THE JOURNAL OF COGNITIVE SYSTEMS
Journal Homepage: https://dergipark.org.tr/en/pub/jcs
Vol.6, No.2, 2021

# Identifying Possible Biomarkers for Early-Stage Hepatocellular Carcinoma using Random Forest Machine Learning Method

[1]Şeyma YAŞAR

[1] Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey. (e-mail: seyma.yasar@inonu.edu.tr).

**ARTICLE INFO**

**Received:**
**Revised:**
**Accepted:**

**Keywords:**
Early-Stage Hepatocellular Carcinoma
Biomarker
Machine learning
Random Forest
Classification

**Corresponding author:** Şeyma YAŞAR
✉ seyma.yasar@inonu.edu.tr
☎ +90 507 639 71 21

**ISSN:** 2548-0650

**DOI:**

**ABSTRACT**

Hepatocellular carcinoma (HCC) is a primary liver tumour arising from hepatocytes, the liver's own cells. It is one of the most common types of cancer in the world. The most important cause is chronic liver disease due to hepatitis B and C infections. In some patients, HCC causes symptoms such as abdominal pain, loss of appetite, anaemia, nausea, fatigue and jaundice and is diagnosed as a result of tests. In some patients, it is detected incidentally by liver ultrasound, tomography or MRI performed for another reason. The most typical finding is an increase in a substance called alpha-fetoprotein (AFP). Although this does not occur in all patients, elevated AFP in a patient with cirrhosis strongly indicates the presence of HCC. HCC can be seen on ultrasound, tomography or MRI films. Especially in tomography and MRI, the rapid and strong retention of the intravenous drug and then its early wash out is a typical finding and if detected in a patient with cirrhosis, HCC can be diagnosed without the need for biopsy. However, in many patients, imaging findings are not typical and a biopsy is required for diagnosis. In this study, a Random Forest machine learning model was created with proteomic data regarding the cancerous tumor tissue and the adjacent non-cancerous tissue of 19 HCC patients. the accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-Score, Matthews correlation coefficient (MCC) and G-Mean values for the Random Forest model were 0.90, 0.88, 0.90, 0.93, 0.82, 0.91, 0.82 and 0.91, respectively. Considering the model-dependent variable significance, SRSF1 and PBLD proteins are suggested as biomarkers that may be clinically useful in the diagnosis of early-stage HCC.

## 1. INTRODUCTION

Liver cancer is the fifth most common type of cancer worldwide and the second most common cause of cancer-related deaths. It accounts for 7% of all cancers, with 854,000 new cases and 810,000 deaths worldwide annually. Hepatocellular carcinoma (HCC) accounts for approximately 90% of primary liver cancers. Most patients are in an advanced stage when diagnosed, and treatment methods and chances of cure are limited [1, 2]. The etiological cause is known in approximately 90% of hepatocellular carcinoma and the most common causes are chronic hepatitis B infection (HBV), chronic hepatitis C infection (HCV) and alcohol use, respectively. More rarely, autoimmune hepatitis, α-1 antitrypsin deficiency, hereditary haemochromatosis and some porphyria may also be the cause [3]. Despite many studies, much remains to be elucidated about the pathogenesis of hepatocellular carcinoma. Therefore, the discovery and validation of new, more specific and cost-effective biomarkers that can more accurately predict the diagnosis and clinical behaviour of early hepatocellular carcinoma on an individual basis is an important research goal. Today, new biomarkers for the diagnosis and treatment of various diseases, especially cancer diseases, are being identified by proteomics, genomics and metabolomics techniques. Proteomics is a branch of science that studies all the proteins of an organism or cell. While genomics focuses on studying the genetic material (DNA) of an organism, proteomics seeks to understand all the proteins produced by that organism or cell. Proteomics studies are of great importance for understanding the physiological functioning of the organism, disease mechanisms, cellular signalling pathways and many other biological processes. Proteomics technologies are of great importance in the fields of biology, medicine and pharmacology, and research in this field contributes to a better understanding of biological processes and the development of new treatment methods [4]. Machine learning algorithms, which have been the subject of many research methods in recent years, are a branch of artificial intelligence. The aim of machine learning algorithms is to create intelligent systems and to ensure that these systems work efficiently and are developed with new data. Machine learning algorithms consist of five steps: collecting and preparing data, training the model, evaluating the model and improving the performance of the model [5]. Recently, machine learning methods have been used frequently in the diagnosis and treatment of diseases in the field of health.

 

The aim of this study is to classify early stage HCC using Random Forest, one of the machine learning methods on open access early stage HCC proteomic data and to identify biomarkers that may be clinically useful in the diagnosis and treatment of early stage HCC.

## 2. MATERIAL and METHODS

### 2.1. Dataset

The data set used in this study consists of 547 proteins obtained as a result of label-free proteomic analyses performed on 19 patients diagnosed with early stage HCC and their adjacent non-tumorous (control) tissues [6]. Descriptive statistics of the subjects forming the data set are given in Table 1.

TABLE I
DESCRIPTIVE FEATURES OF HCC CASES

| Variable | | HCC (n=19) |
|---|---|---|
| **Sex, n (%)** | Male | 13 (68%) |
| | Female | 6 (32%) |
| **Age** | | 61 ± 13.6 |
| **Stage, n (%)** | | |
| | **pT1** | 11 (58%) |
| | **pT2** | 5 (26%) |
| | **pT3** | 3 (16%) |
| **Grading, n (%)** | | |
| | **G1** | 5 (26%) |
| | **G2** | 8 (42%) |
| | **G3** | 6 (32%) |
| **Bloodvessel status, n (%)** | | |
| | **V0** | 14 (78%) |
| | **V1** | 4 (22%) |

**HCC: Hepatocellular carcinoma**

### 2.2. Random Forest

Random Forest is a powerful ensemble machine learning algorithm widely used for both classification and regression tasks. It operates on the premise of bagging, which stands for Bootstrap Aggregating. The fundamental idea behind Random Forest is to create a collection of decision trees, each trained on a random subset of the data and features, and then aggregate their predictions to produce a more accurate and robust outcome.

Random Forest offers several key advantages. First, it helps combat overfitting, a common issue in machine learning, by using a multitude of trees that collectively minimize variance and bias. The algorithm's use of random data subsets and feature subsets for each tree makes it robust against noisy data and irrelevant features.

Additionally, Random Forest provides an essential feature ranking mechanism, allowing you to assess the importance of different features in your dataset. This insight is valuable for feature selection and understanding which factors have the most influence on the model's predictions. The algorithm's versatility and ease of use make it suitable for various applications, from predicting stock prices to diagnosing diseases. Furthermore, its ability to handle large datasets and work well with high-dimensional data contributes to its widespread popularity in the machine learning community.

Overall, Random Forest is a reliable and effective tool for predictive modeling, thanks to its ability to improve accuracy, reduce overfitting, and provide insights into feature importance. It has become a staple in the toolkit of data scientists and machine learning practitioners for tackling a wide range of real-world problems [7].

### 2.3. Data Preprocessing and Performance Evaluation of the Models

Missing values in label-free proteomics analyses are a common problem, especially when working on large proteome datasets. Missing values in label-free proteomics analyses are a common problem, especially when working on large proteome datasets. Many factors, including analytical factors as well as biological factors, contribute to the occurrence of missing values. On the other hand, in machine learning, missing data processing or missing value imputation is a critical pre-processing step to ensure the robustness and accuracy of models. In this study, Random Forest is used as a missing value imputation method [8]. Machine learning, on the other hand, often uses feature selection methods to address an important problem encountered when building models on large and complex datasets. Variable selection is used in data science applications to improve model performance, train the model faster and reduce the risk of overfitting. Again, LASSO variable selection method was used in this study [9]. In the modelling phase, the data set was divided into 70% training and 30% test data. Then, 5-fold cross validation was applied to the test data set. Accuracy, Balanced accuracy, Sensitivity, Specificity, Positive predictive value, Negative predictive value, Matthews correlation coefficient (MCC), G-mean and F1-Score metrics in the performance evaluation of Random Forest machine learning models created to identify candidate biomarkers that can be used in the diagnosis and follow-up of early-stage HCC. R programming language caret package was used for the analyses conducted in the study.

## 3. RESULTS

As a result of the variable selection method applied to 547 proteins in the open source data set used in the study, 13 proteins were included in the study. Classification matrices for training and test stage of Random Forest model created with these 13 proteins obtained to classify early-stage HCC are given in Table 2 and Table 3, respectively.

TABLE II
CLASSIFICATION MATRIX OF THE TRAINING STAGE FOR THE RANDOM FOREST MODEL

| | | Real | | |
|---|---|---|---|---|
| | | **Control** | **Early-Stage HCC** | **Total** |
| **Predicted** | **Control** | 14 | 0 | 14 |
| | **Early-Stage HCC** | 1 | 13 | 14 |
| | **Total** | 15 | 13 | 28 |

**HCC: Hepatocellular carcinoma**

TABLE III

CLASSIFICATION MATRIX OF THE TESTING STAGE FOR THE RANDOM FOREST MODEL

| | | Real | | |
|---|---|---|---|---|
| | | Control | Early-Stage HCC | Total |
| Predicted | Control | 4 | 0 | 4 |
| | Early-Stage HCC | 1 | 5 | 6 |
| | Total | 5 | 5 | 10 |

**HCC: Hepatocellular carcinoma**

The performance metrics calculated through the obtained classification matrices are given in Table 4.

TABLE IV
PERFORMANCE METRICS OF TRAINING AND TESTING RANDOM FOREST MODEL

| METRICS | TRAINING | TESTING |
|---|---|---|
| | VALUE (95% CI) | VALUE (95% CI) |
| ACCURACY | 0.96 (0.89-1.00) | 0.90 (0.71-1.00) |
| BALANCED ACCURACY | 0.97 (0.90-1.00) | 0.90 (0.71-1.00) |
| SENSITIVITY | 1.00 (0.79-1.00) | 0.88 (0.62-0.98) |
| SPECIFICITY | 0.90 (0.55-0.99) | 0.90 (0.55-0.99) |
| POSITIVE PREDICTIVE VALUE | 0.93 (0.66-0.99) | 0.93 (0.68-0.99) |
| NEGATIVE PREDICTIVE VALUE | 1.00 (0.77-1.00) | 0.82 (0.48-0.98) |
| F1-SCORE | 0.96 (0.89-1.00) | 0.91 (0.73-1.00) |
| MCC | 0.93 (0.84-1.00) | 0.82 (0.58-1.00) |
| G-MEAN | 0.97 (0.89-1.00) | 0.91 (0.74-1.00) |

Considering the performance metrics in Table 4, the values for the XGBoost model are higher. Therefore, the importance values of PAS-related proteins determined by this model are shown in Table 5 and Figure 1.

TABLE V
THE IMPORTANCE VALUES OF PAS-RELATED PROTEINS DETERMINED BY XGBOOST MODEL

| EXPLANATORY VARIABLES | IMPORTANCE VALUE | EXPLANATORY VARIABLES | IMPORTANCE VALUE |
|---|---|---|---|
| **Q07955** | 100.00 | **Q96C11** | 48.07 |
| **P30039** | 99.90 | **P33176** | 46.73 |
| **O95954** | 94.83 | **Q9UL12** | 38.20 |
| **Q9UJ68** | 92.68 | **P16219** | 31.54 |
| **Q9H2A2** | 79.10 | **P53396** | 17.76 |
| **Q53FZ2** | 51.98 | **Q96N76** | 14.33 |

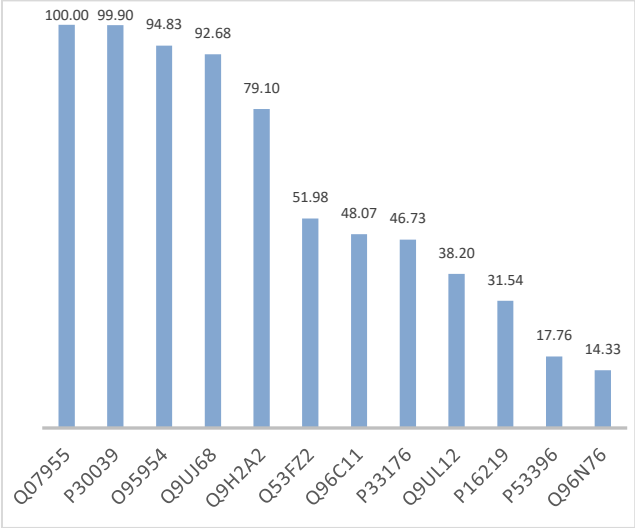


Fig. 1. The importance values for possible biomarkers

## 3. DISCUSSION

The correct understanding of tumour development is based on a comprehensive study of proteins. These are master regulators of vital processes such as signalling pathways that drive the carcinogenic process. Proteomic technologies can be applied to cancer research to detect differential protein expression and assess differential responses to treatment. In recent years, proteomic analysis has become an important tool complementing genetic analysis for cancer biology research. Among the numerous proteomic analysis methods, mass spectrometry techniques enable highly accurate qualitative and quantitative identification of hundreds of proteins in small volumes of various biological samples. Such analyses may soon become the basis for improvement in lung cancer diagnostic procedures. Lung cancer is the most common cause of cancer-related deaths in the world. The late stage of diagnosis and the lack of effective and personalised medicine reflect the need for a better understanding of the mechanisms underlying lung cancer progression. The survival rate of lung cancer patients is highly correlated with the stage of lung cancer. Therefore, improving diagnostic strategies for early lung cancer detection may improve patient survival. In recent years, genomic and proteomic technologies have begun to reveal the molecular complexity of lung tumours, allowing a rapid and complete analysis of the genes and proteins expressed in the context of this disease. New proteomic technologies that allow the analysis of thousands of cancer-associated proteins will improve knowledge of lung cancer biology and pathogenesis and help to develop new early detection biomarkers and identify prognosis and drug-related protein profiles [10].

In this study, early stage HCC was classified by using Random Forest machine learning method with the data obtained after label-free proteomic analyses performed with cancerous tumour tissue of 19 HCC patients and samples taken from cancer-free tumour tissues adjacent to these tissues. According to the experimental results obtained, the accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-Score, MCC and G-Mean values for the Random Forest model were 0.90, 0.88, 0.90, 0.93, 0.82, 0.91, 0.82 and 0.91, respectively.

As a result, the Random Forest model created with the proteomic data set used in the study has a very high

 ISSN: 2548-0650 http://dergipark.gov.tr/jcs

classification performance metrics. Therefore, considering the variable significance obtained from the model result, it can be said that the proteins with access codes Q07955, P30039, O95954 are candidate biomarkers that can be used in the diagnosis of early stage HCC. There are studies showing the relationship between dysregulated expression of Serine/arginine-rich splicing factor 1 protein coded Q07955 and tumour formation. It has been determined that the expression of SRSF1 is significantly increased in different cancers and this is associated with high expression of SRSF1 in cancers [11]. In addition, SRSF1 was found to be up-regulated in colon, kidney, lung, liver, pancreas and breast tumours [12-15]. Another protein proposed as a possible biomarker, P30039-encoded PBLD protein, was found to exhibit low protein levels in HCC tissue compared with normal liver tissue [16]. This indicates that PBLD protein may play an important role in early stage HCC carcinogenesis and development.

In conclusion, these two proteins, which are proposed as possible biomarkers for the diagnosis and diagnosis of early stage HCC cancer based on the Random Forest model, are thought to be very useful in the clinic.

## REFERENCES


[1] T. Akinyemiju, S. Abera, M. Ahmed, N. Alam, M. A. Alemayohu, C. Allen, et al., (2017) The burden of primary liver cancer and underlying etiologies from 1990 to 2015 at the global, regional, and national level: results from the global burden of disease study 2015 JAMA oncology, vol. 3, pp. 1683-1691.

[2] P. R. Galle, A. Forner, J. M. Llovet, V. Mazzaferro, F. Piscaglia, J.-L. Raoul, et al. (2018) EASL clinical practice guidelines: management of hepatocellular carcinoma. Journal of hepatology, vol. 69, pp. 182-236.

[3] H. B. El–Serag and K. L. Rudolph (2007) Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology, vol. 132, pp. 2557-2576.

[4] Z. Ding, N. Wang, N. Ji, and Z.-S. Chen (2022) Proteomics technologies for cancer liquid biopsies. Molecular Cancer, vol. 21, p. 53.

[5] S. Aksoy, M. ÖZAVSAR, A. ALTINDAL (2022) Classification of VOC Vapors Using Machine Learning Algorithms Journal of Engineering Technology and Applied Sciences, vol. 7, pp. 97-107.

[6] W. Naboulsi, D. A. Megger, T. Bracht, M. Kohl, M. Turewicz, M. Eisenacher, et al. (2016) Quantitative tissue proteomics analysis reveals versican as potential biomarker for early-stage hepatocellular carcinoma. Journal of proteome research, vol. 15, pp. 38-47.

[7] M. Schonlau and R. Y. Zou (2020) The random forest algorithm for statistical learning. The Stata Journal, vol. 20, pp. 3-29.

[8] F. Tang and H. Ishwaran (2017) Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 10, pp. 363-377.

[9] V. Fonti and E. Belitser (2017) Feature selection using lasso. vol. 30, pp. 1-25..

[10] T. Kimhofer, H. Fye, S. Taylor-Robinson, M. Thursz, and E. Holmes (2015) Proteomic and metabonomic biomarkers for hepatocellular carcinoma: a comprehensive review. British journal of cancer, vol. 112, pp. 1141-1156.

[11] X. Zheng, Q. Peng, L. Wang, X. Zhang, L. Huang, J. Wang, et al. (2020) Serine/arginine-rich splicing factors: the bridge linking alternative splicing and cancer. International journal of biological sciences, vol. 16, p. 2442.

[12] R. Karni, E. de Stanchina, S. W. Lowe, R. Sinha, D. Mu, and A. R. Krainer (2007) The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nature structural & molecular biology, vol. 14, pp. 185-193.

[13] O. Anczuków, M. Akerman, A. Cléry, J. Wu, C. Shen, N. H. Shirole, et al. (2015) SRSF1-regulated alternative splicing in breast cancer. Molecular cell, vol. 60, pp. 105-117.

[14] F. J. de Miguel, R. D. Sharma, M. J. Pajares, L. M. Montuenga, A. Rubio, and R. Pio (2014) Identification of alternative splicing events regulated by the oncogenic factor SRSF1 in lung cancer. Cancer research, vol. 74, pp. 1105-1115.

[15] C. Ghigna, S. Giordano, H. Shen, F. Benvenuto, F. Castiglioni, P. M. Comoglio, et al. (2005) Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene," Molecular cell, vol. 20, pp. 881-890.

[16] J. Long, Z.-W. Lang, H.-G. Wang, T.-L. Wang, B.-E. Wang, and S.-Q. Liu (2010) Glutamine synthetase as an early marker for hepatocellular carcinoma based on proteomic analysis of resected small hepatocellular carcinomas. Hepatobiliary Pancreat Dis Int, vol. 9, pp. 296-305.


## BIOGRAPHIES

**Şeyma YAŞAR** obtained her BSc. degree in mathematics from GaziosmanPaşa University in 2009. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning, proteomics, bioinformatics.