INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

PubMed Text Data Mining Automation for Biological Validation on Lists of Genes and Pathways

Hui Wen Nies, Zalmiyah Zakaria, Weng Howe Chan, Izyan & Izzati Kamsani
School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia
Email: huiwennies@utm.my

Nor Shahida Hasan
Malaysia-Japan International Institute of Technology (Mjiit)
Universiti Teknologi Malaysia, Kuala Lumpur

Submitted: 13/4/2021. Revised edition: 25/10/2021. Accepted: 21/11/2021. Published online: 16/5/2022
DOI: https://doi.org/10.11113/ijic.v12n1.313

***Abstract*—A prognostic cancer marker is helpful in oncology to identify the abnormal cancer cells from the collected sample. This marker can be used as an indicator to determine a disease outcome, cancer treatment, and drug discovery. Identifying cancer markers is also beneficial to improve cancer patients' survival rate in receiving the treatment decision-making. Cancer markers can be determined by manually testing every gene or pathway in the wet lab or using the text mining automation method. The use of text mining techniques effectively investigates hidden information and gathers new knowledge from many existing sources. Unfortunately, querying relevant text to excavate important information is a challenging task. PubMed text data mining is one of the applications that help explore potential cancer markers as the trend of scientific articles in PubMed is steadily increased. Besides, it can support biologists to concentrate on the identified small set of genes or pathways. PubMed identifiers (PMIDs) are then obtained as evidence to ascertain the relationship between diseases and genes (or pathways) used as biological validation. Thus, this technique can discover the biological relationship between disease and genes or pathways. The existing method is commonly manually curated for the biological validation of genes and pathways. Manual curation takes time in the process and may lead to inconsistency. This study aims to automate the process of biological validation of genes and pathways for PubMed text data mining. Therefore, the PubMed text data mining automation was invented to link to the websites for saving time instead of manually. A list of genes and pathways from breast cancer are used in this study. Using PubMed text data mining automation for biological context verification and validation, p53 signaling pathway and TP53 gene as prognostic cancer markers for breast cancer. Hence, the p53 signaling pathway and TP53 are associated with the development of tumour cells and DNA damage after irradiation in breast cancer.**

***Keywords*—PubMed, text data mining, biological validation, cancer markers, diseases, genes, pathways**

I. INTRODUCTION

The list of genes and pathways identified through an experiment can extract some confidential information on these genes and pathways. The list can either be tested individually in the wet lab to check it as a potential cancer marker or link the existing online sources using computational approaches [1,2]. Text data mining is time-saving and cost-saving, where wet lab analysis needs more time to get the result using devices. Also, all the reagents used in the study are costly. The aim of text data mining is to concern the discovery of new and previously unknown information. This technique is also helped to minimize human input errors which might occur during the manual search.

Bipolar disorder (BP)-Gene literature data analysis is one of the biological validation of genes [3]. This analysis integrated with ResNet data analysis that employed an automated natural language processing-based information extraction system. Scientific articles support it as following the BP_GB→Related Genes and BP_GB→References for Disease-Gene Relation. Functional Annotation tool from the Database for Annotation, Visualization, and Integrated Discovery (DAVID) annotate the identified cancer markers based on the Gene Ontology and KEGG databases [4-6]. They compared the C-index, Fisher test, and p-values of cancer markers reported in the literature. Other than that, Google Scholar literature search has been used for biological data validation [7]. It was performed with a combination of pathway name and experimental condition information. Once a literature association was confirmed, the reference was cited as support.

PubMed is a popular source with biomedical records in the life sciences that provide review facts to disease biology [8,9]. PubMed abstracts are interconnected with the National Center for Biotechnology Information's (NCBI) Entrez Cross-

Database related to the disease's DNA sequence and chemical structure. PubMed text data mining also helps extract information and filter the related keywords to associate the biological conditions. This technique can improve the annotation of a target protein or gene list, extract protein-protein interactions, and predict gene-gene relationships. In this study, PubMed text data mining is applied to automate biological context verification and validation based on the list of genes and pathways. This automation aims to show the relationship between cancer and genes or pathways in order to get a list of cancer markers without manual curation.

This paper is organized following the biological validation related topics as follows: the next section describes the text data mining adopted to mine scientific literature; in Section III, we explain PubMed text data mining; Section IV focuses on experimental design with PubMed text data mining automation; then following with performance measurement; and finally, Section VI presents results that include data collection and frequency of the detected pathways and genes.

II. TEXT DATA MINING

Text data mining usually employs natural language processing (NLP). Text data mining consists of four parts to extract the relationship from the data. The parts consist of information retrieval, named entity recognition, information extraction, and knowledge discovery [1, 8]. Fig.1 presents the flowchart of text data mining.

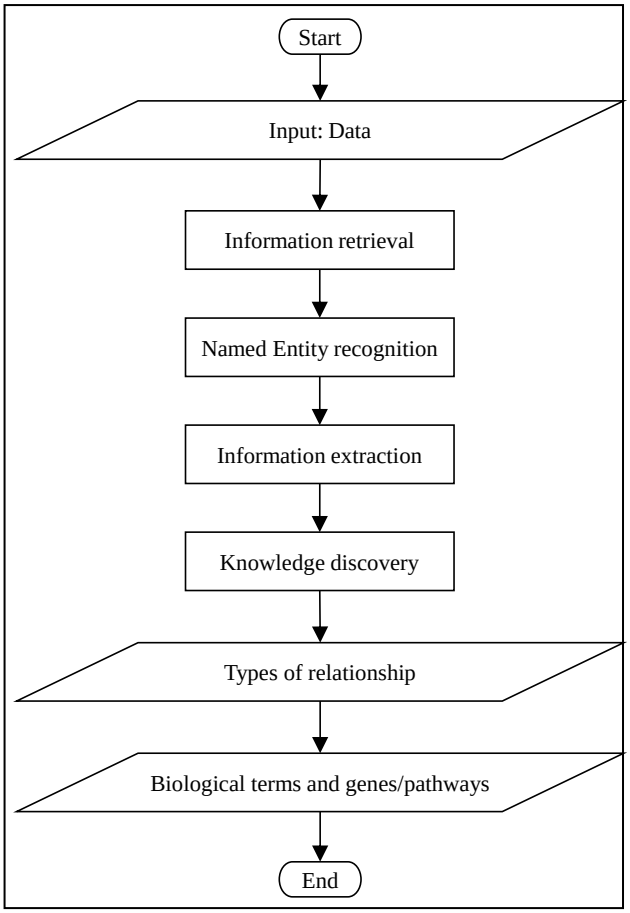

Fig. 1. The flowchart of text data mining

The input is essential for information retrieval because information retrieval will gather all the relevant papers based on interest. For example, the text from articles, abstracts, and others, is available from the user's query.

The second part is named entity recognition. It aimed to identify terms within the gathered text. The example of the collected text can be the biological terms, like gene name (e.g., ERBB4), pathway name (e.g., PI3K/AKT signaling pathway), or disease name (e.g., triple-negative breast cancer). The variations in the entity names that appear can be recognized by the machine learning algorithms like ERBB4, erbb4, and Erbb4.

The next part is information retrieval, which extracts the relationship between the biological terms of the gathered text. This technique can limit and extract the occurrence for the specific types of relationships. For example, the relationship between A and B is A connects to B because A activates B in disease C for organ D.

The last part is knowledge discovery. It is used to generate scientific hypotheses and attempt to discover the biological meaning of facts. For example, knowledge discovery can explore the relationship between biological terms through assumptions generation instead of exploring many resources. This example can be derived from identifying all genes associated with prostate cancer. It can then be seen that gene KLK3 is bound to prostate cancer and, more specifically, the prostate's malignant neoplasm.

DAVID is an integrated data mining environment to extract biological features associated with significant gene lists [4-6]. Mining biomarkers using DAVID can provide biological insights from GO (Gene Ontology) functional enrichment analysis and GSEA (Gene set enrichment analysis). Another biological data validation is using Google Scholar literature search to consider the accuracy of pathway associations with experimental conditions [7]. It combined the pathway name and details of the experimental state to search through Google Scholar. It continued searching until satisfying the association was confirmed or felt reasonably sure that there was not yet literature association confirmed. Once it was established, the most pertinent reference was cited as proof.

III. PUBMED TEXT DATA MINING

PubMed is considered the optimal database for medical and biomedical engineering research [10]. Besides, PubMed searches the MEDLINE database and produces a comprehensive search of articles on abstracts [1]. This text data mining can identify the annotated biomedical terms. For example, the genes' list will be identified with the gene names or Entrez IDs, then verified with controlled vocabularies like disease names. The relationship between the biomedical concepts can be found, such as the gene-gene relationship or gene-disease relationship. Hypothesis generation can occur with support and validation through experimental data. Identifying genes hypotheses with the gene expression data can discover the new relationships between genes. For example, Gene Entrez ID 11260 points to gene Entrez ID 5901 in the directed graph, which can be derived from XPOT (gene Entrez

ID: 11260) inhibiting RAN (gene Entrez ID: 5901) in the RNA transport (KEGG pathway ID: hsa03013) [6].

The rapid accumulation of Covid-19 literature requires tools for data collection that can be implemented by text data mining tools [11]. Around 40,300 Covid-19 related articles were listed in PubMed. The mining of available data and past analysis can provide for currently approved therapeutics and treatment of conditions caused by SARS-CoV-2 infection to use in the treatment of Covid-19. The selection of informative research articles is manual, followed by updating and adding new records to a database that is automated using Perl scripts. Hyperlinks to PubMed are generated automatically by Python scripts. Most text data mining tools did not extract biomedical terms efficiently [12]. Besides that, STRING is a database that collects protein-protein interactions and relies on predictions using automated text mining. However, it used a statistical approach based on OMIM (Online Mendelian Inheritance in Man) and PubMed abstracts. Cancer-related information can extract information for specific biological networks for each type of cancer using text data mining. Investigating cancer markers obtained from the literature consists of a massive amount of confidential information in scientific articles [1]. The number of PubMed articles is steadily increasing. Using text data mining to gather knowledge from existing scientific sources can help investigate the literature for cancer markers to discover the relationship between genes and disease. Also, the PubMed API provided extra information about the articles like keywords, title, abstract, authors, authors' affiliation, publishing date, and journal name. PubMed text data mining is also used in IBM Watson to extract interaction networks of biological entities [13]. There was a rule-based approach for learning syntactic relationships to connect entities through verbs and trigger phrases such as inhibiting and regulating negative.

IV. Experimental Design

Fig. 2 shows the concept of PubMed text data mining. This technique was used to illustrate the relationship between pathways, genes, and cancers. The PubMed text data mining website used in this paper was related to the NCBI website [http://www.ncbi.nlm.nih.gov/pubmed?LinkName=gene_pubmed&from_uid=2066] (accessed on 23 July 2021) [6]. Fig. 3 and Fig. 4 show the flowcharts of PubMed text data mining automation based on the list of genes and pathways. The list of genes and pathways is provided in the number format (Entrez Identifiers or KEGG Identifiers). The genes and pathways are in turn one by one to extract the information about its name from the database (NCBI for gene; KEGG for pathway). Once get the terms, all genes and pathways are then matched to the PubMed database with the keywords.

"Pathway name", "gene name", "prognostic", and "cancer types" are the main keyword terms to be extracted as the concept. This concept is employed to show the pathways and genes exhibiting biological characteristics related to cancers. The keyword for the type of cancer can be specified with breast cancer or luminal A of breast cancer. Hence, a prognostic marker can identify the disease outcome, assisting cancer treatment and drug discovery. Besides, the disease-related text data in the PubMed database has been optimized during the process. The technique ignores the text data that are not related to genes and diseases. PubMed identifiers (PMIDs) are then obtained as evidence to ascertain the relationship between pathways, genes, and diseases [14-15].

Consequently, this technique is automatically linked to the websites for saving time instead of manually. If the genes or pathways do not match the keywords with the PubMed database, the process will look for the following genes or pathways. The entire process is repeated until all genes and pathways are done verified and validated.

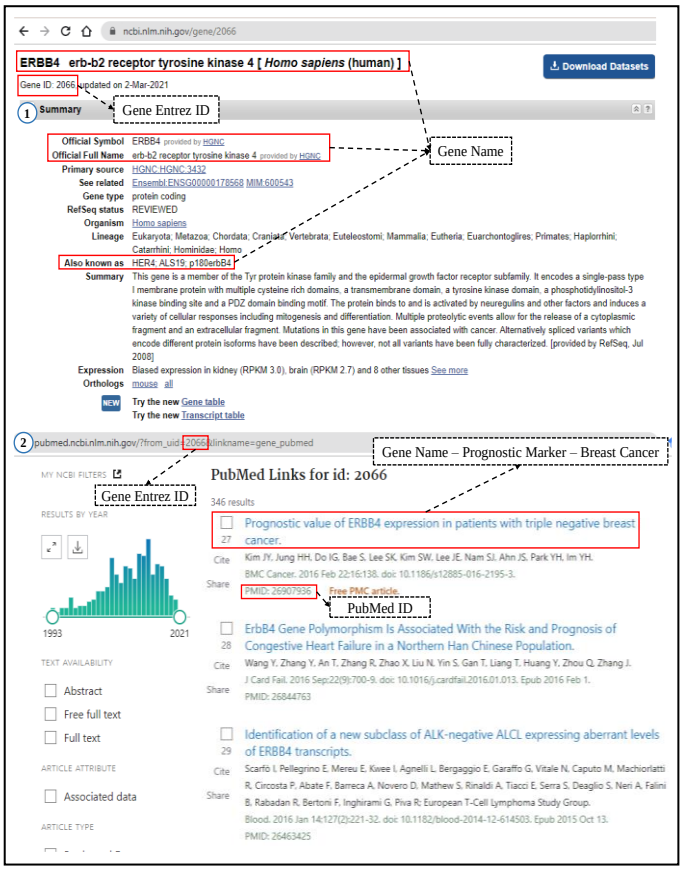

Fig. 2. The concept of PubMed text data mining

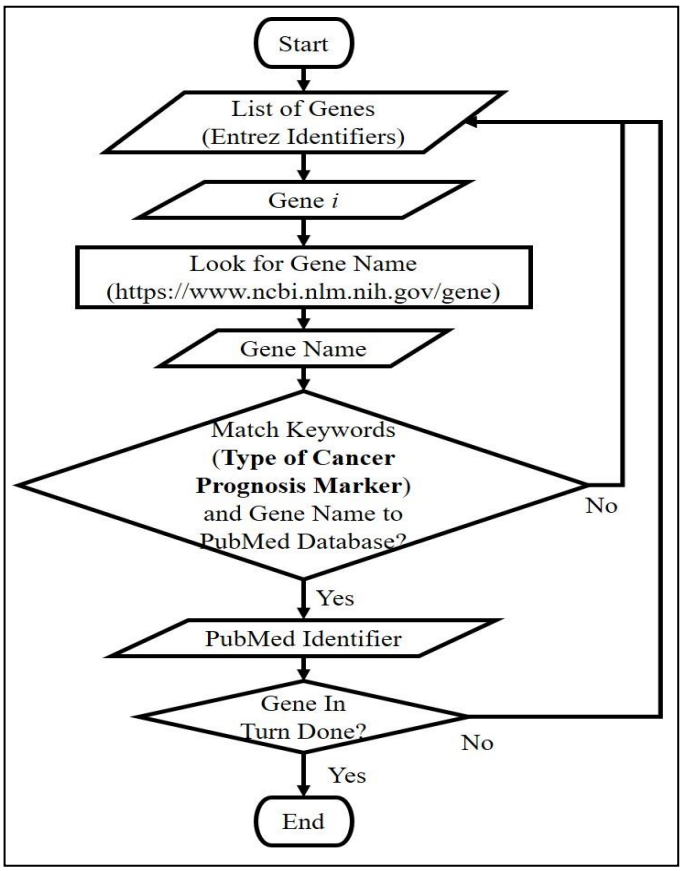

Fig. 3. PubMed text data mining automation based on genes

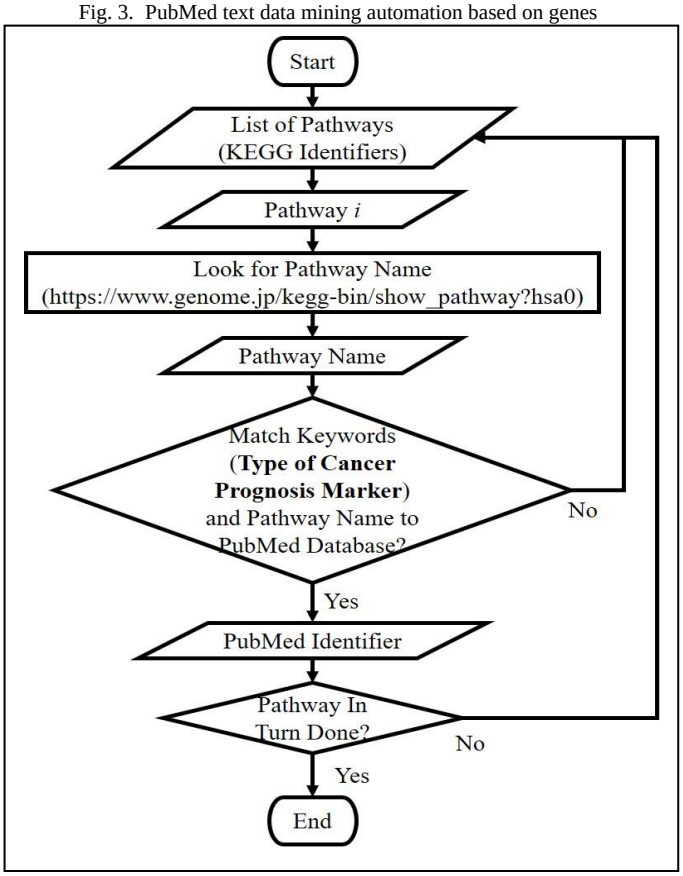

Fig. 4. PubMed text data mining automation based on pathways

V. Performance Measurements

The identified pathways and genes are then analyzed using PubMed text data mining as potential prognostic cancer markers. This technique is used to show the biological relationship between pathways, genes, and cancers. The identified pathways and genes are validated based on the literature published in the PubMed and OMIM databases. PubMed text data mining can mine the data automated to systematically queries with different keywords. With the DAVID tool, the OMIM database is an online catalogue of human genes and genetic disorders being updated daily. The number of cancer pathway markers and cancer gene markers are the main output to be evaluated between PubMed text data mining and the DAVID tool. The keyword of cancer marker is specific cancer type (e.g., breast cancer, breast adenocarcinoma, and breast carcinoma), prognostic marker, pathway names, and gene names.

VI. Results and Discussion

A. Data Collection

The experiment used the gene expression data of breast cancer (GSE1456 and GSE1561), 300 pathways (metabolic and non-metabolic pathways), and a directed graph. GSE1456 is the dataset collected from all breast cancer patients who received surgery at Karolinska Hospital between 1994 and 1996 [16]. For GSE1561, there is a phase III clinical trial dataset, but clinical response data is not yet available. Both datasets consist of 12437 genes [17]. An enhanced Directed Random Walk method (eDRW+) identifies a list of information pathways and genes based on the studied datasets [6]. This method used pathway topology and gene expression to infer a greater reproducibility power of pathway activity. The list of genes and pathways are defined as the identified informative genes and pathways.

B. Biological Validation of Genes and Pathways (Frequency)

Table 1 and Table 2 show the biological validation of the identified genes and pathways using PubMed text data mining and the DAVID tool. In the tables, the number of identified informative pathways (or genes) is the number of pathways (or genes) determined by the computational method (eDRW+ [6]). In contrast, the number of cancer pathway (or gene) markers is the number of identified pathway (or gene) markers validated by PubMed text data mining and DAVID tool associated with breast cancer. Hence, 953 informative genes were identified within 52 informative pathways for the GSE1456 dataset and 536 informative genes within 24 informative pathways for the GSE1561 dataset. PubMed text data mining has validated a more significant number of cancer markers compared to the DAVID tool. The cancer markers detection of the DAVID tool is about 0.7% to 29% based on the identified informative genes and pathways. However, PubMed text data mining detects cancer markers from the identified informative genes and pathways about 21% to 71%.

TABLE I. BIOLOGICAL VALIDATION OF THE IDENTIFIED GENES

Biological Validation	**Number of**	**GSE1456**	**GSE1561**
	Identified Informative Genes	953	536
PubMed Text Data Mining	Cancer Gene Markers	262	117
DAVID Tool	Cancer Gene Markers	9	4

TABLE II. BIOLOGICAL VALIDATION OF THE IDENTIFIED PATHWAYS

Biological Validation	**Number of**	**GSE1456**	**GSE1561**
	Identified Informative Pathways	52	24
PubMed Text Data Mining	Cancer Pathway Markers	37	17
DAVID Tool	Cancer Pathway Markers	13	7

C. Detection of Breast Cancer Markers

Disease outcomes can be identified using prognostic markers. This marker also helps in cancer treatments and drug discovery [18]. In previous studies, most computational methods have ignored the analysis of cancer markers that interact with a pathway or network [19]. With the use of PubMed text data mining for biological validation, cell cycle (HSA04110) and p53 signaling pathway (HSA04115) have been detected in both breast cancer datasets (GSE1456 and GSE1561). The p53 signaling pathway can provoke apoptosis in response to DNA damage after irradiation in breast cancer [20]. Western blot analysis also showed that the expression level of p53 signaling pathway-related proteins was significantly increased in human breast cancer cell line MCF7. Among the identified genes, CFL1 (9244) and BRCA2 (675) were validated as the basal and luminal of breast cancer gene markers [21-23]. RAD21 (5885) was validated in the literature as the luminal, basal, and ERBB2 of breast cancer gene markers [24].

VII. CONCLUSION

This study aims to apply PubMed text data mining in automation mode for biological context verification and validation based on the list of genes and pathways. The use of PubMed text data mining automation for biological validation shows that 37 and 17 informative pathways were validated correctly as breast cancer pathway markers for GSE1456 and GSE1561 datasets. Moreover, 262 and 117 informative genes were confirmed as breast cancer gene markers for the GSE1456 and GSE1561 datasets. The experimental result of PubMed text data mining automation for biological validation is better than the DAVID tool in genes and pathways. Cell cycle, p53 signaling pathway, and TP53 gene have been confirmed as cancer markers for breast cancer. All these cancer markers are significantly associated with the development and invasion of tumour cells. Hence, PubMed text data mining automation can automatically provide the PubMed IDs for each gene and pathway in the lists. This detection of cancer markers can help in earlier diagnosis, treatment, and drug discovery [25]. PubMed text data mining is suggested to apply in biological validation for other cancers. It will be further investigated to filter the keywords for the validation of genes and pathways.

ACKNOWLEDGMENT

The authors acknowledged Universiti Teknologi Malaysia (UTM) for providing the support and facilities for this research.

REFERENCES

[1] Jurca, G., Addam, O., Aksac, A., Gao, S., Özyer, T., Demetrick, D., and Alhajj, R. (2016). Integrating Text Mining, Data Mining, and Network Analysis for Identifying Genetic Breast Cancer Trends. *BMC Research Notes*, 9(1), 1-35.

[2] Steffen, P., Wu, J., Hariharan, S., Voss, H., Raghunath, V., Molloy, M. P., and Schlüter, H. (2020). OmixLitMiner: A Bioinformatics Tool for Prioritizing Biological Leads from 'Omics Data Using Literature Retrieval and Data Mining. *International Journal of Molecular Sciences*, 21(4), 1374.

[3] Xu, Y., Wang, J., Rao, S., Ritter, M., Manor, L.C., Backer, R., Cao, H., Cheng, Z., Liu, S., Liu, Y. and Tian, L. (2017). An Integrative Computational Approach to Evaluate Genetic Markers for Bipolar Disorder. *Scientific Reports*, 7(1), 1-9.

[4] Martinez-Ledesma, E., Verhaak, R. G., and Treviño, V. (2015). Identification of a Multi-cancer Gene Expression Biomarker for Cancer Clinical Outcomes using a Network-based Algorithm. *Scientific Reports*, 5(1), 1-14.

[5] Cai, L., Wu, H. and Zhou, K. (2021). Improved Cancer Biomarkers Identification Using Network-constrained Infinite Latent Feature Selection. *Plos One*, 16(2), p.e0246668.

[6] Nies, H. W., Mohamad, M. S., Zakaria, Z., Chan, W. H., Remli, M. A., and Nies, Y. H. (2021). Enhanced Directed Random Walk for the Identification of Breast Cancer Prognostic Markers from Multiclass Expression Data. *Entropy*, 23(9), 1232.

[7] Haynes, W. A., Higdon, R., Stanberry, L., Collins, D., and Kolker, E. (2013). Differential Expression Analysis for Pathways. *PLoS Computational Biology*, 9(3), e1002967.

[8] Faro, A., Giordano, D., and Spampinato, C. (2012). Combining Literature Text Mining with Microarray Data: Advances For System Biology Modeling. *Briefings in Bioinformatics*, 13(1), 61-82.

[9] Jiang, L., Edwards, S. M., Thomsen, B., Workman, C. T., Guldbrandtsen, B., and Sørensen, P. (2014). A Random Set Scoring Model for Prioritization of Disease Candidate Genes Using Protein Complexes and Data-mining of GeneRIF, OMIM and PubMed Records. *BMC Bioinformatics*, 15(1), 1-13.

[10] Albahri, A. S., Hamid, R. A., Alwan, J. K., Al-Qays, Z. T., Zaidan, A. A., Zaidan, B. B., Albahri, A. O. S., AlAmoodi, A. H., Khlaf, J. M., Almahdi, E. M., Thabet, E., Hadi, S. M., Mohammed, K. I., Alsalem, M. A., Al-Obaidi, J. R., and Madhloom, H. T. (2020). Role of Biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review. *Journal of Medical Systems*, 44, 1-11.

[11] Tworowski, D., Gorohovski, A., Mukherjee, S., Carmi, G., Levy, E., Detroja, R., Mukherjee, S.B. and Frenkel-Morgenstern, M. (2021). COVID19 Drug Repository: Text-mining the Literature in Search of Putative COVID19 therapeutics. *Nucleic Acids Research*, 49(D1), D1113-D1121.

[12] Conceição, S. I., and Couto, F. M. (2021). Text Mining for Building Biomedical Networks Using Cancer as a Case Study. *Biomolecules*, 11(10), 1430.

[13] Hatz, S., Spangler, S., Bender, A., Studham, M., Haselmayer, P., Lacoste, A.M., Willis, V.C., Martin, R.L., Gurulingappa, H. and Betz, U. (2019). Identification of Pharmacodynamic Biomarker Hypotheses through Literature Analysis with IBM Watson. *PloS One*, 14(4), p.e0214619.

[14] Huan, J., Wang, L., Xing, L., Qin, X., Feng, L., Pan, X., and Zhu, L. (2014). Insights into Significant Pathways and Gene Interaction Networks Underlying Breast Cancer Cell Line MCF-7 Treated with 17β-estradiol (E2). *Gene*, 533(1), 346-355.

[15] Zhou, J., and Fu, B. Q. (2018). The Research on Gene-disease Association based on Text-mining of PubMed. *BMC Bioinformatics*, 19(1), 1-8.

[16] Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., c E. T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P. M., Smeds, J., Skoog, L., Wedrén, S. and Bergh, J. (2005). Gene Expression Profiling Spares Early Breast Cancer Patients from Adjuvant Therapy: Derived and Validated in Two Population-based Cohorts. *Breast Cancer Research*, 7(6), R953.

[17] Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., MacGrogan, G., Bergh, J., Cameron, D., Goldstein, D., Duss, S., Nicoulaz, A.-L., Fiche, M., Brisken, C., Delorenzi, M. and Iggo, R. (2005). Identification of Molecular Apocrine Breast Tumours by Microarray Analysis. *Breast Cancer Research*, 7(S2), 1-1.

[18] Obuchowski, N. A., and Bullen, J. A. (2018). Receiver Operating Characteristic (ROC) Curves: Review of Methods with Applications in Diagnostic Medicine. *Physics in Medicine & Biology*, 63(7), 07TR01.

[19] Wang, J., Zuo, Y., Man, Y. G., Avital, I., Stojadinovic, A., Liu, M., Yang, X., Varghese, R.S., Tadesse, M.G., and Ressom, H. W. (2015). Pathway and Network Approaches for Identification of Cancer Signature Markers from Omics Data. *Journal of Cancer*, 6(1), 54.

[20] Liu, H. C., Ma, F., Shen, Y., Hu, Y. Q., and Pan, S. (2015). Overexpression of SMAR1 Enhances Radiosensitivity in Human Breast Cancer Cell Line MCF7 via Activation of p53 Signaling Pathway. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 22(5-6), 293-300.

[21] Quintela-Fandino, M., Arpaia, E., Brenner, D., Goh, T., Yeung, F. A., Blaser, H., Alexandrova, R., Lind, E. F., Tusche, M. W., Wakeham, A. and Ohashi, P. S. (2010). HUNK Suppresses Metastasis of Basal Type Breast Cancers by Disrupting the Interaction between PP2A and Cofilin-1. *Proceedings of the National Academy of Sciences*, 107(6), 2622-2627.

[22] Pécuchet, N., Popova, T., Manié, E., Lucchesi, C., Battistella, A., Vincent-Salomon, A., Caux-Moncoutier, V., Bollet, M., Sigal-Zafrani, B., Sastre-Garau, X., Stoppa-Lyonnet, D. and Stern, M. H. (2013). Loss of Heterozygosity at 13q13 and 14q32 Predicts BRCA2 Inactivation in Luminal Breast Carcinomas. *International Journal of Cancer*, 133(12), 2834-2842.

[23] Dilday, T., Ramos, N., and Yeh, E. (2020). HUNK Signaling in Metastatic Breast Cancer. *Oncoscience*, 7(5-6), 30.

[24] Xu, R., and Wunsch, D. C. (2010). Clustering Algorithms in Biomedical Research: A Review. *IEEE Reviews in Biomedical Engineering*, 3, 120-154.

[25] Eyileten, C., Wicik, Z., De Rosa, S., Mirowska-Guzel, D., Soplinska, A., Indolfi, C., Jastrzebska-Kurkowska, I., Czlonkowska, A. and Postula, M. (2018). MicroRNAs as Diagnostic and Prognostic Biomarkers in Ischemic Stroke-A Comprehensive Review and Bioinformatic Analysis. *Cells*, 7(12), 249.