IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 11 Issue 08, August 2024
ISSN (Online) 2348 – 7968 | Impact Factor – 6.72
www.ijiset.com

Drug reproposing For Melanoma Using Text Analysis

S Divya[a] , Sidesh Sundar[b] , Balu Bhasuran[c], I R Oviya[d*]

[a] Department of Computer Sciences,Amrita School of Engineering ,Amrita Vishwa Vidyapeetham ,Chennai, India,s_divya@ch.students.amrita.edu
[b,d] Department of Computer Science and Engineering , Amrita School of Computing ,Amrita Vishwa Vidyapeetham ,Chennai, India.
[c] Bakar Computational Health Sciences Institute,University of California,San Francisco, USA.

[*] Corresponding Author; iroviya@gmail.com. ORCID ID: 0000-0002-1421-4232

Abstract—Pharmaceutical innovation, encompassing the identification of new treatments using drugs initially intended for different purposes, plays a pivotal role in expediting drug discovery amidst challenges posed by a burgeoning healthcare landscape. Given the substantial expenses and time constraints inherent in conventional drug development, repurposing emerges as a strategic approach to leverage existing resources for identifying candidates with novel therapeutic applications. In this investigation, we used PubMed ID text mining and implemented a pattern-based association extraction method to discern direct correlations between diseases and genes, as well as between genes and drugs, within the scientific literature. These established direct relationships serve as the foundation for determining indirect connections utilizing the ABC model. In contrast to biochemicalcentric methodologies, our approach for assessing drug-target similarity exhibits a robust correlation, with a Pearson coefficient of 0.9777/%. The indirect correlation classification method attains average accuracy scores for the 100 most prevalent diseases. Furthermore, we validated the practicability of utilizing the identified candidates in oncology through a manual review of literature and clinical trials pertaining to cancer and melanoma. Our comprehensive analysis promises significant revelations, particularly in the realm of cancer-related drugs and genes, offering crucial insights into drug therapy. Repositioning existing drugs for novel indications presents a more favorable risk-to-reward ratio compared to alternative drug development strategies, contributing to the enhanced productivity required by the industry while shifting production emphasis towards biotechnology firms.

Index Terms—Text mining, Drug target,Disease target, Named entity recognition, Information extraction

I. INTRODUCTION

Drug development is a complex and difficult process characterized by long time, excessive costs and associated risks. In response to these challenges, drug reformulation has become a strategic approach in the pharmaceutical industry.[1][2] This requires the use of approved drugs for new therapeutic indications, which offers several advantages over traditional drug development approaches. One of the main advantages is the assurance of safety and efficacy, as these drugs have already undergone rigorous clinical trials and regulatory oversight. Take aspirin, for example, which was originally used to relieve pain and was later found to be effective in preventing cardiovascular events and some cancers. Another notable case is thalidomide, which was originally designed as an antinausea drug but was developed to treat skin diseases and certain cancers.[3]

Fig. 1. Drug Development

Drug development is time consuming, expensive and very risky .The Figure 1 shows the various process involved in development of the drug which is time consuming .So this became the reason for the rise of drug reformulation, which involves the adaptation of approved drugs to a new disease. The main advantage ofre-proposing drugs to drug development 1is that the drug has been approved and passed the stages of clinical trials, so it is a sign of safety and therapeutic effect. A well-known example is aspirin, which was once used to relieve pain and was later used to prevent cancer and cancer.

Thalidomide was originally used as an anti-nausea drug and to relieve morning sickness in pregnant women, but it failed because it could cause Phocomelia syndrome.[4] However, thalidomide has recently been found to be effective in the treatment of skin diseases, aphthous stomatitis and multiple mealoma. These reproposed drugs showed that drug re-proposing is a promising feature for drug discovery. Identifying diseases, genes, genes and drugs, and disease-drug relationships is key to identifying and treating new candidates for drug repurposing. Finding and selecting new candidates for medication repurposing requires understanding the links between diseases, genes, and drugs.

The links between different biomedical entities are widely curated in databases , however there may be many more undiscovered correlations hidden in the biomedical literature. Literature-based discovery (LBD) to produce scientific hypotheses for discovering novel indications of already-approved medications appears to be a suitable approach, as mentioned in a review and two research studies . In their assessment of several LBD approaches, indicated that scientists may find it easier to identify hidden connections between biological elements if visualization tools are used.

To find hidden disease-drug links, we concentrated on extracting disease-gene and gene-drug relationships in this work. We discovered that this technique might be used to identify important indirect relationships. In order to enable the detection of indirect linkages that might aid in the discovery of new candidates supporting the curation for drug repurposing, we created a textmining-based ranking technique.The specific goal of this project was to create a drug vector space-based ranking system to find the most promising candidate medications and an intuitive pattern-based learning method to extract relationships from the biomedical literature.

II. LITERATURE REVIEW

We referred paper titled ”Drug Repositioning: Identifying and Developing New Uses for Existing Drugs” which presents an in-depth examination of the challenges and opportunities within the biopharmaceutical industry. It explores the persistent productivity gap despite substantial investments in innovative discovery technologies and the mounting pressures from generics and regulatory hurdles. The paper underscores the advantages of drug repositioning over conventional drug discovery methods, highlighting its potential to streamline development timelines, mitigate risks, and create lucrative prospects for companies.[5]

The significant increase in spending on novel technologies in recent years has not translated into improved research and development (R&D) productivity. Contrary to expectations, R&D productivity has declined since the mid-1990s. This decline is evident when measuring productivity in terms of the number of new drugs approved per dollar spent or the number of original Investigational New Drug (IND) applications received by the US Food and Drug Administration (FDA) from commercial sources per dollar spent.[5] Moreover, the author of this paper provided insightful case studies illustrating successful drug repositioning endeavors and scrutinizes the strategies and methodologies adopted by biotech companies in this domain. It also addresses the intricacies associated with drug repositioning, including navigating intellectual property concerns, conducting comprehensive due diligence, and overcoming internal organizational obstacles.

Despite the substantial investments in innovative technologies, the output in terms of successful drug approvals or IND applications has not kept pace with the financial resources allocated to R&D. This trend underscores the challenges and complexities inherent in the drug development process, highlighting the need for a reevaluation of R&D strategies and investments to enhance productivity and maximize the impact of resources allocated to biomedical research and development. This paper provides valuable insights into the potential of drug repositioning as a strategic tool for tackling productivity challenges in the biopharmaceutical industry. It delivers a nuanced analysis of the hurdles, strategies, and emerging trends in drug repositioning, underscoring its superiority over traditional drug discovery methods and its capacity to generate substantial returns for companies. Through illustrative case studies and successful examples, the paper showcases the transformative impact of drug repositioning in bolstering industry productivity.

We also used article ”Text-Mining Approach to Identify Hub Genes of Cancer Metastasis and Potential Drug Repurposing to Target Them” delves into the crucial task of pinpointing genes associated with cancer metastasis and potential drugs capable of targeting them. Through meticulous text mining of PubMed citations, the study uncovers candidate genes implicated in various metastatic processes, including invadopodia , motility, movement, invasion, and more. These genes are then categorized as driver genes, tumor suppressors, or oncogenes, totaling 185 unique cancer genes pertinent to metastasisrelated activities. Subsequent hub gene analysis identifies 77 key genes pivotal in the metastatic cascade.[6]

Furthermore, the study identifies a cohort of approved drugs—50 in total—that hold promise for repurposing as anti-metastatic agents against 19 hub genes involved in metastasisrelated pathways. Employing a methodology blending text mining, bioinformatics analysis, and virtual screening data, the research unveils a promising avenue for developing anti-metastatic therapies. By shedding light on potential therapeutic targets for impeding metastasis, the findings offer valuable insights into combating one of cancer’s most formidable challenges.[7] The use of complexity inherent in targeting metastasis, which spans multiple intricate steps. It underscores the significance of drug repurposing as a cost-effective and expeditious strategy to advance treatment modalities. However, it acknowledges the challenges posed by metastasis’s multifaceted nature and the necessity for further validation of identified genes and drugs through experimental and computational approaches.

In conclusion, the meticulous exploration—from gene discovery to drug repurposing—highlights its potential to revolutionize cancer treatment strategies, particularly in combating metastasis. It advocates for continued research and validation efforts, suggesting the utilization of Bidirectional Encoder Representations from Transformers (BERT) for future predictive modeling endeavors. Overall, the article stands as a comprehensive investigation into leveraging text mining to unravel cancer metastasis’s molecular intricacies and expedite drug repurposing for improved patient outcomes.[8] In addition to the review they achieved the automated reasoning for the incomplete information of indirect linkages for drug indications using the declarative programming language AnsProlog. Additionally,to uncover alternative pharmacological indications, offered a number of publicly available knowledge resources, including chemical structures, adverse effects, and signaling pathways. As a result, the finding of unknown linkages for drug repurposing may be made possible by text mining scientific literature libraries to identify otherwise hidden relationships. We attempted to create a relationship extraction technique with the understanding that a great deal of information is concealed in the substantial body of literature in order to facilitate the repurposing of drugs. It was possible that traditional methods, which concentrated on identifying the current drug-disease correlations, might not uncover any new ones .[9]

The role of Artificial Intelligence (AI) in drug discovery, encompassing its challenges, opportunities, strategies, and ethical considerations. AI offers immense potential to revolutionize the drug discovery process, promising enhanced efficiency, accuracy, and speed. However, its implementation is not without hurdles. Challenges such as data quality assurance, algorithm transparency, and interdisciplinary collaboration must be addressed to maximize its utility. Despite these obstacles, AI presents numerous opportunities, including accelerated drug repurposing, de-novo drug design, and personalized medicine. By leveraging machine learning algorithms, AI can unlock hidden patterns in biological data, facilitating informed decision-making and targeted therapeutic interventions. To harness the full potential of AI, effective implementation strategies are imperative. This entails robust protocols for data acquisition, algorithm development, and validation to ensure reliable outcomes. Moreover, ethical considerations loom large in the integration of AI in drug discovery. Issues such as data privacy, algorithmic bias, and accountability demand careful scrutiny and regulatory oversight to safeguard against unintended consequences and uphold ethical standards.

Furthermore, while AI holds promise, it is not a standalone solution. Integration with traditional experimental methods and human expertise is essential to complement its capabilities and mitigate inherent limitations. This symbiotic approach optimizes the drug discovery process by leveraging the strengths of both AI and conventional techniques. Additionally, the proliferation of AI-generated content in scientific literature necessitates vigilance against misinformation dissemination and data manipulation. Strict guidelines, transparent reporting standards, and public awareness campaigns are essential to maintain the integrity and credibility of scientific discourse. In conclusion, the responsible and judicious use of AI in drug discovery holds transformative potential for pharmaceutical research. By addressing challenges, seizing opportunities, and upholding ethical principles, AI can usher in a new era of innovation, ultimately benefiting patients and society as a whole.[10][11]

III. METHODOLOGY

In our study, we propose a novel approach for re-proposing drug using Natural Language Processing (NLP), using the data from PubMed and open source data from TTD. The methodology encompasses several key phases, each designed to extract the gene, target and disease and related the relation between them .

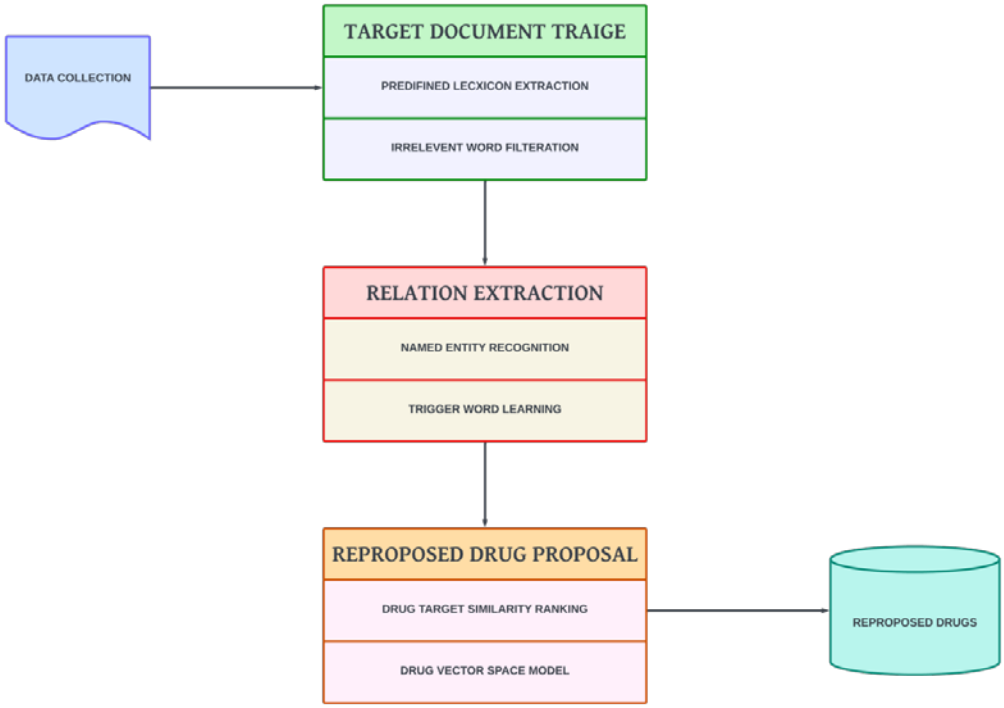

Fig. 3. The method for identifying new candidates for drug re-purposing by text mining

*A. **Data Collection: Retrieving and Organizing Biomedical Abstracts from PubMed***

In this phase we gather essential data from PubMed, a comprehensive repository of biomedical literature and delineates the methodical approach assigned for collecting abstracts associated with specific PubMed IDs, incorporating robust error handling mechanisms to ensure data integrity.

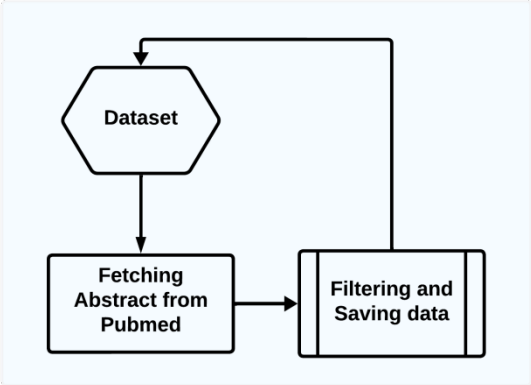

Fig. 4. Data Collection

a. Reading and Fetching Abstracts from PubMed

The figure 4 shows process of data collection from PubMed.The data collection process commences by extracting PubMed IDs listed in Table 1 from a designated CSV file and the data is stored in a list.

Table I
PubMed ID's

S No	Pubmed ID's
1	18771349
2	11774264
3	18312439
4	24595385
5	2484057

The next step involves the providing the email address for identification through the Entrez module, facilitating responsible data retrieval from the PubMed database. Then the abstracts are retrieved for each PubMed ID. The Entrez module is utilized to fetch XML-formatted data for the specified PubMed ID, from which the abstract information is extracted. Robust exception handling is implemented to address potential errors during the fetching process, ensuring the reliability of abstract retrieval.

b. Filtering and Exporting data

The obtained abstracts are organized into a Data-frame, mapping each PubMed ID with its respective abstract. To enhance data quality, a filtering step is introduced to exclude instances where the abstract retrieval process encountered errors, resulting in None values. The finalized DataFrame is saved as a CSV file ('abstract.csv') for ease of access and downstream analysis. This structured approach ensures the seamless integration of retrieved biomedical abstracts into subsequent phases of our study.This meticulous data collection process adheres to best practices, combining efficient data handling with rigorous error handling mechanisms, laying the foundation for a robust and reliable dataset essential for our biomedical research endeavors. The CSV file serves as a valuable resource for researchers seeking access to curated biomedical abstracts for further exploration and analysis.

*B. **Target Document Triage : Prioritizing Biomedical Abstracts based on Length***

The Target Document Triage aims to enhance the efficiency of information retrieval by prioritizing biomedical abstracts based on their lengths. This strategic approach acknowledges the

potential correlation between abstract length and the depth of information provided, allowing for a more targeted exploration of documents. The sample results are shown in Table II

TABLE II
PUBMED ID AND ABSTRACT LENGTH

Pubmed ID's	Abstract Length
20177751	2734
22736149	2537
16681721	2474
25048347	2301
14961576	2134

A quantitative assessment of the abstract's length is done in this step. The abstract lengths are calculated and noted. This metric serves as a proxy for the document's content richness, enabling subsequent prioritization based on length. To optimize the information retrieval process, the data is sorted in descending order based on abstract length. This ensures that abstracts with longer lengths, potentially containing more comprehensive information, are prioritized at the top of the triaged data. The resulting data, reflects a strategic reordering of documents for subsequent analysis. The triaged data is then saved as a CSV file.

C. Gene Extraction : Integrating Genomic Information with Biomedical Abstracts

The next phase of our study, focuses on the extraction of gene-related information from the previously curated biomedical abstracts. Leveraging gene data and the abstracts , we focused to establish connections between genetic entities and the information encapsulated within the abstracts. The figure 5 shows the step by step process of the gene extraction from the biomedical literature set. The gene data, containing gene symbols, is loaded , and the abstracts, associated with PubMed IDs, are also loaded . To establish associations between gene names and biomedical abstracts, a meticulous process is utilized.

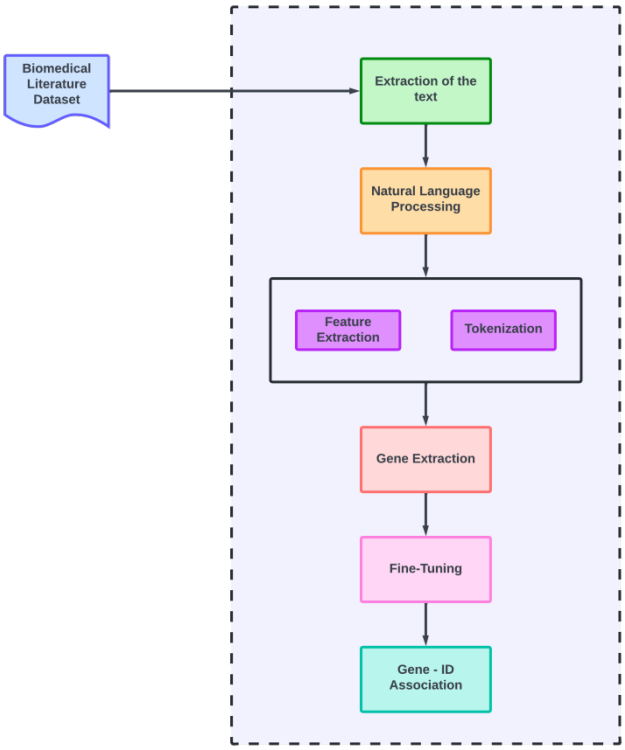

Fig. 5. Gene Extraction

For each gene in the list, the corresponding gene name is tokenized using the BERT tokenizer, allowing for flexible matching within the tokenized abstracts. This step ensures a comprehensive identification of gene mentions within the abstracts, capturing instances where specific genes are

discussed in the context of the biomedical literature. This integrated approach, combining genomic data with abstracts, is pivotal for enriching the contextual understanding of genetic information within the biomedical domain. The result is a subtle dataset that not only prioritizes abstracts based on length but also establishes explicit connections between gene entities and the information present in the abstracts. The subsequent analyses and interpretations benefit from this integrated perspective, paving the way for more informed insights and discoveries in the field of genomics.

D. Gene Tokenization and Mention Extraction: Using BERT for Gene Identification in Biomedical Abstracts

In this phase we use BERT (Bidirectional Encoder Representations from Transformers) to tokenize gene names within the curated biomedical abstracts. This process enhances the precision and comprehensiveness of gene mention extraction, facilitating a more nuanced understanding of genomic information within the textual context. The figure 6 shows the architecture of the BERT model.

a. Loading BERT Model and Tokenizer

The process begins with the loading of a pre-trained BERT model (‘bert-baseuncased‘) and its corresponding tokenizer. BERT, renowned for its contextualized embeddings, proves instrumental in capturing the intricate relationships between words within the abstracts. The tokenizer is used to convert both gene names and abstracts into tokenized representations, ensuring consistency and compatibility for subsequent analysis.

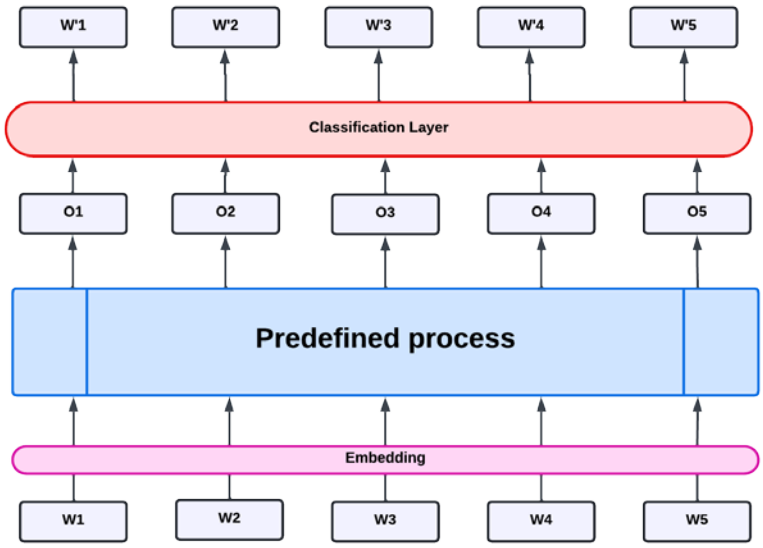

Fig. 6. Architecture of BERT model

b. Tokenization and Encoding of Abstracts

The abstracts, stored in the list, is tokenized and encoded using the BERT tokenizer. The resulting tokenized abstracts, represented as tensors, serve as input for the BERT model. This transformative step prepares the textual data for the subsequent extraction of gene mentions.

TABLE III
PUBMED ID AND EXTRACTED GENE

Pubmed ID	Extracted Gene
20177751	BR ##CA ##2
11668523	CD ##K ##N ##2 ##a ##S1
22080950	CD ##K ##4
15466193	PT ##EN
25974357	BA ##P ##1

The result which we obtained are not in readable format showed in Table III.It also consist some special character which are added to the result during the gene extraction by the BERT model.

c. Gene Mention Identification and Fine-tuning

This section is processed for identification of gene mentions within the tokenized abstracts. For each gene in the list, tokenization and encoding are performed, aligning with the BERT tokenizer's conventions. A nested iteration over PubMed IDs and abstracts facilitates the identification of specific indices where gene mentions are present. The resulting gene mentions, along with their associated PubMed IDs, are recorded and are shown in Table IV.

TABLE IV
PUBMED ID AND EXTRACTED GENE

Pubmed Id	Fine-tuned Gene
20177751	BRACA2
11668523	CDKN2B-AS1
22080950	CDK4
15466193	PTEN
25974357	BAP1

Then we fine-tune the extracted gene mentions to enhance the precision and structure of our genomic analysis. Leveraging a DataFrame containing gene mentions, this process involves refining and aggregating the data to create a more streamlined and interpretable representation.

E. ***Lexicon Extraction : Identifying Cancer-Related Entities in Biomedical Abstracts***

The next step in our methodology uses natural language processing (NLP) techniques to extract cancer-related entities from the biomedical abstracts, providing a targeted exploration of terms associated with cancer and melanoma research. Here are some sample terms used for lexicon extraction. cancer, melanoma, oncology, tumor, carcinoma, neoplasm, malignancy, chemotherapy, radiation therapy, immuno-therapy, oncologist, metastasis,biopsy, lymphoma, leukemia, sarcoma, malignant, benign,radiology, chemo, radiotherapy, hormone therapy. Using the SpaCy model for named entity recognition (NER) and a curated list of cancer-related terms, this approach contributes to a more focused understanding of the oncological landscape within the corpus.

a. Loading NLP Model and Required Libraries

The process begins by loading the SpaCy NLP model *en core web sm* and downloading necessary resources from the Natural Language Toolkit (NLTK). These resources include the WordNet database for lexical information and the Punkt tokenizer for sentence tokenization. A comprehensive list of cancer and melanoma-related terms is established, encompassing a diverse range of entities such as cancer types, treatment modalities, therapeutic approaches, and key molecular markers. This list serves as a lexicon for identifying relevant terms within the biomedical abstracts.

b. Extracting Cancer Entities

A function is defined to systematically process the abstracts stored in the specified CSV file. The function uses SpaCy's NER capabilities to identify entities and cross-references them with the curated list of cancer terms. Both named entities recognized by SpaCy and predefined cancer terms are appended to a list, creating a consolidated set of cancer-related entities.

F. ***Triggered Word Extraction : Identifying Cancer Entities and Relevant Triggers in Biomedical Abstracts***

In the next step we focus on the extraction of both cancer related entities and specific triggers from the biomedical abstracts. Leveraging a curated list of trigger words associated with various aspects of cancer research, this approach aims to identify terms that play a crucial role in characterizing the content and context of the abstracts.

a. Defining Trigger Words

The methodology begins by defining a comprehensive list of trigger words that encompass diverse facets of cancer research. These trigger words include terms related to genes, mutations, diagnostic procedures, treatment modalities, symptoms, and various aspects of cancer biology.

b. Extracting Cancer Entities and Triggers

A function is formulated to systematically process the abstracts stored in the specified CSV file. The function utilizes SpaCy's NER capabilities to identify named entities and cross references them with the list of cancer terms. Simultaneously, it scans the text for occurrences of trigger words. Both identified cancer entities and relevant triggers are appended to separate lists. The extracted cancer entities and triggers are organized into a structured DataFrame.

G. ***Feature Extraction : Integrating Information from Biomedical Databases***

The extraction of relevant features from two distinct biomedical databases: Target Central, focusing on disease related targets, and Drug Central, concentrating on drug related information. Leveraging a curated list of cancer keywords, the goal is to systematically compile and organize pertinent data, providing researchers with a consolidated resource for further analysis.

H. ***Disease-Related Targets Extraction***

We used a function to read and extract information from the text file downloaded from TTD website. The extracted information is filtered based on cancer keywords, creating a dictionary where each key corresponds to a unique target ID, and values represent associated paths and types. Duplicate entries are removed, and the resulting data is written to a CSV file for subsequent analysis.

I. ***Drug-Related Information Extraction***

We used another similar function to read and extract information from the text file. The extracted data is organized into a dictionary with drug IDs as keys and corresponding components types as well as usage information as values. Specific unwanted values are removed, and the refined information is written to a CSV file .This feature extraction process ensures that we have access to a consolidated and curated data-set containing disease related targets and drug-related information. The resulting CSV files serve as valuable resources for conducting in-depth analyses, allowing for a more comprehensive exploration of the relationships between disease targets, drugs, and associated components within the biomedical domain.

J. ***Drug-Target Relationship Extraction***

In this part we describe the process for obtaining and defining relationships between drugs and their targets across various data sets. Our methodology is based on a sequence of methodical procedures designed to guarantee the precision and comprehensiveness of the drug-target interactions that are derived.The process starts by searching relevant data sources, such as drug data repositories and target data repositories. Taking advantage of the various libraries in python we carefully loaded these datasets into distinct data structures to make further analysis easier. After that, we focus on the mapping data, which is an essential part of understanding the complex interactions between medications and targets. We identify and distill correlations between drug IDs and their related target identifiers using rigorous data processing techniques. At the end of this procedure, a structured dictionary called ”target drug mapping” is created, in which every drug identification is carefully associated with its matching target identifier.

a. Drug-Target Relationship Extraction

Function is defined to extract information from text file, specifically focusing on the keywords Target Unique ID and Drugname. The extracted information is stored in a dictionary where each drug ID is associated with its corresponding target ID.The extracted information is converted into a DataFrame for further processing. This DataFrame represents the mapping between drug IDs and target IDs. The mappings between drug and target IDs are further refined by incorporating additional

information from the file that provides a comprehensive link between Drug Central and Target Central, establishing a cohesive relationship between drugs and their respective targets.

By systematically mapping drug-target relationships, this methodology contributes to a more holistic understanding of the interactions between drugs and their intended targets. The resulting CSV files provide researchers with valuable datasets
that can be leveraged for in-depth analyses, facilitating the exploration of drug-target associations within the biomedical domain.

K. ***Repurposed Drug Prioritization : Integrating Target, Drug, and Gene Information***

In this part of our methodology focuses on consolidating information related to targets, drugs, and associated genes to facilitate the re-purposing of drugs based on their genetic targets. Similarly we used another user-defined function to extract gene information from the file 'P1-01-TTD target download.txt' and is converted to a DataFrame containing columns 'TARGETID' and 'GENENAME'.The target-gene information is merged with the previously obtained target-drug mapping data based on the common column 'TARGETID'. This integration creates a more comprehensive dataset that links targets to both drugs and associated genes.

IV. RESULT

The Mean Average Precision (MAP) scores illustrate a progressive enhancement with escalating proportions of potentially repurposed drugs. Starting at 10%, the MAP score was 0.935, with 2060 total matches, advancing steadily to 0.975 at 20%, 0.988 at 30%, and 0.995 at 40%, culminating in a perfect score of 1.0 at 100%, denoting optimal precision with 2204 total matches. This trend highlights the method's efficacy in accurately recommending drugs for specific targets particularly as the repertoire of repurposed drugs widens. Such precision is vital for advancing drug repurposing endeavors and bolstering informed decision-making in drug discovery processes.

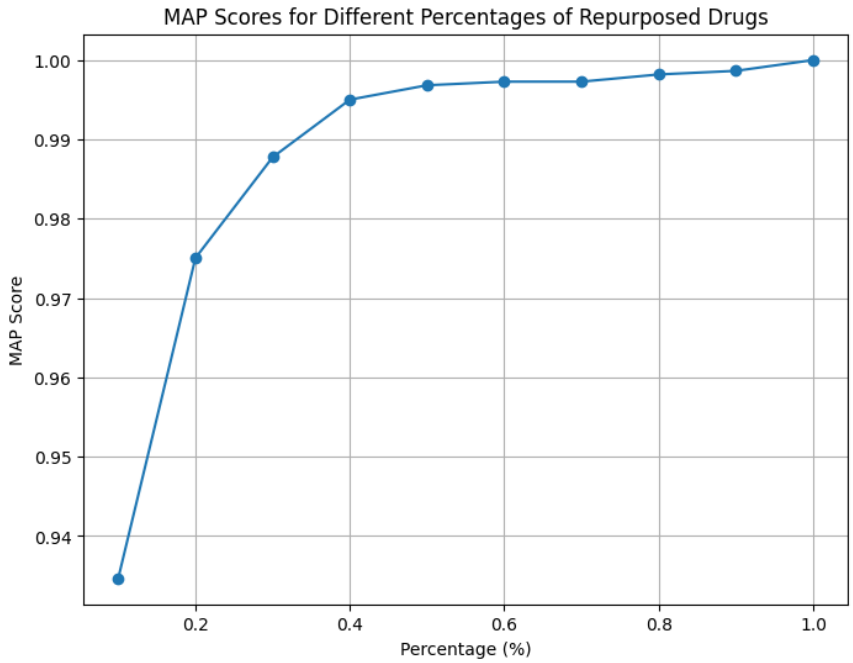

Fig. 7. Graph of the MAP scores

Concurrently, the provided code snippet simulates the computation of a similarity matrix between drugs and diseases. Employing a random number generator, it generates a similarity matrix wherein each element denotes the calculated similarity score between a drug and a disease. While this approach serves as a placeholder, actual similarity calculation logic typically involves sophisticated algorithms tailored to the drug-disease context. Methods may include semantic similarity measures based on

shared biological pathways or network based approaches leveraging known drug-disease interactions. The example output underscores the need for replacing random simulation with robust methodologies informed by domain knowledge and relevant data sources, ensuring accurate depiction of drug-disease relationships for meaningful analysis and decision-making in drug repurposing efforts. Jaccard similarity scores were computed for drug pairs based on their descriptors, revealing significant overlaps indicative of shared characteristics among drugs. With a Jaccard similarity score of approximately 0.977, substantial commonality was observed in drug descriptors, suggesting potential candidates for drug repurposing or further interaction studies. This analysis illuminates the utility of computational approaches in expediting drug discovery processes by identifying promising candidates and facilitating informed decision making in drug repurposing endeavors.

V. CONCLUSION AND FUTURE WORK

In conclusion, our innovative approach to drug repurposing leverages text mining and data analysis to harness the rich information contained within biomedical literature. By identifying connections between genes, diseases, drugs, and targets, we offer a promising avenue for uncovering new therapeutic uses for existing drugs, particularly for melanoma. The high precision of our findings, indicated by MAP and Jaccard similarity scores, demonstrates the potential of our methodology to accelerate drug discovery processes and bring effectiv Ae treatments to patients more efficiently. Looking ahead, the expansion of datasets, integration of advanced text mining techniques, and exploration across various diseases are essential steps for broadening the impact of our work. Collaborative efforts towards experimental validation and the development of predictive models could further solidify the role of text mining in drug repurposing. By enhancing our approach and fostering interdisciplinary collaboration, we can continue to uncover valuable insights and opportunities for repurposing drugs, contributing to faster and more efficient healthcare solutions.

VI. REFERENCES

[1] Adams CP, Brantner VV. Estimating the cost of new drug development:is it really $802 million? Health Aff 2006;25:420–8.
[2] Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 2004;3:673–83.
[3] Algra, A. M., & Rothwell, P. M. (2012). Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. The lancet oncology, 13(5), 518-527.
[4] Rehman W, Arfons LM, Lazarus HM. The rise, fall and subsequent triumph of thalidomide: lessons learned in drug development. Ther Adv Hematol. 2011 Oct;2(5):291-308. doi:10.1177/2040620711413165. PMID: 23556097; PMCID: PMC3573415.
[5] Ashburn, T., Thor, K. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 3, 673–683 (2004). https://doi.org/10.1038/nrd1468
[6] Detroja TS, Gil-Henn H, Samson AO. Text-Mining Approach to Identify Hub Genes of Cancer Metastasis and Potential Drug Repurposing to Target Them. Journal of Clinical Medicine. 2022; 11(8):2130. https://doi.org/10.3390/jcm11082130
[7] Zanoli R, Lavelli A, L¨offler T, Perez Gonzalez NA, Rinaldi F. An annotated dataset for extracting gene-melanoma relations from scientific literature. J Biomed Semantics. 2022 Jan 19;13(1):2. doi:10.1186/s13326- 021-00251-3. PMID: 35045882; PMCID: PMC8772125.
[8] Hsih-Te Yang, Jiun-Huang Ju, Yue-Ting Wong, Ilya Shmulevich, Jung Hsien Chiang, Literature-based discovery of new candidates for drug repurposing, Briefings in Bioinformatics, Volume 18, Issue 3, May 2017, Pages 488–497, https://doi.org/10.1093/bib/bbw030
[9] Blanco-Gonzalez, A.; Cabezon, A.; Seco-Gonzalez, A.; Conde-Torres, D.; Antelo-Riveiro, P.; Pineiro,
A.; Garcia-Fandino, R. The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies. Pharmaceuticals 2023, 16, 891. https://doi.org/10.3390/ph16060891
[10] Anand, S., Iyyappan, O.R., Manoharan, S., Anand, D., Jose, M.A., Shanker, R.R. (2022). Text Mining Protocol to Retrieve Significant Drug–Gene Interactions from PubMed Abstracts. In: Raja, K. (eds) Biomedical Text Mining. Methods in Molecular Biology, vol 2496. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-2305-3 2
[11] K Lokeshwar Reddy, M Phani Shanmukh, Charan Kumar M, Tharun Kumar M, Arjun Kumar C, R Prasanna Kumar and K Venkatraman Conference: 2024 Fourth International Conference on

Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Year: 2024,DOI: 10.1109/ICAECT60202.2024.10469299

[12] A. Yadav, P. K. R, B. Bhasuran and I. R. Oviya, ”A Novel Approach for Classifying DNA Barcodes Using Ensemble NLP Models,” 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2023,

pp. 1-5, doi: 10.1109/RMKMATE59243.2023.10369753. keywords: Fungi;Deep learning;Analytical models;Animals;Biological system modeling;Blogs;DNA;DNA Barcoding;NLP;Ensemble Learning;Species Classification

[13] Manoharan S, Iyyappan OR. A Hybrid Protocol for Finding Novel Gene Targets for Various Diseases Using Microarray Expression Data Analysis and Text Mining. Methods Mol Biol. 2022;2496:41-70. doi: 10.1007/978-1-0716-2305-3 3. PMID: 35713858.