EasyChair Preprint

№ 568

Development of a prediction classifier for the early diagnosis of liver cancer

Di Wu, Jianhua Cao, Wei Li and Xin Wang

October 9, 2018

Development of a prediction classifier for the early diagnosis of liver cancer

Running title: A prediction classifier for early liver cancer

Di Wu[1,*], Jianhua Cao[2], Wei Li[3], Xin Wang[4]

[1]School of Computer Science and Technology, Dalian University of Technology, Da lian City, Liaoning, 116023, China

[2]School of Computer Science and Technology, Dalian University of Technology, Da lian City, Liaoning, 116023, China

[3]Affiliated Hosp 2, medical department, Dalian Medical University, Dalian City, Liaoning, 116023, China

[4]Dalian University of Technology, School of Mechanical Engineering, Dalian City, Liaoning, 116023, China

[*]Corresponding Author: Jianhua Cao

Dalian University of Technology, School of Computer Science and Technology,

No.2 Lingong road, Dalian City, Liaoning, 116023,China

Email: cjh1572318427@mail.dlut.edu.cn; Tel: +8618742507275

Word count(Text only): 2868

What is current knowledge

- HCC is ranked the second among all cancers in China.
- HCC is currently diagnosed by detecting AFP and abdominal ultrasound.
- Abnormal AFP is usually detected at late stages of liver cancer.
- AFP is not detectable in some liver cancer patients.

What is new here

- A set of 22 indicators are closely associated with liver cancer.
- The SVM classifier ingerated with these 22 markers accurately determines the exitence of HCC.
- The SVM classifier accurately distinguishes HCC from other liver diseases.

Abstract

Hepatocellular carcinoma (HCC) is the second cause of cancer-related death worldwide, and the incidence rate of liver cancer has continuously increased, with approximately 750,000 new diagnosed cases each year. Especially in China, both the incidence and mortality rate of HCC have been ranked second among all cancers. Importantly, HCC mortality rate is similar to its incidence rate, indicating that most patients with liver cancer die from HCC. In clinical practice, liver cancers are usually diagnosed by detecting alpha-fetoprotein (AFP) and abdominal ultrasound. However, abnormal AFP is usually detected at late stages of liver cancer, in which most patients are refractory to surgery, radiotherapy and chemotherapy. Moreover, AFP is not detectable in some liver cancer patients. In this study, we aimed to establish an alternative diagnostic method for liver cancer patients by analyzing hidden patterns and relationships among multiple specific markers of liver cancers. By building a predictive classification of liver cancer and the relationship between different markers, a support vector machine (SVM) classifier was developed. Our SVM classifier integrated 22 specific markers. Our results revealed that the input of these 22 markers into the classifier could accurately determine the exitence of HCC in a patient. Our established SVM classifier may achieve the early prediction of liver cancer, thereby improving the accuracy of diagnosis and treatment of live cancer patients.

Keywords: Hepatocellular carcinoma , alpha-fetoprotein , support vector machine classifier, specific markers.

Introduction

Liver cancer is one of the most aggrassive malignancies and the second cause of cancer-related death worldwide (1-5). Due to metastasis at diagnosis, most patients with liver cancers are not suitable for current radiation and chemotherapy treatments (6, 7). Molecular targerted therapy with sorafenib can improve the survival time of liver cancer patients, but drug resistance inevitably occurs at a later life time (8, 9). Liver resection remains the gold standard for the treatment of patients with liver cancer, but its prognosis remains very poor (10, 11). Currently, a number of organizations including the American Association for Liver Diseases, the National Comprehensive Cancer Network of China, and the European Association for Liver Diseases published consensus guidelines for the diagnosis and treatment of liver diseases (12-14). According to these standards, the most effective management of liver cancer patients is early diagnosis based on the known indicators of a patient (3). Therefore, the current challenge is to develop a classifier for the early diagnosis of liver cancer by analyzing a large amount of the patient's data (15).

Currently, alpha-fetoprotein (AFP) has been a widely used biomarker for hepatocellular carcinoma (HCC) (16). In clinical practice, detection of alpha-fetoprotein (AFP) and abdominal ultrasound are usually applied for liver cancers diagnosis (17-23). However, AFT has significant drawbacks for the diagnosis of HCC. First, when an abnormal AFP is detected, the vast majority of patients have already reached an advanced liver cancer stage with an extremely low cure rate. Second, the specificity and sensitivity of AFP is low, since high levels of AFP are

also detected in chronic hepatitis or cirrhosis; while normal AFP levels are sometimes detected in HCC (24, 25). In order to improve the early diagnosis of HCC, a number of candidate tumor markers including Golgi protein 73 and AFPL3 have been proposed. However, these markers have not been applied in clinic due to poor sensitivity and specificity (26). With the advance of precision medicine, it is urgent to develop an alternative strategy for diagnosis and provide prognostic information of HCC (27).

Furthermore, it is currently a challenge to distinguish HCC from other liver diseases such as hepatitis and cirrhosis. A rountine clinic biochemical test includes indicators of blood, urine, kidney function, liver and a series of other indicators. It is necessary for an oncologist to distinguish patients with HCC from hepatitis and cirrhosis based on these indicators (15). However, due to the complexity of cancer, it is hard even for experienced oncologists to precisely distinguish whether a patient is suffering from HCC (28). In order to meet this challenge, a lot of supervised learning approaches for prognosis have been developed. A decision tree analysis of cDNA microarrays has been used for non-small cell lung cancer prognosis (29). Advanced classification algorithms have also been established to predict cancer classification. For example, support vector machines (SVMs) have been used to select highly reliable identified genes to build a cancer classifier (30). Nevertheless, a supervised learning method has not been used in the development of a highly predictive classifier of liver cancer (31). In this study, based on SVM-based methods of surveillance, we successfully developed a SVM-based specific marker classifier for liver cancer by

integrating 22 markers. This established classifier could accurately distinguish liver cancer from other liver diseases and may provide a reference for oncologists to make effective programs in daily decision-making.

Methods

Data Sources

Liver cancer and liver disease datasets including cirrhosis and hepatitis B were collected from the hospital epidemiological surveillance of Dalian City (Dalian, China). The liver cancer datasets contained medical records of 1,879 hospitalized patients from March 2005 to March 2015, while liver disease datasets contained medical records of 195 patients from October 2014 to June 2015. Data elements included patient ID number, gender, age, diagnosis description, discharge cases, admission date, discharge date, doctor's contents, pathology number, pathological diagnosis, coding number, testing indicators, signs, working unit, diagnosis results, specimens serial number, diagnosis date and so on. These elements provided the specific information of patients with liver cancer. Each record represents a particular patient database. Informed intent was obtained from each patient.

Data Preparation

Data preprocessing was first performed by excluding irrelevant attributes such as age, gender, pathological diagnosis, coding and working unit; and including related attributes such as admission date, index, diagnosis results, results of reference, and

diagnosis markers. After the removal of irrelevant attributes of patients, patients were re-indexed and abnormal indicators were filtered out due to personal reasons: severity of the disease, the patient's family's economic status, and the index of each patient was not the same. For example, some patients had more than 17 million items of measure indexes, while some patients were measured for only 30 indicators. Moreover, a number of these indicators have repeated measurements. For an indicator, it might be tested in multiple measurements due to surgery, medication, measurement accuracy and other external factors; especially for specimens with morphology of jaundice, mild hemolysis, or mild chyle. Therefore, these inaccurate indicators were removed and a measured unified index was selected during the first patient admission. The first detected indicators were the most significant, as they were not affected by drugs, surgery and radiation therapy. The latest index data were chosen if the first test was not accurate. This way, 431 indicators were first analyzed and a huge normal index data not related to the analysis of hidden indicators were found. Finally, 63 indicators were selected by analyzing a large number of abnormal indicators. After analyzing the abnormal rate, the consequence of the abnormal indicators on the effects of HCC was explored.

Next, missing indicators were managed. Since not every patient had 63 kinds of abnormal indicators, missing values were filled in by the truncated mean method in SPSS Statistics 19. This method requires the deletion of a much larger than average and much less than average number, analyzes patterns of missing values, and generates multiple versions of data sets; in which each data set contains its own set of

estimates. A complete analysis of each data set results in the average of all data sets (excluding outliers) and generates a single value.

Finally, the selected abnormal indicators were standardized as follows: above normal range was set to 2, below normal range was set to 1, and the index was set to 0 in the normal range. Data sets were imported into the SPSS Modeler for data analysis.

Procedure for the development of SVM predictive modeling

Our classifier was built in accordance to the SVM-based method. In addition, radial basis function was applied as the kernel function, because our classification problem is nonlinear (31). Among the 1,879 samples, 382 had no tests of the above filtered 22 indicators or had less than 10 of these indicators; therefore, the SVM predictive modeling was developed based on the data set of the rest of the 1,497 HCC samples. The 1,497 samples were randomly divided into three groups: training group, testing group and validation group. The training and testing groups were used for the construction of the classifier model, while the validation group was used to predict liver cancer or liver disease using the built classifier model and to test the accuracy of the classifier. During the construction of the classifier model, 1,000 samples from the 1,497 HCC samples and 195 liver samples were applied for training and prediction. In the IBM SPSS Modeler 14.1, the partition ratio were set as 7:3; in which 70% was from the training group (sample number: 1,195 × 0.7 = 835, in which 1,000 × 0.7 = 700 were liver cancer samples and 195 × 0.7 = 135 were liver disease samples) and 30% was from the validation group (sample number: 1,195 × 0.3 = 360, in which

1,000 × 0.3 = 300 were HCC and 195 × 0.3 = 60 were liver diseases). The remaining 497 HCC samples and 195 liver samples were used to test the accuracy of the model and predict liver cancer.

Statistical analysis

Statistical analysis was performed on the 22 markers in the training, test and validation groups using IBM SPSS Statistic v19.0 (IBM Co., Armonk, NY, USA). The training and testing groups were analyzed as a set (Table 1), while the validation group was analyzed as another set (Table 2). In all analyses, a *P*-value <0.01 was considered statistically significant. Data are expressed as the mean ± standard deviation (SD).

Results

Correlation Analysis of Markers with Liver Cancer

Among the 1,879 patients with HCC, 1,511 were male and 371 were female. The Apriori algorithm is capable of reducing the number of sets and comparing the number of valid correlation algorithms, which is easy to implement (32). The correlation analysis of the 63 markers of these 1,879 samples with HCC was performed using an Apriori association algorithm in the SPSS Modeler. The Apriori association algorithm was divided into two steps. In the first step, all frequent item sets were retrieved from the data source through iterations with a support threshold of 30%. In the second step, the minimum rules for confidence of 50% were constructed

from the retrieved frequent item sets in the first step. A total of 22 specific markers that had a strong association with liver cancer were selected based on the support of more than 39% and 100% confidence (Table 3).

Development of the SVM classifier

Our classifier was built in accordance to the SVM-based method using the 22 markers selected via correlation analysis. The SVM classifier model was constructed by entering the 22 indicators of 1,195 samples into the SPSS Modeler model of the SVM modeling. The model predicted that based on the percentage of HCCs, 20% were associated with AFP, 18% were associated with γ- glutamyl transferase, 14% were associated with absolute lymphocyte, 13% were associated with red cell distribution width, 10% were associated with alanine and aspartate, 9% were associated with RBC, 7% were associated with platelet distribution width, 3% were associated with the percentage of eosinophils, 2% were associated with hematocrit and 1% were associated with the percentage of neutrophils. The other indicators added together were related to 3% of HCCs, which is negligible (Supplementary Table 1 and Fig. 1). From these data, these top 10 biomarkers were proposed as key factors for the prediction of HCC, and the remaining 12 biomarkers may serve as supplemental markers.

Diagnostic results were set as 1 for HCC and 0 for liver dieseases. Since this type of binary variables is non-linear, radial basis function was applied as the kernel function.

In the training group, 835 samples were HCC in 700 cases, while 135 cases of patients were with liver diseases. Among the 360 cases of samples in the testing group, 300 cases had HCC and 60 cases had liver diseases. Analysis results revealed that the accuracy for training was 99.4% and the accuracy of the testing was 85.28% (Table 4 and Supplementary Table 2). Results of the training and testing groups were completed together to build a classification model.

A diagnostic value of ≥0.95 can be used for the result of determination, while a value <0.95 can be used as a certain reference combined with other methods for determination. As shown in Table 4, when a diagnosis value ≥0.95 was selected, the precision obtained after training was 99.52%, and the accuracy of the test was 92.18%. The accurancy of training and testing increased as the accuracy of the test increased.

Verification of the prediction model

After training and testing, a SVM classification prediction model was developed. The model with the remaining 497 HCC samples and 195 liver samples were next validated. Results revealed that prediction accuracy was 91.62% (Table 5). However, when diagnosis at ≥0.95 was selected, prediction accuracy was improved to 94.98 (Table 5). Therefore, in the actual diagnosis, it may be better for a doctor set the diagnosis confidence, which would help to improve the diagnostic rate of liver cancer. However, results of the remaining <0.95 credibility may be confirmed by other methods such as imaging.

ROC curves

The ROC curve is a criterion for measuring the sensitivity and specificity of a marker for HCC. The greater the area under the curve, the higher the sensitivity and specificity, and the more easily it is diagnosed with liver cancer. For conventional diagnosis of the liver cancer factor, ROC curves include single factors such as AFP, γ-glutamyl transferase, platelet, aspartate aminotransferase, and lactate dehydrogenase. In this study, the overall SVM classifier integrated with 22 markers was considered as a single factor into the ROC curve to compare the maximum area under the curve between them (AUC). In the training group, the AUC of the SVM classifier was 0.784, and the single marker with a maximum AUC was AFP with 0.747 (Fig. 2A). In the testing group, the AUV of the SVM classifier was 0.807 and the single marker with a maximum AUC was AFP with 0.727 (Fig. 2B). These data demonstrate that the AUC of the SVM classifier was significantly greater than the maximum AUC of all other indicators.

Discussion

The molecular pathogenesis of HCC is heterogeneous and very complex (33). Although a large number of molecules, signaling pathways and genetic alterations have been found to be associated with HCC (34-36), none of these are currently being used for effective screening, early diagnosis, classification, targeted therapy and prognosis (33). For an individual patient, a tumor is not static, but dynamic, during the process of tumorigenesis and treatment over time (33). Therefore, the development of

methods through the integration of multiple specific markers for early diagnosis is a promising approach. In this study, we obtained 22 algorithm-specific indicators through the relevance analysis of HCC-associated indicators of patients spanning 10 years, and constructed a HCC-SVM classifier based on these 22 indicators to distinguish patients with liver cancer and liver diseases. Our results revealed that the HCC-SVM classifier could well-predict patients with HCC with an accuracy of 91.62% and a higher accuracy rate of 94.98% if a greater than 95% confidence was selected for diagnostic results. Importantly, the AUC of the HCC-SVM classifier ROC curve was greater than the traditional markers of liver cancer AFP. Therefore, our established HCC-SVM classifiers may provide clinicians with an efficient and reliable diagnostic tool to predict liver cancer patients.

In 1999, Vapnik introduced the support vector machine (SVM) for data classification and function approximation (37). Among all well-known algorithms, SVM is considered to have the most robust and accurate supervised learning algorithm (38). The SVM optimization process can maximize prediction accuracy and reduce the over-fitting of training data (15). The basis of SVM is to find the optimal decision boundary by finding the maximum achievable distance between the hyperplane to the maximum the edge (39). As a result, a larger decision boundary edge has better generalization error than a smaller decision boundary edge, which leads to the generalization ability of an unknown sample. Accordingly, we built our classifier in accordance to the SVM-based method. In addition, we applied radial basis function as the kernel function, because our classification problem is nonlinear

(31).

Compared with other machine learning algorithms, SVM-based management is more suitable for the classification of high-dimensional data (a limited number of training samples), choosing the most effective of all possible features (37, 40). Previous studies have shown that the expression of individual indicators (such as AFP) for the early diagnosis of HCC have a large number of limitations. In order to improve the accuracy of the early diagnosis of HCC, SVM can combine multiple indicators to predict HCC. To date, our HCC-SVM classifier has ensembled 22 specific indicators. Our results revealed that the association of the expression of individual indexes with HCC had a weak sensitivity and specificity. Among these individual indicators, AFP had the highest sensitivity and specificity. In contrast, the sensitivity and specificity of the HCC-SVM classifier with HCC was higher than any single index. Therefore, the HCC-SVM classifier would provide reliable and effective help for the diagnosis of early HCC patients.

However, there are some limitations in our study. First, only 195 cases of liver disease, which accounted for 11.5 % of the total cases of the experiment, were recruited. Second, only the truncated mean method was applied for handling the missing values. In addition, in the original cases of liver cancer patients from 2005 to 2015, prealbumin was tested only in liver cancer patients from 2011 to 2015, but not for patients from 2005 to 2010. Furthermore, although the developed HCC-SVM classifier ensembled 22 specific indicators and demonstrated high accuracy, new clinical indicators would be found and new technologies would be developed every

year (3). These new indicators and technologies may improve the accuracy of the prediction of the classification. Therefore, the HCC-SVM classifier may further include additional indicators to improve their prediction accurancy.

In summary, in this study, we first discovered and extracted 22 indicators that are closely associated with liver cancer by correlation analysis. Then, based on these indicators, we established a SVM classifier for the early prediction of liver cancer. Our validation analysis revealed that the SVM classifier accurately predicted patients with HCC or liver disease. Our study suggests that the SVM classifier would provide as a reliable and effective tool for the diagnosis of early HCC patients.

Financial support

None

Conflicts of interest

The authors disclose no potential conflicts of interest.

References

1. Jemal A, Bray F, Center MM, et al. Global cancer statistics. CA Cancer J Clin 2011;61:69-90.

2. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA Cancer J Clin 2012;62:10-29.

3. Maluccio M, Covey A. Recent progress in understanding, diagnosing, and treating hepatocellular carcinoma. CA Cancer J Clin 2012;62:394-9.

4. El-Serag HB, Rudolph KL. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology 2007;132:2557-76.

5. Ramkumar N, Prakash S, Kumar S A, et al. Prediction of liver cancer using Conditional probability Bayes theorem[C]// International Conference on Computer Communication and Informatics. IEEE, 2017:1-5..

6. Lopez PM, Villanueva A, Llovet JM. Systematic review: evidence-based management of hepatocellular carcinoma--an updated analysis of randomized controlled trials. Aliment Pharmacol Ther 2006;23:1535-47.

7. Yeo W, Mok TS, Zee B, et al. A randomized phase III study of doxorubicin versus cisplatin/interferon alpha-2b/doxorubicin/fluorouracil (PIAF) combination chemotherapy for unresectable hepatocellular carcinoma. J Natl Cancer Inst 2005;97:1532-8.

8. Llovet JM, Ricci S, Mazzaferro V, et al. Sorafenib in advanced hepatocellular carcinoma. N Engl J Med 2008;359:378-90.

9. Cheng AL, Kang YK, Chen Z, et al. Efficacy and safety of sorafenib in patients in

the Asia-Pacific region with advanced hepatocellular carcinoma: a phase III randomised, double-blind, placebo-controlled trial. Lancet Oncol 2009;10:25-34.

10. Hong J, Hu K, Yuan Y, et al. CHK1 targets spleen tyrosine kinase (L) for proteolysis in hepatocellular carcinoma. J Clin Invest 2012;122:2165-75.

11. Portolani N, Coniglio A, Ghidoni S, et al. Early and late recurrence after liver resection for hepatocellular carcinoma: prognostic and therapeutic implications. Ann Surg 2006;243:229-35.

12. Bruix J, Sherman M. Management of hepatocellular carcinoma: an update. Hepatology 2011;53:1020-2.

13. Benson AB, 3rd, Abrams TA, Ben-Josef E, et al. NCCN clinical practice guidelines in oncology: hepatobiliary cancers. J Natl Compr Canc Netw 2009;7:350-91.

14. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. Eur J Cancer 2012;48:599-641.

15. Lotfnezhad Afshar H, Ahmadi M, Roudbari M, et al. Prediction of breast cancer survival through knowledge discovery in databases. Glob J Health Sci 2015;7:392-8.

16. Liu H, Xu Y, Xiang J, et al. Targeting alpha-fetoprotein (AFP)-MHC complex with CAR T cell therapy for liver cancer.[J]. Clinical Cancer Research, 2017, 23(2):478-488.

17. Bruix J, Sherman M, Llovet JM, et al. Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference. European Association for the Study of the Liver. J Hepatol 2001;35:421-30.

18. Bruix J, Sherman M. Management of hepatocellular carcinoma. Hepatology 2005;42:1208-36.

19. Colombo M. Screening for hepatocellular carcinoma. Digestion 1998;59 Suppl 2:70-1.

20. McMahon BJ, London T. Workshop on screening for hepatocellular carcinoma. J Natl Cancer Inst 1991;83:916-9.

21. Nguyen MH, Keeffe EB. Screening for hepatocellular carcinoma. J Clin Gastroenterol 2002;35:S86-91.

22. Ryder SD. Guidelines for the diagnosis and treatment of hepatocellular carcinoma (HCC) in adults. Gut 2003;52 Suppl 3:iii1-8.

23. Sherman M. Screening for hepatocellular carcinoma. Baillieres Best Pract Res Clin Gastroenterol 1999;13:623-35.

24. AkeyamaT K, KamadaT. A lphafeto proteinina cute viral hepatitis. N Engl J Med 1972;287:989.

25. Di Bisceglie AM, Hoofnagle JH. Elevations in serum alpha-fetoprotein levels in patients with chronic hepatitis B. Cancer 1989;64:2117-20.

26. Rich N, Singal AG. Hepatocellular carcinoma tumour markers: current role and expectations. Best Pract Res Clin Gastroenterol 2014;28:843-53.

27. Schutte K, Schulz C, Link A, et al. Current biomarkers for hepatocellular carcinoma: Surveillance, diagnosis and prediction of prognosis. World J Hepatol 2015;7:139-49.

28. Fieschi M, Dufour JC, Staccini P, et al. Medical decision support systems: old

dilemmas and new paradigms? Methods Inf Med 2003;42:190-8.

29. Boutros PC, Lau SK, Pintilie M, et al. Prognostic gene signatures for non-small-cell lung cancer. Proc Natl Acad Sci U S A 2009;106:2824-8.

30. Spinosa EJ, Carvalho AC. Support vector machines for novel class detection in Bioinformatics. Genet Mol Res 2005;4:608-15.

31. Wang HY, Sun BY, Zhu ZH, et al. Eight-signature classifier for prediction of nasopharyngeal [corrected] carcinoma survival. J Clin Oncol 2011;29:4516-25.

32. Xindong Wu VK, J. Ross Quinlan. Top 10 algorithms in data mining. Knowl InfSyst 2008;14:1-37.

33. Wang J, Gong L, Zhu SJ, et al. The Human Homolog of Drosophila Headcase Acts as a Tumor Suppressor through Its Blocking Effect on the Cell Cycle in Hepatocellular Carcinoma. PLoS One 2015;10:e0137579.

34. Nikolaou K, Sarris M, Talianidis I. Molecular pathways: the complex roles of inflammation pathways in the development and treatment of liver cancer. Clin Cancer Res 2013;19:2810-6.

35. Shin JW, Chung YH. Molecular targeted therapy for hepatocellular carcinoma: current and future. World J Gastroenterol 2013;19:6144-55.

36. Nakagawa H, Shibata T. Comprehensive genome sequencing of the liver cancer genome. Cancer Lett 2013;340:234-40.

37. Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw 1999;10:988-99.

38. V V. The nature of statistical learning theory. Springer 1995.

39. Amna A, Umer, K., Ali, T., Minkoo, K. Analyzing Potential of SVM based Classifiers for Intelligentand Less Invasive Breast Cancer Prognosis. In: Paper presented at the Second International Conference onComputer Engineering and Applications; 2010.

40. Vapnik V, Chapelle O. Bounds on error expectation for support vector machines. Neural Comput 2000;12:2013-36.

Table 1. The expression of 22 Markers for Patients from the Training and Testing Cohorts

Characteristic	**No. of Patients (*n*=1,195)**	**95% CI (%)**	***P***
AFP expression level		0.701 to 0.772	<0.001
Normal	532		
High	663		
γ- glutamyl transferase expression level			
Low	2		
Normal	405	0.601 to .688	<0.001
High	788		
Abumin expression level			
Low	616		
Normal	577	0.413 to 0.501	0.058
High	2		
Alanine aminotransferase expression level			

Low	7		
Normal	670	0.578 to 0.659	<0.001
High	518		
Aspartate and alanine expression level			
Low	63		
Normal	687	0.513 to 0.599	0.013
High	485		
Aspartate aminotransferase expression level			
Low	10		
Normal	475	0.558 to 0.644	<0.001
High	710		
RBC expression level			
Low	541		
Normal	639	0.451 to 0.540	0.838
High	15		
RDW expression level			

Low	20		
Normal	428	0.517 to 0.606	0.006
High	747		
Hematocrit expression level			
Low	524		
Normal	648	0.445 to 0.535	0.659
High	23		
Alkaline phosphatase expression level			
Low	22		
Normal	707	0.501 to 0.587	0.053
High	466		
Lymphocyte percentage expression level			
Low	457		
Normal	651	0.521 to 0.605	0.005
High	87		
Absolute lymphocyte			

expression level			
Low	693		
Normal	497	0.512 to 0.602	0.012
High	5		
Prealbumin expression level			
Low	930	0.491 to 0.581	0.111
Normal	265		
Lactate dehydrogenase expression level			
Low	20		
Normal	536	0.505 to 0.592	0.032
High	639		
Eosinophils percentage expression level			
Low	292		
Normal	709	0.625 to 0.697	<0.001
High	194		
Hemoglobin expression level			

Low	520		
Normal	656	0.488 to 0.575	0.163
High	19		
Blood platelet expression level			
Low	440		
Normal	680	0.428 to 0.513	0.193
High	75		
Platelet distribution width expression level			
Low	532		
Normal	628	0.515 to 0.603	0.009
High	35		
Direct bilirubin (DB) expression level			
Normal	455	0.431 to 0.519	0.264
High	740		
Neutrophils percentage expression level			

Low	110		
Normal	674	0.559 to 0.643	<0.001
High	411		
Total bilirubin expression level			
Low	1		
Normal	643	0.433 to 0.521	0.307
High	551		
TBA expression level			
Normal	530	0.457 to 0.546	0.944
High	665		

Table 2. The expression of 22 Markers for Patients from the Validation Cohort

Characteristic	**No. of Patients (*n*=1,195)**	**95% CI (%)**	***P***
AFP expression level		0.755 to 0.830	<0.001
Normal	291		
High	401		
γ- glutamyl transferase expression level			
Low	2		
Normal	281	0.574 to 0.667	<0.001
High	409		
Abumin expression level			
Low	319		
Normal	371	0.367 to 0.461	<0.001
High	2		
Alanine aminotransferase expression level			
Low	5		

Normal	358	0.625 to 0.712	<0.001
High	329		
Aspartate and alanine expression level			
Low	16		
Normal	428	0.486 to 0.581	0.168
High	248		
Aspartate aminotransferase expression level			
Low	6		
Normal	263	0.583 to 0.676	<0.001
High	422		
RBC expression level			
Low	238		
Normal	442	0.377 to 0.473	0.002
High	12		
RDW expression level			
Low	10		

Normal	272	0.500 to 0.595	0.053
High	410		
Hematocrit expression level			
Low	209		
Normal	471	0.346 to 0.442	<0.001
High	12		
Alkaline phosphatase expression level			
Low	16		
Normal	419	0.493 to 0.587	0.103
High	257		
Lymphocyte percentage expression level			
Low	258		
Normal	378	0.529 to 0.620	0.002
High	56		
Absolute lymphocyte expression level			

Low	410		
Normal	279	0.527 to 0.623	0.002
High	3		
Lactate dehydrogenase expression level			
Low	15		
Normal	318	0.499 to 0.594	0.055
High	359		
Eosinophils percentage expression level			
Low	181		
Normal	407	0.650 to 0.730	0.000
High	104		
Hemoglobin expression level			
Low	166		
Normal	472	0.415 to 0.509	0.123
High	54		
Blood platelet expression level			

Low	225		
Normal	421	0.393 to 0.486	0.014
High	46		
Platelet distribution width expression level			
Low	294	0.519 to 0.612	0.007
Normal	370		
High	28		
Direct bilirubin (DB) expression level			
Normal	254	0.433 to 0.528	0.415
High	438		
Neutrophils Percentage expression level			
Low	77		
Normal	390	0.569 to 0.661	0.000
High	225		
Total bilirubin expression level			

Low	1		
Normal	363	0.435 to 0.530	0.468
High	328		
TBA expression level			
Normal	320	0.441 to 0.536	0.643
High	372		

Table 3. The 22 predictors selected via correlation analysis

	Sample	Marker	Support %	Confidence %
1	HCC	γ- glutamyl transferase	65.247	100.0
2	HCC	RDW	61.469	100.0
3	HCC	Aspartate aminotransferase	60.245	100.0
4	HCC	Lactate dehydrogenase	60.192	100.0
5	HCC	Direct bilirubin (DB)	58.329	100.0
6	HCC	Absolute lymphocyte	57.477	100.0
7	HCC	Platelet distribution width	55.934	100.0
8	HCC	AFP	55.455	100.0
9	HCC	TBA	51.623	100.0
10	HCC	Prealbumin	50.665	100.0
11	HCC	Alanine aminotransferase	46.354	100.0
12	HCC	Lymphocyte percentage	45.929	100.0
13	HCC	Abumin	45.929	100.0
14	HCC	Eosinophils percentage	44.864	100.0
15	HCC	Neutrophils percentage	44.811	100.0
16	HCC	Total bilirubin	43.481	100.0
17	HCC	RBC	41.724	100.0
18	HCC	Alkaline phosphatase	40.021	100.0
19	HCC	Hemoglobin	39.968	100.0
20	HCC	Aspartate and alanine	39.915	100.0

21	HCC	Hematocrit	39.755	100.0
22	HCC	Blood platelet	39.542	100.0

Table 4. Accuracy of the prediction model with and without diagnostic confidence ≥0.95

	Without diagnostic confidence ≥0.95				**With diagnostic confidence ≥0.95**			
	Training group		**Testing group**		**Training group**		**Testing group**	
	No. Patients	**Accuracy %**	**No. Patients**	**Accuracy %**	**No. Patients**	**Accuracy %**	**No. Patients**	**Accuracy %**
Correct	830	99.4%	307	85.28%	829	99.52%	271	92.18%
Wrong	5	0.6%	53	14.72%	4	0.48%	23	7.82%
Total	835		360		833		294	

Table 5. Verification of prediction accuracy with and without diagnostic results ≥0.95 forecast accuracy

Verification group	**Without diagnostic confidence ≥0.95**		**With diagnostic confidence ≥0.95**	
	No. Patients	**Accuracy %**	**No. Patients**	**Accuracy %**
Correct	634	91.62%	586	94.98%
Wrong	58	8.38%	31	5.02%
Total	692		617	

Figure legends

Fig. 1. The relative importance of predictor biomarkers identified by the HCC-SVM classifier in predicting HCC.

Fig. 2. Logistic regression was used in the SPSS Statistics analysis. The SVM classifier for liver cancer was built by integrating the 22 kinds of markers as covariates into a single variable. The ROC curve was obtained by using the 22 markers and SVM classifier as input variables, and diagnostic results as output variables. (A) ROC curve of the training set. (B) ROC curve of the testing set.