International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. Comp. Info. Vol. 5 (2024) 28–43

Paper Type: Review Article

From Data to Insights: A Survey on Biomedical Text Summarization Approaches and Challenges

Nabil M. AbdelAziz [1], Aliaa A. Ali [1,*], Soaad M. Naguib [1] and Lamiaa S. Fayed [1]

[1] Department of Information Systems, Faculty of Computer and Informatics, Zagazig University, Zagazig 44519, Egypt.
Emails: nmabedelaziz@fci.zu.edu.eg; aliaam@fci.zu.edu.eg; smnagieb@fci.zu.edu.eg; lsfayed@fci.zu.edu.eg.

Received: 31 Aug 2024 **Revised**: 30 Oct 2024 **Accepted**: 13 Nov 2024 **Published**: 15 Nov 2024

Abstract

The explosive growth of biomedical literature and clinical data is increasing the difficulty for healthcare professionals to continuously access new information. This survey summarizes the state-of-the-art studies in biomedical text summarization (extractive, abstractive, and hybrid) and their implementations. In this regard, the survey delves into the influence of deep learning and natural language processing (NLP) methods on enhancing summarization capabilities while simultaneously highlighting ongoing challenges related to domain-specific jargon, truthfulness, and explainability. The paper further describes some evaluation metrics and datasets specific to biomedical problems that are important for both training and evaluating summarization models. Finally, we highlight three of these considerations that we believe are worth pursuing in future work toward designing more effective and grounded summarization systems. Biomedical text summarization, which extracts useful information from vast quantities of data, might help render scientific knowledge more reachable, assist clinical decision-making and policy, and potentially push research forward.

Keywords: Biomedical Text Summarization; Natural Languge Processing; Deep Learning.

1 | Introduction

The rapid expansion of the Internet and related technologies has led to a dramatic growth in the amount of biomedical electronic texts. Various resources, such as electronic health record systems, online clinical reports, and biomedical literature databases, make a variety of medical information and publications, such as research articles and patients' health records, available in various formats [1, 2]. There are almost 27 million citations to scientific papers in PubMed alone. It is worth noting that the MEDLINE bibliographic database, maintained by the US National Library of Medicine, has over 24 million citations from over 5500 biomedical publications [2].

Information overload is a common issue researchers face when trying to come up with new ideas, mastering the state-of-the-art in a specific area, evaluating recent progress in a research discipline, designing their studies and understanding their outcomes [3, 4]. Therefore, given the vast amount of biomedical data and the regular updates made to them, Extracting required and pertinent information from such data is difficult and hard for clinical researchers. This leads to a long duration of time being spent by clinical researchers to obtain desired materials. They cannot read every line of text in search results and understand them. Consequently, it is essential to summarize and condense the textual resources with urgency and heightened significance.

Corresponding Author: aliaam@fci.zu.edu.eg

Summarizing manually requires significant resources and can be quite time-consuming. It is quite complicated for individuals to manually condense this vast volume of textual information. Automatic Text Summarization (ATS) is a crucial aspect to tackle this issue [5, 6].

ATS is the method of distilling the source material by choosing the essential elements that appear inside the text, referred to as text summarization. Automatic biomedical text summarization is an effective and dependable technique designed to condense a whole biomedical document while maintaining its most critical elements. Consequently, ATS is essential in addressing the challenge of obtaining precise and current information pertinent to the requirements of biological researchers and practitioners[7, 8]. Automated biomedical text summarization broadly falls into extractive or abstractive methods. Extractive summarization picks important sentences or quotes from the original text, and abstractive summarization rephrases the main ideas using new sentences to resemble summaries written by humans.

The state of the art for most NLP tasks, including biomedical text summarization, has seen significant improvements with the advent of deep learning architectures, specifically with transformer-based models, including BERT, BioBERT, and T5. Typically, these models have been pre-trained on large-scale biomedical datasets, which contributes to the ability of the summarization process to be more context-aware and relevant to capturing domain-specific input text with specific language and terminologies. Nonetheless, biomedical summarization is still challenging and needs to develop effective summarization methods to deal with the complexity and specificity of biomedical language, to meet the need for interpretation, as well as the conciseness and informativeness of summaries.

In this sense, this survey outlines existing techniques and approaches in biomedical text summarization by reviewing state-of-the-art models, introducing benchmark datasets and evaluation metrics. This paper seeks to discuss current trends and challenges and provide guidance on future research directions to facilitate the development of more sound and reliable summarization solutions to better serve the needs of biomedical research and clinical practice.

This paper is organized as follows: Section 2 discusses ATS and different factors used to classify it. Section 3 explores different datasets used in the medical domain, recent summarization methods, and evaluation methods used for assessing the quality of summaries. In Section 4, discussion was provided. Section 5 presented the closing comments and outlined potential areas for future work.

2 | Automatic Text Summarization

Automatic text summarization (ATS) is a subset of natural language processing (NLP) where the computer generates a brief summary of one or multiple documents while preserving the important parts and facts from the input text [9]. ATS can be categorized based on different criteria, as shown in Figure 1:

2.1 | Based on Input Type

Text summarization can be either single document summarization or multi-document summarization. In multi-document summarization, it summarizes more than one document together to produce one summary that covers all topics represented in the input documents. on other hand, single document summarization involves summarizing each document independently [8, 10].

2.2 | Based on Purpose of Summarization

Another criterion for categorizing automatic text summarization is the purpose of summarization. The purpose can be generic or query-focused. Generic text summarization doesn't require any specified information, while query-focused summarization produces a summary that addresses the user's query and needs [11-13].

2.3 | Based on Content of Summary

The content of the generated summary may be informative or indicative. The indicative summary provides a concise overview of the topic and the concerns addressed in the document, while the informative summary provides a complete summary and more information contained in the input document [1, 12].

2.4 | Based on Summarization Approach

The summarizing methods can be classified as extractive, abstractive, and hybrid. The extraction technique involves sorting and determining the most salient phrases in the material to provide a concise, representative summary [14]. Conversely, abstractive summarization reformulates the principal concepts conveyed in the material rather than extracting specific phrases [15]. Hybrid summarizing starts with the selection of salient sentences from the text to produce an extractive summary, followed by the application of abstractive techniques to reformulate and transform the extractive summary into an abstractive one [16].

2.5 | Based on Language

Summarization can be classified as monolingual or multilingual, depending on whether the document's text is in the same language [12].

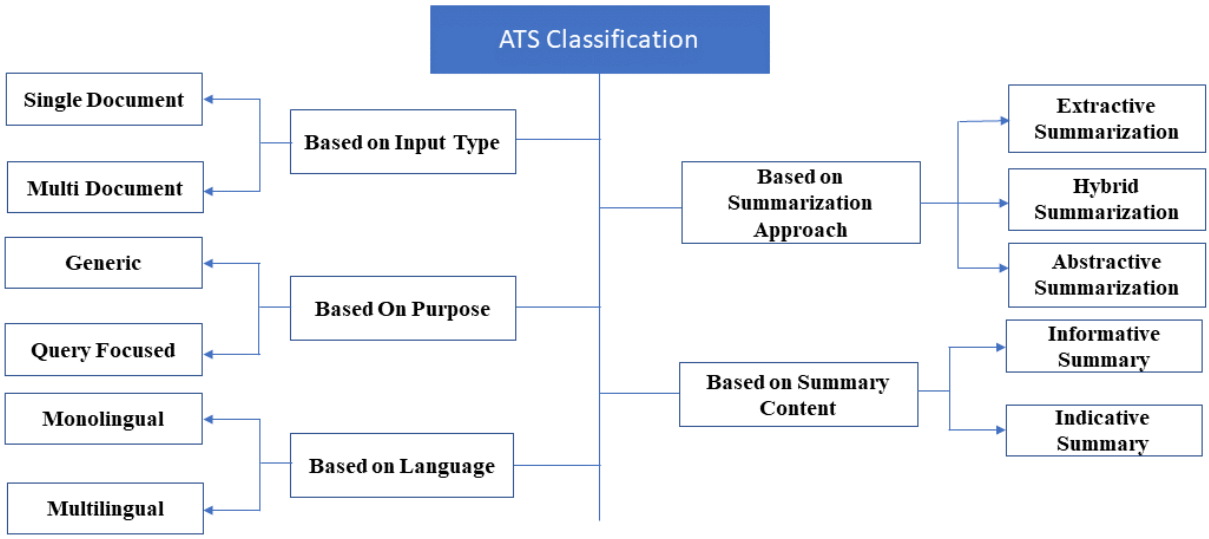

Figure 1. Different Criteria to classify ATS systems [5].

3 | Medical Text Summarization

3.1 | Datasets

Numerous datasets have been developed in natural language processing for the purpose of automated summarization over the years. An extensive review of the literature uncovers a variety of datasets utilized to tackle the issues associated with extractive, abstractive, and hybrid summarizing jobs.

3.1.1 | PubMed Dataset

The PubMed dataset comprises XML files from the open-access collection of the PubMed Central (PMC) repository. The collection comprises 133K documents with abstracts. The mean length of the abstract is 214 words, whereas the mean length of the whole text is 3224 words [17].

3.1.2 | arXiv Dataset

The arXiv dataset comprises LATEX files from the arXiv library that contains digital preprints. The collection contains 215K abstracted documents. The mean length of the whole text is 6,913 words, whereas the mean length of the abstract is 292 words [17, 18].

3.1.3 | CORD-19 Dataset

CORD-19 is an open research dataset comprising over 140,000 publications, including more than 72,000 full texts. Since the start of 2020, around 47,000 publications and 7,000 preprints concerning COVID-19 and coronaviruses have been published, accounting for roughly 40% of the dataset [19].

3.1.4 | SUMPUBMED Dataset

The SUMPUBMED dataset comprises 26 million biomedical research articles sourced from PubMed. The documents originate from several sources, including digital books, MEDLINE, and life science publications. The dataset is divided into three categories: training (93%), testing (3%), and validation (4%) [20].

3.1.5 | BMC Dataset

BioMed Central (BMC) is an open-access publisher offering more than 250 scientific publications. It presently disseminates all its journals online. BioMed Central is the first and largest open data science publisher. Founded in 2000, it has been known as Springer Nature since 2008 [21].

3.1.6 | BioRED dataset

The Biomedical Relation Extraction Dataset (BioRED) facilitates automated relation extraction from biomedical research articles. This is the first biomedical relation extraction dataset, encompassing diverse entity kinds such as gene-protein, chemical, and health issues, as well as relation pairs including chemical-chemical and gene-disease inside the text. A collection of 600 PubMed abstracts [22].

3.1.7 | EBMSummariser Corpus Dataset

The Evidence-Based Medicine Summarization (EBMSummariser Corpus) dataset is a publicly available collection of 2707 individual document summaries. The dataset comprises data from the Journal of Family Practice (JFP). It contains 1,388 training records and 1,319 assessment data [23].

3.1.8 | PQR Dataset

The Prognosis Quality Recognition (PQR) dataset is derived from scientific publications inside the PubMed dataset that are suitable for summarization. It comprises 2,686 papers and 697 positive records (scientifically sensitive) [9].

3.1.9 | S2ORC Dataset

The Semantic Scholar Open Research Corpus (S2ORC) is the largest repository of publicly available scientific publications in English, encompassing several academic disciplines. It comprises 1.5 million LATEX source files, 8.1 million open-access PDFs, 81.1 million publications, and 380.5 million resolved citation connections. The corpus encompasses several academic disciplines, including biological and computer science domains [24].

3.1.10 | CCA Dataset

The Clinical Context-Aware (CCA) dataset is generated by integrating the National Center for Biotechnology Information's (NCBI) PubMed with the Biomedical Natural Language Processing (BioNLP) dataset. The collection comprises 173,000 documents, including 131,000 sourced from BioNLP and 42,000 from NCBI PubMed [9].

3.1.11 | Custom Dataset

Numerous investigations have developed their own tailored datasets for the purpose of biomedical summarization [25-27] . For instance, Davoodijam et al. constructed their assessment corpus by randomly picking450 scientific research papers from BioMed Central for extractive summarization tasks [28].

3.2 | Summarization Methods

There are three methods of text summarization: extractive summarization, abstractive summarization, and hybrid summarization.

Extractive methods are categorized into (1) statistical-based techniques, (2) concept-based techniques, (3) topic-based techniques, (4) graph-based techniques, and (5) machine learning methods [28, 29]. Numerous research studies in the medical field employed extractive methods to improve the quality of the produced summaries. For example, Hark et al. introduced a novel model named BioGraphSum for extractive summarization, which utilizes graph-based methodologies to find the most relevant sentences in a text for summary purposes. BioGraphSum was evaluated on a medical corpus consisting of 450 research papers collected from the PubMed database. The proposed method was compared to three methods: Leveraging BERT, LexRank, and MultiGBS. The BioGraphSum outperformed the comparison methods according to the ROUGE metric [30].

Xie et al. incorporated a graph neural topic model and domain-specific knowledge from the UMLS into a transformer-based pre-trained language model (PLM) for biomedical summarization [31]. Moradi et al. presented a Bayesian summarizer that correlated the text with Unified Medical Language System (UMLS) concepts, including six distinct criteria to delineate the essential concepts. This approach was assessed on a medical corpus including 400 biomedical papers. The efficacy of the Bayesian summarizing approach was assessed against various biomedical summarizers that utilize concept frequency, domain-independent summarizers, and baseline methodologies. The results indicate that the Bayesian summarizer, when employing the meaningfulness or CF-IPF measure for feature selection instead of relying on the frequency of individual ideas, has superior performance compared to alternative techniques [2].

Kirmani et al. presented an innovative extractive summarizer that maintains text semantics through the application of biosemantic models. Bio-semantic models were used to transform sentences into (Big vectors) through the concatenation of word vectors to ensure semantic representations. The proposed method then employed k-means clustering on big vectors followed by a ranking algorithm to select the ranked sentences to form the final summary. The results indicate that the usage of biosemantics models can enhance the performance and generate better summaries compared to baseline methods [32].

Furthermore, Moradi et al. developed a graph-based approach that employs the Helmholtz principle to extract essential concepts from text, subsequently creating a graph-based method to capture the key phrases for the summary. The developed method was evaluated using the ROUGE metric and tested on a biomedical corpus comprising 300 articles sourced from BioMed Central.[33].

A domain-specific method called MultiGBS is proposed that represents a document as a multi-layer graph, facilitating the simultaneous processing of multiple text features. The study utilizes three distinct features: word similarity, semantic similarity, and co-reference similarity, each represented as separate layers in the model. The unsupervised approach utilizes the MultiRank algorithm to select sentences from the multi-layer graph, taking into account the quantity of concepts involved. The MultiGBS algorithm utilizes UMLS to extract concepts and relationships through various tools, including SemRep, MetaMap, and OGER. Comprehensive assessment utilizing ROUGE and BERTScore indicates an enhancement in F-measure values [28].

Rouane et al. [12] employed UMLs to present biomedical articles as a combination of concepts. Sentences with similar content are clustered together utilizing the K-Means clustering technique. The Apriori approach

was subsequently employed to determine the prevalent itemsets inside the categorized phrases. Ultimately, significant sentences were selected from each group to provide an extracted summary.

Recently, numerous studies have integrated pre-trained language models (PLMs) for the extractive summarization of biomedical documents. For biomedical extractive summarization, Due et al. proposed a new model named BioBERTSum. The model employs a domain-specific bidirectional language model as an encoder, pre-trained on extensive biomedical corpora, and subsequently fine-tunes it for the extractive text summarization job on each individual biomedical document. Trials conducted on the PubMed dataset demonstrate that the suggested model surpasses the current state-of-the-art SOTA model by a margin of ROUGE-1/2/L [34]. Also, Kanwal et al. [35] finetuned BERT on the MIMIC-III dataset for extractive summarization of Digital Health Records. Moradi et al. [36] utilized a hierarchical clustering method to group contextual embeddings of phrases using the BERT encoder. The most important sentences from every category are selected to construct the final summary.

Padmakumar et al. [37] introduced an unsupervised extractive summarization approach that encodes phrases using the GPT-2 model and uses Pointwise Mutual Information (PMI) to determine semantic similarity between texts. This method was assessed on a medical journal dataset. A new approach that combines graph-based and domain-specific word embedding BioBERT for summarizing biomedical articles was proposed by Moradi et al. [38]. Xie et al. [39] introduced a KeBioSum framework for biomedical extractive summarization tasks. It improved the performance of PLMs by incorporating fine-grained domain knowledge (PICO components) and employing sophisticated training approaches. CovSumm is an unsupervised approach that leverages the strengths of both transformer-based models and graph-based methods for summarizing COVID-19 literature [40].

Overall, there are many PLMs pre-trained specifically for biomedical texts, such as BioBERT [41], PubMed BERT [42], SciBERT [43] , BlueBERT [44], ClinicalBERT [45], and ALBERT [46]. Meng et al. [47] suggested splitting the knowledge graph into subgraphs and injecting them with several PLMs like BioBERT, SciBERT, and PubMed BERT.

Abstractive summarization in the biomedical domain focuses on generating concise, coherent summaries that capture the core insights of medical and scientific texts. In contrast to extractive summarization, which identifies and picks significant sentences from the source material.

Hu et al. developed a method to enhance the summarization of radiology findings. The method utilizes graph encoders to extract relational information from medical entities and dependency structures. By comparing positive and negative instances, the integration of contrastive learning enhances the model's capacity to discern between important and non-essential elements. The efficacy of this method is confirmed by experimental results on benchmark datasets: OPENI and MIMIC-CXR, which establish new benchmarks for precision and thoroughness in the summarizing of radiology findings. This study highlights how contrastive learning and graph-based techniques may be used to improve medical text summarization [48].

Du et al. introduced a new model named UGDAS. UGDAS integrates an auto-regressive generator with an unsupervised graph network for sentence-level denoising. To further enhance the quality of the produced summaries, the model denoises the original text using domain knowledge and sentence position information. The performance of the proposed model was evaluated using the CORD-19 (COVID-19 Open Research Dataset) and the PubMed dataset. The experimental findings indicate that the model attains state-of-the-art findings on the CORD-19 dataset and surpasses the relevant baseline models on the PubMed Abstract dataset [49].

Table 1 provides an extended literature review of the most relevant biomedical text summarization studies. The literature review indicates the main objectives, key findings, future work, and the type of summarization for the mentioned studies.

Table 1. Biomedical summarization: study objectives, key findings, and future work.

Ref	Type of summarization	The objective of the study	Key Findings	Recommendations/Future work
[28]	Extractive summarization	• Propose a MultiGBS, an innovative biomedical text summarizer that models three distinct kinds of sentence-to-sentence relationships using multi-layer graphs. • Utilizes the Unified Medical Language System (UMLS) knowledge source in MultiGBS method to define concepts and relationship between them. • Rank the input document using the MultiRank algorithm which applied to the multi-layer graph in the suggested approach.	According to the metrics ROUGE and BERTScore, the presented summarizer performs better than previous baseline approaches when compared to MultiGBS.	• More similarity metrics tailored to document context might be the subject of future study. • Complementing the multi-layer graph model, user needs may also be expressed.
[12]	Extractive summarization	• Integrate two data mining methodologies: clustering and frequent itemset mining, to generate individual summaries (one summary per document), treating each text as a collection of biomedical concepts rather than terms.	• The findings indicate that this combination effectively improves summarizing performance, and the suggested system surpasses other evaluated summarizers such as (TextRank, TextTeaser, Itemset based summarizer, and Microsoft AutoSummarize). • The proposed method achieves (0. 23840) for Rouge-1, (0.08715) for Rouge-2 and (0.11456) for Rouge-SU4.	• Expand the current method to incorporate word embedding into the text representation in order to conduct a deeper semantic analysis of biomedical literature in future work. • Aim to add a new anti-redundancy approach to decrease the amount of duplicate material.
[8]	Extractive summarization	• • Utilize graph-based methods for representing biomedical information, as this approach effectively captures the relationships among various pieces of information. • Employ itemset mining to uncover major patterns within the text, facilitating the identification of crucial concepts to be included in the summary. • The method also includes clustering phrases to group comparable concepts, making the summary clearer and more understandable.	• The integration of domain-specific knowledge into biomedical summarization systems enhances the creation of a comprehensive and more precise semantic model for the biomedical field. • Utilizing itemset mining can boost the effectiveness of a summarization system by allowing it to recognize various connections among several ideas and provide greater semantic depth. • The findings indicated that the suggested technique surpassed baseline methods, achieving a Rouge-1 score of 0.7648 and a Rouge-2 score of 0.3524.	• Integrating diverse information sources may enhance the summarizer's effectiveness by addressing ideas overlooked in a singular knowledge base. • Evaluate the efficacy and utility of the summarizer in multi-document summarizing. • Execute the system for query-driven biomedical text summarization.

Table 1. (Continued).

Ref	Type of summarization	The objective of the study	Key Findings	Recommendations/Future work
[34]	Extractive summarization	• An encoder-based, domain-specific language model, pretrained on extensive biomedical dataset, is used to incorporate external knowledge to enhance the fine-tuning process for extractive summarization in biomedical domain. • A sentence position embedding technique is introduced to acquire the positional information of sentences and attain the structural characteristics of a text.	• The experimental findings indicate that the suggested model surpasses the previous state-of-the-art model, achieving scores of 0.4313, 0.19, and 0.3747 for Rouge-1, Rouge-2, and Rouge-L, respectively. • Through comparing various decoders, it was demonstrated that the attention mechanism excels at summary tasks. Additionally, previous domain knowledge significantly enhances summarization task performance in the biomedical subject.	• Develop a hybrid method which merges extractive and abstractive summarization. • Use more expert knowledge source.
[50]	Abstractive Summarization	• Introduce a new model for summarizing scientific articles, which incorporates SciBERT trained on an extensive corpus of scientific literature and a graph transformer that leverages the relational features of the knowledge graph without the need for linearization or hierarchical restrictions. • Introduces a graph-based methodology for extensive document summarization. • The efficacy of summarization models in condensing both short and lengthy documents was evaluated against the suggested model.	• Experimental findings indicate that the proposed approach surpasses baseline methods in summarizing lengthy scientific papers, achieving ROUGE-L scores of 34.96. • The findings from human evaluation indicate that the produced summary is typically informative, fluent, and aligns with the ground-truth summary. •	• The model cannot effectively summarize papers with mathematical equations, pictures, and tables. These portions were removed from the research during preprocessing. To improve future study, it may be beneficial to compress the papers. • Articles in science are regularly published. Different meanings of terminology may be used in scientific literature. Regularly updating the dataset will enhance future studies.
[1]	Hybrid Summarization	• Introducing a novel approach for extractive summarization of biomedical literature utilizing graph generation and frequent itemset mining. Transform the extractive summaries into abstractive ones.	• The proposed approach super passed state-of-art methods by 17% in terms of the ROUGE score. • The proposed approach addresses two main problems in abstractive summarization: defining the significant concepts in the text and creating new sentences represents the core of the text.	• Expanding one's knowledge base beyond UMLS can lead to a more precise understanding of concepts. • Future study can focus on modifying the suggested method to express the summary in abstract and extractive forms according to inquiries.

Table 1. (Continued).

Ref	Type of summarization	The objective of the study	Key Findings	Recommendations/Future work
[49]	Abstractive Summarization	• An improved model for capturing sentence relationships is suggested, UGDAS, which combines an auto-regressive generator with a graph-network based sentence-level denoiser. • UGDAS's denoiser utilizes domain knowledge to convey the phrase's extensive information in the biomedical field, while the ranking process is improved by using sentence position data.	• The proposed model attains the state-of-the-art result on the new CORD-19 dataset and surpasses comparable models on the PubMed Abstract dataset. • The ablation experiments demonstrate the requirement of noise reduction prior to summary generation, the significance of domain knowledge for representation, and the efficacy of sentence position information for ranking. • The model attains scores of 0.3303, 0.1351, and 0.2930 for Rouge-1, Rouge-2, and Rouge-L, respectively, on the PubMed dataset. The model attains scores of 0.3368, 0.2256, and 0.3284 for Rouge-1, Rouge-2, and Rouge-L, respectively, on the CORD-19 dataset.	• In future study, they will concentrate on modeling real text alongside graph neural networks (GNN) and investigating more effective methods to incorporate domain knowledge and pre-trained language models for the job of abstractive summarization.
[51]	Extractive summarization	• As linguistic aspects of sentences, word co-occurrence graphs are employed, and heuristic sentence extraction algorithms based on prior knowledge are established. • Provide a new technique to improve the SciBERT-based summarization model by adding linguistic knowledge to the contextual embeddings of scientific publications. • This model makes use of pre-trained language models, graph neural networks, and highway networks.	• The experimental findings reveal that the proposed COVIDSum outperforms competing summarizing techniques on the COVID-19 open research dataset. • The suggested COVIDSum would aid researchers' investigations into COVID-19 by speeding up the research process, and it highlights the potential and promise of customizing certain NLP approaches to the domain of COVID. • COVIDSum achieves 0.4456 for Rouge-1, 0.1889 for Rouge-2, and 0.2653 for Rouge-L.	Not Mentioned
[35]	Extractive Summarization	• The objective of this work is to develop a multi-head attention-based method that can improve clinical note extractive summarization.	• A multi-head attention-based technique for extractive summarization of clinical notes successfully identifies important phrases and sentences from thick medical data. • Use of the model's output in a heat-mapping tool improves the visual representation of important information, which in turn makes electronic health records easier to understand and use for humans.	• Future research may concentrate on refining the multi-head attention mechanism to augment the precision and pertinence of the retrieved summaries from clinical notes, perhaps integrating more sophisticated deep learning methodologies. • Explore the implementation of this summary approach across other medical record forms, including radiology reports and pathology notes, to evaluate its adaptability and efficacy in varied contexts.

3.3 | Evaluation Methods

Evaluation methodologies are essential tools for determining whether automated system summaries adequately convey the main points of the source material. Over the past decade, many assessment tools for automatically generated summaries have been created. There are two forms of evaluation methods: Quantitative Analysis, and Qualitative Analysis. The most recent methodologies in text summarization and the evaluation methods employed to evaluate their performance are captured in Table 3.

3.3.1 | Quantitative Analysis

A. Precision, Recall, and F-Measure scores

Precision and Recall: The evaluation of extractive summaries can be effectively conducted using these two established metrics. Precision and Recall evaluate the summaries produced by automated systems against those created by humans, serving as the benchmark, and assess the degree of lexical overlap. Recall is the percentage of human-selected sentences that were also appropriately recognized by the algorithm. As illustrated in Eq. (1). Recall is computed by dividing the total number of sentences in both the reference and candidate summaries by the total number of sentences in the reference summary. Precision denotes the proportion of accurately executed system statements. The calculation, as per Eq. (2), involves dividing the total number of sentences in both the reference and candidate summaries by the total number of sentences in the candidate summary [52, 53].

$$\text{Recall} = \frac{\text{summary reference} \cap \text{summary Candidate}}{\text{summary reference}} \tag{1}$$

$$\text{Precision} = \frac{\text{summary reference} \cap \text{summary Candidate}}{\text{summary Candidate}} \tag{2}$$

F-Measure: The F-score, or F1-score, denotes a balanced assessment of Recall and Precision.F-score integrates recall and precision into a single metric. The formula of f-score is represented in Eq. (3).

$$\text{F1} - \text{score} = \frac{2\,x\,(\text{Precision x Recall})}{\text{Precision}+Recall} \tag{3}$$

B. ROUGE Metric

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) was introduced by Lin [54]. ROUGE is considered the most metric used for automatically evaluating the summaries that are created automatically [55]. When it comes to natural language processing (NLP), ROUGE is a set of programs and collection of metrics used to assess automated summarization and machine translation systems. It evaluates the AI-generated summaries in comparison to a number of reference summaries written by humans [56]. The core premise of ROUGE is to assess the frequency of overlapping units, such as shared n-grams, between reference summaries and candidate summaries (or system summaries)[10]. The following are some variants of ROUGE metrics:

ROUGE-N: It's based on the uni-gram metric that compares reference summaries with candidate summaries. ROUGE- N calculates the number of unigrams (individual words) that appear in both the generated and the reference summaries [26].

ROUGE-L: It employs the Longest Common Subsequence (LCS), which denotes the longest sequence of words that occurs in both the candidate and reference summaries in the same order, albeit not necessarily consecutively [57].

ROUGE-W: The Weighted Longest Common Sub-sequence is an improved version of ROUGE-L, in which the sequence words can be either consecutive or non-consecutive, with intervening words included. ROUGE-W governs the extent of the subsequent phrases.

Despite the extensive usage of ROUGE in text summarization evaluation, the ROUGE metric suffers from significant limitations. (1) It relies on n-gram overlap and discards the semantic meaning of the summary; (2) it lacks coherence and readability of the summary; (3) it requires human-written reference summaries for evaluation; (4) it can't determine if the information represented in the summary is correct or not [54, 58-59].

C. BLEU Metric

Bilingual Evaluation Understudy (BLEU) measures the similarity between a candidate text (machine output) and one or more reference texts (human-written). It calculates this similarity based on n-gram precision (matching sequences of n consecutive words) [60]. Eq. (4) and Eq. (5) represent the formula of BLEU [61].

$$\text{BLEU} = BP \,.\, exp\left(\sum_{n=1}^{N} Wn \ \log Pn\right) \tag{4}$$

$$\text{BP} = \begin{cases} 1 & if \ \ c > r \\ e^{(1-r)/c)} & if \ c \leq r \end{cases} \tag{5}$$

- r is the length of the effective reference corpus.
- c is the candidate translation's length.
- Pn is an n-gram precision, employing n-grams up to length N, and positive weights Wn.
- BP is the brevity penalty.

D. BERTScore

The BERT Score (BS) is a measure for evaluating text-generation systems. The BERTScore method uses contextual embeddings retrieved using the BERT model as its foundation for evaluating language creation approaches. Its primary use cases are sentence-level machine translation and picture captioning, although it may be modified for use in summary evaluation. Evaluation with BERTScore is done using a greedy matching of cosine similarity between candidate and reference summaries' embeddings. For each token in the candidate summary, the matching process looks for its closest comparable counterpart in the reference summary. The use of contextual embeddings has the benefit of assigning a distinct embedding to a word in different contexts [62].

3.3.2 | Qualitative Analysis

This evaluation approach involves human experts carefully reviewing and analyzing the summaries generated by automatic systems, comparing them with source texts or reference summaries to assess how well the generated summaries capture the key information and maintain coherence and readability [63]. There are different criteria used to evaluate the quality of the generated summaries when qualitative evaluation is involved. Table 2 explains some of the most popular criteria used for qualitative evaluation. The process begins by asking the human experts to evaluate a set of generated summaries against some criteria by providing them their reference summaries or source texts. For example, two experts are asked to assess 40 random summaries generated by the proposed method using Likert scales 1-5 to evaluate coherence, fluency, consistency, and relevance [6].

Table 2. Different Criteria for assessing the quality of produced summaries.

Criteria	Description	Reference
Coherence	Describes how well the summary' sentences well organized and connected to present the main information and ideas.	[6]
Consistency	Represents to what extent the facts supplied in the source text are presented in the produced summary	[6]
Relevance	Describes how well the summary contains only the relevant facts.	[6, 63]
Informativity	Defines how well the generated summary extracts important points from the reference summary.	[64]
Conciseness	Represents how well the generated summary is clear, short, and captures the essential information without unnecessary details.	[64, 65]
Readability	Describe how the summary is easy to read and underatand	[48, 64, 65]
Completeness	Represents how well the summary covers all the key points and essential information from the reference summary.	[48, 63]
Repetition	Represents how well the produced summary avoids repeating the same ideas in different sentences.	[63]
Contradiction	Describe if there are any conflicting ideas or information represented in the summary.	[63]

4 | Discussion

As mentioned before, there has been a lot of interest in studying summarizing algorithms recently because of the rapid expansion of biological text data from many sources. This is why several systematic reviews have been conducted and why summarizing techniques are so important in the medical field. In this survey, we conduct a comprehensive study on text summarization in the medical field. Text summarizing algorithms, medical datasets, and several strategies for assessing summary quality are thoroughly examined.

Text summarization can be classified based on different factors. When it comes to the input factor, researchers investigated different methods to summarize single documents and multiple documents. It is still believed that extractive approaches are best suited for summarizing individual documents, as stated in the Afantenos et al. [66] systematic study. Additionally, abstractive methods provide the basis of multiple-document summarization [67]. In recent years, there has been an increased focus on techniques related to single-document summarization. Furthermore, a significant portion of the research focuses on summarizing biomedical literature. This could be attributed to several factors, such as the rapid increase in scientific literature published across various databases in recent years and the improved accessibility of this information compared to the availability of patient clinical records.

Numerous datasets have been presented in the context of the biomedical domain (discussed in Section 3.1). Such datasets include PubMed, a large collection of scientific abstracts and full-text articles; CORD-19, a dataset created based on COVID-19 research needs; and BioMed, a dataset that consisted of various biomedical documents. However, it is noteworthy that many researchers create a custom dataset for the text summarization task by fetching biomedical documents from established medical databases such as PubMed Central and BioMed Central. The trend of creating custom datasets highlights the unique demands of various summarization tasks as well as the shortcomings of current datasets in meeting specific research objectives. Rohil et al. conducted a study and explored which types of documents were suitable for various text-summarization methods. The explorations show that news publications, experimental work, and medical research papers are suitable for extractive summarization tasks, while EHRs and clinical documents are suited for abstractive summarization [68].

Recent studies incorporated pretrained models for biomedical text summarization. For example, Meng et al. proposed to inject several PLMs, such as BioBERT, SciBERT, and PubMed BERT, into the knowledge graph by dividing it into subgraphs [47]. Also, Du et al. presented BioBERTSum, a PLM encoder that has been fine-tuned and optimized for extractive summarization tasks in the medical domain [34]. The BERT algorithm for

extractive summarization of digital health records was fine-tuned using the MIMIC-III dataset by Kanwal et al. [35].

Table 3. Comprehensive review: Methods, Dataset, and Evaluation Metrics.

Ref	Methodology	Dataset	ROUGE			Qulitative Evaluation
			R-1	R-1	R-L	
[30]	BioGraphSum	450 biomedical papers from PubMed Database	0.2942	0.1031	0.1829	Not used
[2]	Bayesian summarizer	400 biomedical papers from BioMed Central	0.7886	0.3529	------	Not used
[28]	MultiGBS	450 biomedical articles from BioMed Central	0.1640	0.0520	0.1460	Not used
[33]	Graph-based summarizer	300 biomedical articles from the BioMed Central	------	0.3321	------	Not used
[12]	Clustering and frequent itemset mining summarizer	100 biomedical manuscripts from the BioMed Central	0.2384	0.08715	------	Not used
[8]	Graph-based using an itemset mining and sentence clustering approach	Compilation of 400 biomedical publication from BioMed Central	0.7648	0.7648	------	Not used
[1]	Graph-based abstractive summarizer	A random selection of 400 biomedical articles were made from BioMed Central.	0.5613	0.2790	------	Not used
[51]	COVIDSum	CORD-19	0.4456	0.1889	------	Used
[49]	UGDAS	PubMed	0.3303	0.1351	------	Not used

The most used metric used for summary evaluation is ROUGE. However, there are four major issues with the Rouge metric: (1) it ignores the summary's semantic meaning in favor of n-gram overlap, (2) the summary isn't coherent or easy to read, (3) it can only be used to evaluate reference summaries that have been written by humans, and (4) it can't tell if the information presented is accurate. Due to these issues, many studies conduct experts to evaluate the quality of summaries generated by summarization systems against various criteria like readability, coherence, relevance, and informativity.

Several factors make the task of text summarization challenging in general and even more complex when it comes to medical applications, which have a serious impact on the research and the practical applications. The main difficulty is the technical language of experts and clinical accuracy. While medical texts include many dense, technical vocabularies where exact meaning really matters for patient care, even a small error in the summarization process could have grave clinical implications. Medical documents have some inherent complexity in having non-linear narratives and non-linear and connected information from multiple sources, which gives many structural challenges for summarization systems. Finally, keeping context is also an important challenge for the medical summarization task because the information of patients is very much tied to the temporal relationships and patient contexts and cannot be misrepresented in the summary [69, 70].

Table 3 provides a comprehensive overview of recent methodologies, datasets, and different evaluation methods used for assessing the effectiveness of the proposed methodologies.

5 | Conclusion and Future Work

Since healthcare and life sciences are fields where publications continually grow, it explains that biomedical text summarization is an important approach to manage the huge amount of scientific literature in these fields of study. This survey has outlined the main approaches starting from the initial extractive methods up till the recent state-of-the-art abstractive models, while showing their key advantages along with the drawbacks. While NLP and machine learning approaches have greatly improved summarization quality over time, there is still much academic grounds to cover, including adapting to various biomedical domains, maintaining factuality, and overcoming differences in fine-grained terminology. The adaptability of this approach could be further

developed, particularly with stronger, more interpretable, and more context aware models that can accommodate the unique challenges posed by biomedical data. Lastly, future directions should also focus on the multi-modal information fusion and biomedical domain-specific evaluation metrics improvement. In conclusion, effective biomedicine summarization may help improve the dissemination of knowledge, facilitate the implementation of evidence-based practices, and fast-forward the research by providing accurate, brief, contextual insights from complex information.

Acknowledgments

The authors are grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Author Contribution

All authors contributed equally to this work.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors

References

[1] Givchi, A., R. Ramezani, and A. Baraani-Dastjerdi, Graph-based abstractive biomedical text summarization. Journal of Biomedical Informatics, 2022. 132: p. 104099.
[2] Moradi, M. and N. Ghadiri, Different approaches for identifying important concepts in probabilistic biomedical text summarization. Artificial intelligence in medicine, 2018. 84: p. 101-116.
[3] Mishra, R., et al., Text summarization in the biomedical domain: a systematic review of recent research. Journal of biomedical informatics, 2014. 52: p. 457-467.
[4] Chaurasia, S., D. Dasgupta, and R. Regunathan, T5LSTM-RNN based Text Summarization Model for Behavioral Biology Literature. Procedia Computer Science, 2023. 218: p. 585-593.
[5] El-Kassas, W.S., et al., Automatic text summarization: A comprehensive survey. Expert systems with applications, 2021. 165: p. 113679.
[6] Searle, T., et al., Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models. Journal of Biomedical Informatics, 2023. 141: p. 104358.
[7] Antony, S. and D.D.S. Pankaj. Survey on Automatic Text Summarization methods and techniques. in Proceedings of the International Conference on Systems, Energy and Environment. 2022.
[8] Azadani, M.N., N. Ghadiri, and E. Davoodijam, Graph-based biomedical text summarization: An itemset mining and sentence clustering approach. Journal of biomedical informatics, 2018. 84: p. 42-58.
[9] Afzal, M., et al., Clinical context–aware biomedical text summarization using deep neural network: model development and validation. Journal of medical Internet research, 2020. 22(10): p. e19810.
[10] Singh, S., J.P. Singh, and A. Deepak, Supervised weight learning-based PSO framework for single document extractive summarization. Applied Soft Computing, 2024. 161: p. 111678.

[11] Sanchez-Gomez, J.M., M.A. Vega-Rodríguez, and C.J. Pérez, An indicator-based multi-objective variable neighborhood search approach for query-focused summarization. Swarm and Evolutionary Computation, 2024. 91: p. 101721.
[12] Rouane, O., H. Belhadef, and M. Bouakkaz, Combine clustering and frequent itemsets mining to enhance biomedical text summarization. Expert Systems with Applications, 2019. 135: p. 362-373.
[13] Alanzi, E. and S. Alballaa, Query-Focused Multi-document Summarization Survey. International Journal of Advanced Computer Science and Applications, 2023. 14(6).
[14] Onan, A. and H.A. Alhumyani, FuzzyTP-BERT: Enhancing extractive text summarization with fuzzy topic modeling and transformer networks. Journal of King Saud University-Computer and Information Sciences, 2024: p. 102080.
[15] Jiang, X. and M. Dreyer. CCSUM: A large-scale and high-quality dataset for abstractive news summarization. in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024.
[16] Zhang, H., P.S. Yu, and J. Zhang, A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models. arXiv preprint arXiv:2406.11289, 2024.
[17] Gidiotis, A. and G. Tsoumakas, A divide-and-conquer approach to the summarization of long documents. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020. 28: p. 3029-3040.
[18] Cohan, A., et al., A discourse-aware attention model for abstractive summarization of long documents. arXiv preprint arXiv:1804.05685, 2018.
[19] Wang, L.L., et al., Cord-19: The covid-19 open research dataset. arXiv preprint arXiv:2004.10706, 2020.
[20] Gupta, V., et al. SumPubMed: Summarization dataset of PubMed scientific articles. in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop. 2021.
[21] Gupta, S., A. Sharaff, and N.K. Nagwani, Biomedical Text Summarization Based on the Itemset Mining Approach, in New Opportunities for Sentiment Analysis and Information Processing. 2021, IGI Global. p. 140-152.
[22] Luo, L., et al., BioRED: a rich biomedical relation extraction dataset. Briefings in Bioinformatics, 2022. 23(5): p. bbac282.
[23] Mollá, D. and M.E. Santiago-Martinez. Development of a corpus for evidence based medicine summarisation. in Proceedings of the Australasian Language Technology Association Workshop 2011. 2011. Australian Language Technology Association.
[24] Lo, K., et al., S2ORC: The semantic scholar open research corpus. arXiv preprint arXiv:1911.02782, 2019.
[25] Plaza, L., A. Díaz, and P. Gervás, A semantic graph-based approach to biomedical summarisation. Artificial intelligence in medicine, 2011. 53(1): p. 1-14.
[26] Moradi, M. and N. Ghadiri, Quantifying the informativeness for biomedical literature summarization: An itemset mining method. Computer methods and programs in biomedicine, 2017. 146: p. 77-89.
[27] Moradi, M., CIBS: A biomedical text summarizer using topic-based sentence clustering. Journal of biomedical informatics, 2018. 88: p. 53-61.
[28] Davoodijam, E., et al., MultiGBS: A multi-layer graph approach to biomedical summarization. Journal of Biomedical Informatics, 2021. 116: p. 103706.
[29] Sharma, G. and D. Sharma, Automatic text summarization methods: A comprehensive review. SN Computer Science, 2022. 4(1): p. 33.
[30] Hark, C., The power of graphs in medicine: Introducing BioGraphSum for effective text summarization. Heliyon, 2024.
[31] Xie, Q., P. Tiwari, and S. Ananiadou, Knowledge-enhanced graph topic transformer for explainable biomedical text summarization. IEEE journal of biomedical and health informatics, 2023.
[32] Kirmani, M., et al., Biomedical semantic text summarizer. BMC bioinformatics, 2024. 25(1): p. 152.
[33] Moradi, M., Small-world networks for summarization of biomedical articles. arXiv preprint arXiv:1903.02861, 2019.
[34] Du, Y., et al., Biomedical-domain pre-trained language model for extractive summarization. Knowledge-Based Systems, 2020. 199: p. 105964.
[35] Kanwal, N. and G. Rizzo. Attention-based clinical note summarization. in Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. 2022.
[36] Moradi, M., G. Dorffner, and M. Samwald, Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. Computer methods and programs in biomedicine, 2020. 184: p. 105117.
[37] Padmakumar, V. and H. He, Unsupervised extractive summarization using pointwise mutual information. arXiv preprint arXiv:2102.06272, 2021.
[38] Moradi, M., M. Dashti, and M. Samwald, Summarization of biomedical articles using domain-specific word embeddings and graph ranking. Journal of Biomedical Informatics, 2020. 107: p. 103452.
[39] Xie, Q., et al., Pre-trained language models with domain knowledge for biomedical extractive summarization. Knowledge-Based Systems, 2022. 252: p. 109460.
[40] Karotia, A. and S. Susan, CovSumm: an unsupervised transformer-cum-graph-based hybrid document summarization model for CORD-19. The Journal of Supercomputing, 2023. 79(14): p. 16328-16350.
[41] Lee, J., et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020. 36(4): p. 1234-1240.
[42] Gu, Y., et al., Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 2021. 3(1): p. 1-23.

[43] Beltagy, I., K. Lo, and A. Cohan, SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676, 2019.
[44] Peng, Y., S. Yan, and Z. Lu, Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474, 2019.
[45] Huang, K., J. Altosaar, and R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342, 2019.
[46] Chen, Y.-P., et al., Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): development and performance evaluation. JMIR medical informatics, 2020. 8(4): p. e17787.
[47] Meng, Z., et al., Mixture-of-partitions: Infusing large biomedical knowledge graphs into BERT. arXiv preprint arXiv:2109.04810, 2021.
[48] Hu, J., et al., Graph enhanced contrastive learning for radiology findings summarization. arXiv preprint arXiv:2204.00203, 2022.
[49] Du, Y., et al., UGDAS: Unsupervised graph-network based denoiser for abstractive summarization in biomedical domain. Methods, 2022. 203: p. 160-166.
[50] Ulker, M. and A.B. Ozer, Abstractive Summarization Model for Summarizing Scientific Article. IEEE Access, 2024.
[51] Cai, X., et al., COVIDSum: A linguistically enriched SciBERT-based summarization model for COVID-19 scientific papers. Journal of Biomedical Informatics, 2022. 127: p. 103999.
[52] Moratanch, N. and S. Chitrakala. A survey on extractive text summarization. in 2017 international conference on computer, communication and signal processing (ICCCSP). 2017. IEEE.
[53] Nenkova, A. Summarization evaluation for text and speech: issues and approaches. in Ninth International Conference on Spoken Language Processing. 2006.
[54] Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. in Text summarization branches out. 2004.
[55] Ganesan, K., C. Zhai, and J. Han, Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. 2010.
[56] Lloret, E., L. Plaza, and A. Aker, The challenging task of summary evaluation: an overview. Language Resources and Evaluation, 2018. 52: p. 101-148.
[57] Rhazzafe, S., et al., Hybrid Summarization of Medical Records for Predicting Length of Stay in the Intensive Care Unit. Applied Sciences, 2024. 14(13): p. 5809.
[58] Schluter, N. The limits of automatic summarisation according to rouge. in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017. Association for Computational Linguistics.
[59] Ng, J.-P. and V. Abrecht, Better summarization evaluation with word embeddings for ROUGE. arXiv preprint arXiv:1508.06034, 2015.
[60] Yuan, H., et al., Revisiting Automatic Question Summarization Evaluation in the Biomedical Domain. arXiv preprint arXiv:2303.10328, 2023.
[61] Papineni, K., et al. Bleu: a method for automatic evaluation of machine translation. in Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
[62] Zhang, T., et al., Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019.
[63] Zhang, L., et al., Leveraging pretrained models for automatic summarization of doctor-patient conversations. arXiv preprint arXiv:2109.12174, 2021.
[64] Zhang, N., et al. Summarizing chinese medical answer with graph convolution networks and question-focused dual attention. in Findings of the Association for Computational Linguistics: EMNLP 2020. 2020.
[65] Hu, J., et al., Word graph guided summarization for radiology findings. arXiv preprint arXiv:2112.09925, 2021.
[66] Afantenos, S., V. Karkaletsis, and P. Stamatopoulos, Summarization from medical documents: a survey. Artificial intelligence in medicine, 2005. 33(2): p. 157-177.
[67] Nguyen, Q.-A., et al. A Hybrid Multi-answer Summarization Model for the Biomedical Question-Answering System. in 2021 13th International Conference on Knowledge and Systems Engineering (KSE). 2021. IEEE.
[68] Rohil, M.K. and V. Magotra, An exploratory study of automatic text summarization in biomedical and healthcare domain. Healthcare Analytics, 2022. 2: p. 100058.
[69] Wang, M., et al., A systematic review of automatic text summarization for biomedical literature and EHRs. Journal of the American Medical Informatics Association, 2021. 28(10): p. 2287-2297.
[70] Pawar, D., et al., Survey on the Biomedical Text Summarization Techniques with an Emphasis on Databases, Techniques, Semantic Approaches, Classification Techniques, and Similarity Measures. Sustainability, 2023. 15(5): p. 4216.