ALGORITHM FOR EARLY DIAGNOSIS OF HEPATOCELLULAR CARCINOMA BASED ON GENE PAIR SIMILARITY

Zarifa Jabrayilova
Doctor of Technical Sciences, Associate Professor
Chief Researcher[1]
djabrailova_z@mail.ru

Lala Garayeva
Junior Researcher[1]

[1]*Department of Number 11*
Institute of Information Technology
MInstry of Science and Education Republic of Azerbaijan
9 B Vahabzada str., Baku, Azerbaijan, AZ1141

Abstract: The article proposes an algorithm based on intelligent methods for the early diagnosis of hepatocellular carcinoma (HCC), known as liver cancer, which is rated third cause of cancer deaths in the world. Initial diagnosis of HCC is based on laboratory studies, computer tomography and X-ray examination. However, in some cases, identifying cancerous tissues as similar non-cancerous tissues (cirrhotic tissues and normal tissues) made it necessary to perform gene analysis for the diagnosis. To predict HCC based on such numerous, diverse and heterogeneous unstructured data, preference is given to the method of artificial intelligence, i.e., machine learning. It shows the possibility of applying machine learning methods to solve the problem of accurate identification of HCC due to the compatibility of HCC tissues with identical CwoHCC non-cancerous tissues. The technology of gene pair profiling using relevant peer databases is described and the Within-Sample Relative Expression Orderings (REO) technique is used to determine the gene pair's similarity. The article also presents a new approach based on The Within-Sample Relative Expression Orderings technique for determining the gene pair's similarity, Incremental feature selection method for feature selection, and Support Vector Machine methods for gene pair classification. The proposed approach constitutes the methodological basis of a decision support system for the early diagnosis of HCC, and the development of such a system may be beneficial for physician decision support in the relevant field.

Keywords: hepatocellular carcinoma, gene pairs, HCC cancer tissue, machine learning algorithms, similarity of gene pairs.

1. Introduction

The most successful work towards the creation of knowledge-based intelligent systems is currently observed in the medical segment [1]. Medical decision support systems have become a more effective tool for collecting, storing, and manipulating the knowledge of qualified medical experts, as well as for determining the disease and making adequate decisions for each specific data set. These systems are based on the diseases in the specific subject area of medicine, their possible causes, development periods, clinical manifestations, observed signs, symptoms, etc. These systems are applied to make a diagnosis, choose a more effective treatment method, predict, search for suitable situations (precedents), control and plan therapy, recognize and interpret images, monitor the clinical-pharmacological properties (toxicity) of drugs, etc.

This article shows the possibilities of applying machine learning algorithms in the decision support system to be created for physicians regarding the early diagnosis of Hepatocellular Carcinoma (HCC).

2. Methods

The number of patients who die from liver cancer ranks third compared to those diagnosed with malignant cancer [2]. HCC, which accounts for about 90 % of all liver cancer cases, is often diagnosed in the late stages of the disease and therefore causes a high risk of death. Consequently, early diagnosis of HCC is very important in the disease prevention and increases the patient's survival probability.

Currently, the diagnosis of HCC is based on laboratory studies and computed tomography (CT), and X-ray examination [3]. Liver biopsy is estimated to be a good diagnostic option in the clinical condition when CT and X-ray examination cannot provide accurate identification of HCC [4]. Sometimes, non-cancerous tissues (cirrhotic tissues and normal tissues) containing some common molecular features of cancerous tissues are recognized as cancerous tissues [5]. In such cases, gene analysis signatures are included among the available diagnostic signatures to eliminate the risk factor. Gene analysis signatures have a batch effect and are difficult to determine in clinical conditions [6]. Optimizing treatment strategies and choosing a more effective treatment method requires the development of precise standardized methods. Thus, the application of artificial intelligence technologies, or rather, machine learning methods to solve the considered problem, can be vital in early diagnosis and treatment of HCC.

Presently, many studies on the development of intelligent systems for the detection, diagnosis and treatment of liver diseases are available in scientific literature [7–9] shows the importance of using machine learning and deep learning methods for HCC prediction. [10, 11] propose a fuzzy rule-based technique for developing the HCC staging system and the principles of forming the knowledge base.

The present article considers the importance of gene analysis along with the laboratory studies, CT, X-ray examination for more accurate identification and correct diagnosis of HCC. The question of determining the similarity of HCC cancer tissues with identical CwoHCC non-cancerous tissues for the HCC identification, and proposes a machine learning-based solution technique.

3. Results

Based on the gene analysis of HCC cancer tissue, for its early diagnosis a solution algorithm in the following stages is proposed.

*Gene Expression of HCC tissues***.** Differentiation of HCC cancerous tissue and similar non-cancerous (CwHCC and NwHCC) tissues according to selected genes is used for early HCC diagnosis. [12] performs the solution to this problem on the basis of 10 genes included in the range of risk factors. *LAMC*1*, UBE4B, HSPH*1*, HNF*1*A, SF3B*1*, APC*2*, CHST*4*, HGF, MTHFD*2, and *AGO*3 are the key cancer genes. The association of each gene with cancerous tissue and similar non-cancerous tissue is determined. For example, *UBE4B* can be used as a potential prognostic marker for HCC treatment due to the carcinogenic effect of primary HCC cancer. Furthermore, the gene *HNF*1*A* is closely related to HCC cancer case, because the number of *HNF*1*A* increases when non-cancerous tissues turn into HCC cancer tissues. Its Gene Expression is significantly increased in liver HCC tissues. Gene expression of *SF3B1* is significantly increased in HCC cancer tissues during disease progression. Additionally, the genes *HSPH*1*, APC*2*, CHST*4*, HGF, MTHFD*2 and *AGO*3 are closely related to HCC cancer.

Gene Expression profiling. The Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases can be used for a gene expression profiling. Initially, a database storing relevant HCC biopsy samples (*D*1), HCC surgical samples (*D*2), CwoHCC biopsy samples (*D*3), and CwoHCC surgical samples (*D*4) is generated. To ensure the objectivity of the created model, the samples of each mentioned type are divided into two data subsets: training data set (80 % samples from each type) and test data set (20 % samples from each type). The training dataset

uses both HCC sample (HCC biopsy sample and HCC surgical sample) and CwoHCC sample (CwoHCC biopsy sample and CwoHCC surgical sample).

Determination of the Within-Sample Relative Expression Ordering. Using the Within-Sample Relative Expression Orderings (REO) technique, gene expression profiling can be more reliable. REO technique is used for feature extraction. According to the REO technique, if the gene *a* has a higher analysis level than the gene *b* (or vice versa) in a given sample, they are analyzed as *Ea>Eb* (or *Ea<Eb*). If at least 95 % of the samples for a gene pair have the same gene pair ordering, then that gene is considered to be stable according to the REO technique. Gene pair analysis may differ according to REO ordering in both cirrhotic tissues within a sample. Thus, when analyzing gene pairs, the value of the gene *a* may be high in non-HCC patients (i.e., with CwoHCC tissue) but low in HCC patients (*Ea<Eb* or *Ea>Eb* in CwoHCC samples, but *Ea>Eb* or *Ea<Eb*). Reversal gene pairs are selected as candidate REOs signatures in the REO technique for the HCC identification. Common genes between the training dataset and the test dataset and their corresponding gene expression are then obtained. Based on the gene expression profiles and reversal gene pairs, new profiles are generated by specifying the cases *Ea>Eb, Ea<Eb,* and (*Ea* or *Eb* unavailable) as 1, 0 and –1, respectively.

Feature selection with mRMR (minimum Redundancy Maximum Relevance) and IFS (Incremental feature selection) methods. Based on the new profiles, the mRMR method is applied to rank the HCC cancer and non-cancer gene pairs within minimum Redundancy Maximum Relevance conditions [12].

Interaction between genes *I* is defined as:

$$I(g_i,T)=\int p(g_i,T)\ln\left(\frac{p(gi,T)}{p(gi)p(T)}\right)dg_i\mathrm{d}T. \quad (1)$$

Here, $I(g_i, T)$ denotes the interaction between the gene pair g_i and type disease *T*. The following formula is used to determine the relevance by all gene pairs:

$$mRMR=\frac{1}{\Omega}\sum_{g_i\in\Omega}I(g_i,T)-\frac{1}{\Omega^2}\sum_{g_i,g_j\in\Omega}I(g_i,g_j). \quad (2)$$

Here Ω represents all given gene pairs, g_i – one of given gene pairs, $I(g_i, g_j)$ – interaction between the genes g_i and g_j.

In the next step, optimal gene pairs are selected from the mRMR gene pair in the sample given as a candidate signature, and the IFS method is used for this purpose [13]. According to a certain evaluation criterion, optimal subsets of features (genes) are selected from the whole feature set. This ensures the elimination of irrelevant and unnecessary features. The importance of a feature subset is evaluated by its relevance and redundancy parameters. This aspect predicts decisions, otherwise features are considered relevant. A feature is redundant when it is highly correlated with some other feature. In the application of IFS, the main goal is to define unrelated genes and to find the optimal subset of signatures associated with the decision feature.

4. Discussion

Support Vector Machine (SVM) classification method can be used for feature classification [14]. This method is widely used in the biological data classification [15, 16]. Radial basis function (RBF) is used due to its good performance in solving non-linear problems. The RBF kernel function calculates the similarity for the genes pair g_i and g_j, or rather how close they are to each other. The following function is used for this:

$$K(X_1,X_2)=\exp\left(-\frac{\|X_1-X_2\|^2}{2}\right), \quad (3)$$

$$K(gi,gj)=\exp\left(-\frac{\|g_i-g_j\|^2}{2}\right). \quad (4)$$

Here 'σ' denotes the discrepancy and the hyperparameter, $||X_1–X_2||$ denotes the interrelation of the genes g_i and g_j.

Hyperparameter setting methods such as *Grid Search Cross Validation and Random Search Cross Validation* can be used to find the correct σ for the dataset used.

Determination of classification and confusion matrix criteria. The *SVM* machine learning algorithm is considered appropriate to be used to perform the classification by the gene pair similarity in the database used for the early diagnosis of HCC. Evaluating the classifiers' detection performance is of great importance in machine learning. The criteria of precision, recall, false positive rate (FPR), true positive rate (TP), f-measure and accuracy are used in the evaluation of detection performance.

Precision (*P*) is defined as the number of true positives and determined as:

$$P=\frac{T_p}{T_p+F_p}. \quad (5)$$

F_p denotes the number of data not associated with misclassified prediction.

Here T_p denotes the number of data associated with correctly classified prediction.

Recall (*R*) is defined as the number of true positives and calculated using the following formula:

$$R=\frac{T_p}{T_p+F_n}, \quad (6)$$

where F_n denotes the number of data not associated with misclassified prediction.

*F*1-Score (*F*1) is defined as the harmonic mean of precision and recall.

$$F1=2\times\frac{P\times R}{P+R}. \quad (7)$$

Accuracy is defined as follows:

$$A=\frac{T_p+T_n}{T_p+T_n+F_p+F_n}. \quad (8)$$

Model training is an important step to achieve good performance and to associate the model's extreme data with each other, as well as to avoid non-associated data.

5. Conclusions

Gene-based approaches attract attention in order to address the problems arisen in early HCC diagnosis through laboratory studies, CT and X-ray examination. The methodological basis of such approaches is the determination of the similarity of gene pairs of HCC cancer and similar non-cancerous tissues. This article selected a gene expression profile taking into account the availability of relevant databases and proposed a new approach by referring to the mREO method for determining the

gene pair's similarity of cancer and non-cancerous tissues, the IFS method for feature refinement, and the SVM method for classification.

The proposed algorithm was aimed at improving the HCC diagnosis. In the further studies, it is planned to perform experiments with the application of other methods along with the SVM classification method based on the similarity of gene pairs and to select the method with better performance according to the performance evaluation, and to develop the methods and algorithms for HCC diagnosis taking into account other characteristics.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

The study was performed without financial support.

Data availability

Manuscript has no associated data.

References

1. Mammadova, M., Jabrayilova, Z. (2019). Electronic medicine: formation and scientific-theoretical problems. Baku: "Information Technologies" publishing house, 319. Available at: https://ict.az/uploads/files/E-medicine-monograph-IIT-ANAS.pdf
2. Indhumathy, M., Nabhan, A. R., Arumugam, S. (2018). A Weighted Association Rule Mining Method for Predicting HCV-Human Protein Interactions. Current Bioinformatics, 13 (1), 73–84. doi: https://doi.org/10.2174/1574893611666161123142425
3. El-Serag, H. B. (2011). Hepatocellular Carcinoma. New England Journal of Medicine, 365 (12), 1118–1127. doi: https://doi.org/10.1056/nejmra1001683
4. Russo, F. P., Imondi, A., Lynch, E. N., Farinati, F. (2018). When and how should we perform a biopsy for HCC in patients with liver cirrhosis in 2018? A review. Digestive and Liver Disease, 50 (7), 640–646. doi: https://doi.org/10.1016/j.dld.2018.03.014
5. Budhu, A., Forgues, M., Ye, Q.-H., Jia, H.-L., He, P., Zanetti, K. A. et al. (2006). Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. Cancer Cell, 10 (2), 99–111. doi: https://doi.org/10.1016/j.ccr.2006.06.016
6. Guan, Q., Yan, H., Chen, Y., Zheng, B., Cai, H., He, J. et al. (2018). Quantitative or qualitative transcriptional diagnostic signatures? A case study for colorectal cancer. BMC Genomics, 19 (1). doi: https://doi.org/10.1186/s12864-018-4446-y
7. Singh, A., Pandey, B. (2016). An Efficient Diagnosis System for Detection of Liver Disease Using a Novel Integrated Method Based on Principal Component Analysis and K-Nearest Neighbor (PCA-KNN). International Journal of Healthcare Information Systems and Informatics, 11 (4), 56–69. doi: https://doi.org/10.4018/ijhisi.2016100103
8. Gorunescu, F., Belciug, S., Gorunescu, M., Badea, R. (2012). Intelligent decision-making for liver fibrosis stadialization based on tandem feature selection and evolutionary-driven neural network. Expert Systems with Applications, 39 (17), 12824–12832. doi: https://doi.org/10.1016/j.eswa.2012.05.011
9. Calderaro, J., Seraphin, T. P., Luedde, T., Simon, T. G. (2022). Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma. Journal of Hepatology, 76 (6), 1348–1361. doi: https://doi.org/10.1016/j.jhep.2022.01.014
10. Mammadova, M., Bayramov, N., Jabrayilova, Z. (2021). Development of the principles of fuzzy rule-based system for hepatocelular carcinoma staging. EUREKA: Physics and Engineering, 3, 3–13. doi: https://doi.org/10.21303/2461-4262.2021.001829
11. Mammadova, M. G., Bayramov, N. Y., Jabrayilova, Z. G., Manafli, M. I., Huseynova, M. R. (2022). Knowledge transformation in the intelligent system for hepatocellular carcinoma staging. 8th Conference on Control and Optimization with Industrial Applications-COIA'2022, Azerbaijan, vol.1, pp. 318–320. http://coia-conf.org/upload/editor/files/COIA2022_V1.pdf
12. Zhang, Z.-M., Tan, J.-X., Wang, F., Dao, F.-Y., Zhang, Z.-Y., Lin, H. (2020). Early Diagnosis of Hepatocellular Carcinoma Using Machine Learning Method. Frontiers in Bioengineering and Biotechnology, 8. doi: https://doi.org/10.3389/fbioe.2020.00254
13. Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., Zou, Q. (2017). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics, 34 (3), 398–406. doi: https://doi.org/10.1093/bioinformatics/btx622
14. Cao, R., Wang, Z., Wang, Y., Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. BMC Bioinformatics, 15 (1). doi: https://doi.org/10.1186/1471-2105-15-120
15. Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget, 8 (44), 77121–77136. doi: https://doi.org/10.18632/oncotarget.20365
16. Meng, C., Jin, S., Wang, L., Guo, F., Zou, Q. (2019). AOPs-SVM: A Sequence-Based Classifier of Antioxidant Proteins Using a Support Vector Machine. Frontiers in Bioengineering and Biotechnology, 7. doi: https://doi.org/10.3389/fbioe.2019.00224

Received date 02.10.2022
Accepted date 07.11.2022
Published date 29.11.2022

How to cite: *Jabrayilova, Z., Garayeva, L. (2022). Algorithm for early diagnosis of hepatocellular carcinoma based on gene pair similarity. Technology transfer: fundamental principles and innovative technical solutions, 11–13. doi: https://doi.org/10.21303/2585-6847.2022.002670*