Hindawi
Computational and Mathematical Methods in Medicine
Volume 2022, Article ID 4004130, 9 pages
https://doi.org/10.1155/2022/4004130

Research Article

Big Data Analysis and Application of Liver Cancer Gene Sequence Based on Second-Generation Sequencing Technology

Chaohui Xiao,[1] Fuchuan Wang,[2] Tianye Jia,[3] Liru Pan,[1] and Zhaohai Wang[1]

[1]*Faculty of Hepato-Biliary-Pancreatic Surgery, Chinese People's Liberation Army (PLA) General Hospital, Beijing 100853, China*
[2]*Faculty of Hepatology Medicine, Chinese People's Liberation Army (PLA) General Hospital, Beijing 100039, China*
[3]*Department of Laboratory, Fifth Medical Center, Chinese People's Liberation Army (PLA) General Hospital, Beijing 100039, China*

Correspondence should be addressed to Zhaohai Wang; 161849064@masu.edu.cn

Received 24 May 2022; Revised 3 July 2022; Accepted 14 July 2022; Published 16 August 2022

Academic Editor: Shimin Wang

In big data analysis with the rapid improvement of computer storage capacity and the rapid development of complex algorithms, the exponential growth of massive data has also made science and technology progress with each passing day. Based on omics data such as mRNA data, microRNA data, or DNA methylation data, this study uses traditional clustering methods such as kmeans, K-nearest neighbors, hierarchical clustering, affinity propagation, and nonnegative matrix decomposition to classify samples into categories, obtained: (1) The assumption that the attributes are independent of each other reduces the classification effect of the algorithm to a certain extent. According to the idea of multilevel grid, there is a one-to-one mapping from high-dimensional space to one-dimensional. The complexity is greatly simplified by encoding the one-dimensional grid of the hierarchical grid. The logic of the algorithm is relatively simple, and it also has a very stable classification efficiency. (2) Convert the two-dimensional representation of the data into the one-dimensional representation of the binary, realize the dimensionality reduction processing of the data, and improve the organization and storage efficiency of the data. The grid coding expresses the spatial position of the data, maintains the original organization method of the data, and does not make the abstract expression of the data object. (3) The data processing of nondiscrete and missing values provides a new opportunity for the identification of protein targets of small molecule therapy and obtains a better classification effect. (4) The comparison of the three models shows that Naive Bayes is the optimal model. Each iteration is composed of alternately expected steps and maximal steps and then identified and quantified by MS.

1. Introduction

Next-generation sequencing (NGS), also known as high-throughput sequencing or massively parallel sequencing, is a technology that can sequence thousands to billions of DNA fragments simultaneously and independently. The dideoxynucleoside end-termination sequencing method began in the 1970s. In the follow-up continuous improvement, the Sanger method caused a sequencing boom and became the mainstream due to its simplicity and rapidity. In order to meet the increasingly complex research needs, next-generation sequencing technology emerges from time to time [1–3]. Using NGS technology to detect a variety of cancers, and compared the results with the Sanger method, it was found that in addition to common gene mutations, NGS technology can also detect many gene mutations that were ignored by real-time quantitative PCR detection technology. It may play a prompting and guiding role in the occurrence and development of cancer and the diagnosis and treatment of patients and also reflects the value of using NGS technology in clinical work. There are currently three mainstream NGS platforms: Roche454, Ion Torrent, and Illumina platforms. The Roche454 platform is based on the pyrosequencing method, that is, bases are incorporated in the order of T, A, C, and G during sequencing, and pyrophosphate is released after pairing. The Ion Torrent platform is the semiconductor sequencing technology. The ion sensor can detect the pH change caused by proton release during the synthesis process and then judge the sequence of the base. NGS technology detection programs have different focuses, including whole genome sequencing (WGS), which can detect all genetic changes and conduct a comprehensive analysis of tumor-related genes, but it is costly and time-consuming. Whole exome sequencing (WES), which only detects the coding gene regions, is more economical and can detect already known mutant coding genes and discover new gene mutations in cancer. Whole transcriptome sequencing, based on cDNA sequence sequencing, can detect information about the overall transcriptional activity [4]. Targeted target sequencing can select some genes required for disease research for higher sequencing efficiency, but it is not suitable for detecting unknown mutations [5–7]. The techniques of experimental manipulation (wet experiment) and bioinformatics analysis (dry experiment) have been developed continuously. NGS technology is widely used in solid tumors, and more new gene mutations have been discovered, providing new ideas for the detection of genetic susceptibility and the guidance of individualized precision medicine, and have played an extremely important role in the study of the genetic pathways of human malignant tumor mutations effect. Liver cancer is one of the most common cancers in cancer patients today [8–10]. According to the 2020 report by the American Cancer Society, there are an estimated 42,810 new cases and 30,160 deaths from the liver and intrahepatic cholangiocarcinoma in the United States throughout the year. Statistics at home and abroad show that liver cancer is an important cause of cancer death worldwide, and the treatment of liver cancer is also an urgent problem to be solved. Liver cancer is divided into two types: primary and secondary. Primary liver cancer (PLC) is the most common. From a histological point of view, primary liver cancer can be divided into different subtypes according to the cell origin, hepatocellular carcinoma (HCC) (about 75-85% of all cases), intrahepatic cholangiocarcinoma (ICC) (about 75% of all cases), intrahepatic cholangiocarcinoma (ICC) (10-15%), and other rare forms. Hepatocellular carcinoma has become the main type of liver cancer research. The main known carcinogens of liver cancer are hepatotropic virus: mainly chronic infection with hepatitis B (HBV) and hepatitis C (HCV) virus; chemical stimulation: such as alcohol abuse and aflatoxin; metabolic abnormalities: diabetes and nonalcoholic fatty liver disease, hereditary hemochromatosis, etc.; immune-related causes: such as cirrhosis-related immune dysfunction syndrome (CAID) and autoimmune hepatitis; etc. Among them, viral infection is the main factor causing liver cancer [11–13]. Hepatocellular carcinoma cells have extensive heterogeneity from undesired lesions caused by a small number of mutations to eventually develop into an advanced form of the disease. Because the factors that induce liver cancer are diverse and the distribution in different countries and regions is different, the molecular mechanism of liver cancer is complicated. In a broad sense, liver cancer is divided into two categories: proliferative and nonproliferative. The proliferative type is common in HBV-induced liver cancer, with low degree of differentiation, high alpha-fetoprotein (AFP) expression, more vascular invasion, and worse prognosis; this type of liver cancer is characterized by increased inactivating mutations in TP53 and AXIN1, and at the same time, cell cycle, mTOR, RAS-MAPK, and MET signaling pathways that promote survival are all activated. The nonproliferative class is commonly seen in HCV and alcohol-related hepatocellular carcinoma, with moderate or high differentiation, low AFP expression, less aggressiveness, and chromosomal stability. This type of hepatocellular carcinoma is characterized by more heterogeneity, higher frequency of CTNNB1 (β-catenin) activating mutations, and TERT promoter mutations, as well as activation of WNT and IL6/JAK-STAT signaling pathways. However, these commonly mutated genes TP53, AXIN1, CTNNB1, and TERT in liver cancer proved to be difficult to target [14, 15]. At present, liver resection and liver transplantation have become the main treatment methods for patients with early-stage liver cancer, and patients with intermediate-stage liver cancer are often treated with hepatic arterial chemoembolization and radioembolization, which can greatly prolong the survival of patients. However, due to the lack of specific symptoms and tumor biomarkers, most HCC patients are diagnosed at an advanced stage, so these curative treatments are not suitable. Sorafenib, a multi-receptor tyrosine kinase inhibitor, was identified as a therapeutic drug with survival benefits for patients with advanced liver cancer. Multiple drugs have since been shown to have clinical efficacy, including other RTK inhibitors such as lenvatinib, regorafenib, and cabozantinib. The liver is an important organ that removes toxins and regulates blood sugar, fat, and amino acid uptake. Similar to all cancers, the gradual accumulation of genetic and epigenetic changes in the liver, accompanied by a large number of metabolic changes, leads to abnormal proliferation of mature hepatocytes and the evolution of liver cancer.. Due to the high heterogeneity of liver cancer cells and the complex pathogenic factors caused by the involvement of various signaling pathways, it is clinically found that using a unified treatment regimen to treat all patients may have different curative effects and may even exacerbate symptoms. Therefore, "personalized medicine" is the development direction of contemporary treatment of liver cancer, and different therapeutic methods based on molecular and cell therapy have also been developed. The emerging molecular-level therapeutic strategies include molecular targeted therapy, targeted radionuclide therapy, and epigenetic modification-based therapy, which provide new strategies for the treatment of liver cancer.

2. Big Data Analysis of Liver Cancer

2.1. Disease Diagnosis of Omics Big Data. Cancer subtype classification methods based on omics data mainly include subtype classification methods based on single omics data and subtype classification methods based on multiomics data fusion. The former is based on an omics data such as mRNA data, microRNA data, or DNA methylation data and uses traditional clustering methods such as kmeans, K-nearest neighbors, hierarchical clustering, affinity propagation, and nonnegative matrix decomposition to classify samples. The results of subtype classification of cancer are obtained, as shown in Figure 1. With the development of related technologies, the collection of omics data has

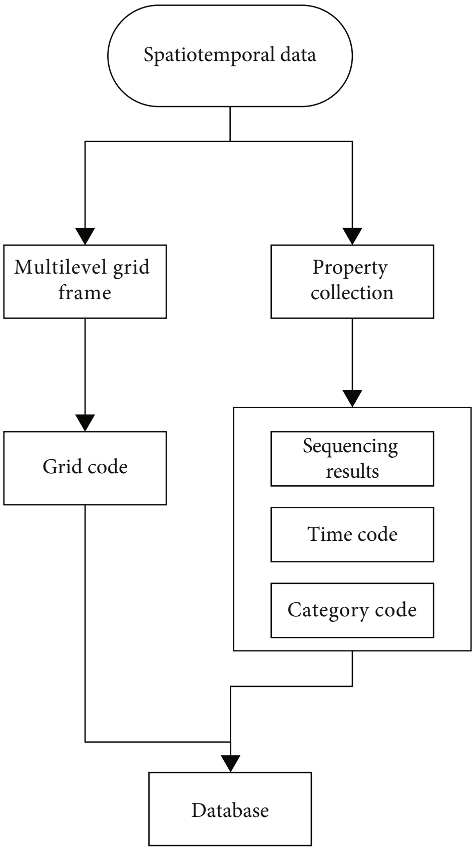

Figure 1: Conceptual model.

shown an explosive trend, and its acquisition cost has been greatly reduced. A large number of genomics, transcriptomics, proteomics, and other data of different cancer patients are given in the database headed by TCGA omics data. Since different omics data can describe the complex life process in cancer cells and the interactions between various molecules from different perspectives, the information is complementary, and the integrated analysis of multiomics data can identify more accurate and reasonable subtype results. In recent years, research has mainly focused on the field of multiomics data integration analysis methods. Most existing integrative analysis methods need to address problems closely related to biological data. The data sample size is small, and the dimensionality is high (the so-called curse of dimensionality problem). When the data scope and data type are not consistent, the underlying omics-specific and between-group data structures are easily overlooked in multi-omics data. Divided from the data supported by integrative analysis methods, existing integrative analysis methods include general methods that can analyze any multiomics data and specialized methods designed only for specific data types. The former can be applied to any multiomics data and can be easily extended to the analysis of more omics data, while the latter requires the use of known biological relationships (such as the association between copy number changes and gene expression profiles), which can only be analyzed specific data types.

2.2. Epigenetics of Liver Cancer. Abnormal epigenetic changes are important etiologies for the occurrence, development, and metastasis of liver cancer. Epigenetics, the heritable modification of gene function without altering the DNA sequence, is caused by many different factors. Epigenetic alterations are often present in liver cancer. The screening process is shown in Figure 2. Epigenetic processes include, but are not limited to, chromatin remodeling, histone modification, DNA methylation, and expression of noncoding RNAs. Unlike the irreversible nature of genomic alterations, epigenetic changes are reversible, opening a promising avenue for the development of new therapeutic modalities. Therefore, epigenetic changes associated with cancer and liver cancer are gradually being widely used in the development of biomarkers. Hepatocellular carcinoma (HCC) is one of the most common liver tumors and has become the leading cause of cancer-related death in many regions and countries. Although many measures have been taken to prevent, early screen, diagnose, and treat liver cancer, the current situation of liver cancer in my country is still not optimistic.

3. Algorithm Model

3.1. Naive Bayes [16–20]. Semiconductor sequencing

$$\begin{aligned} X' &= \left\{ \left(x'_1, y'_1\right), \left(x'_2, y'_2\right), \cdots, \left(x'_n, y'_n\right) \right\} X' \\ &= \left\{ \left(x'_1, y'_1\right), \left(x'_2, y'_2\right), \cdots, \left(x'_n, y'_n\right) \right\}, \\ x_{\text{new}} &= x_i + \text{rand}\,(0,1) \times \left(x'_j - x_i\right). \end{aligned} \tag{1}$$

Ion sensor, D is the matrix reconstruction function, x is the group sparse constraint function, and y is the weight of the group sparse constraint term.

$$\begin{gathered} D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}, \\ h_i = \varepsilon(N, m). \end{gathered} \tag{2}$$

Unknown genome sequence

$$\begin{gathered} H(x) = \arg_{\max}^{y \in Y} \sum_{t=1}^{T} (I(h(x) = y), \\ P_i = (x_i, y_i, z_i)\ i = 1, 2, 3, 4 \cdots. \end{gathered} \tag{3}$$

Nucleic acid fragments are sequenced, *Row* is the number of clusters, y_i is the index of the sample belonging to the ith category, and r is the sparse constraint function.

$$\begin{gathered} \text{Row} = \frac{2y_i}{r}, \\ \text{Column} = \frac{2x_i}{r}\ i = 1, 2, 3, 4 \cdots, \\ Z = \sum_{i=1}^{n} z_i. \end{gathered} \tag{4}$$

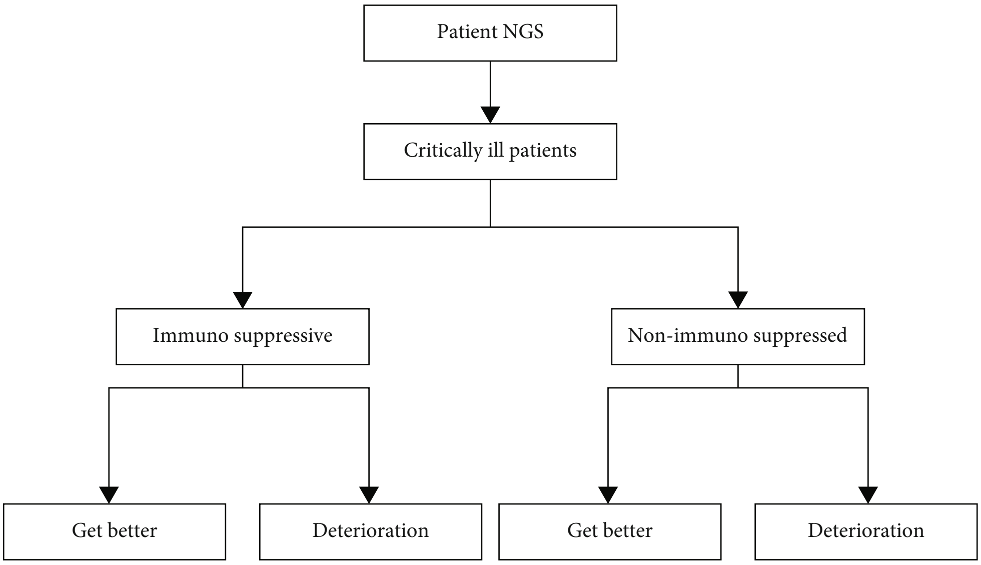

Figure 2: Screening process.

Table 1: Differences in the expression of hepatocellular carcinoma-related genes.

Gene	GeneBank	LogFC	FDR	Gene	GeneBank	LogFC	FDR
CXCL11	NM_005409.5	1	2.51	IDO1	NM_002164.6	4	2.44
KLK3	NM_001648.2	10	2.07	SLAM	NM_020125.3	2	2.01
WARS	NM_004184.4	2	2.8	OLFM	NM_006418.5	8	2.48
GBP4	NM_052941.5	8	2.77	APOC1	NM_001645.5	10	2.02
CD300LF	NM_139018.5	2	2.29	KRT81	NM_002281.4	8	2.22
ZNF683	NM_001114759.3	9	2.61	FCGR1B	NR_164759.1	5	2.06
CXCL10	NM_001565.4	8	2.35	OR2B2	NM_033057.2	7	2.8
LAM	NM_014398.4	3	2.98	CETN1	NM_004066.3	10	2.78
GBP5	NM_052942.5	1	2.9	RUNX1	NM_001754.5	5	2.65
ELANE	NM_001972.4	3	2.69	DNASE2B	NM_021233.3	4	2.29
EPSTI1	NM_001002264.4	8	2.35	IFI30	NM_006332.5	3	2.1
HAPLN3	NM_001307952.2	9	2.6	DM	NM_021951.3	1	2.02

3.2. Okumura-Hata [21–23]. Assembly and splicing

$$X = \frac{\sum_{i=1}^{n} x_i z_i}{\sum_{i=1}^{n} z_i}. \tag{5}$$

Coding gene regions for detection

$$\begin{aligned} Y &= \frac{\sum_{i=1}^{n} y_i z_i}{\sum_{i=1}^{n} z_i}, \\ L(db) &= 69.55 - 13.82 \log h_b + (44.9 - 6.55 h_b) \\ &\quad \times \log d + 26.16 \log f_c - \alpha(h_m), \\ \alpha(h_m) &= (1.11 \log f_c - 0.7) h_m - (1.56 f_b - 0.8), \\ \gamma(f_c) &= 8.29 (\log 1.54 h_m)^2 - 1.1. \end{aligned} \tag{6}$$

Capture probe hybridization

$$f = 3.2 (\log 11.75 h_m)^2 - 4.97. \tag{7}$$

Quality control

$$C_{\text{cell}} = -2 \left[\log \left(\frac{f_c}{28} \right) \right]^2 - 5.4. \tag{8}$$

Data filtering

$$\begin{aligned} \sigma_{\text{cell}} &= -4.78 (\log f_c)^2 - 18.33 \log f_c - 40.98, \\ P(dB) &= 13.82 \log h_b + (44.9 - 6.55 \log h_b) \\ &\quad \times \log d + 33.9 \log f_c - \alpha(h_m). \end{aligned} \tag{9}$$

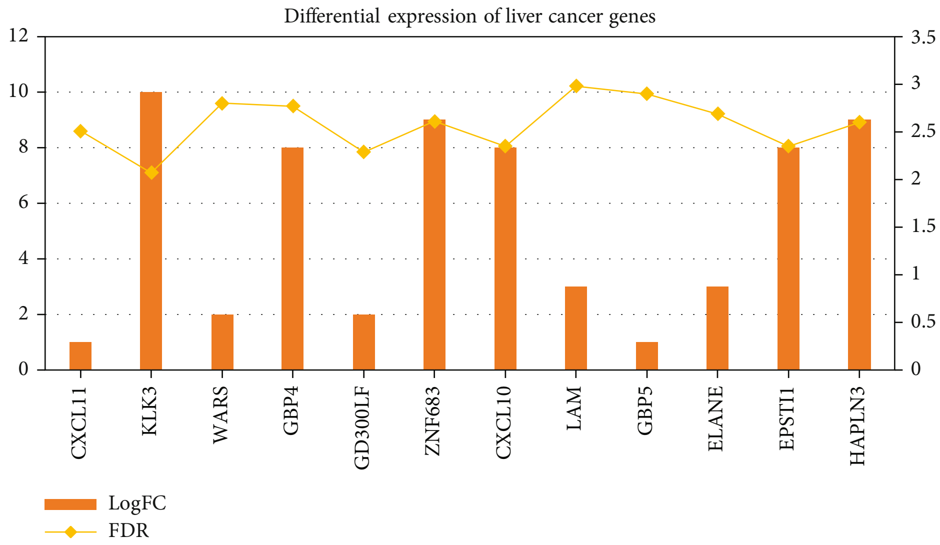

Figure 3: Differences in the expression of hepatocellular carcinoma-related genes.

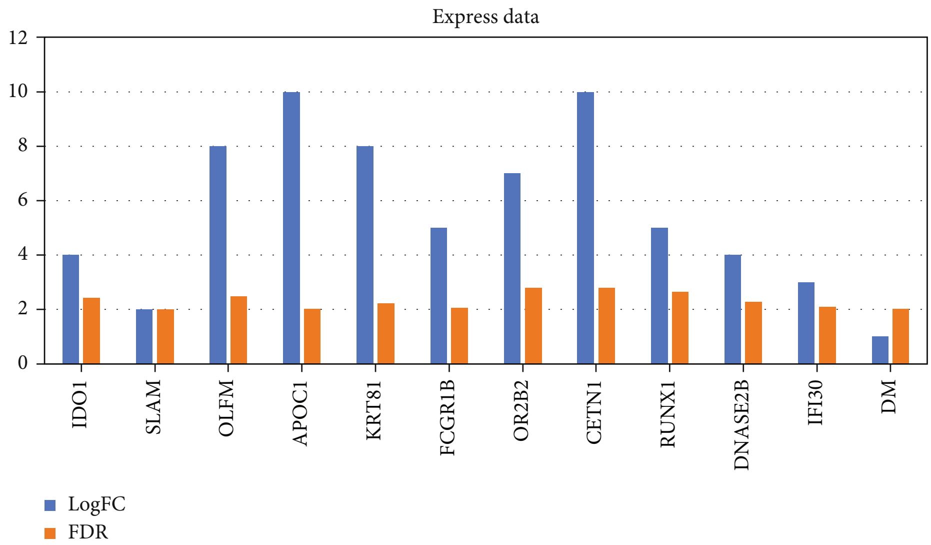

Figure 4: Expression data.

Table 2: Test point properties.

	EARFCN_1	PCI_1	RSRP_1	EARFCN_2	PCI_2	RSRP_2
Collect_Time	2.32	12.13	12	5.15	13	9
IMEI	2.45	3.37	15	4.24	3	4
LAT	2.79	18.84	11	3.23	19	5
LNG	2.52	10.18	14	2.32	20	4
ECI	2.55	11.95	15	2.27	20	19
EARFCN	2.52	13.83	17	8.64	13	10
PCI	2.31	7.31	17	5.35	2	3
RSRP	2.63	6.85	14	9.77	19	17

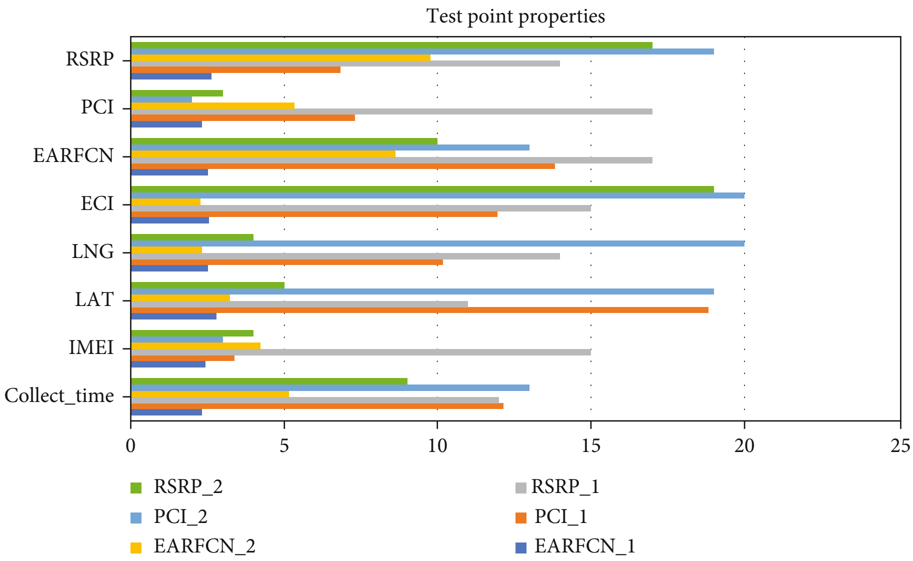

Figure 5: Test point properties.

Table 3: Classification algorithm training.

	CRS	EPSG	Extent	Level
M2SMF	3	8	15	2.43
SNF	1	5	14	7.86
PAM50	4	5	15	2.15
iCluster	4	10	16	4.71
kmeans	4	5	13	5.44
pins	2	6	17	2.59
MCCA	2	5	12	7.63

3.3. AdaBoost [24–27].

$$L(dB) = k_1 + k_2 \log d + k_3 \log h_b + k_4 Diff_{\text{loss}}, \quad x_i = \max\left(B\_loc_i, B\_Bore_i\right). \tag{10}$$

For sequence alignment, f_i is a shared implicit expression matrix, and x, k are modality-specific basis matrices.

$$f_i = \mathrm{SPM}(x_i, k_1, k_2), \quad \langle k_1, k_2 \rangle = \operatorname{argmin} \sum_{i=1}^{N} (y_i - f_i)^2. \tag{11}$$

Subtype classification of omics data

$$\mathrm{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - y_i'}{y_i} \right|. \tag{12}$$

Multiomics data fusion

$$\mathrm{MAE} = \frac{1}{n} \left(\sum_{i=1}^{n} \left| y_i - y_i' \right| \right). \tag{13}$$

Subtype classification results

$$A_H(\varphi) = -\min \left[12 \left(\frac{\varphi}{\varphi_{3dB}} \right)^2, A_m \right]. \tag{14}$$

Category division

$$A_v(\theta) = -\min \left[12 \left(\theta - \frac{\theta_{\text{etilt}}}{\theta_{3dB}} \right)^2, \mathrm{SLA}_v \right], \quad A(\varphi, \theta) = -\min \left\{ -\left[A_H(\varphi), A_v(\theta) \right], A_m \right\}. \tag{15}$$

4. Simulation Experiment

4.1. Big Data Analysis of Liver Cancer Sequencing Results. The coding of the multilevel grid adopts a simple construction by assuming that the attributes of a given target value are conditionally independent of each other. It can realize the space filling of the data dimensionality reduction map, and then learn the joint probability distribution from the input to the output from the training data set. In the calculation process, it is represented by binary values 0 and 1, and input the feature data set of the unknown category to obtain the output category vector that maximizes the posterior probability. Indexing is also commonly used in Geohash encoding algorithms. There are shown in Table 1, Figures 3 and 4. The assumption that the attributes are independent of each other reduces the classification effect of the algorithm to a certain extent. According to the idea of multilevel grid, the mapping from high-dimensional space to one-dimensional space is one-to-one. The complexity is greatly simplified by encoding the one-dimensional grid of the hierarchical grid. The logic of the algorithm is relatively simple, and it also has a very stable classification efficiency. The grid division and coding rules are all calculated from the grid definition. The calculation process is to bisect the longitude

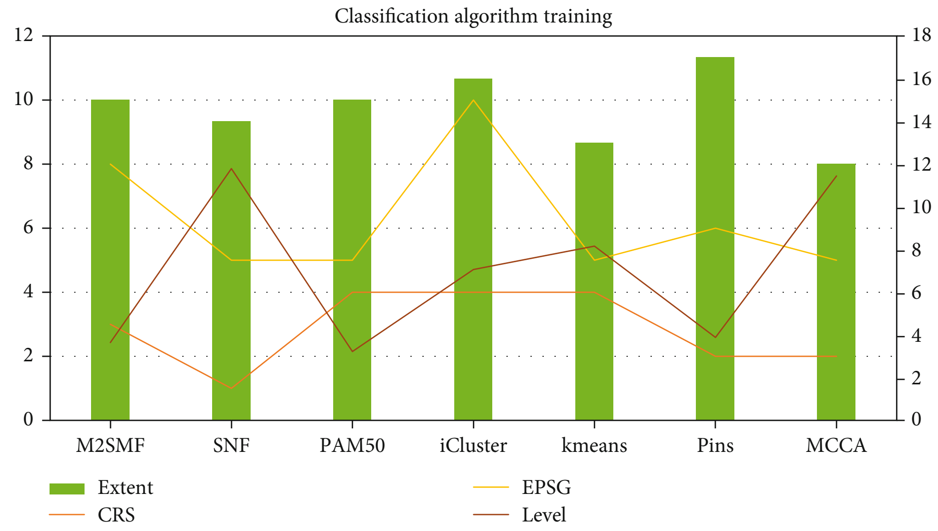

Figure 6: Classification algorithm training.

Table 4: Algorithm optimization.

	Naive Bayes	Okumura-Hata	AdaBoost
LAML	0.92	0.03	0.38
KIRC	0.96	0.37	0.07
LIHC	0.67	0.5	0.55
M2SMF	0.71	0.32	0.54
SNF	0.52	0.49	0.27
MCCA	0.67	0.37	0.51
PINS	0.68	0.8	0.36

values of the grid, and the data sets containing missing values are not sensitive.

4.2. Encoding Process of Liver Cancer Data. Big data has three invisible connotations of space, time, and semantics. As shown in Table 2 and Figure 5, convert the actual spatial position of the data to the position in the global multilevel grid, so that the two-dimensional representation of the data is converted into a binary one-dimensional representation, Collect_Time = 5.15, IMEI = 4.24, LAT = 3.23, LNG = 2.32, ECI = 2.27, EARFCN = 8.64, and PCI = 5.35. It realizes the dimensionality reduction processing of data and improves the efficiency of data organization and storage. Grid coding expresses the spatial position of data, maintains the original organization method of data, and does not abstract data objects. Collect_Time = 13, IMEI = 3, LAT = 19, LNG = 20, ECI = 20, EARFCN = 13, and PCI = 2. Instead, the expression is converted again, that is, the way of grid code identification. According to the actual area range of the object, the regional characteristics of the data are expressed as grid units, and the final grid code is composed of the codes of the grid units.

4.3. Naive Bayes Algorithm Training on Sequencing Data. The Naive Bayes algorithm is a classification algorithm that, in a genome screening-based approach, sequenced the human genome and many model organism genomes. As shown in Table 3 and Figure 6, for data processing that is not discretized and contains missing values, M2SMF = 3, SNF = 1, PAM50 = 4, iCluster = 4, kmeans = 4, pins = 2, and MCCA = 2. It provides new opportunities for the identification of protein targets for small molecule therapeutics. With better classification results, new chemical genomics and genomics approaches link small molecules to their protein targets. Chemical proteomic methods may also facilitate the identification of protein targets, M2SMF = 8, SNF = 5, PAM50 = 5, iCluster = 10, kmeans = 5, pins = 6, and MCCA = 5. It is used for scenarios such as efficient classification of multidimensional feature data. It will favor feature data with more attribute values and use drug affinity chromatography combined with mass spectrometry and computational analysis to classify whole protein small molecule-protein interactions. In the compound-centric chemical proteomics method, which has an impact on the construction of the decision tree and the final classification effect, the molecules are fixed on a substrate to maintain their activity and improve the accuracy of the algorithm.

4.4. Iterative Optimization. The EM expectation-maximum algorithm is an iterative optimization algorithm followed by incubating the cell lysate of interest with an affinity matrix. Looking for parameter maximum likelihood estimates, eluted proteins were processed without gel. As shown in Table 4 and Figure 7, the comparison of the three models shows that Naive Bayes is the best model, LAML = 0.92, KIRC = 0.96, LIHC = 0.67, M2SMF = 0.71, SNF = 0.52, MCCA = 0.67, and PINS = 0.68. Each iteration consists of alternating expected and maximal steps, which are then identified and quantified by MS. An advantage of chemical proteomics is the ability to probe the entire proteome until convergence ends. In the Okumura-Hata model, LAML = 0.03, KIRC = 0.37, LIHC = 0.5, M2SMF = 0.32, SNF = 0.49, MCCA = 0.37, and PINS = 0.8. Small molecules that encounter and interact with these proteins in their natural state and environment serve as a data addition algorithm. Another advantage of this is that proteomics can be tested

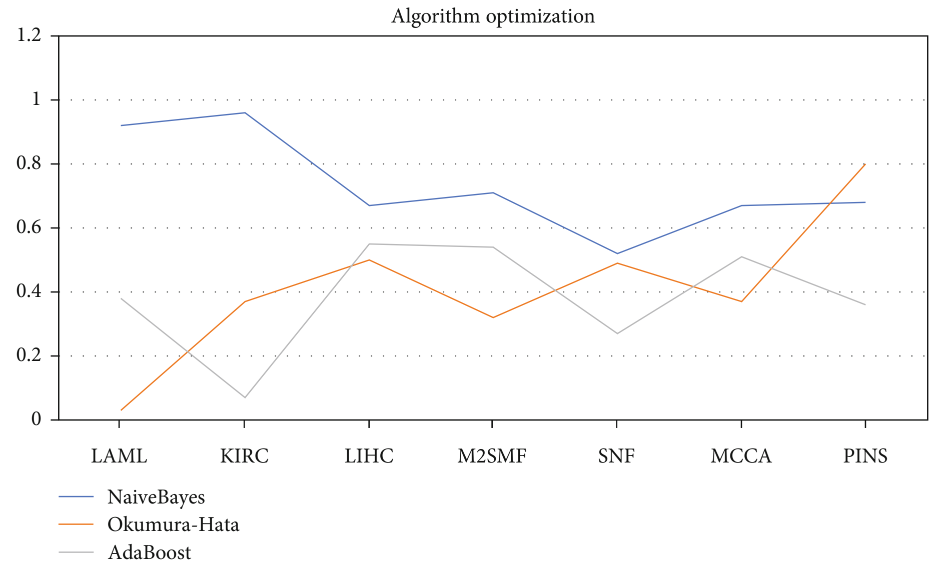

Figure 7: Algorithm optimization.

in any cell type or tissue of interest, guaranteeing a steady rise in parameter value estimates over an iterative process. AdaBoost model has the worst effect, LAML = 0.38, KIRC = 0.07, LIHC = 0.55, M2SMF = 0.54, SNF = 0.27, MCCA = 0.51, and PINS = 0.36. Iterative optimization can analyze the trend of information gain rate from a large number of classification algorithm data. Therefore, the decision tree model is constructed according to the selection attributes, which is the core key to solving complex problems. The top-down recursive solution is accurate and complete. The rules for mapping attribute values to categories are a series of clear instructions for solving problems.

5. Conclusion

In big data analysis with the rapid improvement of computer storage capacity and the rapid development of complex algorithms, the exponential growth of massive data has also made science and technology progress with each passing day. Based on omics data such as mRNA data, microRNA data, or DNA methylation data, this study uses traditional clustering methods such as kmeans, K-nearest neighbors, hierarchical clustering, affinity propagation, and nonnegative matrix decomposition to classify samples into categories, and obtained: (1) The assumption that the attributes are independent of each other reduces the classification effect of the algorithm to a certain extent. According to the idea of multilevel grid, the mapping from high-dimensional space to one-dimensional space is one-to-one correspondence. The complexity is greatly simplified by encoding the one-dimensional grid of the hierarchical grid. The logic of the algorithm is relatively simple, and it also has a very stable classification efficiency. (2) Convert the two-dimensional representation of the data to the one-dimensional representation of binary, Collect_Time = 5.15, IMEI = 4.24, LAT = 3.23, LNG = 2.32, ECI = 2.27, EARFCN = 8.64, and PCI = 5.35. It realizes the dimensionality reduction processing of data and improves the efficiency of data organization and storage. The grid coding expresses the spatial position of the data, maintains the original organization method of the data, and does not make the abstract expression of the data object. (3) For data processing that is not discretized and contains missing values, M2SMF = 3, SNF = 1, PAM50 = 4, iCluster = 4, kmeans = 4, pins = 2, and MCCA = 2. It provides a new opportunity for the identification of protein targets of small molecule therapy and obtains a better classification effect. Chemical proteomics methods may also facilitate the identification of protein targets, M2SMF = 8, SNF = 5, PAM50 = 5, iCluster = 10, kmeans = 5, pins = 6, and MCCA = 5, for multidimensional feature data analysis. It will favor feature data with more attribute values. (4) The comparison of the three models shows that Naive Bayes is the optimal model, LAML = 0.92 , KIRC = 0.96, LIHC = 0.67, M2SMF = 0.71, SNF = 0.52, MCCA = 0.67, and PINS = 0.68. Each iteration consists of alternating expected and maximal steps, which are then identified and quantified by MS. An advantage of chemical proteomics is the ability to probe the entire proteome until convergence ends.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

Authors' Contributions

Chaohui Xiao and Fuchuan Wang contributed equally to this work.

References

[1] Y. Deng, Y. Wang, and F. Jia, “Tailoring supramolecular prodrug nanoassemblies for reactive nitrogen species-potentiated chemotherapy of liver cancer,” *ACS Nano*, vol. 19, no. 4, p. 1, 2020.

[2] Y. Zhu and W. Siqi, “Integrative analysis of long extracellular RNAs reveals a detection panel of noncoding RNAs for liver cancer,” *Theranostics*, vol. 11, no. 1, pp. 181–193, 2021.

[3] F. Ramadhani, S. Kang, R. Kawala, B. Chung, H. W. Bai, and B. Kang, “γ-irradiated prednisolone promotes apoptosis of liver cancer cells via activation of intrinsic apoptosis signaling pathway,” *Molecular Medicine Reports*, vol. 23, no. 6, 2021.

[4] X. Chen, X. Wu, and G. Liu, “Structure analysis of ethyl ferulate from Rubus corchorifolius L.f. leaves and its inhibitory effects on HepG2 liver cancer cells,” *Food Bioscience*, vol. 3, article 101340, 2022.

[5] D. Gnocchi, F. Castellaneta, G. Cesari, G. Fiore, C. Sabbà, and A. Mazzocca, “Treatment of liver cancer cells with ethyl acetate extract of Crithmum maritimum permits reducing sorafenib dose and toxicity maintaining its efficacy,” *Journal of Pharmacy and Pharmacology*, vol. 73, no. 10, pp. 1369–1376, 2021.

[6] T. Shang, Z. Zhao, X. Ren, and J. Liu, “*Differential identifiability clustering algorithms for big data analysis*,” *Science China Information Sciences*, vol. 64, no. 5, p. 18, 2021.

[7] G. Zhu, A. Murshed, H. Li et al., “O-GlcNAcylation enhances sensitivity to RSL3-induced ferroptosis via the YAP/TFRC pathway in liver cancer,” *Cell Death Discovery*, vol. 7, no. 1, p. 83, 2021.

[8] Y. Cheng and Y. Song, “*Sports big data analysis based on cloud platform and its impact on sports economy*,” *Mathematical Problems in Engineering*, vol. 2021, 12 pages, 2021.

[9] Y. F. Sun, A. P. Lu, L. Zhuo et al., “Quality big data analysis and management based on product satisfaction index,” *IOP Conference Series: Materials Science and Engineering*, vol. 1043, no. 3, article 032004, 2021.

[10] J. Bashingwa, D. Mohan, S. Chamberlain et al., “Assessing exposure to Kilkari: a big data analysis of a large maternal mobile messaging service across 13 states in India,” *British Medical Journal Global Health*, vol. 6, article e005213, Suppl 5, 2021.

[11] M. T. Odenkirk, D. M. Reif, and E. S. Baker, “Multiomic big data analysis challenges: increasing confidence in the interpretation of artificial intelligence assessments,” *Analytical Chemistry*, vol. 93, no. 22, pp. 7763–7773, 2021.

[12] G. P. Diller and H. Baumgartner, “Impact of adequate provision of care models and big data analysis for adults with congenital heart disease,” *Aktuelle Kardiologie*, vol. 10, no. 5, pp. 403–407, 2021.

[13] G. Brancaccio, R. Salpini, L. Piermatteo et al., “An increase in the levels of middle surface antigen characterizes patients developing HBV-driven liver cancer despite prolonged virological suppression,” *Microorganisms*, vol. 9, no. 4, p. 752, 2021.

[14] L. Li, L. Niu, N. Guo et al., “*Identification of driver genes in primary liver cancer by integrating NGS and TCGA mutation data*,” *Open Journal of Gastroenterology*, vol. 12, no. 1, pp. 1–18, 2022.

[15] M. Ramamurthy, I. Krishnamurthi, S. Vimal, and Y. H. Robinson, “Deep learning based genome analysis and NGS-RNA LL identification with a novel hybrid model,” *Biosystems*, vol. 197, article 104211, 2020.

[16] J. Chen, M. Quan, Z. Yan et al., “Distribution of BRCA1/2 germline and somatic alterations across cancer type,” *Journal of Clinical Oncology*, vol. 39, 15_suppl, pp. 10590–10590, 2021.

[17] B. J. Munarriz, Z. Dantes, and J. D. Castro, “Patient-derived lung cancer organoids for the selection of therapeutic options in an ALK-rearranged tumor,” *Journal of Clinical Oncology*, vol. 39, no. 15, pp. e21014–e21014, 2021.

[18] K. A. Shastry and H. A. Sanjay, “Cancer diagnosis using artificial intelligence: a review,” *Artificial Intelligence Review*, vol. 55, no. 4, pp. 2641–2673, 2021.

[19] C. F. Teng, H. W. Tsai, T. C. Li et al., “Detection of hepatitis B virus pre-S mutants in plasma by a next-generation sequencing-based platform determines their patterns in liver tissues,” *PLoS One*, vol. 15, no. 6, article e0234773, 2020.

[20] H. Sulaiman, “Perbandingan Algoritma Decision Tree C4.5 Dan Naive Bayes pada Analisa Tekstur Gray Level Co-Occurrence Matrix Menggunakan Citra Wajah,” *SISTEMASI*, vol. 10, no. 2, Article ID 470, 2021.

[21] F. X. Goudot, S. Msadek, T. Boukertouta, P. O. Schischmanoff, and C. Meune, “Routine use of natriuretic peptides: lessons from a big data analysis,” *Annals of Clinical Biochemistry*, vol. 58, no. 5, pp. 481–486, 2021.

[22] M. Wang and M. Qian, “Research and application of the network learning platform in based on big data analysis and mobile communication,” *Journal of Physics Conference Series*, vol. 1952, no. 4, article 042054, 2021.

[23] J. W. Park, S. Lee, and T. Hong, “The implementation and performance analysis of container-based HPC cluster system for big data analysis and machine learning services,” *KIISE Transactions on Computing Practices*, vol. 27, no. 5, pp. 228–233, 2021.

[24] J. Bai, “Research on brand design and promotion strategy of health food in Changbai Mountain region based on the big data analysis under the background of big health industry,” *Journal of Physics Conference Series*, vol. 1744, no. 3, article 032039, 2021.

[25] T. G. Kim and S. Yu, “Big data analysis of the risk of intracranial hemorrhage in Korean populations taking low-dose aspirin,” *Journal of Stroke and Cerebrovascular Diseases*, vol. 30, no. 8, article 105917, 2021.

[26] B. Pellini, N. Pejovic, P. Harris et al., “Urinary cell-free DNA analysis for tumor mutation detection in patients with oligometastatic colorectal cancer,” *Journal of Clinical Oncology*, vol. 38, 15_suppl, pp. e15545–e15545, 2020.

[27] H. Guan and X. Zhao, “*Study on the prediction system of shrimp field distribution in the East China Sea based on big data analysis of fishing trajectories*,” *Journal of Ocean University of China*, vol. 20, no. 1, pp. 228–234, 2021.