Machine Learning Based Cancer Classification Using Gene Expression Data

Ruiyi Li*

College of Science, Master student, Northeastern University, Boston, United States of America
Email: li.ruiy@northeastern.edu (R.L.)

*Corresponding author

Manuscript received October 12, 2023; revised November 28, 2023; accepted December 10, 2023; published May 15, 2024

***Abstract*—Cancer is a common severe disease today, and this type of disease has a high mortality rate. Therefore, a cancer diagnosis is an essential tool because most patients die due to a lack of early diagnosis and treatment. To better diagnose cancer, people use gene sequencing techniques and microarray techniques to replace traditional tumor morphology because cancer is a genomic disease. As people widely utilize gene sequencing technology, many gene data aggregately form a cancer gene database to help people manage these data better. Through large-scale cancer genomics datasets, people have used different Machine Learning algorithms to create cancer prediction models. The prediction model created by these algorithms has higher accuracy and a lower error rate than other cancer classification methods. Therefore, a large amount of genetic data can be obtained using different gene detection techniques to detect cancer gene expression. People will perform unique analyses and processing of these data to obtain valuable information. These large amounts of data are aggregated into specialized cancer gene databases such as TCGA that can be used to train ML algorithms to obtain the best predictive models.**

***Keywords*—machine learning, TCGA, cancer genomics**

I. Introduction

Cancer is a global disease that has a significant impact on human society. According to the World Health Organization (WHO) statistics, cancer is the first or second leading cause of death in most countries, so a large number of people in the world die from cancer [1]. According to research statistics, there will be 19.3 million new cancer cases and nearly 10 million deaths in 2020. At the same time, the diversity of cancer has a specific impact on the treatment and diagnosis of cancer [1]. Uncontrolled cell growth, genetic mutation, and the spread of human cells lead to different types and effects of cancer. Therefore, cancer cells are uncontrolled cells because they do not follow the usual trajectory of cell movement. This type of cell has unlimited proliferation and spread, forming tumors in the human body. Research on cancer cells has shown that most cancer cells' abnormal or uncontrolled behavior is caused by genetic mutations and changes in gene expression patterns, so mutations in some functional protein genes can lead to uncontrolled proliferation [2]. Therefore, gene mutations are a significant marker of cancer gene expression. In cancer research and development, understanding different gene mutations and gene expressions can provide a good understanding of the mechanisms and causes of cancer formation and distinguish cancer from other genetic diseases. From this information, it appears that cancer is a genomic disease and that such diseases can occur in different parts of the body. One of the reasons why cancer cells cause cancer is because they interfere with normal cells' processes of proliferation, repair, communication, or apoptosis [3]. In cancer cells, genes exhibit abnormal phenotypes by altering gene expression, and such alterations can contribute to tumor initiation or progression. An abnormally mutated gene is often thought to be a type of oncogene, and changes in gene expression contribute to the cancer phenotype [3]. Therefore, studying cancer gene expression can help cancer treatment and diagnosis.

With the in-depth study of cancer cells, it has been found that gene mutations play an important role in cancer. However, cancer is still a disease with high mortality because traditional diagnostic methods do not have high accuracy in the early stages of cancer. This method uses microscopy, histology, morphology, or immunoassays to detect cancer cells or tissues [4]. This is similar to a biopsy, so researchers can use microscopy and cell morphology to detect cancer cells in a patient's tissue. However, this approach can miss specific cancer cells in the early stages of cancer, leading to misdiagnosis [4]. Another more common traditional test is an immunoassay, which uses cancer biomarkers to detect the presence and type of cancer. This method has high sensitivity and selectivity; however, it is expensive and not sensitive to the concentration of early markers [4]. Therefore, these traditional cancer detection methods all have certain limitations. This way of classifying tumors based on morphology and staining has significant inaccuracies because different observers obtain different results, creating specific human errors [5]. Because of these factors, the researchers turned their attention to molecular classification. In contrast to traditional approaches, this approach is based on cancer genomics and proteomics, and it uses tools to create molecular signatures of disease [4].

Some researchers are combining machine learning with cancer genomics so that molecular methods can be better used to classify and diagnose cancer. This is a way of combining cancer genomics, computer technology, and bioinformatics technology. Machine learning (ML) is a data analysis method to build predictive models and detect key features from complex datasets [6]. ML is not a simple single technology, so it includes different algorithm types such as Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), and Decision Trees (DT). Large-scale data is needed to train an ML model to achieve its goal. In this way, the algorithm can be improved through the training process and by adjusting for its numerical parameters into obtain the optimization [7]. Therefore, ML can be designed to a computational path network, and it can determine the probability distribution and predict the result based on the

input data [7]. When algorithms and cancer genomics are combined, the predictive models formed by ML can classify and diagnose cancers. This depends on the accumulation of genetic data on cancer and the development of bioinformatics. Therefore, machine learning techniques and the accumulation of cancer genetic data facilitate the development of cancer diagnosis or classification research.

II. Cancer Genomics Drives Modern Cancer Research

Cancer genomics is a way to systematically analyze gene expression and mutations in cancer cells by using human genome sequences and high-throughput technology [5]. When algorithms and cancer genomics are combined, the predictive models formed by ML can classify and diagnose cancers. This depends on the accumulation of genetic data on cancer and the development of bioinformatics. Therefore, machine learning techniques and the accumulation of cancer genetic data facilitate the development of cancer diagnosis or classification research [5]. Traditionally, human error is unavoidable because humans cannot judge and observe things from a completely objective perspective. Therefore, people began to use instruments and data to analyze the genome to increase objectivity and accuracy. Genomic tools are generally divided into three categories: identifying areas in the genome that have been altered by cancer, areas where specific gene mutations have occurred, and areas where gene expression has changed [5]. Gene expression changes in cancer cells can be measured and screened by these techniques. The data obtained can be used to make a database of cancer mutated genes or gene expression, which can help people study homologous cancers.

High-throughput genome sequencing is a common way to measure cancer mutation genes. This is also called next-generation sequencing (NGS), which can perform both targeted and whole genome DNA and RNA analysis [8]. Compared with traditional sequencing methods, next-generation sequencing can read and analyze millions of sequences using one instrument, avoiding specific cloning bias problems in genome expression [9]. With the development of technology, NGS only needs to input a small number of gene fragments to get the sequencing results of the gene, and the instrument can sequence many gene sequences in a short time. One of the most common methods of genome sequencing is Illumina sequencing technology. With this technology, researchers can generate sequencing reads from multiple surface-amplified DNA fragments simultaneously, and the analysis pipeline can filter out erroneous data to reduce the error rate [9, 10]. In Illumina sequencing technology, cancer genome analysis is typically measured using RNA sequencing. This technique allows for a more detailed analysis of gene expression profiles in cells and can also be used to compare differences between normal and cancer cells. Researchers can analyze RNA transcription characterization using this technology, including transcription sites, small RNA characterization, gene fusion detection, splicing event detection, etc [11]. Compared with other sequencing technologies, RNA sequencing has higher coverage and resolution of the dynamic properties of the transcriptome, so this can provide a deeper understanding of the gene expression of the transcriptome in the cell [12]. On the other hand, biology shows that the activities of cells are influenced by functional proteins whose functions are controlled by genes. According to the central dogma of biology, the production of proteins requires the transcription of DNA and RNA translation. The genes in which they are transcribed into complementary RNA molecules are called the transcriptome, and these transcriptomes have specific effects on the function of proteins [12]. Therefore, RNA sequencing is a good tool for studying cells by understanding transcriptome gene expression.

Another technique used to analyze genomes is microarray technology. Microarrays can be used to study molecular interactions that conventional methods cannot analyze, and they can simultaneously detect many gene expressions [13]. This technology can help people screen the right genes and compare them with normal cells to generate good gene expression profiles for cancer diagnosis. In this regard, microarray technology and next-generation sequencing technology have similar capabilities. At the same time, microarray technology can also aid cancer research by identifying genomes associated with metastasis or treatment response [13]. This advantage plays a crucial role in cancer diagnosis and prediction because it reduces the number of genetic tests. However, this has the obvious drawback: the technique requires physical disruption of cells to obtain gene expression patterns. This causes loss of cells [13]. At the same time, since RNA samples are extracted from more complex tissues or environments, maintaining the quality and quantity of RNA samples is a challenge. In the case of microarrays, incomplete or low-quality RNA can produce erroneous data and prevent the analysis of multi-tissue samples [13]. By comparison, it can be found that NGS technology is more suitable for cancer genomics detection and research than microarray technology.

A. The Cancer Genome Atlas

The cancer genome atlas (TCGA) is a publicly funded database of large-scale cancer genomics dataset, and this database contains research on individual cancer types and comprehensive pan-cancer analysis [14]. The advent of such databases provides broad access to information for cancer research and treatment, reducing the time researchers spend on genetic sequencing. The cancer information in TCGA comes from large-scale genome sequencing and comprehensive multidimensional analysis of different types of tumors, thus providing information on the significant oncogene of nearly 30 human tumors [14]. The information in TCGA comes from different organizations, which have a standardized process. Tissue Source Sites (TSS) is responsible for collecting tissue samples from cancer patients and forwarding them to Biospecimen Core Resource (BCR), and BCR processes the samples to obtain clinical data [14]. However, these data are not stored directly in the TGCA database. Until then, the Data Coordinating Center (DCC) will use this information for genome characterization and high-throughput sequencing. The final data will be imported into the TCGA database for researchers to access [14]. In the process, TCGA uses different types of high-throughput technologies to obtain more comprehensive information on the Cancer Genome

Atlas.

In TCGA, RNA sequencing is a technique for high-precision analysis of the transcriptome. This technology can provide transcriptome information on sequence coverage, sequence variation, and gene expression through sequence comparison [14]. Other more common TCGA applications include DNA sequencing, array-based DNA methylation sequencing, and reverse-phase protein arrays (RPPA). These techniques can detect and analyze DNA sequence changes, epigenetic changes in the genome, aberrant DNA methylation, and protein expression [14]. Therefore, TCGA is a public database that stores many cancer genome information. Its existence provides an excellent aid for cancer research, diagnosis, and treatment. And researchers can use this data in conjunction with bioinformatics to create predictive models for cancer classification and diagnosis.

III. Data Analysis for RNA-SEQ

The genetic data obtained by high-throughput technology is large and complex, so it is challenging for cancer researchers to analyze and use it correctly. TCGA is a good platform that reflects the importance of data analysis because the analyzed data can better reveal the information of cancer genes. Among them, RNA sequencing data analysis has a commonly used standardized process.

RNA-seq analysis can be used as a standard for analyzing gene expression at the transcriptome level, and it can also be used to study molecular events in cells and tissues [15]. Through this analysis, people can have a good understanding of gene expression or regulatory changes in cells and the response of cancer cells. In RNA analysis, polyadenylated RNA transcripts will contain large oligonucleotide primers, and this will be fragmented and size selected [16]. These fragmented cDNAs generate many sequence reads, requiring data analysis techniques to interpret and apply these data. This analysis roughly includes four steps. This is because the gene sequence information obtained by RNA sequencing technology is raw, so people cannot quickly analyze this unprocessed information. Quality checking and information preprocessing of the original sequence information are required to solve this problem [16]. In this way, high-quality sequence data can be obtained for cancer research. After obtaining high-quality data, these data will be mapped to a reference genome or transcriptome and counted reads are mapped to a single gene or transcript [16]. The mapping can compare the experimental data with the reference genome so that the position of the sample in the genome can be found. This can well show the expression of the measured gene. At the same time, RNA sequencing data analysis can help identify differences in gene expression between different biological conditions because this technology can be used to compare the differences in gene expression between cancer cells and normal cells. [16]. After clarifying the differences in gene expression between two different organisms or cells, people can better explain the gene expression.

With the development of technology, people have invented different analysis software to apply to RNA sequencing data analysis, which provides a certain degree of convenience for genome research. FastQC can be used to check the quality of sequencing data, and it can quickly determine whether there are contaminated sequences or adaptor sequences in the original data [16]. TopHat is a traditional software for mapping. For this software, the position of the sequence in the reference genome can be determined, and this can also determine the position of the read across the exon-exon junction [16]. On the other hand, features such as featureCounts can infer gene and isotype abundance, and the computational memory required for this is relatively small [16]. There are many software's like this for RNA sequencing data analysis, and these systems significantly reduce the shortcomings of RNA sequencing data.

IV. Principles of Machine Learning

Machine learning (ML) is an artificial intelligence that can predict results through algorithms and input data [6]. This technique is widely used in cancer classification and diagnostic research because it can analyze biological samples through different techniques and algorithms. Through the basic definition of ML, it can be determined that the algorithm model thus constructed can be used for the diagnosis and classification of cancer. The more common types are supervised, unsupervised, and reinforcement learning. Supervised learning refers to using labeled training data to estimate or map input data to desired outcomes. Unsupervised learning refers to obtaining results without labeling examples and learning the output concepts of the process [3]. At the same time, by using ML technology, it can be roughly classified according to tag types and feature types. In cancer research, people can diagnose the nature of tumors by using collected data and algorithms.

Supervised learning includes naive Bayes classification, support vector machines (SVM), and random forests. It can predict the difference between the predicted and known labels and reduce errors [17]. Unsupervised learning like k-means clustering can classify the samples most according to the characteristics of the training data, and this does not generate labels. Cancer research can identify histone modifications in cancer cell samples stained by immunohistochemistry to predict the risk of recurrence [17]. Reinforcement learning is an algorithm that acts and predicts the features of future steps based on past and present features and gives results, so it can be used to build predictive models [17]. Therefore, these algorithms can combine mathematics, data, computing, and bioinformatics. The cancer information database can be used to classify and diagnose cancer through the design of an ML algorithm. The advent of such algorithms or techniques can reduce human errors, allowing for analyzing large amounts of data.

V. Application of ML in Cancer Classification

Cancer classification is an integral part of cancer research related to cancer diagnosis and treatment. People do in-depth research and treatment by understanding the type of cancer in a sample. In the past, microscopy and staining have been the most commonly used methods of classifying cancer, but limitations of this approach have also been found. As a result, ML techniques are gaining popularity in cancer classification. Researchers have designed different

algorithms to classify cancer since there are many types of ML techniques.

A. ML Classification Based on TCGA Database

For different purposes, researchers have invented different ML algorithms for cancer classification. Yuanyuan Li, Kai Kang, and others classified cancers using the RNA sequencing data of tumors in the TCGA database and the KNN method. They also conducted specific research on the gender-non-specific types and pseudogenes tumors [18]. In this study, the researchers selected the genetic information of 9,096 patients with 31 tumor types in the TCGA database and 602 " normal " samples adjacent to tumors representing 17 tumor types. Researchers hope to design algorithms that can classify non-sex type tumor genes, genes from standard tissue samples near the tumor, and gender non-specific tumor genes based on these data [18]. The classification algorithm they designed is based on the extension of the k-nearest neighbor's algorithm (KNN), and it uses the genetic algorithm (GA) as the engine for gene selection. In this system, the data is randomly selected according to the training set (75%) and the test set (25%) for classification, eventually forming an optimal classifier [18]. Comparing the prediction results with actual samples can determine the accuracy of the prediction model. For non-gender, the proportional multiple (π cc) of the sample displayed by the classification model is between 90-100%. Most tumor types can be easily distinguished from other types by this model, but the similarity of the original tissue of some tumors leads to indistinguishable types [18]. At the same time, the presence of pseudogenes in tumors is relatively high, so the expression of pseudogenes can distinguish tumor types well [18]. This algorithm also clearly shows a certain degree of correctness for gender-nonspecific tumors through the π cc value. Generally speaking, gender does not affect prediction accuracy because most genes have the same contribution in different genders. Although some genes have different results on tumor types due to genders, such as FOXA1 and BNC1 [18]

For the classification of cancer, there are other algorithm applications besides the KNN classification algorithm. In 2020, Ricardo Ramirez *et al.* developed a graph convolutional neural network (GCNN) to classify cancer by modeling data [19]. This study primarily used all tumor and standard samples from the TCGA gene expression dataset. These data were generated into four GCNN models, including expression network, co-expression + singleton network, PPI network, and PPI + singleton network [19]. To improve the model's efficiency, the GCNN model uses categorical cross-entropy as the loss function and Adam optimizer. Therefore, this system consists of three steps. The comparisons were made by removing error samples by sample screening and calculating the contribution score for each gene. Finally, all data were normalized and summarized [19]. In this process, it is found that the prediction accuracy of the PPI GCNN model is the lowest, while the prediction accuracy of the co-expression GCNN model and the PPI+singleton GCNN model will be higher [19]. When using the TCGA dataset for cancer classification, the average accuracy of the classification was 99.34%. The GCNN model does not use biomarkers associated with specific tissues to classify cancer samples in this process. This method utilizes the FPKM unit and the TPM unit [19]. Meanwhile, gene regulation is usually linear or monotonic, so this study uses correlations to construct co-expression networks to contain more information [19]. This information indicates that the link between genes and gene databases can be used to build GCNN models. GCNN models can also be used to build literature-derived graphs or cancer research networks, which is very helpful for cancer classification research.

B. ML Classification Based Microarray Data

In addition to using RNA sequencing data to build predictive models, ML technology can also be done using DNA microarray data. Due to the large and complex DNA microarray data, Sara Tarek et al. developed an integrated system of multiple classifiers [20]. The system includes gene expression datasets to define the system and preprocessing modules to process the datasets. A gene selection module is also included to remove irrelevant features, reducing the complexity of classification. Traits include wrappers, filters, and embedded methods. This can be run through three feature selection algorithms, including Backward Elimination Hilbert-Schmidt Independence Criterion "BAHSIC," Extreme Value Distribution Based Gene Selection " EVD, " and Singular Value Decomposition Entropy Gene Selection "SVDEntropy." [20]. The ensemble system consists of 5 basic classifiers with the 3-NN algorithm, and sBRE can reduce the bias to improve the ensemble's diversity. Therefore, the whole system can analyze cancer through a multi-faceted algorithmic process, and the plug-in in it can reduce the error of classification. A 3-NN classifier will classify samples in this ensemble system, and semi-BRE will generate training samples to reduce computational errors based on correctly classified samples. The classifier's predictions generate the final decision, and the computed ensemble error is compared to the sensory side to compute the confusion matrix. This can be used to plot the receiver operating characteristic (ROC) [20]. The study used this system to classify leukemia, colon cancer, and breast cancer. The results show that the integrated system can quickly classify three types of oncogene data, and the classification error is less than 5%. Therefore, the system can overcome the shortcomings brought by microarray data and achieve high precision, high coverage, and low impact [20]

Some cancer classification algorithms have certain limitations in microarray cancer gene expression data due to the microarray dataset's small sample size and the data's complexity. To solve this problem, Sitan Yang and Daniel Q. Naiman presented a Top Scoring Pair (TSP) classifier of the "Top Scoring Set" (TSS) [21]. TSP is a binary classification method and can make class predictions on expression levels in multiple pairs of genes. The TSS in this classifier is a multi-class classification method and can be robust to standard microarray preprocessing transformations [21]. Compared with other standard classifiers, TSS has better classification and simpler decision-making results. The TSP classifier aggregates the predictions for each class over all the top-scoring gene pairs and then uses the majority rule to produce the final result. On the other hand, TSS augments

this ability to multi-class cases and finds the set with the highest score [21]. To test the function of this classifier, the researchers classified seven cancer gene expression data. Types include cancer subtype, tumor stage, and response to treatment. The gene dataset with the highest score obtained by the greedy search algorithm in the classifier contains 73 groups of leukemia, seven groups of MLL, three groups of bladder cancer, two groups of SRBCT and NSCLC, and 1 group of lung adenocarcinoma and ChildALL [21]. These sets have higher class separability, and it has a higher relative frequency to provide better discriminability for cancer classification. At the same time, by comparing the classification accuracy of TSS and other classifiers in the sample gene dataset, it can be found that TSS has higher classification accuracy and stability for more cancer types. The highest can reach more than 90%, and three datasets have the highest accuracy [21]. On the other hand, the system can also sequence or combine genes to obtain statistical results to apply this approach to larger gene sample sizes. The researchers combined the TSS system with gene pathway analysis to obtain complete biological information. This can be a good representation of the composition of different complex genetic material networks [21]. Therefore, this classifier can handle the constraints imposed by microarray data well and reveal a complete genetic network. Another obvious problem with using microarray data is biological noise, which can cause certain deviations in the results. In order to avoid and deal with the impact of biological noise on itemized results, scientists have proposed several microarray data analysis packages to handle heavy-tailed noise, and these applications rely on the Gaussian hypothesis [22]. These data packets and models can handle most of the biological noise generated by microarray data, thereby reducing the error of the results. This greatly reduces the result errors and limitations in microarray data, and enhances its applicability.

C. Artificial Neural Network (ANN) Technology

In addition to the several ML techniques mentioned earlier, another more traditional ML algorithm is the artificial neural network (ANN). This algorithm can classify cancers based on their gene expression signatures. Javid Khan et al. used an ANN to classify small round blue cell tumors (SRBCT) to determine the impact of this algorithm on cancer diagnosis [23]. The researchers used gene expression data from cDNA microarrays containing 6567 genes to form 63 training samples in this study. These data were processed by principal component analysis (PCA) and filtering to reduce dimensionality and gene counts. All data can be divided into 3750 models using artificial neural network technology [23]. Meanwhile, this study ranks the calibration genes and identifies 96 genes with a minimum classification error of 0%. The researchers calibrated the ANN model by using these 96 genes and used multidimensional scaling (MDS) to classify cancer samples into four categories. To determine the diagnostic ability using the ANN model, some samples were used for classification diagnosis and showed that 20 SRBCT samples and others were correctly classified. However, five non-SRBCT samples do not belong to the four types of ANN model outputs [23]. Therefore, this technique can be an adjunct to standard cancer classification and diagnosis. This study demonstrates the application of artificial neural network technology to cDNA microarray gene expression data and demonstrates its accuracy in cancer classification. This has important implications for cancer diagnosis, gene expression detection, and the discovery of new targets.

VI. Conclusion

With the development of science and technology, the gene expression information of cancer will become more and more complex. Some genes cause cells in an organism to behave out of control to cause different types of cancer, sometimes depending on where the cancer is. Because different cells have different functions, each type of cancer has different symptoms. With the in-depth study of cancer, it can be found that most patients die without timely treatment, so the correct cancer diagnosis is an important way for cancer treatment. The move from traditional microscopy to artificial intelligence (AI) is a significant advance for cancer diagnosis because it reduces the limits of human error. Machine learning (ML) is a standard classification method in artificial intelligence technology, so different algorithms have different applications in cancer classification. This development is based on generating large amounts of genetic data because ML techniques require large amounts of data to train and refine. On the other hand, cancer is a hereditary disease, so gene expression data or other genetic data are mainstream data in cancer research. NGS and microarrays are the most commonly used technologies for cancer genetic testing. These two technologies can detect gene sequences to determine gene expression and reveal information in gene expression through data analysis. At the same time, since different genes have different carcinogenic factors in cancer, this will lead to a large amount of cancer genetic data. To better use this vast and complex information, some organizations establish public cancer gene databases for human use, such as TCGA. With the support of these databases, bioinformatics researchers have created predictive models of cancer using different algorithms, such as KNN and ANN. The advent of these algorithms is of great help in cancer classification and diagnosis because they can help patients get faster and more accurate diagnoses. The advent of predictive models is an important advance for cancer research, diagnosis, and treatment because it is a convenient and error-free method and can be improved with big data.

Conflict of Interest

The author declares no conflict of interest.

References

[1] H. Sung *et al*., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, 2021. vol. 71, no. 3, pp. 209–249.

[2] E. Sahai, "Mechanisms of cancer cell invasion," *Current Opinion in Genetics & Development*, vol. 15, no. 1, pp. 87–96, 2005.

[3] R. Sager, "Expression genetics in cancer: shifting the focus from DNA to RNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 3, pp. 952–955, 1997.

[4] I. E. Tothill, "Biosensors for cancer markers diagnosis," *Seminars in Cell & Developmental Biology*, vol. 20, no. 1, pp. 55–62, 2009.

[5] B. L. Weber, "Cancer genomics," *Cancer Cell*, vol. 1, no. 1, pp. 37–47, 2002.
[6] K. Kourou *et al.*, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.
[7] I. El Naqa and M. J. Murphy, "What is machine learning?" *Machine Learning in Radiation Oncology: Theory and Applications*, I. El Naqa, R. Li, and M. J. Murphy, eds. 2015, Springer International Publishing: Cham. pp. 3–11.
[8] J. S. Reis-Filho, "Next-generation sequencing," *Breast Cancer Research*, vol. 11, no. 3, pp. S12, 2009.
[9] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends in Genetics*, vol. 24, no. 3, pp. 133–141, 2008.
[10] M. A. Quail *et al.*, "A large genome center's improvements to the Illumina sequencing system," *Nature Methods*, vol. 5, no. 12, pp. 1005–1010, 2008.
[11] F. Ozsolak and P. M. Milos, "RNA sequencing: Advances, challenges and opportunities," *Nature Reviews Genetics*, 2011, vol. 12, no. 2, pp. 87–98.
[12] K. R. Kukurba and S. B. Montgomery, "RNA sequencing and analysis," *Cold Spring Harbor Protocols*, 2015, no. 11, top084970.
[13] G. Russo, C. Zegar, and A. Giordano, "Advantages and limitations of microarray technology in human cancer," *Oncogene*, vol. 22, no. 42, pp. 6497–6507, 2003.
[14] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge," *Contemporary Oncology (Poznan, Poland)*, vol. 19, no. 1A, pp. A68–A77, 2015.
[15] P. L. Auer and R. W. Doerge, "Statistical design and analysis of RNA sequencing data," *Genetics*, vol. 185, no. 2, pp. 405–416, 2010.
[16] S. Zhao *et al*., "Bioinformatics for RNA-Seq data analysis," *Bioinformatics—Updated Features and Applications: InTech*, 2016, p. 125–149.
[17] S. L. Goldenberg, G. Nir, and S. E. Salcudean, "A new era: Artificial intelligence and machine learning in prostate cancer," *Nature Reviews Urology*, vol. 16, no. 7, pp. 391–403, 2019.
[18] Y. Li *et al*., "A comprehensive genomic pan–cancer classification using The Cancer Genome Atlas gene expression data," *BMC Genomics*, vol. 18, no. 1, p. 508, 2017.
[19] R. Ramirez *et al.*, "Classification of cancer types using graph convolutional neural networks," *Frontiers in Physics*, vol. 8, p. 203, 2020.
[20] S. Tarek, R. Abd Elwahab, and M. Shoman, "Gene expression based cancer classification," *Egyptian Informatics Journal*, vol. 18, no. 3, pp. 151–159, 2017.
[21] S. Yang and D. Q. Naiman, "Multiclass cancer classification based on gene expression comparison," *Statistical Applications in Genetics and Molecular Biology*, vol. 13, no. 4, pp. 477–496, 2014.
[22] A. Posekany, K. Felsenstein, and P. Sykacek, "Biological assessment of robust noise models in microarray data analysis," *Bioinformatics*, vol. 27, Issue 6, pp. 807–814, March 2011.
[23] J. Khan *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.