Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

NLP Framework to Summarize p53–FBW7 Pathway Alterations and Associated miRNAs in CRC

Swati Kulshrestha
Celltheon Cooperation, Union City, CA, USA
Corresponding Author: Swati Kulshrestha, **E-mail**: swati@celltheon.com

| ABSTRACT

The increasing intricacies of molecular pathways involved in CRC, especially focusing on the p53-FBW7 pathway network and the microRNAs (miRNAs) that regulate them, present significant difficulties when synthesizing knowledge from the exploding body of biomedical literature. Conventional methods of manual literature review have proven inadequate for capturing and analyzing such multidimensional interactions in a timely fashion. In this paper, we describe an innovative NLP-based approach for mining CRC-related pathway changes and miRNA interactions. The architectural framework uses domain-driven data extraction methods along with sophisticated pre-processing, named entity recognition using transformers, and relation extraction techniques to extract important biological entities and their relationships. The relations extracted by the framework are then encoded within a knowledge graph capturing pathway-level dynamics of gene-protein-miRNA relationships, which can provide an integrated perspective on gene-protein-miRNA relationships. Furthermore, the transformer-based summarizer also provides summarized insights related to p53-FBW7 axis. The model aims to bring together computational intelligence and molecular oncology through a comprehensive, automated, and interpretable process for knowledge integration in biomedical sciences. Expected benefits will involve greater efficiency in research, the ability to identify key regulatory mechanisms, and contributions to precision medicine and drug discovery initiatives. Despite being conceptual, the paper introduces a novel method that combines NLP with cancer pathways, holding promise for extensive development in the future.

| KEYWORDS

Natural Language Processing, p53–FBW7 Pathway, Colorectal Cancer, microRNA Regulation, Biomedical Text Mining

| ARTICLE INFORMATION

ACCEPTED: 01 April 2026 **PUBLISHED:** 15 April 2026 **DOI:** 10.32996/jcsts.2026.8.6.3

1. Introduction

Colorectal cancer (CRC) continues to be one of the most common and lethal cancers around the world owing to complicated molecular changes that lead to abnormal regulation within cells. One such mechanism that regulates cell behavior is the p53 tumor suppressor pathway, which regulates genome stability, controls the cell cycle, and induces apoptosis due to damaged DNA. Mutations within the gene that encodes for the p53 protein have been shown to play a significant role in CRC development, prognosis, and resistance to therapies. At the same time, the F-box and WD repeat domain-containing 7 (FBW7) acts as an important E3 ubiquitin ligase by promoting the degradation of several oncogenes through the ubiquitination process. Abnormal functioning of the FBW7 protein increases the levels of these oncogenes, leading to uncontrolled cellular proliferation and oncogenesis [1].

Apart from changes at the genetic level, microRNA molecules (miRNAs) have also been identified as important post-transcriptional regulators of both p53 and FBW7 signal transduction mechanisms. miRNAs control gene expression by acting on messenger RNAs, thus influencing cell behavior like differentiation, growth, and death. Regulatory relationships between miRNAs and the p53-FBW7 pathway add another level of complexity to research endeavors for synthesizing information gleaned from the vast amount of biomedical knowledge being generated today [2].

The increase in scientific literature makes it impossible to rely on traditional manual techniques when identifying and incorporating relevant information concerning molecular interaction. This is where natural language processing (NLP) comes in; NLP is an area of study within artificial intelligence. Though there have been considerable developments in biomedical NLP, the process has not managed to effectively combine pathways and miRNAs [3].

The current study addresses this challenge by developing a framework using natural language processing (NLP) techniques to extract, analyze, and summarize modifications in the p53-FBW7 pathway and associated miRNAs in colorectal cancer (CRC). This research is intended to promote knowledge discovery and enhance understanding in cancer studies.

2. Background and Literature Review

Colorectal cancer (CRC) has attracted considerable attention in the context of molecular studies during the past few decades with respect to pathways that mediate cellular proliferation, apoptosis, and genome integrity. In this regard, the p53 tumor suppressor pathway emerges as a relatively well-understood example within the realm of cancer biology. The TP53 gene is responsible for encoding a transcription factor that senses cellular stress and regulates either DNA repair mechanisms or apoptosis. Alterations within the TP53 gene have been found in a significant number of patients with CRC, especially those in advanced stages and with poor prognoses. Meanwhile, the FBW7 gene, as a substrate recognition subunit within the SCF (SKP1-CUL1-F-box protein) E3 ubiquitin ligase, serves an important role in the ubiquitination and proteolysis of oncoproteins, including c-Myc, cyclin E, and Notch proteins. Deficiency or mutations in FBW7 lead to the stabilization of oncoproteins [4].

The most recent studies have also highlighted the importance of microRNAs (miRNAs) as key players in controlling the signaling pathways of p53 and FBW7. MicroRNAs are post-transcriptional regulators that interact with target mRNAs in order to promote either their degradation or translational repression. Several miRNAs have been found that target FBW7 and down-regulate it leading to oncogenesis. P53 is known to be able to regulate specific miRNAs and create feedback loops, thus affecting tumor progression. The complexity of the interaction between genes and miRNAs considerably increases the challenges in studying the etiology of CRC.

The emergence of biomedical literature characterized by increasing complexity and quantity has led to the use of NLP techniques aimed at obtaining valuable knowledge from unstructured data. Initially, there was an emphasis on rules and keyword matching approaches, which turned out to be too simplistic for the detection of context-dependent relationships. However, further improvements made with regard to deep learning and transformers (such as BERT and BioBERT), have contributed to improving named entity recognition and relationship extraction within biomedical documents [5].

However, despite all the progress made, current NLP systems tend to operate on a fragmented basis, focusing only on the extraction of individual entities instead of the integration of pathway interactions with miRNA regulation. In addition, automatic summarizers are rarely tailored to biological pathways, thus limiting their use in cancer studies. Therefore, there is a clear need for an integrated system combining modern NLP techniques with biological knowledge to generate summaries of complex molecular interaction networks, especially those associated with the p53-FBW7 pathway in CRC [6][7].

2.1 Problem Statement and Research Gap

Although there have been many advancements in the field of understanding the molecular mechanism behind CRC, integrating all information relating to pathway modifications and microRNA involvement is an area that still poses great difficulty. While the regulation axis between p53-FBW7 has been studied extensively, its interactions are quite complicated and involve not only changes in genes but also their degradation, in addition to interactions through microRNA regulation. This complexity is compounded by the vast body of biomedical information where relevant data are scattered across thousands of papers, making curating this information more difficult than ever [8].

Presently available NLP methods for computational analysis in biomedicine largely focus on performing individual tasks such as named entity recognition and basic relation extraction. While these models have improved the capability to identify genes, proteins, and disease mentions in texts, they often struggle to bring together biological processes occurring on multiple levels in an effective manner [9]. In particular, current models have difficulty in representing the interdependency of signaling pathways and miRNAs involved in regulating them to understand how CRC progresses. Another major issue with the currently available approaches is the use of generic summarizers that produce biologically irrelevant information.

An additional limitation of current studies involves the inadequacy of approaches that can convert free-text information into pathway knowledge representations like knowledge graphs that properly represent real biological systems. This problem impairs the capacity of scientists to derive practical knowledge and discover potential therapeutic targets effectively.

Thus, it is evident that there should be a specialized NLP system that will allow scientists to extract, summarize, and organize interactions of the p53-FBW7 pathway and associated miRNAs in CRC. Filling this research gap becomes the core aim of the current study.

3. Proposed NLP Framework Architecture

In order to solve problems that are associated with retrieving and integrating complex information about molecular interactions through the large body of literature available, the paper introduces a multilayered NLP model, which aims at summarizing any changes in the p53-FBW7 interaction pathway and miRNAs involved in colorectal cancer (CRC). This model is created in such a way that it allows integration of all processes into one single platform.

The proposed framework begins with the stage of data acquisition where relevant biomedical literature concerning the subject matter is acquired in a systematic way from reliable databases such as PubMed, MEDLINE, and medical studies. The literature is carefully filtered through the application of a filter specific to the subject matter so that only those publications discussing CRC, p53, FBW7, and miRNA relationships will be included. Preprocessing of textual data involves normalizing text strings, tokenizing texts, and eliminating unimportant linguistic elements. Domain ontologies and vocabularies play an important role during this stage.

The NER module forms the heart of this pipeline. It utilizes state-of-the-art transformer models like BioBERT to extract important biological terms such as genes, proteins, microRNAs (miRNAs), and diseases. These terms are then fed into a relation extraction module where both semantic and syntactic approaches are used to identify relationships, such as gene–miRNA regulation, protein degradation pathways, and mutations that cause changes in function. This step is crucial in identifying biological relationships associated with the p53-FBW7 pathway.

Finally, after the extraction of these entities and their interactions, all of these are organized using a graph structure to represent pathway-level interactions. This kind of graph model helps in visualizing the complex networks to understand the molecular biology involved in CRCs. Further to this representation, a transformation summarization engine is developed, which will summarize all important features from the pathways.

The last piece of the framework includes a visualization and user interface layer which renders insights and knowledge graphs in a visually appealing form, allowing users to study interactions within pathways more easily. Through the combination of all the discussed layers, the developed approach is able not only to automate the identification of important biological insights but also to ensure their accessibility through a unified architectural design.

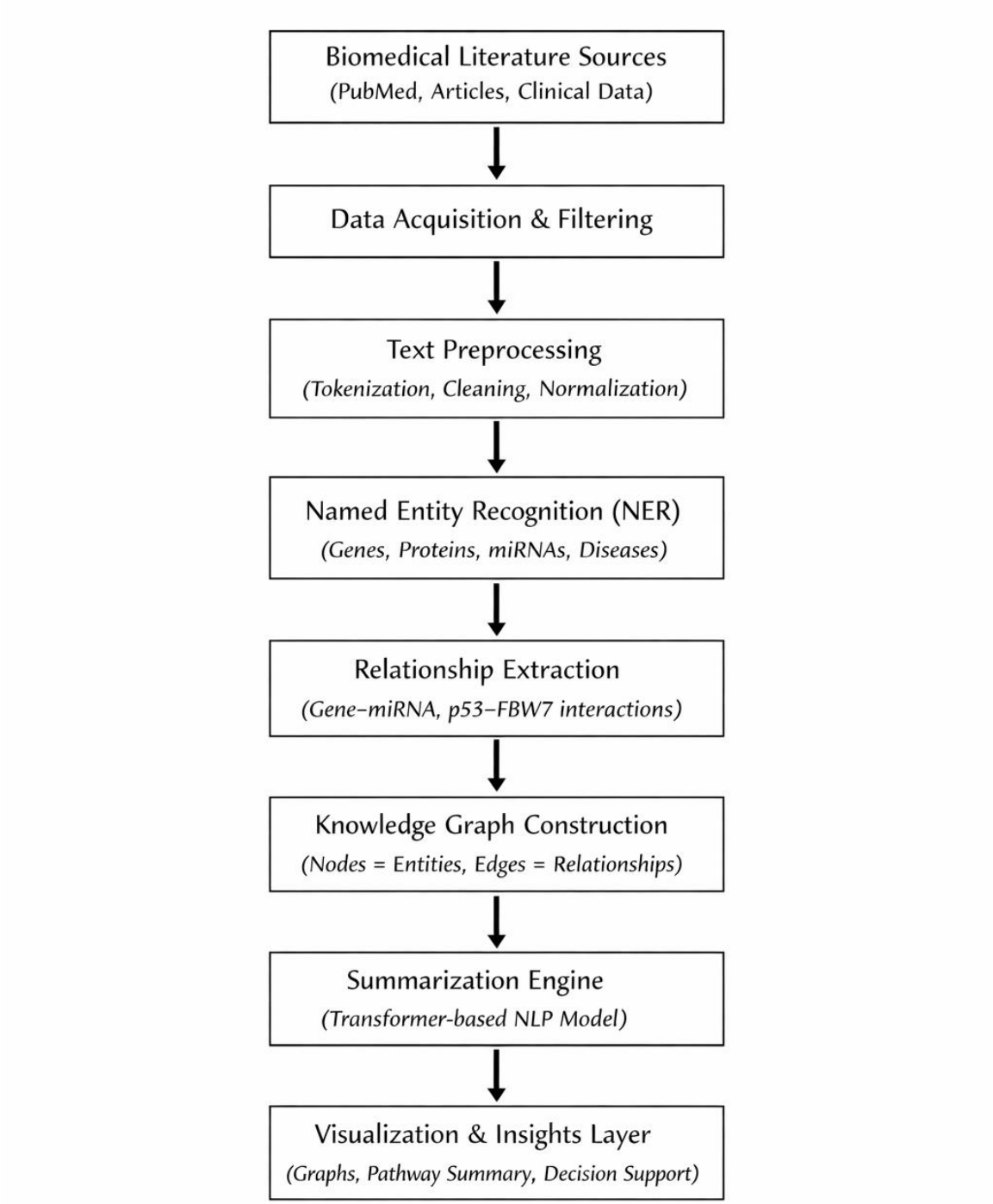

Fig 1: Proposed Framework

The diagram above demonstrates an organized methodology for Natural Language Processing (NLP) in order to extract and synthesize information related to changes in p53-FBW7 pathway and the corresponding microRNAs (miRNAs) involved in CRC. The process starts off with collecting data from the biomedical literature found on resources like PubMed along with relevant clinical databases. Data collection and filtering will then take place based on the context of relevance. Text preprocessing is done in order to clean, tokenize, and normalize the text data. Named Entity Recognition (NER) is performed in order to identify the key biological entities in the literature, which include the genes, proteins, miRNAs, and the diseases. Relationship extraction is performed to analyze the interaction between various entities identified through NER. The extracted information will be represented using a knowledge graph that will capture the complex biological relationships between different entities.

Transformer model-based summarization will provide succinct information from the knowledge graph in order to understand the biological processes better.

4. Methodology

The suggested study utilizes a modular computing technique to develop and test an NLP-based model for identifying changes in the p53-FBW7 signaling pathway and miRNA interactions in colorectal cancer (CRC). The methodology involves creating a domain-specific corpus by collecting relevant articles that can be obtained by conducting searches using keywords such as CRC, TP53, FBW7, and miRNA interactions. The pre-processing step entails cleaning the text from irrelevant content and formatting it in accordance with the input requirements of biomedical NLP techniques.

Post preprocessing, deep neural network-based transformers like BioBERT are applied for Named Entity Recognition (NER), which helps recognize important biological entities like genes, proteins, miRNAs, and disease entities. The NER output is aligned to existing biomedical ontologies to ensure domain consistency in identifying the entities. The next step is to perform relationship extraction through dependency parsing and context embedding methods to identify relationships between the recognized entities.

After extraction, the entities and relations are further structured into a knowledge graph that serves as the foundation for modeling the p53-FBW7 regulatory network. Through such an approach, key entities and relations can be identified, facilitating better understanding of the underlying mechanisms. Following that, a transformer-based summary generation method is utilized to generate informative summaries about the changes in the pathways, capturing critical gene-miRNA relations and their biological significance.

In order to test the efficiency of the proposed framework, traditional performance parameters such as Precision, Recall, and F1 score will be used during entity recognition and relation extraction tasks. The process of validating the model using experimental data is not possible; therefore, we rely on comparing the findings with those obtained by summarizing the literature in the field.

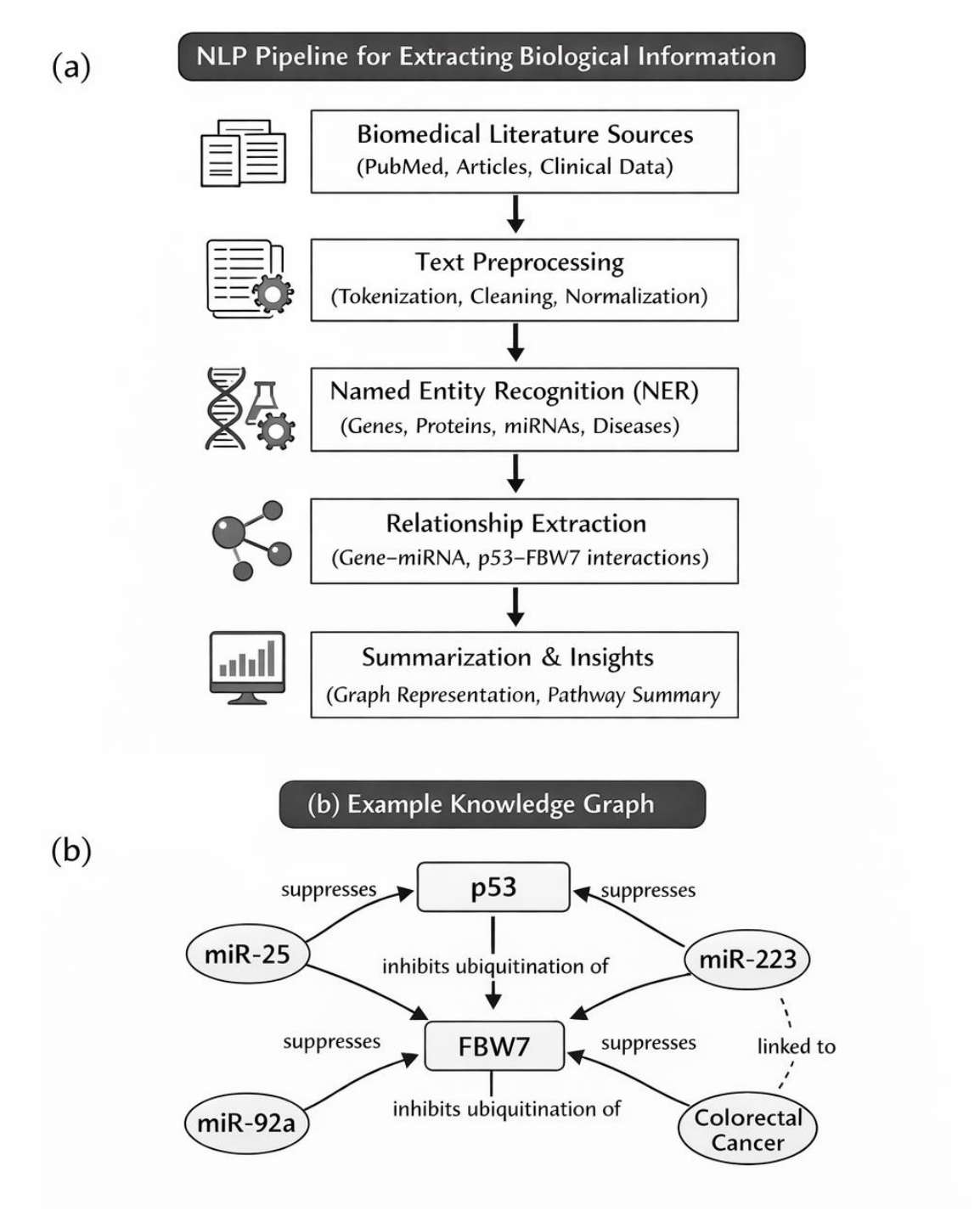

Fig 2: Proposed methodology

The following figure provides details about the overall approach used within the proposed NLP framework to analyze the p53-FBW7 pathway and associated miRNAs in colorectal cancer. Figure (a) illustrates the approach starting from the acquisition of medical literature from databases like PubMed. Following the initial acquisition phase, a preprocessing phase follows, including tasks such as tokenization and cleaning the data before the actual analysis. Next, the approach involves using NER technology to recognize important entities from biology (such as genes, proteins, miRNAs, and diseases). These entities are then processed in the phase of extracting relationships among entities (gene-miRNA regulatory relations or p53-FBW7 pathway associations). This information is finally used to produce summaries and graphical depictions of the information gathered. Figure (b) displays one of the examples of knowledge graphs created by the proposed method and showing the interactions among entities such as p53, FBW7, and different miRNAs like miR-25, miR-223, and miR-92a contributing to colorectal cancer.

5. Expected Results and Contributions

It is envisaged that the suggested NLP-based framework would serve as an efficient approach towards automatically acquiring and analyzing the intricate biological information related to the p53-FBW7 pathway and miRNAs linked with it in the context of CRC patients. The framework would be capable of identifying and linking the relevant molecular entities using entity recognition and relationship extraction approaches. Through this process, the unstructured textual data would be converted into a structured knowledge base. This, in turn, would decrease the dependence on a manual review of the literature.

Another key expected outcome will be the creation of a complete knowledge graph that will represent dynamics at the level of pathways, taking into account mutations in genes, protein degradation, and miRNA regulation. Such a knowledge graph is supposed to make it easier to recognize important nodes and their interactions, which could not be easily determined using conventional methods. Additionally, the inclusion of a transformer-based summarization module will likely generate concise summaries highlighting biological importance.

In terms of contributions to scientific research, the work under discussion is distinguished by its unified model combining NLP techniques and molecular oncology. Whereas previous research focuses on single extraction processes, the presented methodology stresses the significance of pathway-based integration and summarization specific to the problem domain. From this point of view, the described system can be seen as a contribution to intelligent computing technologies for medical knowledge discovery. Furthermore, the suggested model offers a flexible platform which can be implemented to cover other signal pathways and illnesses for different biomedical applications such as precision medicine, drug target discovery, and clinical decision support.

6. Applications

In this regard, the suggested framework, based on natural language processing technology, presents considerable prospects in various fields of biomedical investigation and medicine, with special attention paid to improving the comprehension and treatment of colorectal cancer (CRC). One of its important uses includes precision medicine, where the accurate analysis of changes in the signaling pathway between p53 and FBW7 and related microRNAs (miRNAs) may contribute to the creation of tailor-made approaches to therapy.

For drug discovery studies, the proposed methodology aids in finding new drug targets via the discovery of important genes, proteins, and miRNAs involved in CRC development. The structured representation of knowledge generated from the system assists in pinpointing regulatory nodes along the pathway that can act as possible drug targets, thus facilitating the early stages of drug discovery studies. Moreover, the updating mechanism of the knowledge database based on the latest published information guarantees that drug discovery activities remain current with the latest research advancements.

The proposed system has the potential for applications in clinical decision support systems where the summarization of important molecular facts will assist medical practitioners in making decisions. The system helps overcome the cognitive burden of reading large amounts of biomedical documents. In addition, for academic purposes, the proposed framework can speed up the process of knowledge discovery by rapidly synthesizing complex interactions between different biological entities that may help formulate new hypotheses and design experiments.

In addition to CRC, the proposed model has the flexibility of being extended to other types of cancers, thus making it a useful tool for research in data-driven medicine and computational biology.

7. Challenges and Limitations

Although there are many advantages in using NLP techniques for the development of an information extraction tool, some challenges and difficulties have to be mentioned. The first difficulty is the complexity of language, specifically in biomedicine, where the same term can be represented by different terms and abbreviations. It can influence both named entity recognition and relationship extraction, especially if specific issues, such as the interaction between genes and miRNA, have to be addressed. The second issue refers to the lack of quality annotated biomedical data.

Another significant limitation relates to the difficulty of understanding the full biological environment in which molecular interactions take place based purely on text-based information. While natural language processing tools can identify correlations and trends, they cannot necessarily convey the dynamic nature of biology or the experimental conditions mentioned in scientific texts. This results in incomplete and context-dependent pathways. Furthermore, although transformers are quite effective, they

tend to be highly opaque, making it difficult to understand how certain outputs have been generated, which is important for biomedical use cases.

The issues of scalability and computing requirements limit practical applications, particularly with large collections of literature. Also, the fact that there is no clinical validation of this framework in the current environment limits its application in clinical practice. Solving these issues will be key to enhancing the system's effectiveness and usefulness in the future.

8. Future Scope

This suggested framework provides a great base for biomedical knowledge synthesis, but there is still a lot to do and improve in the field. The primary direction of further research would be to include the information from clinical studies and from various omics fields such as genomics, transcriptomics, and proteomics in order to make the analysis more complete regarding the studied pathway and its associated miRNAs.

Further areas of improvement include the implementation of sophisticated machine learning algorithms and predictive analytics that would enable moving past mere descriptive summarization to prediction and prescription. With the aid of past data and discovered patterns, such an approach might be able to predict the future progression of the illness or reaction to a particular treatment regime. Explainable AI technologies would also be required for increased understandability of the model.

The model may be enhanced further to enable analysis of multiple pathways or even diseases. In addition, interactive tools or interfaces, and AI-powered research assistants can be developed to make the model easily accessible to non-experts such as medical practitioners and biomedical researchers. Given the recent advances in natural language processing and biomedical informatics, it is possible that the proposed system will transform into an all-encompassing decision support tool incorporating computational intelligence and translational medicine.

9. Conclusion

In our research work, we have proposed a unique framework of Natural Language Processing (NLP) that would help us to extract, analyze, and summarize information related to modifications that take place within the p53-FBW7 signaling pathway and miRNAs associated with colorectal cancer (CRC). It is important to recognize that there is an urgent need in bio-medical research where huge amounts of text can be converted into structured forms of knowledge using the proposed framework.

From a theoretical standpoint, the design of the framework highlights the potential it possesses to increase efficiency in research, reduce reliance on traditional manual literature review, and improve the accessibility of complicated molecular knowledge for use by scientists and doctors alike. The ability to identify the relationship between genes, proteins, and miRNAs in one unified model can be seen as an advancement in comparison to the currently fragmented efforts in biomedical NLP.

Though not accompanied by any empirical evidence at the current stage of development, the framework does serve as a scalable and flexible groundwork for further implementation with clinical and multi-omics datasets. In conclusion, the present study is another step towards the growing interaction between artificial intelligence and molecular oncology research.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] S. Yan, F. Zhan, Y. He, Y. Zhu, and Z. Ma, "p53 in colorectal cancer: from a master player to a privileged therapy target," 2025. doi: 10.1186/s12967-025-06566-4.
[2] J. Xiao, H. Lin, X. Luo, X. Luo, and Z. Wang, "MiR-605 joins p53 network to form a p53:miR-605:Mdm2 positive feedback loop in response to stress," *EMBO Journal*, vol. 30, no. 3, 2011, doi: 10.1038/emboj.2010.347.
[3] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed. Tools Appl.*, vol. 82, no. 3, 2023, doi: 10.1007/s11042-022-13428-4.
[4] L. Yan, J. Shi, and J. Zhu, "Cellular and molecular events in colorectal cancer: biological mechanisms, cell death pathways, drug resistance and signalling network interactions," 2024. doi: 10.1007/s12672-024-01163-1.
[5] C. A. Withers *et al.*, "Natural language processing in drug discovery: bridging the gap between text and therapeutics with artificial intelligence," 2025. doi: 10.1080/17460441.2025.2490835.

[6] B. Bhasuran, "BioBERT and Similar Approaches for Relation Extraction," in *Methods in Molecular Biology*, vol. 2496, 2022. doi: 10.1007/978-1-0716-2305-3_12.
[7] C. H. Yeh, M. Bellon, and C. Nicot, "FBXW7: A critical tumor suppressor of human cancers," 2018. doi: 10.1186/s12943-018-0857-2.
[8] M. Tufail, C. H. Jiang, and N. Li, "Wnt signaling in cancer: from biomarkers to targeted therapies and clinical translation," 2025. doi: 10.1186/s12943-025-02306-w.
[9] Scienta Team *et al.*, "EVA: Towards a universal model of the immune system," Feb. 2026, [Online]. Available: http://arxiv.org/abs/2602.10168