Selecting a Machine Learning Algorithm for the Prediction of Hepatocellular Carcinoma with the Determination of Key Indicators

Masuma Mammadova
Department of Number 11
Institute of Information Technology
Baku, Azerbaijan
0000-0002-2205-1023

Zarifa Jabrayilova
Department of Number 11
Institute of Information Technology
Baku, Azerbaijan
djabrailova_z@mail.ru

Lala Garayeva
Department of Number 11
Institute of Information Technology
Baku, Azerbaijan
0000-0003-2109-4280

2023 5th International Conference on Problems of Cybernetics and Informatics (PCI) | 979-8-3503-1906-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/PCI60110.2023.10325997

***Abstract*—Hepatocellular carcinoma, also known as liver cancer, is characterized by a large number of several types of indications. The HCC Dataset, data visualization and GridSearchCV hyper-parameter on the Kaggle website are used to determine the most important indicators. Using Logistic Regression, Support Vector Machine, K-nearest neighborhood, Naive Bayes and Random Forest machine learning algorithms, the method with the best prediction results is selected.**

Keywords—hepatocellular carcinoma, machine learning algorithms, data visualization

I. Introduction

Hepatocellular carcinoma (HCC) accounts for 80% (90% in the US) of liver cancer, the second leading cause of cancer-related deaths worldwide [1], [2]. HCC often manifests as chronic liver disease or is detected in patients with cirrhosis. HCC is annually found in about one million people in the world [1], [3]. According to recent data, HCC is estimated to be one of the most common fatal cancers in the world, causing more than 600,000 deaths every year [4], [5]. from the quantity of the interval of admissible controls are given.

HCC is represented with a large number of clinical indicators, critical situations defined by clinical signs, and there are no clear, unambiguous criteria for its diagnosis and treatment [1]. The fact that numerous indicators characterizing the critical situations of HCC are of different types and unstructured create the possibility of errors in making decisions regarding its diagnosis and forecasting. HCC-related data analysis shows that in terms of abundance of information, the doctor has to make a decision by referring to some of this information. As a result, errors occur in physicians' decisions determined by certain combinations of a large number of indicators and clinical symptoms.

To eliminate the problem, it is necessary to create the knowledge-based intelligent systems (health decision support systems) as a more effective tool for the collection, storage, manipulation of the knowledge of experienced medical experts, as well as for the HCC diagnosis on each specific data set and for making adequate decisions. These systems are used for HCC diagnosing and staging, choosing a more effective treatment method, HCC prediction, etc. In recent times, along with the creation of clinical decision support systems based on the knowledge of medical experts, there has been an increased effort on research in the field of disease prediction by referring to databases collected about clinical patients. These systems perform disease diagnosis and prediction with reference to machine learning and deep learning methods.

This paper proposes an algorithm for applying machine learning methods for predicting HCC based on more important indicators by visualizing the indicators characterizing HCC.

II. Model and Methods

Currently, scientific literature includes a certain number of studies on the development of intelligent systems for the detection, diagnosis and treatment of liver diseases [6]–[8], however little attention has been paid to the diagnosis and prediction of HCC. Nevertheless, these studies are conducted in two directions as in a number of other diseases, in accordance with the trend of applying artificial intelligence methods in this segment. The first direction includes the development of traditional medical expert systems for the diagnosis and treatment of HCC. The development of these systems is based on decision-making with reference to the knowledge of medical experts, or rather, evaluation of the patient's condition based on evidence-based medicine and selection of the appropriate outcome. Such systems, while representing all the advantages inherent in traditional ESs, prevent physician errors as a desktop tool. Within the framework of the development of such a system, we proposed a conceptual model of the intelligent system for the HCC diagnosis by referring to the knowledge of a physician-expert in [9], and elaborated the principles of creating a system for HCC diagnosis based on fuzzy rules in [10]. [11] presents the working principle of the blocks of the HCC diagnosis system, including the knowledge transformation and the knowledge base functionalization.

The second line of research on the diagnosis and prediction of HCC is based on collected clinical databases of received patients. By using such data, the forecasts according to various indicators, knowledge extraction, and new rules creation are performed on the basis of machine learning methods. With reference to such a base, [5] performs clustering to evaluate the main HCC patient groups, and evaluates the survival prognosis for these groups based on the K-means cluster and the SMOTE algorithm. [3] is dedicated to the possibilities of complex application of artificial intelligence methods for the prevention of HCC risk. In addition to the importance of applying machine learning and deep learning methods for the diagnosis and prediction of HCC, it shows the significance of referring to various data sources, including electronic health record data, visualization methods, histopathology and molecular biomarkers, to increase the accuracy of the prediction. It also provides the possibilities of complex application of artificial intelligence methods to standardize various data and generalize results, improve interpretability.

We highlight the results of our research using machine learning methods such as Logistic Regression (LR), Support

Vector Machine (SVM), Random Forest (RF) and the HCC Dataset taken from the Kaggle website for the HCC prediction in [12]. The use of the RF method, which provide better performance due to the prediction accuracy, in the creation of the HCC prediction system is justified.

It should be noted that a number of problems still need to be solved in the prediction of HCC by machine learning methods. One of them is related to the large number of clinical signs, which is the main obstacle in the process of creating new rules and extracting new knowledge from the collected data. This paper considers the solution to the problem of "removal" of more scattered parameters to reduce the amount of data in the selected database, proposes an algorithm for predicting HCC based on machine learning methods with reference to the main parameters, and compares the results.

III. Problem Solving

49 characteristic/attribute data of 165 patients in the HCC Dataset retrieved from the Kaggle platform are used to select the machine learning algorithm to build the HCC prediction system (figure 1). The database is formed from the data of 165 clinical patients suffering from HCC at the Hospital of the University of Portugal. The database includes 49 characteristics (clinical signs) recommended by *the European Association for the Study of the Liver - European Organization for Research and Treatment of Cancer*.

```
1 liver_df.head()
```

	Gender	Symptoms	Alcohol	HBsAg	HBeAg	HBcAb	HCVAb	Cirrhosis	Endemic	Smoking	...	ALP	TP	Creatinine	Nodule	Major_Dim	Dir_Bil	Iron	Sa
0	1	0	1	0	0	0	0	1	0	1	...	150	7,1	0,7	1	3,5	0,5	52,5	3
1	0	0	0	0	0	0	1	1	0	1	...	120	7	0,58	1	1,8	0,85	32	1
2	1	0	1	1	0	1	0	1	0	1	...	109	7	2,1	5	13	0,1	28	
3	1	1	1	0	0	0	0	1	0	1	...	174	8,1	1,11	2	15,7	0,2	131	7
4	1	1	1	1	0	1	0	1	0	1	...	109	6,9	1,8	1	9	0,1	59	1

5 rows × 50 columns

Fig. 1. HCC Dataset fragment [13]

Thus, the determination of the accuracy of the machine learning algorithms by reducing the number of indicators for the prediction of HCC and the selection of the algorithm performing a more accurate result are implemented according to the following steps:

A. .Data Preprocessing.

This step checks the relationships among the data in the HCC Dataset, and performs the cleaning of unrelated and scattered data by user intervention (figure 2). Therefore, this process is sometimes called base cleaning.

```
1 liver_df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 204 entries, 0 to 203
Data columns (total 50 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Gender          204 non-null    int64
 1   Symptoms        204 non-null    int64
 2   Alcohol         204 non-null    int64
 3   HBsAg           204 non-null    int64
 4   HBeAg           204 non-null    int64
 5   HBcAb           204 non-null    int64
 6   HCVAb           204 non-null    int64
 7   Cirrhosis       204 non-null    int64
 8   Endemic         204 non-null    int64
 9   Smoking         204 non-null    int64
 10  Diabetes        204 non-null    int64
 11  Obesity         204 non-null    int64
 12  Hemochro        204 non-null    int64
 13  AHT             204 non-null    int64
 14  CRI             204 non-null    int64
```

Fig. 2. Types of data in the HCC Dataset

Pandas [14] and NumPy [15] libraries are used to make the database useful. Out of 49 attributes in this database, 23 are quantitative and 26 are qualitative. The data characterizing the target class of the base takes the values 0 (death) and 1 (survival). Data of different types (object, int) in the base are brought to the same type (float) (figure 3).

```
1 liver_df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 204 entries, 0 to 203
Data columns (total 50 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Gender          204 non-null    float64
 1   Symptoms        204 non-null    float64
 2   Alcohol         204 non-null    float64
 3   HBsAg           204 non-null    float64
 4   HBeAg           204 non-null    float64
 5   HBcAb           204 non-null    float64
 6   HCVAb           204 non-null    float64
 7   Cirrhosis       204 non-null    float64
 8   Endemic         204 non-null    float64
 9   Smoking         204 non-null    float64
 10  Diabetes        204 non-null    float64
 11  Obesity         204 non-null    float64
 12  Hemochro        204 non-null    float64
 13  AHT             204 non-null    float64
 14  CRI             204 non-null    float64
```

Fig. 3. Data standardization in the HCC Dataset

Using correlation heatmaps, both linear and non-linear relationships between data in the HCC Dataset are determined (figure 4).

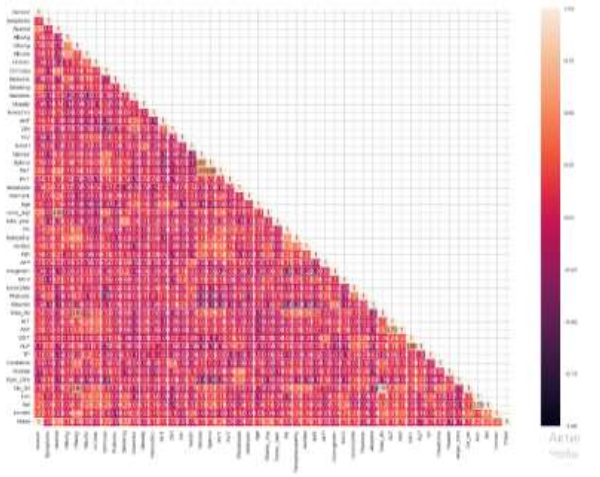

Fig. 4. Correlation heatmaps illustrating the data relationship in the HCC Dataset

B. Data Visualization

In the reference database, a match is determined between the data according to the target class (0-"death", 1 - "survival"). In this regard, data visualization is performed. Figure 5 shows a visual representation of the correlation among "Age", "Albumin", "Total_Bil" attribute.

```
sns.pairplot(df2, hue='Class', vars=['Age','Albumin', 'Total_Bil' ])
plt.show()
```

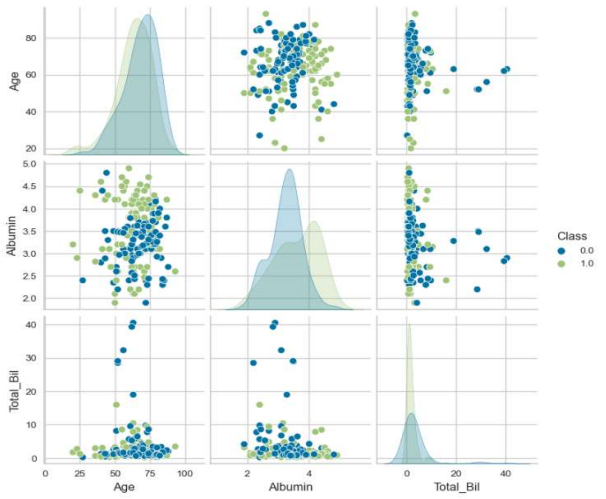

Fig. 5. Visual representation of correlation among "Age", "Albumin", "Total_Bil" attribute.

According to the visualization result, unrelated data are determined, they are "removed" from the database with the participation of the user (doctor), and the problem solution continues on 22 attribute (figure 6).

```
X = X[:,[5,6,7,10,11,13,14,15,16,17,18,19,20,21,22,23,26,27,28,29,31,32,35,36,44]]
```

```
#X_train,X_test,y_train,y_test=train_test_split(X,y,random_state=0)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state = 42)
```

Fig. 6. Relationship visualization through Correlation heatmaps among the 22 most important attribute in the HCC Dataset

Training the model is an important step to get good results and to find correlations among data and to "avoid" unrelated data. To avoid overfitting and underfitting of the model, 80% of the data are selected as training data and the remaining 20% as testing data.

It is assumed that a ratio of 80/20 of training and test data will be sufficient to achieve good classifier accuracy. In order to properly use the database and obtain the highest accuracy, we use the evaluation criteria presented in [14].

C. Classification

This step classification and error matrix criteria determination**.** KNN, Naive Bayes, SVM, RF and LR machine learning algorithms are used to perform classification based on GridSearchCV parameters. For evaluating the classifiers' detection performance in machine learning, precision, recall, false positive rate (FPR), true positive rate (TP), f-measure, and accuracy criteria are used.

Precision (P) denotes the proportion of the number of true positives to whole predicted positives as follows:

$$P = \frac{T_p}{T_p + F_p}$$

Here: Tp denotes the number of data related to correctly classified prediction.

Fp denotes the number of data unrelated to a misclassified prediction.

Recall (R) is defined as the proportion of the number of true positives to all actual positives is calculated using the following formula:

$$R = \frac{T_p}{T_p + F_n}$$

Where: Fn is the number of data unrelated to the prediction classified as errors.

F1-Score is defined as the harmonic mean of the recall and precision, and is calculated by the following formula:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

Accuracy is defined as follows:

$$A = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

D. Analysis of the Results

Analysis of the results obtained by applying machine learning algorithms. Prediction results based on KNN, Naive Bayes, SVM, RF and LR machine learning algorithms are analyzed. Jupiter program is used in the Anaconda environment for this.

Figure 7 illustrates the values of the accuracy matrix criteria and the results of the accuracy criteria obtained from the application of the classifiers.

Machine learning algorithms	F1 Score	Recall	Accuracy	Mean
KNN(K=5)	0.631579	0.6	0.666667	0.632749
Logistic Regression	0.600000	0.6	0.619048	0.606349
SVM	0.571429	0.6	0.571429	0.580952
Random Forest	0.900000	0.9	0.904762	0.901587
Naive Bayes	0.615385	0.8	0.523810	0.646398

Fig. 7. Illustration of the accuracy criteria obtained from the classification

E. Comparison of Results.

According to the results obtained, RF algorithm performs better in terms of accuracy, and it is appropriate to use it in the HCC prediction system. It should be noted that in [12], that is, in the experiment conducted without visualization in the selected base, the RF algorithm performed better results for HCC prediction too.

To check the visualization accuracy, Table 1 presents the comparison of the results obtained in the present study with the results obtained without visualization, more precisely, those obtained in [12].

TABLE I. PREDICTION ACCURACY (NON-VISUALIZATION AND VISUALIZATION) BASED ON DATA RETRIEVED FROM KAGGLE COMPANY HCC DATASET

Machine learning algorithms	Non-visualization [12]	Visualization
Logistic Regression	76, 19	60,63
SVM	76,19	58,09
Random Forest	90,48	90,18

As the table shows, the result obtained by the RF machine learning algorithm with respect to others is almost the same as the result obtained in [12]. This shows that the RF algorithm is more tolerant to visualization in terms of prediction accuracy, it shows the precision of the visualization, and justifies the feasibility of using RF in the HCC forecasting system once again.

IV. CONCLUSION

The article presented an algorithm for selecting a machine learning method performing better by "removal" more "scattered" data and selecting more significant indicators in the Kaggle company HCC Dataset. 22 attribute out of the 49 attribute were selected as more significant based on the visualization of their correlation with one another according to the target class (0-death, 1-survival) in the HCC Dataset. The application of KNN, Naive Bayes, RF, SVM and LR machine learning methods to predict HCC based on 22 more important parameters was given in stages. The use of the RF algorithm in the HCC forecasting system was justified and it was estimated to be more tolerant to the "removal" of "scattered" data from the database.

Selection of more important indicators from the HCC Dataset, reducing the number of indicators characterizing clinical patients allowed obtaining new rules for predicting HCC, which is important for supporting physician decisions. Development of new prediction rules based on data from the HCC Dataset is one of the issues to be solved in our further research.

REFERENCES

[1] Nuru Y. Bayramov, "Surgical diseases of the liver", Baku: Qismet, 2012, 327 p..

[2] S. Xiaopu, W. Fenfang, W. Di, L.Shan, L. Jingyi, Z. Nan, C. Xiaoni, and X. Anlong, "Human Hepatic Cancer Stem Cells (HCSCs) Markers Correlated With Immune Infiltrates Reveal Prognostic Significance of Hepatocellular Carcinoma", Frontiers in Genetics, 28 February 2020. https://doi.org/10.3389/fgene.2020.00112

[3] J. Calderaro, T. P. Seraphin, T. Luedde, and T. G. Simon, "Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma", Journal of Hepatology, 2022, vol. 76, pp. 1348-1361.

[4] D. Shetty, K. Rit, S. Shaikh, and N. Patil, "Diabetes disease prediction using data mining", International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1–5.

[5] M. Santos, P. Henriques Abreu, P. Garc´ıa-Laencina, A. Simao, and A. Carvalho, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients", Journal of biomedical informatics, 2015, vol. 58, pp. 49–59.

[6] S. Aman, and P. Babita, "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, 2016, vol.11, no.4, pp.56–61.

[7] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)", Computer Methods and Programs in Biomedicine, 2015, vol.108, no.2, pp.570–579.

[8] F. Gorunescu, S. Belciug, M. Gorunescu, and R. Badea, "Intelligent decision-making for liver fibrosis stadialization based on tandem feature selection and evolutionary-driven neural network", Expert Systems with Applications, 2012, vol.39, no.17, pp. 12824–12832.

[9] Masuma H. Mammadova, Nuru Y. Bayramov, Zarifa G. Jabrayilova, Minara I. Manafli, and Mehriban R. Huseynova, "Principles for the development of an intelligent hepatocellular carcinoma staging system", Problems of information technology, 2021, vol. 11, no. 1, pp. 3-14.

[10] Masuma H. Mammadova, Nuru Y. Bayramov, and Zarifa G. Jabrayilova, "Development principles of fuzzy rule-based system for hepatocelular carcinoma staging", Eureka:physics and engineering, 2021, no.3, pp. 3-13.

[11] Masuma H. Mammadova, Nuru Y. Bayramov, Zarifa G. Jabrayilova, Minara I. Manafli, and Mehriban R. Huseynova, "An algorithm for the decision synthesis in the remote monitoring system of physiological state of workers employed at high-risk facilities", 8th Conference on Control and Optimization with Industrial Applications-COIA'2022, 24-26 August 2022, Baku, Azerbaijan, pp. 318-320.

[12] Masuma H. Mammadova, Zarifa G. Jabrayilova, Lala A. Garayeva, and Ayten A. Ahmadova, "Prediction of hepatocellular carcinoma using a machine learning", The 16th IEEE International Conference Application of Information and Communication Technologies (AICT-2023), Washington DC, 12-14 Oct 2022, INSPEC Accession Number: 22541899, DOI: 10.1109/AICT55583.2022.10013575, https://ieeexplore.ieee.org/document/10013575

[13] www.kaggle.com

[14] S. Yadav, and S. Shukla, "Analysis of k-fold cross-validation over holdout validation on colossal datasets for quality classification," in 2016 IEEE 6th International Conference on Advanced Computing (IACC), 2016, pp. 78–83.

[15] C. R. Harris, K. J. Millman, and et.al., "Array programming with NumPy," Nature, 2020, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2