5-29-2021

Bibliometric Analysis of Machine Learning and Text Mining Algorithms for Diagnosis of Leukemia

Sonali Kothari
Symbiosis Institute of Technology, sonali.tidke@sitpune.edu.in

Rohanraje Bhosale
Symbiosis Institute of Technology, bhosale.rohanraje.btech2018@sitpune.edu.in

Sanket Gadakh
Symbiosis Institute of Technolgoy, sanket.gadakh.btech2018@sitpune.edu.in

Nikhil Sawant
Symbiosis Institute of Technology, nikhil.sawant.btech2018@sitpune.edu.in

Vijayshri Khedkar
Symbiosis institute of Technology

Bibliometric Analysis of Machine Learning and Text Mining Algorithms for Diagnosis of Leukemia

Rohanraje Bhosale[1], Sanket Gadakh[1], Nikhil Sawant[1], Dr. Sonali Kothari Tidke[1*], Vijayshri Khedkar[1]

Symbiosis Institute of Technology, Pune, Maharashtra, India[1]
sonali.tidke@sitpune.edu.in

Abstract
Bibliometric analysis of leukemia area was carried out using information about publications from Scopus. The most productive journals, countries and authors were determined. The most frequently cited article and its citation history was described. A bibliometric map based on a citation network among countries was constructed.

Keywords
Bibliometric analysis; leukemia; text mining; NLP.

I. Introduction

For scientific researchers determining of research trends and progress of the area of their interest is very important. The field of science is concerned with the classification of information. Bibliometric study, the extraction of statistics on written scientific papers and the research areas they discuss, includes some of the tools used in library science. Different disciplines were analysed using bibliometric methods. [11]

The statistical study of bibliometric characteristics and data for - instance research, citations, and publications outputs is known as bibliometrics. All researchers may use bibliometrics to learn about the patterns, structure, and properties of research activities.
The research activities are integrated into a research domain through the review process. The analytic method includes analyses of scientific activity in different conditions, for instance citations, organisations, writers, publications. [11]
Leukemia is a category of blood cancers that start in the bone marrow and cause an excess of irregular blood cells. Blasts, also known as leukaemia cells, are cells that have not yet fully developed. Bleeding and injury, exhaustion, fever, and an elevated risk of infection are all possible symptoms. These symptoms are caused by a lack of regular blood cells. [12]

A blood test or a marrow biopsy are used to make the diagnosis. Leukaemia's specific causes are unclear. A mixture of hereditary and non-inherited factors is thought to be involved. Leukemia is mainly of four types and mentioned as followed, chronic myeloid leukaemia, acute myeloid leukemia, chronic lymphocytic leukaemia and acute lymphoblastic leukaemia. [12]

There are several forms of leukaemia. Acute leukaemia can grow quickly in some cases, while other chronic forms of the disease may require less vigorous therapy. The first step in developing a leukaemia care plan is to get a full and correct cancer diagnosis. Imaging and laboratory experiments are used to monitor treatment response and adjust the treatment strategy as appropriate. A biopsy is performed to determine the type of leukaemia, the tumour's growth rate, and whether the disease has spread. The following are some of the most common leukaemia biopsy procedures: A sample of bone marrow is taken during a bone marrow biopsy. The removal of all or part of a lymph node is needed for a lymph node biopsy. Imaging tests: These techniques will reveal the amount of leukaemia present in the body, as well as the existence of infections or other issues. The following imaging studies may be used to aid in the diagnosis of leukaemia: CT scan with X-rays, CT/PET scan, Ultrasound MRI, a two-dimensional echocardiogram Test of pulmonary function Lab tests: The number of RBC, WBC, and platelets in the blood is determined by these tests are often smaller than average as leukaemia develops. This examination, also known as a spinal tap, may be essential to assess the degree of leukaemia. Lumbar punctures may also be used to administer medicines to cure the condition, such as chemotherapy drugs. [12]

The term Text mining refers to approach in Artificial Intelligence (AI) that mainly focuses on using natural language processing (NLP) to transform unstructured text in database into structured and well-formed data that can be analysed or used to drive machine learning (ML) algorithms. Text mining is a method of analysing large collections of documents to find new information or help address specific research questions. It is often used in knowledge-driven organisations. Text mining discovers facts, links, and assertions that would otherwise be hidden in structured bigdata. The data is then translated into a standard format that can be analysed or interpreted in a number of ways. It includes mind maps, charts, clustered HTML tables and other visual aids. To process the text, text mining employs a number of methodologies, one of the most well-known of which is Natural Language Processing (NLP). [13]

Natural Language Recognition (NLR) simulates the human ability to learn a natural language such as English, French, or Russian, allowing machines to “read” input text/speech.
NLP encompasses both Understanding and Generation, which allows a person's ability to produce natural language text, such as to review information or participate in a conversation. [13]

II. **Research Method**

Eligibility Criteria:

Scopus was used to perform a systematic search for English- language peer- reviewed publications. We discovered Leukaemia-related search terms. The following is a list of the keywords that were used:
Scopus (all years, journal articles): TITLE-ABS-KEY (“leukaemia*” OR “nlp*” OR “textmining*” OR “text AND analysis*” OR “named AND entity AND recognition*” OR “lingpipe*” OR “opennlp*”) AND (LIMIT-TO (DOCTYPE , "ar")).

Keywords that resulted in search results that were well outside the reach of the search were refused. For example, the terms "disease" and "illness" may be used in a search, but they're more often used in unrelated ways, adding to the search's noise. Such words were excluded because they contributed to a large number of insignificant findings. We included papers written since Scopus' inception.

Articles that did not specifically concentrate on leukaemia or any forms of leukaemia, as well as reviews and meta-analyses, were excluded. Before the preliminary title and abstract screening, inclusion and exclusion criteria were established. The eligibility requirements were purposefully vague in order to obtain a comprehensive image of current applied research. We performed an initial pilot screening of 100 papers to improve our trust in the inclusion criteria.

Consecutive Clustering and Trend Report:

The number of papers published each year, and thus the number of authors per article, were analysed chronologically. We grouped papers by the number of publications per year, author name, subject region, country, funding sponsor, and other factors. The number of citations and the publishing journal of the included articles was identified.

Text Analysis:

We removed terms with high frequencies that were popular in research papers but weren't unique to our topic (e.g., "paper," "using," and "results") after analysing all of the titles and abstracts. We have combined words like "identified entity recognition" and "NER" to form "named entity recognition".

Using VOS viewer, we analysed text titles and abstracts to create a high-level idea map composed of specifics and their connections. The programme began by using an unsupervised machine learning method to extract a network of meaning from the data, then created a heat map to visualise the top results. “Themes” are expressed by bubbles, and “concepts” are represented by lines. The inventory of comparable words that coalesce into a monothematic idea is also equated to concepts, and themes are clusters of such concepts. [3]

The lines between the dots imply a strong connection between the two definitions. For a better understanding, some of the graphs were created using data from Scopus and Visualized. It offers the audience a visual representation of the study and makes it easier to comprehend.

III. Literature Review

Table 1. Findings of some reference papers

SR. NO	TITLE	AUTHORS	YEAR	JOURNAL	FINDINGS
1	Automated Text Mining and Ranked List Algorithms for Drug Discovery in Acute Myeloid Leukemia [1]	Tran, Damian	2019	Cell Discovery 7(1),2	By T2F system which is connected to a series of ranked list consolidation algorithms, a concise group of candidate genes were identified for further investigation in the acute myeloid leukemia context. The relatively condensed list of candidates, stratified by their ranks on a larger consolidated list, was manageable enough to be more thoroughly assessed using bioinformatics tools. The power of automated

					literature searches enabled rapid and concise therapeutics discovery on a scale not previously reported. These tools offer transparency in the way they operate.
2	miRCancer: a microRNA–cancer association database constructed by text mining on literature [2]	Boya Xie , Qin Ding, Hongjin Han and Di Wu	2013	Bioinformatics , Volume 29, Issue 5	Using an automated extraction process by rule matching and text mining on numerous articles & tiles, Mir Cancer will provide detailed collections of miRNA in Human Cancers. and this can be used for targeting any specific genes of microRNA.
3	Gene expression–based classification and regulatory networks of paediatric acute lymphoblastic leukemia [3]	Zhigang Li, Wei Zhang, Minyuan Wu, Shanshan Zhu, Chao Gao, Lin Sun, Ruidong Zhang, Nan Qiao, Huiling Xue, Yamei Hu, Shilai Bao, Huyong Zheng, Jing-Dong J. Han	2021	Blood (2009) 114 (20)	Analysis of first developed gene interpretation according to every subtypes and mining an outsized summary of examples trying another marker strategy, which commenced to a huge efficient analysis which can be implemented to a unique specimen.
4	LeukmiR: a database for miRNAs and their targets in acute lymphoblastic leukemia[4]	Abdul Rawoof, Guruprasadh Swaminathan, Shrish Tiwari, Rekha A Nair, Lekha Dinesh Kumar	2020	Database, Volume 2020, 2020, baz151	LeukmiR is curated to target just the miRNA which acts as a bridge among the databases available publicly by giving our microRNAs and their target data only for acute lymphoblastic leukaemia.
5	Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and	Ayush Singhal, Michael Simmons, Zhiyong Lu	2016	PLoS Computational Biology, November 2016	For creating a database, the non-curated triplets obtained by applying text mining are great applicants and by using the same text mining related genes with 80% and more

	Precision Medicine [5]				efficiency which gives constant best results.
6	A text-mining system for knowledge discovery from biomedical documents[6]	N. Uramoto H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, K. Takeda	2004	IBM Systems Journal (Volume: 43, Issue: 3, 2004)	MedTAKMI system which uses a specialized method of deriving knowledge discovery from thousands and lakhs of biomedical documentations using text-mining. with the ability to parse from over 11 million MEDLINE citations. this also will provide a toolkit with interactive viewing options to held in the discovery of any underlying knowledge.
7	Harder to treat than leukemia-opioid use disorder in survivors of cancer[7]	Loren, A.W.	2018	New England Journal of Medicine 379(26), pp. 2485-2487	The results show the purpose of individual medical options in differentiating the uncertainty of opioid abuse in blood cancer from the non-blood cancer community preferably in individual diagnoses. More attempts are required to assure a continued supply of opioid distribution. Instructions on the administration of opioids to disease-free survivors and incurable cancer stages, particularly those associated with the treatment of new metastatic diseases.
8	Rare double-hit with two translocations involving IGH both, with BCL2 and BCL3, in a monoclonal B-cell lymphoma/ leukemia[8]	Kayabasi, C., Okcanoglu, T.B., Yelken, B.O., (...), Saydam, G., Gunduz, C.	2017	Gene 637, pp. 173-180	It is found out that, genomic instability is backed by large translocation junctions and microduplication events in the transmission analysis and cytogenetically defined resistance pathways which include modified regression.

9	Automatic characterization of leukemic cells with 2D light scattering static cytometry[9]	Wang, L., Liu, Q., Xie, L., Shao, C., Su, X.	2017	Proceedings - 2017 Chinese Automation Congress, CAC 2017-January, pp. 5925-5928	The Analysis of HL-60 & standard granulocytes are achieved by using machine learning algorithms. which could be then used as an automatic way to analyse leukemia.
10	Effect of probiotics on diarrhoea secondary to chemotherapy for leukemia[10]	Zhou, X.-F., He, Y., Wang, S.-J., Wang, J.-M., Hong, W.-Y.	2017	World Chinese Journal of Digestology 25(36), pp. 3248-3252	The occurrence of CID among cancer patients could be prevented by using the right probiotics during or before chemotherapy. which also shows an improvement in therapeutic effects on CID.

IV. Results and Discussion

Search Results:

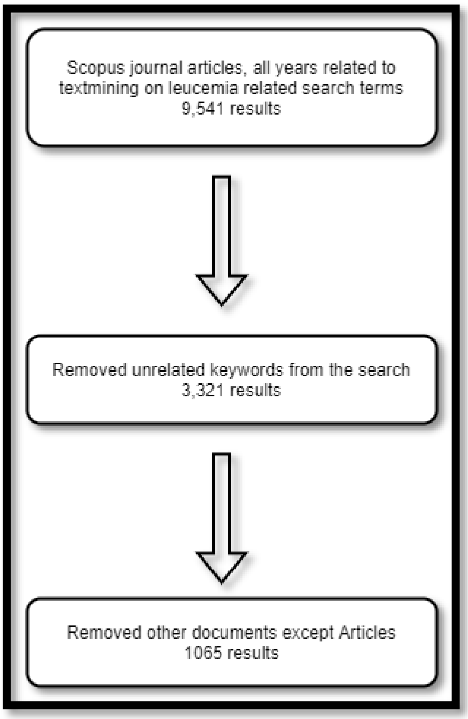
Scopus journal articles, all years related to textmining on leucemia related search terms 9,541 results
Removed unrelated keywords from the search 3,321 results
Removed other documents except Articles 1065 results

Figure 1: Search method and results

The main search on Scopus for papers mentioning about the terms related to "Leukaemia" gave 9,541 articles. Based on the filtered criteria, 3,321 articles were included in the first screening for full-text review. Figure 1 is about Search method and results. We excluded the papers which includes unrelated keywords from the search. We also removed the books and other stuffs from it and just kept articles. So, at the end we were left with 1065 articles.

Consecutive Clustering and Trend Report:

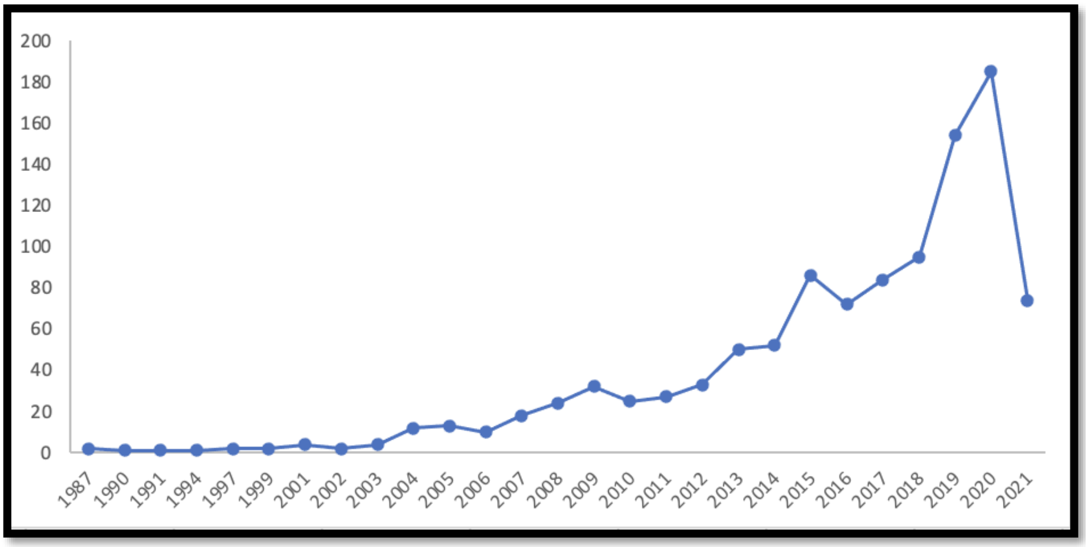

Figure 2: Cumulative number of publications by year

Figure 2 represents the trend for all publications that happened over the time from 1987 to 2021; the very first featured article was officially published in 1987. Figure 2 reveals a constant rise in the number of reports published about text mining on Leukaemia on Scopus.

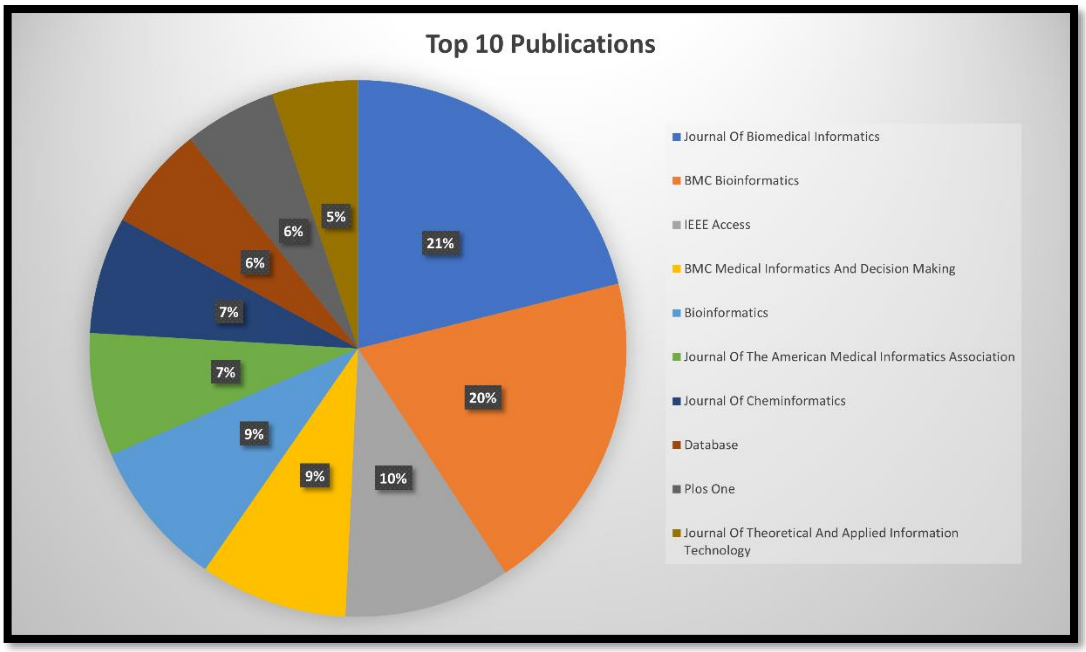

Figure 3: Top 10 Publications

Figure 3 shows the top 10 publications which are published on Scopus. Top 5 positions are taken by Journal of Biomedical Informatics (25%) followed by BMC Bioinformatics (20%) followed by IEEE Access (10%) followed by BMC Medical Informatics and Decision Making (9%) followed by Bioinformatics (9%).

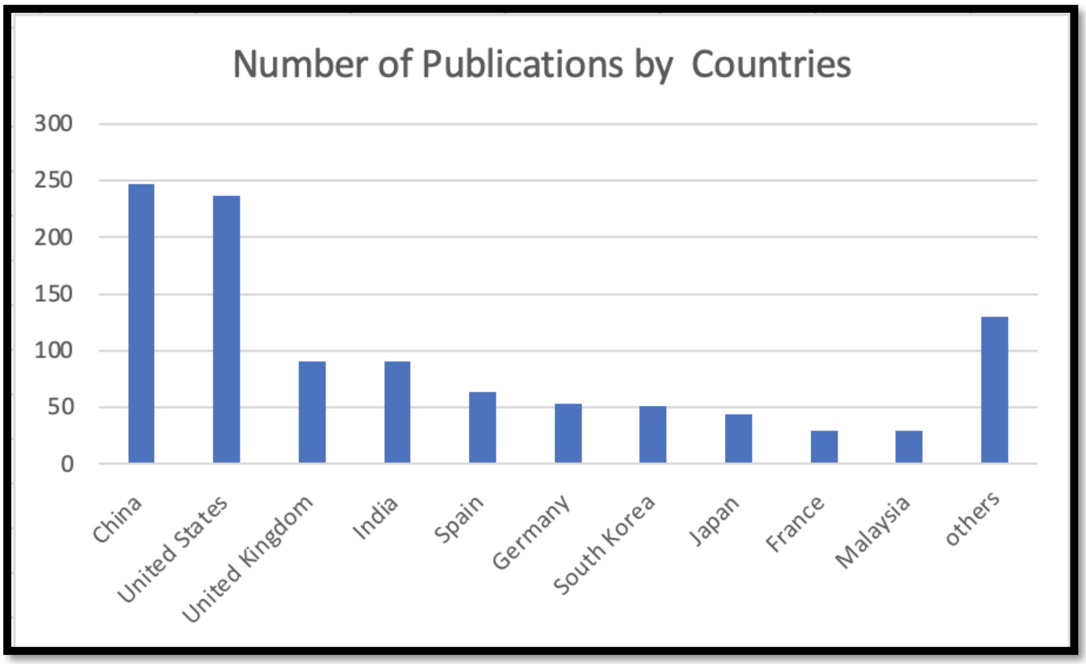

Figure 4: Number of Publications by Countries

Figure 4 shows the number of publications made by particular counties. Top 5 countries to publish paper on Leukaemia on Scopus are China, United States, United Kingdom, India, and Spain. China has published over 247 papers, United States has published over 237 papers, United Kingdom has published over 91 papers, India has published over 90 papers and Spain has published over 64 papers.

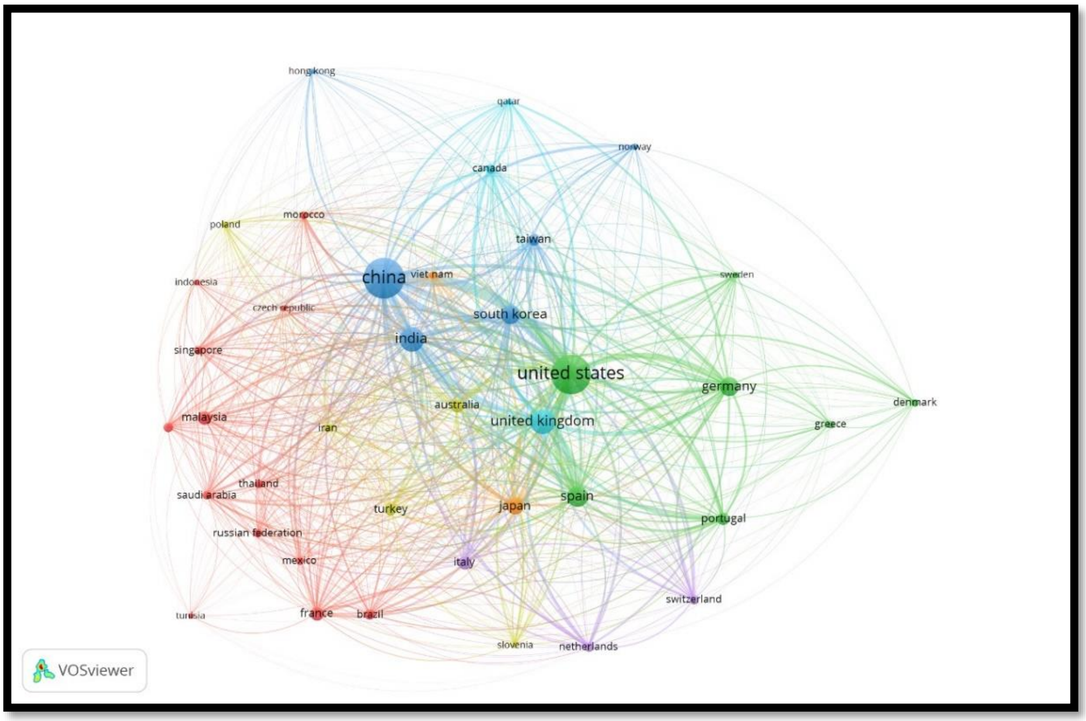

Figure 5: Bibliometric citation network map based on countries

Figure 5 shows the Bibliometric map based on citation network. Figure 5 is used for determining the Leukaemia research distribution in the world created bibliometric map using Vos viewer mapping software from citations network. It shows which countries have done more research about Leukaemia. Table 21 shows the information of the Top 10 Author and affiliation of him/her who have published in Scopus. It also shows how many Publication are done by respective author.

Table 2: Top 10 Most Productive Authors

No	Author	Affiliation	Number of papers
1	Xu, H.	Key Laboratory of Advanced Process Control for Light Industry	25
2	Ananiadou, S.	Istituto di Scienza e Tecnologie dell'Informazione	16
3	Lu, Z.	Istituto di Scienza e Tecnologie dell'Informazione	16
4	Collier, N.	Laboratorio de Innovación en Humanidades Digitales	11
5	Leaman, R.	Institute of Marine Biology, Biotechnology and Aquaculture	11
6	Wei, C.H.	National Center for Biotechnology Information	10
7	Wu, Y.	Jordan University of Science and Technology, Irbid, Jordan	10
8	Lin, H.	Division of Haematology and Medical Oncology	9
9	Tang, B.	Departament Ciències Mèdiques Bàsiques	9
10	Ekbal, A.	Department of Computer Science and Engineering, SSN College of Engineering, Chennai	8

V. Conclusion

In medical sciences these data research is citied as prominent and important information extracted from thousands of articles and published papers a visual representation of majority of citied information in the world created bibliometric map using Vos viewer mapping software from citations network. It creates a better view of idea for implementation.

References

1] Tran & Damian, “Automated Text Mining and Ranked List Algorithms for Drug Discovery in Acute Myeloid Leukemia”, Cell Discovery, 2019, Volume 7.

2] Boya Xie , Qin Ding , Hongjin Han and Di Wu, “miRCancer: a microRNA–cancer association database constructed by text mining on literature”, Bioinformatics, Volume 29.

3] Zhigang Li, Wei Zhang, Minyuan Wu, Shanshan Zhu, Chao Gao, Lin Sun, Ruidong Zhang, Nan Qiao, Huiling Xue, Yamei Hu, Shilai Bao, Huyong Zheng and Jing-Dong J. Han , “Gene expression–based classification and regulatory networks of paediatric acute lymphoblastic leukemia ”, Blood, 2009, Volume 114.

4] Abdul Rawoof, Guruprasadh Swaminathan, Shrish Tiwari, Rekha A Nair, Lekha Dinesh Kumar, “LeukmiR: a database for miRNAs and their targets in acute lymphoblastic leukemia” Database, Volume 151.

5] Ayush Singhal, Michael Simmons & Zhiyong Lu , “Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine” , PLoS Computational Biology, 2016.

6] N. Uramoto H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi & K. Takeda, “A text-mining system for knowledge discovery from biomedical documents” IBM Systems Journal, 2004 , Volume 43.

7] Loren, A.W., “Harder to treat than leukemia-opioid use disorder in survivors of cancer”, New England Journal of Medicine, 2018, Volume 379.

8] Kayabasi, C., Okcanoglu, T.B., Yelken, B.O, Saydam, G & Gunduz, C., “Rare double-hit with two translocations involving IGH both, with BCL2 and BCL3, in a monoclonal B-cell lymphoma/ leukemia”, Gene, 2017, Volume 637.

9] Wang, L., Liu, Q., Xie, L., Shao, C & Su, X. “Automatic characterization of leukemic cells with 2D light scattering static cytometry” , Proceedings - 2017 Chinese Automation Congress, CAC 2017 , 2017.

10] Zhou, X.-F., He, Y., Wang, S.-J., Wang, J.-M., Hong & W.Y. , “Effect of probiotics on diarrhoea secondary to chemotherapy for leukemia”, World Chinese Journal of Digestology , 2017, Volume 25.

11] https://en.wikipedia.org/wiki/Leukemia

12] https://www.cancercenter.com/cancer-types/leukemia/diagnosis-and-detection

13] Abhinav Rai, “What is Text Mining: Techniques and Applications”, Upgrad Blog, June 2019, www.upgrad.com.