ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

Latent Topic Modeling of Cancer Hallmark for Analyzing Biomedical Literature

[1] Supriya Gupta, [2] Aakanksha Sharaff, [3] Naresh Kumar Nagwani

[1] [2] [3] Department of Computer Science and Engineering, National Institute of Technology, Raipur, India
Corresponding Author Email: [1] sgupta.phd2018.cs@nitrr.ac.in, [2] asharaff.cs@nitrr.ac.in, [3] nknagwani.cs@nitrr.ac.in

***Abstract*— The Cancer hallmarks represent an important impression for new information regarding cancer and to unravel the complexities of cancer. The limitations of frameworks for the analysis of clearly developed topics of cancer knowledge, study of topic modeling in cancer research remains a major challenge. The methods provided were successfully used to study the targeted semantic and contextual data in scientific texts using embedding words according to the hallmarks of cancer.**

***Index Terms*— Multi-task learning, Biomedical domain analysis, Cancer Hallmark, Topic analysis.**

I. INTRODUCTION

Cancer is the next leading source of death in 2018 and is incredibly complex [1]. The hallmarks of cancer are the qualities which are used to identify disease cells from ordinary cells [2, 3]. The cancer hallmark is a basic target for malignant cancer cell mutations and is helpful in diagnosing tumor pathogenesis. It retrieves very important information in cancer research [4-6]. There are genetic markers which can relay information regarding malignant growth and are proposed to provide a regulatory framework to determine the extent of biological processes leading to cancer [3]. These hypertensive stressors (SPS), growth retardants (EGSs), cell death resistance (RCD), replicative immortality (ERI), inducing angiogenesis (IA), active attack and metastasis (AIM)) are the hallmarks of cancer and shown in Figure 1.

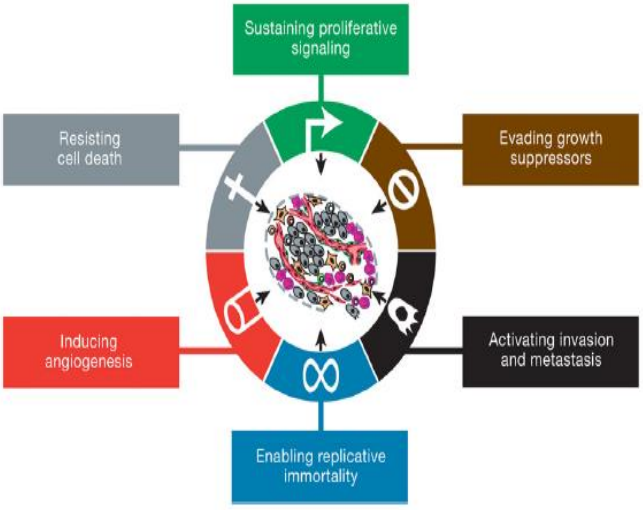

Figure 1. The Hallmark of Cancer

The biomedical literature contains the majority of cancer research findings, and the PubMed database has recognized around 30 million quotes and 4 million cancer-related literatures as of 2019. In the same field, authors have studied to gather more precise and historical data [7, 8]. That database's extensive collection of biological literature presents a fantastic chance to gather crucial data for cancer research. Researchers can retrieve relevant literature from that database by using a keyword-based search. The complexity of cancer necessitates the use of numerous keywords, synonyms, and combinations, making it time-consuming to find pertinent information using just the keyword. It's crucial to model a topic in order to retrieve data from many cancer-related papers at PubMed in a somewhat more efficient manner. Most of the cancer research results are found in the biomedical literature with the PubMed database.

II. LITERATURE REVIEW

The technique of topic modeling helps cancer researchers express biomedical knowledge more effectively. However, there aren't many experiments looking into biomedical themes and the literature that surrounds them, thus it hasn't been estimated how to use a topic model to quantify the hallmark of cancer. The topic model has been used by researchers to analyze biomedical literature on genomes, including studies of genetic research and protein-protein interactions. In order to find protein-protein interactions in biomedical literature, Andrzejewski et al. developed an automatic extraction model [9]. They measured vocabulary differences from Medline abstracts using the LDA model. In order to successfully synthesize protein interactions in biomedical literature, Wang et al. model the LDA-producing subject [10]. They claim that the topic model reflects the intricate relationships between the various procedures as well as the words that are associated with them.

Wang et al. [11] proposed a strategy to remove shared features between quality-related archives in a way that was not utilized using a point model. They used the LDA model to remove the contents of the record. They trace that points are usually logically indicated in the statistics of the currently

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

used theme. One of the key issues with biomedical testing is the drug discovery. Bis-gin et al. [12] evaluated the effectiveness of a point that demonstrates the disclosure of hidden models and their implications from the diet and drug association approved for drug marks. The follow-up on visual cues with error settings that are directly linked to events highlight relations or corrective actions. Bisgin et al. [13] additionally constructs a potential topic model in terms of reference for clinical drug re-branding.

The topic model is divided into 52 genres, each containing a number of words in this review. Be aware that medications that are considered comparable may have the same effect. Obtaining relevant data from biomedical text information is an effective test site. Chen et al. [14] have proposed a method that uses the LDA topic production model to improve the variability of bio-clinical data acquisition status [15]. In addition, Song et al. [16] investigated the design of knowledge and patterns in bioinformatics using textual extraction processes including the presentation of themes in PubMed Central full text articles. As a result, using the topic model, they discovered that significant topic modeling revealed that key themes were more focused on natural ideas than computer components of bioinformatics. Wang et al. [17] promoted a draft-based mining framework called the Bio Topic. They tracked that the preliminary refined analysis associated with the title model shows preferred results than previous literature mining frameworks.

The topic model is then used to relate a few diseases and their investigations. In their assessment of the top five malignant cancer research trends, Cui et al. [18] extended the LDA Gibbs experimental model to include the cosine coefficient using a vector space model.

III. METHOD

The title model is a feasible model, used to obtain subject details from multiple choruses. Conventional theme models such as latent Dirichlet allocation (LDA) have been successfully used in various texts. It plays a map of the recurring space of the highest magnitude in the space of the subtitle heading. In addition, the topic model can capture semantic data, which can cover the hidden relationships within the archive and can adequately address polisemy, similar terms, and a variety of issues, which have significant implications for content analysis. The purpose of this paper is to highlight the analysis of the cancer hallmark, to scrutinize the semantic data from a large number of papers, to exclude the complexity of cancer, to introduce the critical structure of dependency analysis and cancer hallmark, and to advance the modeling result of each hallmark category.

A. Natural Language Processing

We have designed and constructed a controlled NLP pipeline shown in Figure 2 that separates the semantic and syntactic supply to the biomedical corpus: Empty words: A small complex section uses each word that occurs in input texts. We lemmatize words to minimize including sparsity. Named Objects (NE): Specific concepts of business holdings in the text, which provide another way of combining words into key categories.

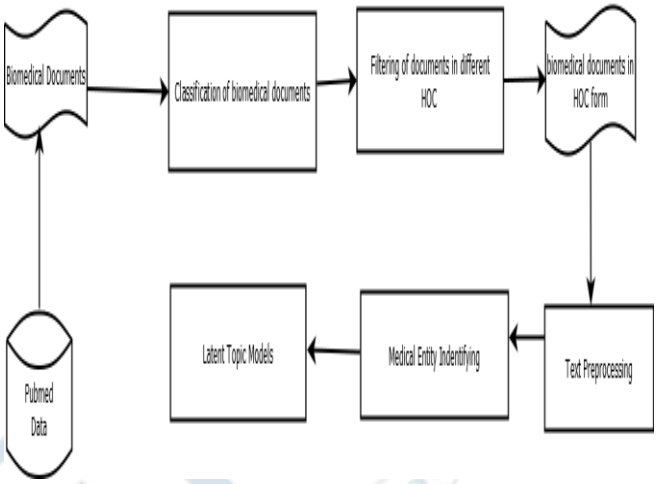

Figure 2. The overview of the categorization of Cancer hallmark literatures and latent topic model

B. Medical Entity Identifier

The procedure used to identify the medical business is described in this section. In order to identify systemic terms of the integrated medical language (UMLS) system [20] in the clinical context and, accordingly, performed the role of key records to be retrieved, the suggested method utilized MetaMap materials [19]. The US National Library of Medicine created the UMLS Metathesaurus, a vocabulary for biomedicine. It is comprised of 14 million original ideas and an estimated 3.67 million ideas. Our framework will benefit greatly from this identification. The modification of the biomedical text in the UMLS Metathesaurus can be achieved by MetaMap, which is a very adaptable program. NLP and language integration techniques like tokenization, POS marking, syntactic study, and others are used by MetaMap.

C. Latent Topic Models

The model showed the standard LDA-based subtle models used in this paper. Basically, it shows the steps in a topic comparison that includes the word bag, model training, and exit model title. In the LDA, the record is seen as a combination of different themes in which each report was adopted to compile a set of selected themes to report using the LDA. The two distribution of opportunities are $P(T|D)$ and $P(E|T)$ as a mixed distribution. The distribution of topics is shared with the previous standard document dirichlet D. Parameter θ_D is the detailed distribution over the Q-headers which makes the distribution of Dirichlet $Dir(\theta_D|\alpha)$ the same as the title Q β_Q parameter of mixed distribution over R entities; the effect of the distribution of Dirichlet $Dir(\theta_D|\eta)$. As a pre-multinomial conjugate, Dirichlet

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

distribution is as simple as the previous one and can facilitate mathematical considerations in the LDA.

There are titles, organizations, and papers on the corpus. The word bag template is used for retrieving information and processing the native language. In the vocabulary, it represents the text by ignoring its order and grammar.

IV. EXPERIMENTAL RESULT AND DISCUSSION

In this section, the model used biological data sets to evaluate the performance of the subject modeling framework, particularly cancer-related texts and evaluated the effectiveness of the LDA model in documents with weak labels with UMLS.

A. Topic Models

LDA which has been used from the Gensim Python library [21], has all the default parameter settings, for typical subject models. Adjusting test parameters can lead to better test outcomes. Using semantic integration and guiding principles, we evaluated the models. The gear used for testing in this article was an Intel E3, 32G memory, and GTX 1070 Ti. The Windows operating system was used as the software platform, and Python 3.5 was employed for the environment development. The Python libraries Scikit-lern [23] and Pytorch Library [22] were used to develop the framework that is being described.

B. Topic Modeling Result

Topic modeling's impacts are demonstrated in this section where in Tables 1-6, we have listed the top 10 topics of each hallmark. Using the LDA model, the concepts of all the notable symbols were examined. We followed up on associated subjects that were frequently linked to unfavorable information from lung cancer information. The names are noted in common non-clinical form. According to Table 1, the "Epidermal Growth Factor Receptor" CUI discovered is mostly worried about the mark "maintaining a growing signature" (SPS) in a variety of circumstances. Similarly EGFR deficiency affects the receptor for the growth factor. Ten primary articles discuss the idea of "EGFR genetics, EGFR proteins, and EGFR measurements." According to Liu et al.'s explanation in their study [24], the activation of EGFR-tyrosine kinases is crucial.

According to Table 2, TP-53, the gene TP-53 wt Allele, and TP-53, was the word most focused on the symptoms of avoiding growth stressors (EGS) in many hypotheses. TP-53 was the most frequently mutated gene in a recent Cancer Genome Atlas (TCGA) test result for squamous cell lung cancer, according to Amin et al. [25]. According to our examination of the theme modeling, TP53 was the dominant topic for the EGS brand.

As shown in Table 3, we were able to identify "Apoptosis" as a significant contributor to cell death (RCD). Apoptosis was studied by Liu et al. [26] in relation to non-small cell lung cancer (NSCLC). They report that the cycle of apop-tosis, which causes pollution of proteins and organelles or cell death over cell pressure.

As shown in Table 4, we tracked down the fact that the concepts of "senescence", "age" and "old personality", were the main themes in (ERI) that allowed the recurring sign of immortality. Senescence or organic maturing is a progressive deterioration of functional and physical characteristics indicating that a person is experiencing dementia leading to aging. Yaswen et al. [27] spoke of a corrective focus on repetitive immortality. They reported that the protective function of senescence was collected in murine models of lung adenomas, T-cell lymphomas, prostate cancer, and pituitary growth.

We identified "Vascular Endothelial Growth Factors (VEGF)" as a key concept for identifying income angiogenesis (IA), as indicated in Table 5. VEGF is crucial for angiogenesis, vascular porousness, and metastasis during tumor growth, according to Shimoyamada et al. [28].

Table 6 demonstrates how we discovered a connection between the terms "Neoplasm" and "Metastasis" and the activation of the attack and metastasis (AIM) sign. The liver is one of the most often affected organs by metastatic infection, according to Martin et al. [29], and in the United States and Europe, selective liver neoplasms are unquestionably more prevalent than major hepatic neoplasms. As opposed to significant liver illness, persistent lung cancer is referred to as metastatic lung cancer.

Table 1. The top topics on the sustaining proliferative signaling (SPS) hallmark

Patients; Non-Small Cell Lung Carcinoma; Epidermal Growth Factor Receptor; EGFR proteins; Epidermal Growth Factor Receptor Consolidation Tryptophanase, the EGFR gene, a combination of items, and the therapeutic-peutic process.

Table 2. The **top** topics on the evading growth suppressor (EGS) hallmark

Humans; Apoptosis; Tryptophanase; TP53 wt Allele; gene TP53; Prohibition; Increase (activity); Speech (fundamental metadata notion); Display procedure; Result.

Table 3. The top topics on the resisting cell death (RCD) hallmark

Patients; Tryptophanase; Neoplasms; Treating; Therapeutic Techniques; Less than Two PSA Level, 2+ WHO Score, therapeutic features, and administration process

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

Table 4. The top topics on the enabling replicative immortality (ERI) hallmark.

Cellular senescence, old age, induce (activity), fibroblasts, cell count, and homo sapiens are associated with Increase.

Table 5. The top topics on **the** inducing angiogenesis (IA) hallmark.

Laboratory rats, a community group, a party object, a tumour mass, recombinant vascular endothelial growth factors, neoplasms, tryptophanase, and the angiogenic process.

Table 6. The top topics on the activating invasion and metastasis (AIM) hallmark.

Neoplasms, metastatic qualification, metastatic clinical trial setting, cell count, metastatic of neoplasm, second neoplasm, tryptophanase, and meta-static.

C. Visualization

This section presents a word cloud that represents the top topics and concepts for each cancer hallmark. Word clouds have become a popular and understandable method of seeing literature. As seen in Figure 3-6, they are employed in a number of circumstances as a means of conveying an overall impression by narrowing the text down to the terms that occur most frequently.

Figure 3: Word cloud for SPS and EGS Cancer hallmark

Figure 4: Word cloud for RCD and ERI Cancer hallmark

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

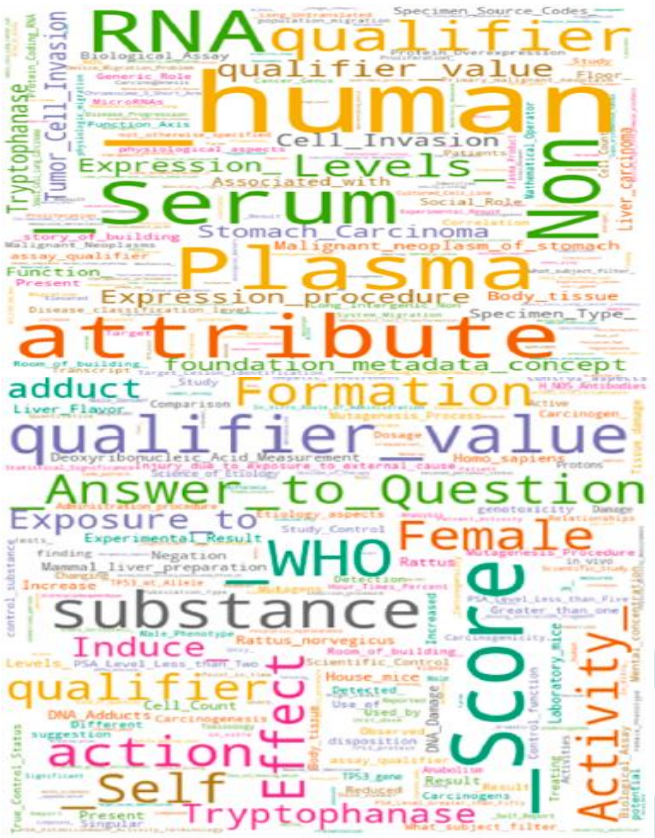

Figure 5: Word cloud for IA and AIIM Cancer hallmark

V. CONCLUSION AND FUTURE WORK

This paper has introduced a topic analysis framework with a biomedical archive for topic analysis. The worker used models of common LDA topics in analyzing cancer symptoms. Topic models use just biomedical concepts recognized by the UMLS sentence. The presented framework was highly variable in a large number of unpublished repositories and standard topic models were effective in analyzing the articles in a volume labeled weak according to the symptoms of cancer. In future, the analysis of features during topic modeling can be improved to improve overall topic analysis.

REFERENCES

[1] Mehmet Sitki Copur, M.D. State of Cancer Research around the Globe. Oncology 2019, 14, 33.

[2] Hanahan, D.;Weinberg, R.A. The hallmarks of cancer. Cell 2000, 100, 57–70.

[3] Hanahan, D.;Weinberg, R.A. Hallmarks of cancer: The next generation. Cell 2011, 144, 646–674.

[4] Gutschner, T.; Diederichs, S. The hallmarks of cancer: A long non-coding RNA point of view. RNA Biol. 2012 9, 703–719. [PubMed]

[5] Piao, Y.; Piao, M.; Ryu, K.H. Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles. Comput. Biol. Med. 2017, 80, 39–44. [PubMed]

[6] Li, F.; Piao, M.; Piao, Y.; Li, M.; Ryu, K.H. A New direction of cancer classification: Positive effect of Low-ranking MicroRNAs. Osong Public Health Res. Perspect. 2014, 5, 279–285. [PubMed]

[7] Munkhdalai, T.; Li, M.; Batsuren, K.; Park, H.A.; Choi, N.H.; Ryu, K.H. Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. J. Chemin. 2015,

[8] Munkhdalai, T.; Namsrai, O.E.; Ryu, K.H. Self-training in significance space of support vectors for imbalanced biomedical event data. BMC Bioinform. 2015, 16, 6.

[9] Andrzejewski, D. Modeling Protein–Protein Interactions in Biomedical Abstracts with Latent Dirichlet Allocation;CS 838-Final Project; University ofWisconsin–Madison: Madison, WI, USA, 2006.

[10] Wang, H.; Huang, M.; Zhu, X. Extract interaction detection methods from the biological literature. BMC Bioinform. 2009, 10, 55.

[11] Wang, V.; Xi, L.; Enayetallah, A.; Fauman, E.; Ziemek, D. GeneTopics-interpretation of gene sets via literature-driven topic models. BMC Syst. Biol. 2013, 7, 10.

[12] Bisgin, H.; Liu, Z.; Fang, H.; Xu, X.; Tong, W. Mining FDA drug labels using an unsupervised learning technique-topic modeling. BMC Bioinform. 2011, 12, 11. [PubMed]

[13] Bisgin, H.; Liu, Z.; Kelly, R.; Fang, H.; Xu, X.; Tong,W. Investigating drug repositioning opportunities in FDA drug labels through topic modeling.BMC Bioinform. 2012,13,6. [PubMed]

[14] Chen, Y.; Yin, X.; Li, Z.; Hu, X.; Huang, J.X. A LDA-based approach to promoting ranking diversity for genomics information retrieval. BMC Genomics 2012, 13, 2. [PubMed]

[15] 15. Hersh, W.R.; Cohen, A.M.; Roberts, P.M.; Rekapalli, H.K. TREC 2006 Genomics Track Overview; TREC:Gaithersburg, MD, USA, 2006.Appl. Sci. 2020, 10, 834 24 of 25

[16] Song, M.; Kim, S.Y. Detecting the knowledge structure of bioinformatics by mining full-text collections.Scientometrics 2013, 96, 183–201.

[17] Wang, X.; Zhu, P.; Liu, T.; Xu, K. BioTopic: A topic-driven biological literature mining system. Int. J. Data Min. Bioinform. 2016, 14, 373–386.

[18] Cui, M.; Liang, Y.; Li, Y.; Guan, R. Exploring Trends of Cancer Research Based on Topic Model. IWOST-1 2015, 1339, 7–18.

[19] Aronson, A.R. E_ective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In Proceedings of the AMIA Symposium American Medical Informatics Association, Chicago November 2001

[20] Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Res. 2004, 32, 267–270.

[21] R˘ ehu° r˘ek, R.; Sojka, P. Gensim—Statistical Semantics in Python. Statistical Semantics; Gensim; EuroScipy: Paris, France, 2011.

[22] Ketkar, N. Introduction to Pytorch; Apress: Berkeley, CA, USA, 2017.

[23] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python.

[24] Liu, T.C.; Jin, X.;Wang, Y.;Wang, K. Role of epidermal

growth factor receptor in lung cancer and targeted therapies. Am. J. Cancer Res. 2017, 7, 187. [PubMed]

[25] Amin, A.R.; Karpowicz, P.A.; Carey, T.E.; Arbiser, J.; Nahta, R.; Chen, Z.G.; Dong, J.T.; Kucuk, O.; Khan, G.N.; Huang, G.S. Evasion of anti-growth signaling: A key step in tumorigenesis and potential target for treatment and prophylaxis by natural compounds. In Seminars in Cancer Biology; Elsevier: Amsterdam, , 2015; Volume 35, pp. 55–77

[26] Liu, G.; Pei, F.; Yang, F.; Li, L.; Amin, A.D.; Liu, S.; Buchan, J.R.; Cho, W.C. Role of autophagy and apoptosis in non-small-cell lung cancer. Int. J. Mol. Sci. 2017, 18, 367.

[27] Elsevier: Amsterdam, The Netherlands, 2015; Volume 35, pp. 104–128.

[28] Yaswen, P.; MacKenzie, K.L.; Keith, W.N.; Hentosh, P.; Rodier, F.; Zhu, J.; Firestone,G.L.; Matheu, A.; Carnero, A.; Bilsland, A. Therapeutic targeting of replicative immortality. In Seminars in Cancer Biology;

[29] Shimoyamada, H.; Yazawa, T.; Sato, H.; Okudela, K.; Ishii, J.; Sakaeda, M.; Kashiwagi, K.; Suzuki, T.; expression in lung cancer cells. Am. J. Pathol. 2010, 177, 70–83.

[30] Martin, T.A.; Ye, L.; Sanders, A.J.; Lane, J.; Jiang, W.G. Cancer Invasion and Metastasis: Molecular and Cellular Perspective. Available online: https://www.ncbi.nlm.nih.gov/books/NBK164700

[31] Ninomiya, H.; Nomura, K.; Satoh Y.; Okumura, S.; Nakagawa, K.; Fujiwara, M.; Tsuchiya, E.; Ishikawa, Y. Genetic instability in lung cancer: Concurrent analysis of chromosomal, mini-and microsatellite instability and loss of heterozygosity. Br. J. Cancer 2006

[32] Melkamu, T.; Qian, X.; Upadhyaya, P.; O'Sullivan, M.G.; Kassie, F. Lipopolysaccharide enhances mouse lung tumorigenesis: A model for inflammation-driven lung cancer. Vet. Pathol. 2013, 50, 895–902

[33] Harmey, J.H.; Bucana, C.D.; Lu, W.; Byrne, A.M.; McDonnell, S.; Lynch, C.; Bouchier-Hayes, D.; permeability and tumor cell invasion. Int. J. Cancer 2002, 101, 415–422

[34] Min, H.Y.; Lee, H.Y. Oncogene-driven metabolic alterations in cancer. Biomol. Amp Ther. 2018, 26, 45.

[35] Gwin, J.L.; Klein-Szanto, A.J.; Zhang, S.Y.; Agarwal, P.; Rogatko, A.; Keller, S.M. Loss of blood group antigen A in non-small cell lung cancer. Ann. Surg. Oncol. 1994, 1, 423–427.

[36] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.

[37] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.

[38] Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.