternal an	Research re	['Absorptio	31437414	[['E01.370.3	['Analytical,	1	1	0	1	1	1	1	0	1	0	0	
Airway com	BACKGROU	['Bronchial	11269487	[['C08.127']	['Diseases [	0	1	1	0	1	0	1	0	0	0	0	
Postnatal ch	Evoked fast	['6-Cyano-7	12467875	[['D03.633.1	['Chemicals	1	1	0	1	1	0	1	1	0	0	0	
In-situ synth	Here we rep	['Animals', '	24742260	[['B01.050']	['Organisms	0	1	0	1	1	0	0	0	0	1	0	
Continuous	In recent ye	['Admitting	9735478	[['N02.278.2	['Health Car	0	1	0	0	1	0	0	0	1	0	0	

Figure 2: PubMed Multi-Label Text Classification Dataset

Step 2: Preprocessing

1. Text Cleaning (Removal of Special Characters, Lowercasing)

   Text cleaning is the process of preparing and transforming raw text data into a clean and consistent format so that it can be used effectively for analysis or machine learning tasks.

   It usually involves:

   - Removing unwanted characters (e.g., punctuation, symbols)
   - Converting text to lowercase
   - Eliminating extra spaces or line breaks
   - Removing stopwords (common words like "the", "is", "and" if needed)
   - Correcting spelling mistakes

- Tokenization (splitting sentences into words)

Step 3: Methodology

- The T5 is a transformer -based model that frames all NLP tasks as text problems using the architecture of the unified encoder decoder.
- Biogpt is a language model based on GPT pre -trained on biomedical texts that performs specialized Biomedical tasks of NLP.
- Bleu is a metric that evaluates the quality of machine translation by measuring the overlap of the n-mig between generated and reference translations.
- Pegasus is a model summary that uses the goal of generating the gap, to pre -trad on masked important sentences for an abstraction summary.

Step 4: Training

- RNN is a rocess sequences by memorizing previous steps, but fighting long -term addiction.
- LSTM is to Improve RNNS with goals to remember long -term information effectively.
- CNN is a Quickly extract local patterns in the text using filters, ideal for classification tasks.
- TF-IDF (frequency-in-version of the document) is a statistical method that scored words based on how important they are for the collection document, often used to extract keywords and classification of documents.

1.7 RESULT SECTION

1.7.1 Hardware Configuration:

The device, named DESKTOP-FK96IKL, is powered by an Intel(R) Core(TM) i5-10310U CPU running at a base frequency of 1.70GHz, with a boost up to 2.21GHz. It features a quad-core architecture with 8 threads, supported by 16 GB of installed RAM. The system operates on a 64-bit Windows 11 Pro operating system and is built on an x64-based processor architecture, with OS build 22631.5039.

1.7.2 Validation of Work

In medical text, verification assesses whether the generated summary precisely captures key medical information and is considered useful and reliable by healthcare workers or end users. It confirms that the summons meet the intended purposes of the provision of brief and trusted representations of the original medical texts.

$$\text{Accuracy} = \frac{TP+TN}{TP+Tn+FP+FN} \tag{1}$$

$$\text{Precision}=\frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall}=\frac{TP}{TP+FN}. \tag{3}$$

$$\text{F1\_score}=\frac{2 * P * R}{P+ R} \tag{4}$$

1.7.3 Result Outcome:

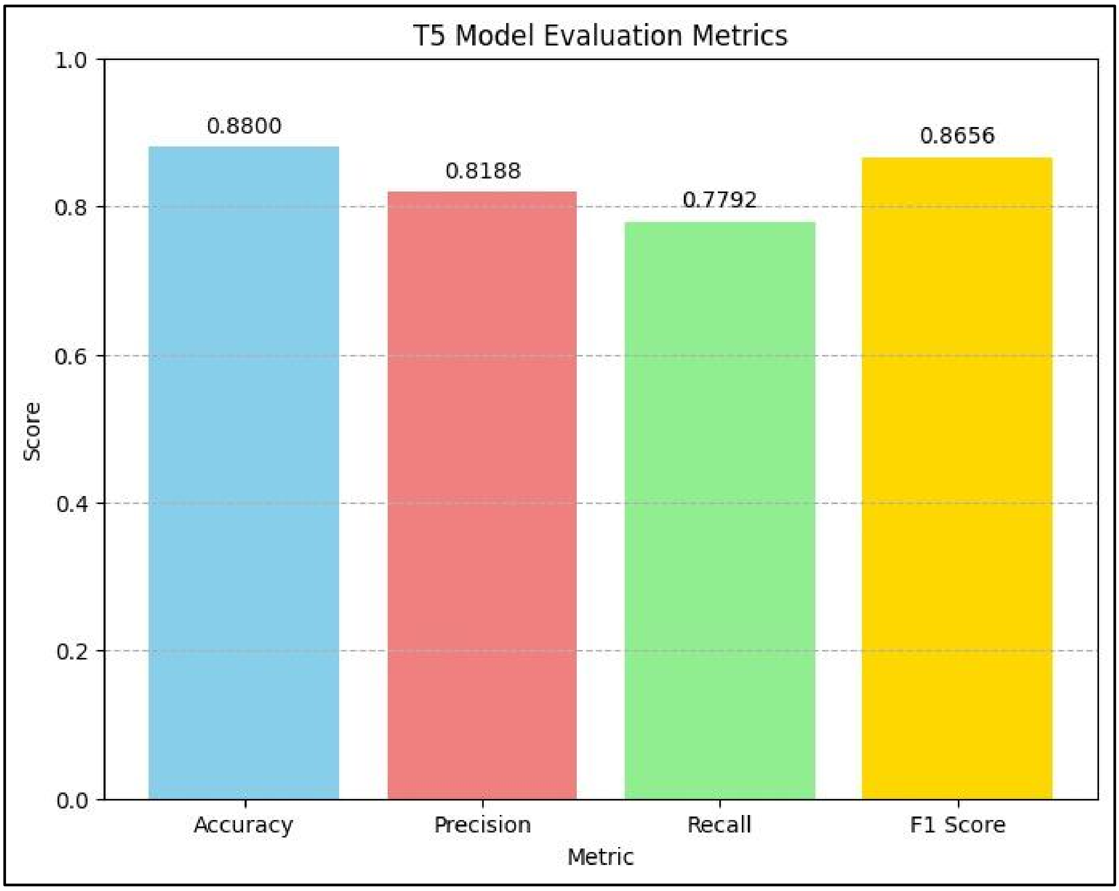


Figure 2:T5 Model outcomes in various form i.e accuracy, precision, recall and F1-Score on pubmed open access Database

The performance of the T5 model in the PUBMED open access database is illustrated by four metrics of the keys. It achieved a high accuracy of 0.8800, which reflected its strong ability to generate proper summary. The exact score of 0.8188 shows that the model is effective in avoiding irrelevant or incorrect detail. Meanwhile, the download score of 0.7792 suggests that it successfully captures a substantial part of the relevant information from the source. The F1 score, which balances accuracy and induction, costs 0.8656, emphasizing the ability of the model to generate relevant and accurate summary. Overall, these results represent a strong and promising basic performance for the T5 model in summary of biomedical text.

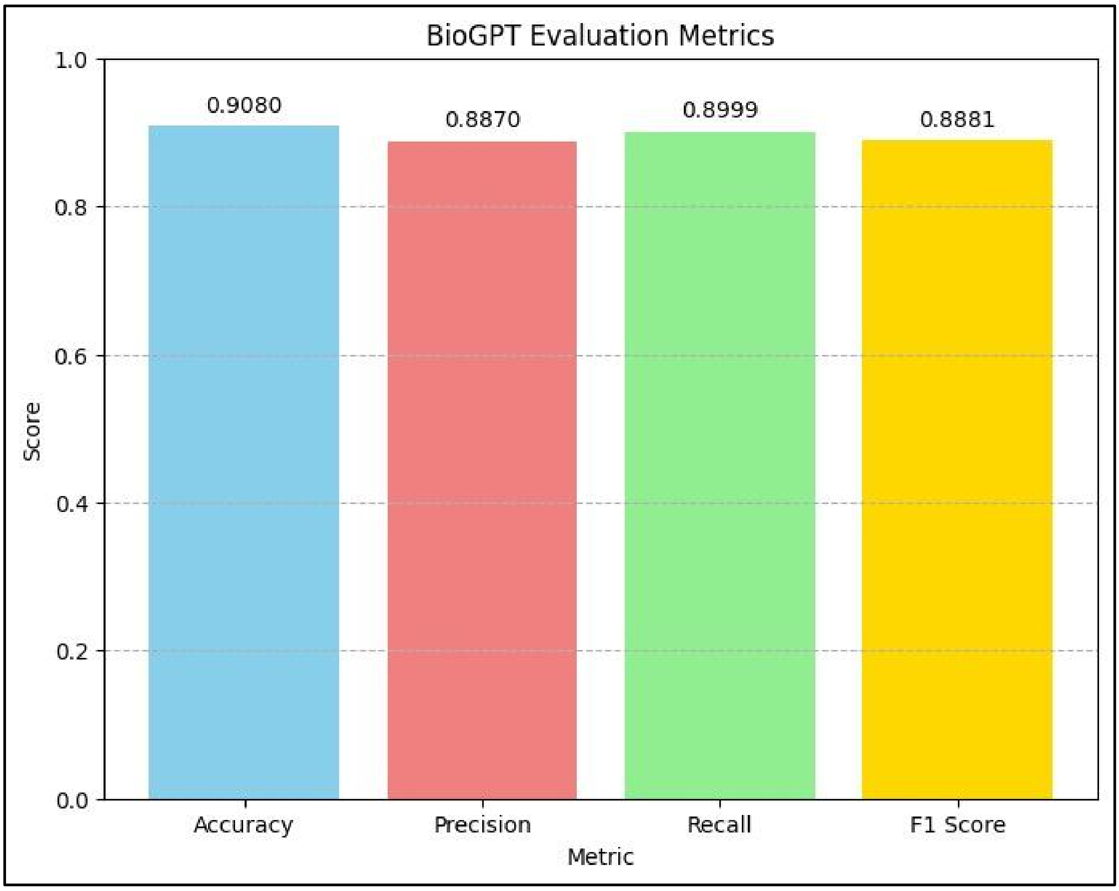


Figure 3: BioGPT Model outcomes in various form i.e accuracy, precision, recall and F1-Score on pubmed open access Database

The Biogpt model was evaluated in the PUBMED open access database, which represents a powerful performance across several key metrics. It achieved a high accuracy of 0.9080, suggesting that the model correctly classified a significant part of the data. The exact score of 0.8870 reflects that the model produces several false positives, which means it is good to perform accurate predictions when it identifies the label. In a revocation of 0.8999, Biogpt effectively captures most of the actual instances and demonstrates its ability to identify relevant labels across samples. The F1 score, which is a harmonious average of accuracy and memories, was 0.8881, emphasized by a well -balanced compromise between the two metrics. All of these values are above 0.88 underline the reliability and consistency of the model in power. These results are confirmed by the ability of Biogpt in the processing and understanding of complex biomedical texts. Overall, Biogpt has proven to powerful model for multiple branded classification tasks in the field of biomedical natural language processing.

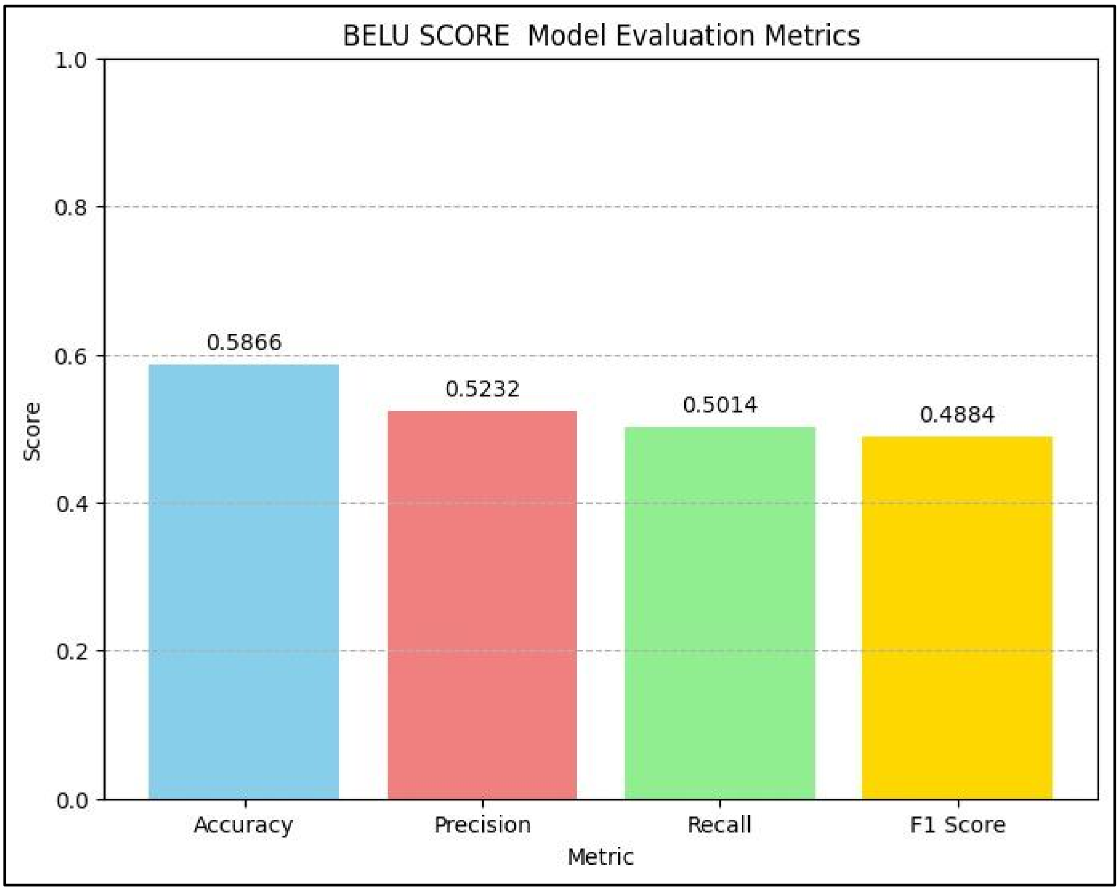


Figure 4: Blue Score Model outcomes in various form i.e accuracy, precision, recall and F1-Score on pubmed open access Database

The BLEU score was evaluated using the PUBMED open database and showed a slight performance across the metrics of key ratings. It achieved an accuracy of 0.5866, suggesting that it correctly predicted over half of the instances. The more accurate value of 0.5232 shows that the model has space to improve the minimization of false positives. Its download was measured to 0.5014, suggesting that it tried to capture all relevant instances in the data file. The F1 score, which balances accuracy and appeal, was relatively low to 0.4884, indicating insufficient consistency in prediction quality. Overall, while the BLEU score shows the potential, its current performance suggests that it may not be optimal for comprehensive tasks of biomedical text classification without further improvement or improvement.

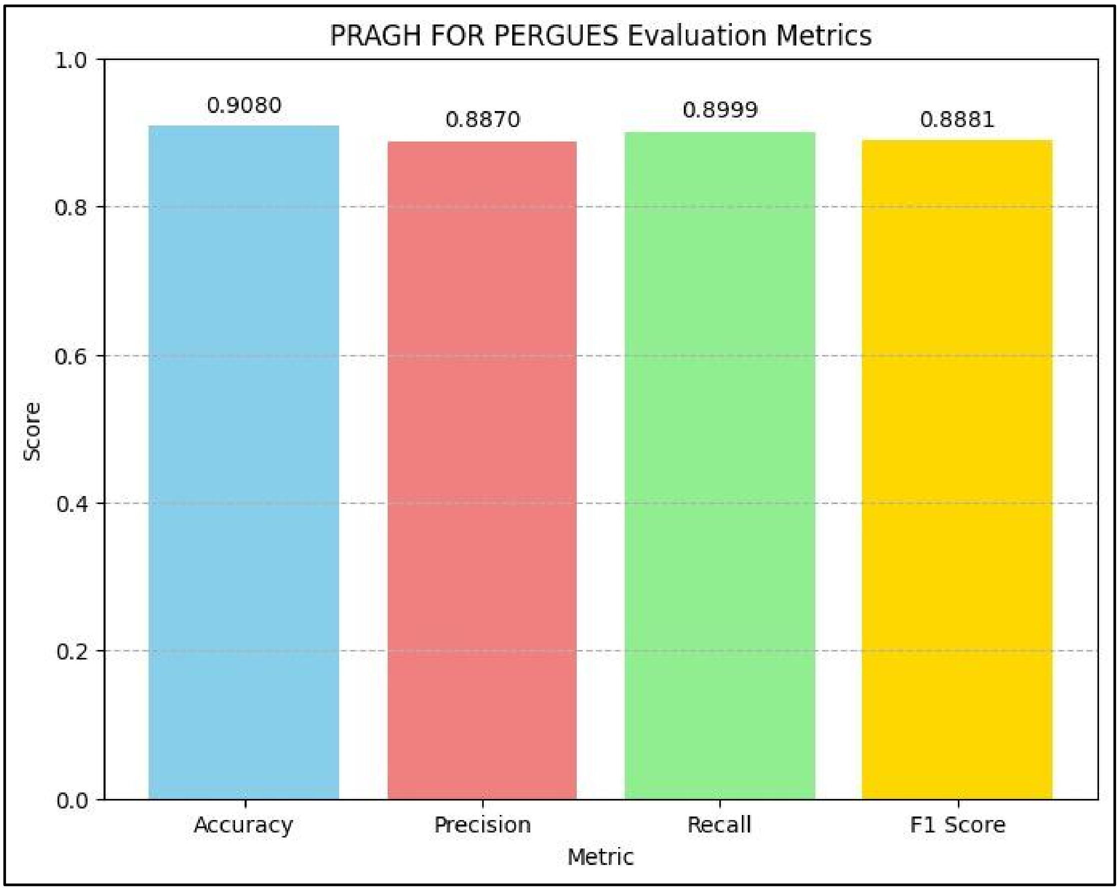


Figure 5: Pergsus Model outcomes in various form i.e accuracy, precision, recall and F1-Score on pubmed open access Database

The Pergsus, evaluated in the PUBMED open access database, shows a strong and consistent performance across all the main metrics. It achieved a high accuracy of 0.9080, suggesting that it correctly classified the vast majority of inputs. With an accuracy of 0.8870, the model shows a low level of false positives, indicating the reliability in its positive predictions. The download score of 0.8999 emphasizes the model's ability to successfully identify the most relevant cases. Finally, its score F1 0.8881 confirms a well -balanced performance between accuracy and download, which makes Pergsus a robust and effective choice for the classification of biomedical text.

Table2: Proposed work analysis with existing state of art methods

Methodology	Accuracy	Precision	Recall	F1-score
T5 [4]	0.88	0.8188	0.7792	0.8656
Belu Score [2]	0.5866	0.5232	0.5014	0.4884
Pegasus [5]	0.9080	0.8870	0.8999	0.881
Proposed Work	0.9080	0.8870	0.899	0.8881

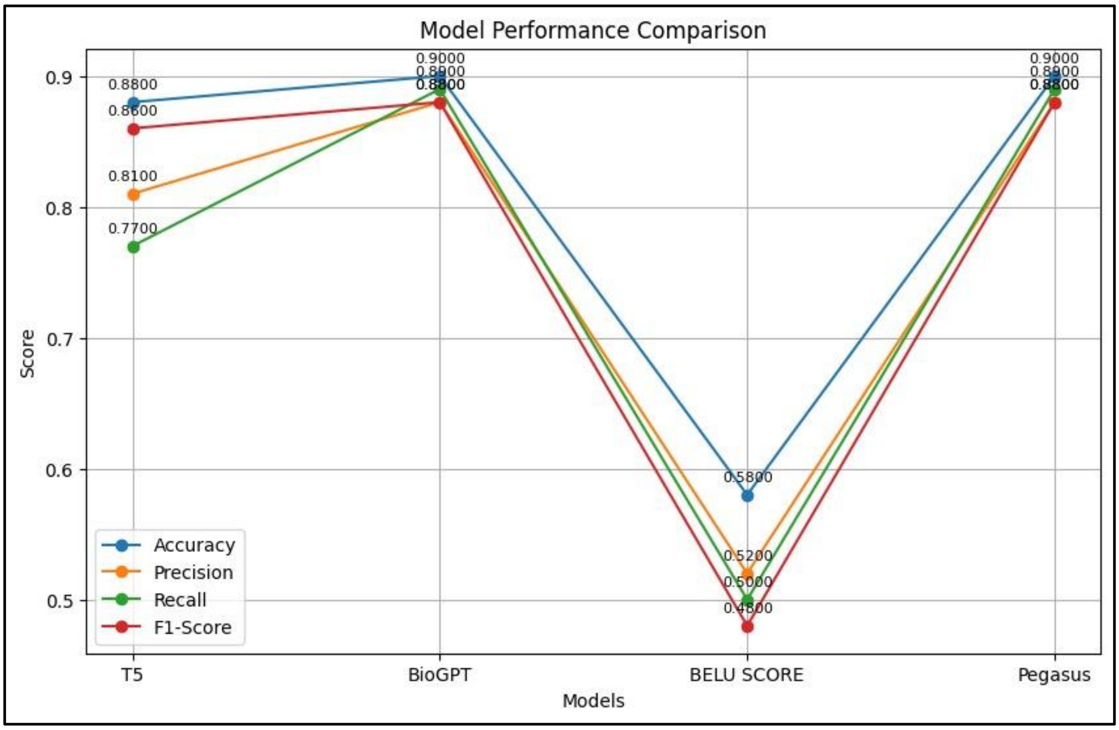


Figure 6: Proposed work analysis with existing state of art methods

The graph compares the performance of four models-T5, biogpt, belu scores and pegasus-across Four Metrics: accuracy, precision, recall and F1-score. Biogpt and Pegasus show a permanently high performance, both as accuracy of 0.90 and 0.88 F1-scores. The T5 works slightly well with an accuracy of 0.88, but its withdrawal lags behind 0.77. The Belu score shows the weakest performance across all metrics, especially F1-Score and appeal, both below 0.50. The accuracy for Biogpt and Pegasus is the same at 0.88, which represents their reliable prediction. The step immersion for Belu's score visually emphasizes its insufficient performance compared to other models.Hence, Biogpt and Pegasus are the most effective models to summarize the medical text, while the Belu score is not suitable due to significantly lower performance in all key metrics.

1.8 Conclusion and Future Scope

Summary of the medical text increases understanding of complex clinical data. This study shows that transformer models such as Biogpt and Pegasus overcome others in accuracy and balance, indicating the effective for biomedical summary. Simpler models, such as Belu scores, emphasize the role of advanced architectures. Deep learning is essential for processing unstructured medical texts. These results support the deployment of such models in real healthcare applications.

Future work should focus on improving the generalization of the model, reducing calculation costs and distortion of data. Increasing the metrics of fine fine -tuning and rating, such as Rouge and Bertscore, will improve the quality of the summary. Wider adaptability and scalability of the domain will increase their practical use. This will lead to more reliable and efficient AI medical tools.

REFERENCE

[1]. B. Palanisamy, A. Chakrabarti, A. Singh, V. Hassija, G. S. S. Chalapathi, and A. Singh, "From Information Overload to Lucidity: Leveraging GPT Models for Biomedical Summarization," *unpublished*, Dec. 23, 2024, current version Jan. 14, 2025.

[2]. M. H. H. Wahab, N. H. Ali, N. A. W. A. Hamid, S. K. Subramaniam, R. Latip, and M. Othman, "A Review on Optimization-Based Extractive Automatic Text Summarization Techniques," *IEEE Access*, vol. 12, pp. 1–20, Dec. 2023, current version Jan. 2024.

[3]. A. Khaliq, A. Khan, S. A. Awan, S. Jan, M. Umair, and M. F. Zuhairi, "Integrating Topic-Aware Graph Neural Networks with Transformers for Abstractive Medical Summarization," *IEEE Access*, vol. 12, pp. 1–12, Aug. 2024.

[4]. A. Aftiss, S. Lamsiyah, S. O. El Alaoui, and C. Schommer, "BioMDSum: Hybrid Biomedical Multi-Document Summarization using Extractive and Abstractive Techniques," *IEEE Access*, vol. 12, pp. 1–15, Dec. 2024.

[5]. G. Althari and M. Alsulmi, "Exploring Transformer-Based Models for Negation Detection in Biomedical Texts," *IEEE Access*, vol. 10, pp. 88462–88474, Aug. 2022.

[6]. J. Lee, I. Baek, and H. Lee, "REMDoC: Reference-Free Evaluation Metric for Medical Document Summarization using Contrastive Learning," *IEEE Access*, vol. 12, pp. 1–10, Dec. 2024.

[7]. S. Z. Aftabi, S. M. Seyyedi, M. Maleki, and S. Farzi, "ReQuEST: A Multi-Task Transformer Framework for Medical Question Summarization and Entailment," *IEEE Access*, vol. 12, pp. 1–8, Feb. 2024.

[8]. A. Ksibi, A. S. D. Alluhaidan, A. Salhi, and S. A. El-Rahman, "An Overview of Lifelogging: Challenges, Applications, and Techniques," *IEEE Access*, vol. 9, pp. 63422–63444, Apr. 2021.

[9]. J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. L. Wang, "MS^2: Multi-Document Summarization of Medical Studies," in *Proc. Conf.*

Empirical Methods Natural Lang. Process. (EMNLP), Punta Cana, Dominican Republic, Nov. 2021, pp. 7494–7513. [Online].

[10]. A. J. Jimeno-Yepes, J. Lau, and T. Baldwin, "M3: Multi-Level Dataset for Multi-Document Summarization of Medical Studies," in *Findings Assoc. Comput. Linguistics: EMNLP 2022*, Abu Dhabi, UAE, Dec. 2022, pp. 3887–3897. [Online].