# Applying the Shapley Value Method to Predict Mortality in Liver Cancer Based on Explainable AI

Lun-Ping Hung[1(✉)], Chong-Huai Xu[1], Ching-Sheng Wang[2], and Chien-Liang Chen[3]

[1] Department of Information Management, National Taipei University of Nursing and Health Sciences, Taipei 112303, Taiwan
lunping@ntunhs.edu.tw

[2] Department of Computer Science and Information Engineering, Aletheia University, Taipei 25103, Taiwan

[3] Department of Innovative Living Design, Overseas Chinese University, Taichung 40721, Taiwan

**Abstract.** Hepatocellular carcinoma (HCC) is the sixth-leading cause of death worldwide and has the highest mortality rate among all types of cancers. In most cases, the patient has entered the terminal phase of a cancer disease when hepatocellular carcinoma occurs. Therefore, if the cause of cancer can be identified, disease deterioration can be prevented. With the rise of artificial intelligence (A.I.) technology in recent years, many scholars have used machine learning technology to predict the probability of dying from hepatocellular carcinoma and have obtained good results. However, the studies lack interpretability and do not facilitate the further analyses of medical experts. Therefore, this study proposes a deep learning model based on XGBoost and uses the data evaluation method of Shapley value to study the characteristics of machine learning and verify the results using the hepatocellular carcinoma dataset. The proposed model delivered strong prediction performance, with an accuracy of 92.68%, and accurately interpreted the dataset features, supporting analyses by medical experts.

**Keywords:** Machine learning · Hepatocellular carcinoma · Risk factors · SHapley Additive exPlanations (SHAP) · Extreme gradient boosting (XGBoost)

## 1 Introduction

Hepatocellular carcinoma is the sixth-leading cause of death worldwide and has the highest mortality rate among all cancer types. In most of the cases, there is no significant symptom of HCC in the early phase of cancer, and the patient has entered the terminal phase when HCC occurs [1]. Therefore, if HCC is identified at an earlier stage for prevention and treatment, the mortality can be reduced. With scientific and technological advancement, scholars began to explore the application of machine learning in the medical care of diseases [2], such as lung [3], breast [4] and liver cancer [5]. Also, machine learning can be used to conduct clinical research in a virtual environment [6].


D.-J. Deng et al. (Eds.): SGIoT 2022, LNICST 497, pp. 133–143, 2023.
https://doi.org/10.1007/978-3-031-31275-5_14

The goal of machine learning research is mostly to improve algorithm accuracy, which is vital to medical diagnosis. However, in clinical practice, physicians often encounter complex situations that cannot be solved only with accuracy. For example, in deep learning methods with high precision, the characteristics of black box techniques will cause difficulties for physicians in understanding and evaluation, and they cannot accept the advice without a scientific basis when faced with serious diseases. Additionally, many medical institutions lacked the medical data to satisfy the criteria of a training set for deep learning.

Many researchers have focused on how to interpret complex and powerful models such as CNN [7], XNN [8] and LSTM [9], which were all common and practical machine learning tools. However, the explanations given by researchers regarding the information of the models were still incomprehensible for physicians; some new studies attempt to provide a more comprehensible model interpretation for physicians, to ease their worries and support their decision-making process. For example, Bas H.M. et al. used the Explainable AI (XAI) standard framework for classification in medical image analysis and conducted investigations and classification in a paper about XAI techniques based on frameworks and anatomical location [10]. Andreas et al. proposed using explainable AI via multi-modal and multi-center data fusion to address the lack of interpretability and transparency [11].

A challenge remains in how to strike a balance between accuracy, interpretability, and other AI factors in the medical field. Therefore, this study attempts to integrate the methods of solving these problems into the prediction of liver cancer mortality by proposing a method of predicting liver cancer mortality incorporating Shapley value and ensemble learning. Compared with the traditional method of machine learning that trains a single classifier, ensemble learning, or the method that combines multiple classifiers, can improve generalizability or accuracy [12]. We use the eXtreme Gradient Boosting (XGBoost) machine learning framework to predict the risk of liver cancer mortality and use Shapley value to interpret the causes of the prediction results. This approach has the following advantages: 1. Allowing physicians to understand the importance of each dataset attribute, compared with other attributes. 2. Effectively improving physicians' disease diagnosis and prognosis decision-making. 3. No necessity of having a large amount of training data, and the accuracy of predicting liver cancer mortality is relatively high.

## 2 Related Works

As the techniques designed for machine learning are increasingly popular, the studies of applying machine learning to cancer diagnoses are also maturing. The intervention of machine learning has successfully improved research efficiency and generated results with low error rates, effectively helping cancer diagnosis [13]. Except for this, the subfield of machine learning known as XAI can interpret complex artificial intelligence models [14], and studies have found that the machine learning systems that have been interpreted can support clinical cancer diagnosis [15]. Therefore, in the following paragraphs, this study categorizes and explains the relevant literature about the interpretable prediction of liver cancer diagnosis:

### 2.1 Shapley Additive Explanation (SHAP)

SHAP (SHapley Additive exPlanations) is a game theoretic approach to interpret the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions [16]. The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction. Compared with the traditional method that analyzes the differences in the features' importance, SHAP only clarifies which of the features are more important to the model, without explaining how the features affect the results. The positive and negative values of SHAP correspond to the effect of each sample's features, which reveals the effects of which features in the dataset are the most important. The interpretation of SHAP itself is an additive feature attribution method and is similar to the linear regression method [17].

### 2.2 XGBoost

XGBoost (eXtreme Gradient Boosting) is a framework based on gradient boosting that integrates the advantages of bagging and boosting and is mainly used for monitoring and learning, while it can also be applied to classification and regression analysis. XGBoost is composed of a set of classification and regression trees (CART). Each leaf of the regression trees is assigned a set of scores, which are the basis for the subsequent classification. The trees are interrelated and the goal is to generate news trees that can correct the mistakes of the previous tree [18]. As shown in Fig. 1, each person is assigned to

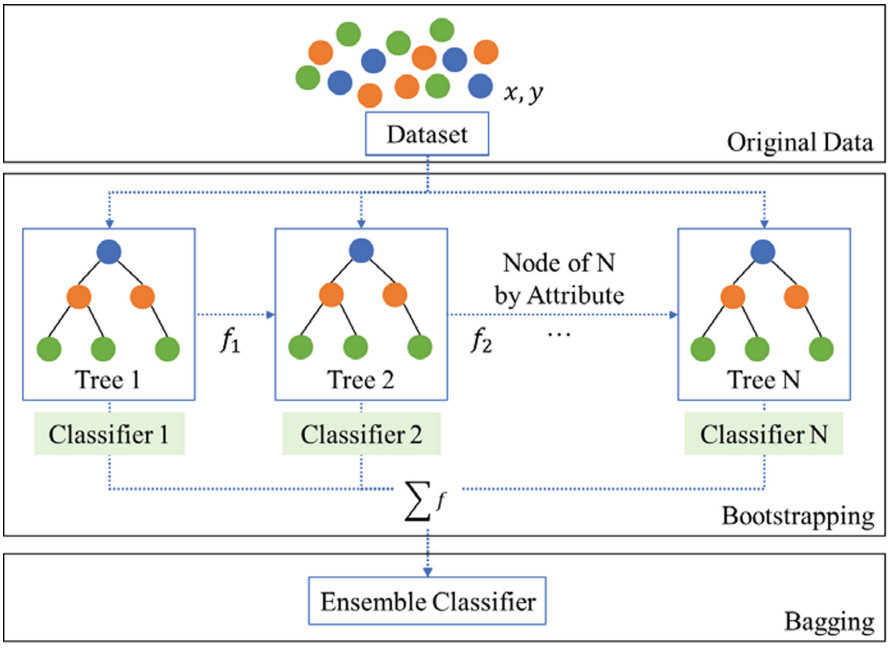


**Fig. 1.** XGBoost architecture.

different leaves, and is assigned a score based on the leaf the person belongs to. The difference between CART and decision trees is that the leaves of decision trees only contain decision values, while the scores of CART are related to all leaves, and the final individual score of each person is obtained by summing the scores obtained from each tree.

## 3 Research Methods and Results

This study proposes a new interpretable prediction system based on Shapley value that uses XAI technology to combine XGBoost to predict the death risks by focusing on liver cancer cells, effectively improving physicians' disease diagnosis and prognosis decision-making. This research analyzes and generate ROC curves, ROC performance evaluation and SHAP visualization results, which are explained as follows:

### 3.1 Patients Data Gathering

The database of liver cancer deaths provided by the UCI Machine Learning Repository is used here, and the data is obtained from a university hospital in Portugal. There are a total of 204 records in the data [19], and the dataset includes 50 variables selected according to EASL-EORTC (European Association for the Study of the Liver-European Organization for Research and Treatment of Cancer), among which the target-dependent variables are: Class (1 denotes death and 0 denotes survival); Table 1 displays the categories and attributes of its data. After reading the data, it is found that there is no missing value.

**Table 1.** Dataset attribute table.

| Attribute | Type | Possible values |
|---|---|---|
| Gender | int64 | Male or female |
| Symptoms | int64 | Type of Symptoms - True or False (1 or 0) |
| Alcohol | int64 | Alcohol Usage - Yes or No (1 or 0) |
| HBsAg | int64 | Hepatitis Markers - Present or Absent (1 or 0) |
| HBeAg | int64 | Hepatitis Markers - Present or Absent (1 or 0) |
| HBcAb | int64 | Hepatitis Markers - Present or Absent (1 or 0) |
| HCVAb | int64 | Hepatitis Markers - Present or Absent (1 or 0) |
| Cirrhosis | int64 | Liver Cirrhosis - Present or Absent (1 or 0) |
| Endemic | int64 | Specific Endemic disease - (Like Malaria - Present or Absent (1 or 0) |
| Smoking | int64 | Smokes or not (1 or 0) |
| Diabetes | int64 | Is the patient Diabetic (1 or 0) |
| Obesity | int64 | Is the patient Obese (1 or 0) |

(*continued*)

**Table 1.** (*continued*)

| Attribute | Type | Possible values |
|---|---|---|
| Hemochro | int64 | Body loads too much of Iron - Yes or No (1 or 0) |
| AHI | int64 | AHT Present or not |
| CRI | int64 | Chronic Renal Insufficiency - Yes or no (1 or 0) |
| HIV | int64 | Does the patient have HIV -Yes or No (1 or 0) |
| NASH | int64 | Non-Alcoholic Fatty Liver steatohepatitis (NASH) - Yes or No (1 or 0) |
| Varices | int64 | Presence of Esophageal Varices - (1 or 0) |
| Spleno | int64 | Presence of Gastric Varices like bleeding in upper intestinal tract - (1 or 0) |
| PHT | int64 | Parathyroid Hormone Test - Present or Absent - (1 or 0) |
| PVT | int64 | other Pathology test for HCC confirmation (1 or 0) |
| Metastasis | int64 | Presence of Cancer in Bones and other organs - Present or Absent - (1 or 0) |
| Hallmark | int64 | Cancer Markers Test -Present or Absent (1 or 0) |
| Age | int64 | Age of the patient |
| Grams_day | int64 | Doses given -Grams per day |
| Packs_year | float64 | No of Cigar Packets per Year |
| PS | int64 | Staging of HCC |
| Encephalopathy | int64 | End Stage liver disease |
| Ascites | int64 | Poor Outcome in the absence of Transplantation |
| INR | float64 | Used to assess coagulation function |
| AFP | float64 | Biolevel Markers for early HCC |
| Hemoglobin | float64 | 12 to 17.5 gms per deciliter |
| MCV | float64 | 80 to 96 |
| Leucocytes | float64 | 4500 to 11000 WBC per microliter |
| Platelets | float64 | 150,000 to 450,000 platelets per microliter |
| Albumin | float64 | 3.5 to 5.5 g/dL |
| Total_Bil | float64 | 0.1 to 1.2 mg/dL |
| ALT | int64 | 7 to 56 Units |
| AST | int64 | 10 to 40 units – Normal Range |
| GGT | int64 | 9–48 units per liter |
| ALP | int64 | 44 to 147 |

(*continued*)

**Table 1.** (*continued*)

| Attribute | Type | Possible values |
|---|---|---|
| TP | float64 | 6 and 8.3 Range |
| Creatinine | float64 | 0.6 to 1.2 |
| Nodule | int64 | The size of the nodules determines the liver disease |
| Major_Dim | float64 | Dimension of the Tumor |
| Dir_Bil | float64 | Upto 1.2 mg/dl |
| Iron | float64 | 13.5 to 17.5 |
| Sat | float64 | Iron related test (in Numerical value) |
| Ferritin | float64 | 12 to 300 Range |
| Class | int64 | Present or Absent (1 or 0) |

### 3.2 Predictive Assessment

The following cross-validation method is used to evaluate liver cancer mortality. As the true classification of each condition is known, the values of the absolute performance indicators of a classifier are calculated with a confusion matrix:

- TP (true positive) – died from liver cancer in reality and has been predicted to die from liver cancer.
- TN (true negative) – did not die from liver cancer in reality and has been predicted to die from liver cancer.
- FP (false positive) – died from liver cancer in reality but has not been predicted to die from liver cancer.
- FN (false negative) – did not die from liver cancer in reality and has not been predicted to die from liver cancer.

The corresponding relative indicators: Accuracy is the ratio of correctly classified samples by the classifier to the total sample in the tested dataset, and is a comprehensive score that reflects the overall performance of the classifier.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Recall is the prediction performance on positive cases of the classification model. In the prediction of liver cancer deaths, if the performance of confirming positive cases can be improved, the survival rate can also be improved, as the patients can be treated earlier with earlier diagnosis. Therefore, recall is one of the measurement indicators of performance in machine learning.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$F_\beta$ is a score value that combines two types of score values, accuracy and recall. Additionally, in the combination process, the weight of recall is $\beta$ times that of accuracy,

which indicates that β represents the relative importance of accuracy and recall. Considering that cancer treatment may have a negative impact on the patient, β is set to 1 (i.e., F1), while an F1 score represents that accuracy and recall are considered to be equally important.

$$F_\beta = \left(1+\beta^2\right)\frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \textit{Precision} + \textit{Recall})}, \text{where Precision} = \frac{TP}{TP+FN} \tag{3}$$

The receiver operating characteristic (ROC) curve is used, and a decision threshold is needed for other measurement methods (accuracy, recall and F1 score). In a ROC curve, the model's continuous outputs of probabilities and analogous probabilities are collapsed to become one set of classification prediction results. The ROC curve may come from the continuous outputs of probabilities, and is an efficient method to evaluate the model's performance at the decision threshold. AUC is the area under the ROC curve, and is the most commonly used summary indicator for ROC curves. In general, a higher AUC represents a better performance of the classifier.

### 3.3 Results

Table 2 compares the accuracy, recall and F1 score of the classifier. The values shown in Table 2 result from 10-fold cross-validation. The results show that the accuracy of XGBoost is 92.68%, and the recall and the F1 score are better than that of other classifiers. Furthermore, the performance of XGBoost is more stable than other classifiers. Although the accuracy of neural networks is relatively high, their performance may be poor when the amount of training data is low, because neural networks can be affected by the characteristic of overfitting.

Figure 2 presents the ROC curves of 10 classifiers, and it shows that the performance of XGBoost is significantly better than most of the other classifiers, and is not inferior to other classifiers in terms of AUC.

**Table 2.** Predicted performance of the classifier.

| Classifier | Accuracy (%) | Recall (%) | F1-score (%) |
|---|---|---|---|
| KNN | 85.36 | 85.23 | 85.28 |
| Decision Tree | 75.60 | 75.35 | 7524 |
| Random Forest | 90.24 | 90.11 | 90.19 |
| Naive Bayes | 56.09 | 56.78 | 52.69 |
| Linear Regression | 87.80 | 87.85 | 87.80 |
| Support Vector Machine | 92.68 | 92.85 | 92.66 |

(*continued*)

**Table 2.** (*continued*)

| Classifier | Accuracy (%) | Recall (%) | F1-score (%) |
|---|---|---|---|
| Gradient Boosting Classifier | 85.36 | 85.23 | 85.28 |
| AdaBoosting Classifier | 75.60 | 75.59 | 75.59 |
| Neural Network | 92.68 | 92.73 | 92.68 |
| XGBoost | 92.68 | 92.5 | 92.61 |

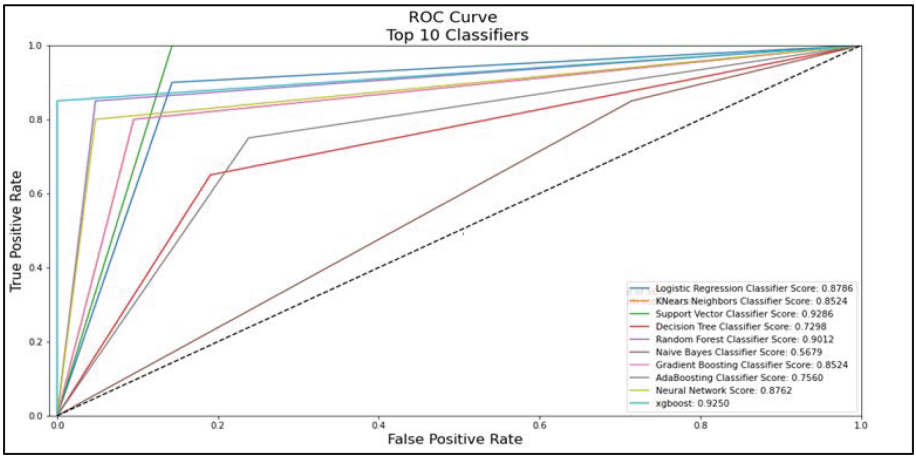


**Fig. 2.** ROC Curve.

### 3.4 Using SHapley Additive Explainable

**SHAP Summary Chart**

The features are ranked according to the sum of SHAP values of all samples to obtain the 20 features that have the greatest impact on the model output, as shown in Fig. 3, while the distribution of the influence of each feature is presented with SHAP values, in which different colors represent different feature values (red represents a high value and blue represents a low value). For example, a higher AFP feature is more likely to affect the probability of dying from liver cancer.

**SHAP Feature Map**

The SHAP values show how each feature affects the model's output. As SHAP values represent the key to the changes in the model caused by the features, the research results show that there is a great impact on the prediction of liver cancer death as the Albumin feature parameters changes. The vertical discrete value of a single Albumin value has an interactive effect with other features. Different colors are used to help distinguish this feature, as shown in Fig. 4. For example, areas with higher PHT values have a lower effect on the Albumin values that affect the probability of dying from liver cancer.

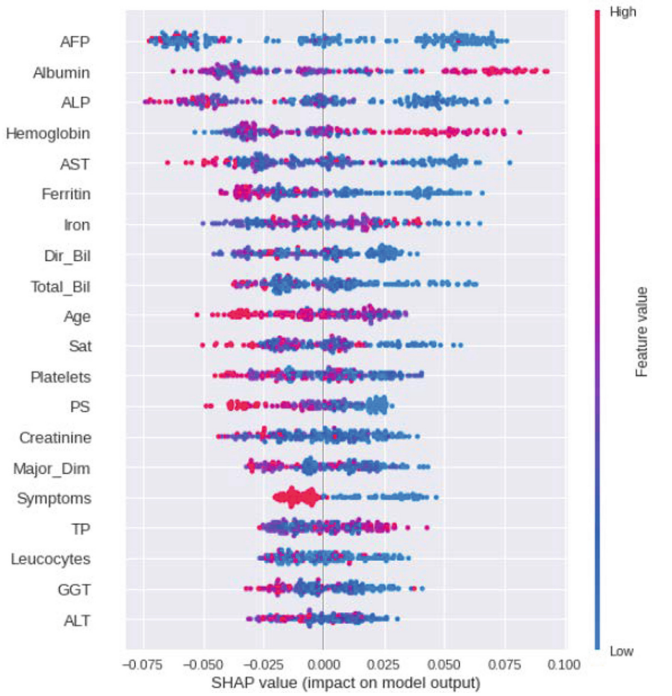


**Fig. 3.** SHAP values Summary Chart. (Color figure online)

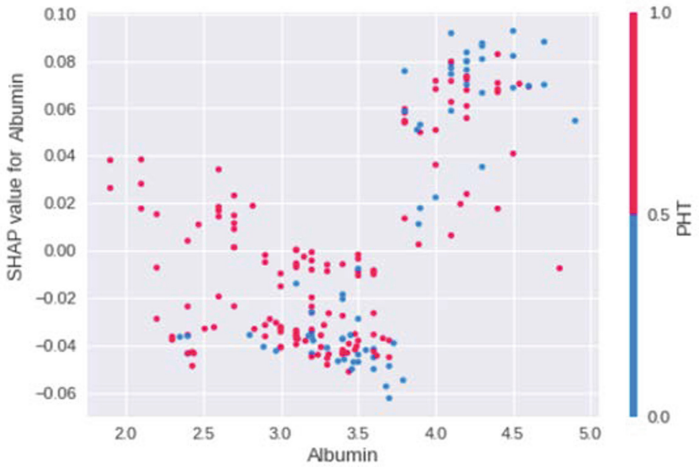


**Fig. 4.** SHAP dependence plot.

**Personal Risk Explanation**

The SHAP heatmap is used to explain the individuals' risks, and the outputs are the features of a person's risks of dying from liver cancer, interpreting each risk feature

that helps the model to output the average predicted values from the training dataset. Figure 5 shows the values predicting the death of someone suffering from liver cancer. In the figure, the features pushing up the prediction values are marked in red, which indicates that they increase the risk of death, while the features lowering the prediction values are marked in blue, which indicates that they decrease the death risk. For example, the negative effect of Albumin is the largest, because when the Albumin level is equal to 3.2 g/dL, it falls in a normal range, and thus reduces the risk of death.

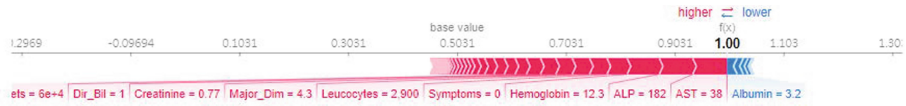


**Fig. 5.** SHAP force plot.

## 4 Conclusion

As machine learning technology is matures, many medical decision-making systems tend to improve accuracy and ignore physicians' professional competence. When physicians need to make a decision with the patient's life at stake, they still find it difficult to trust the prediction result analysis generated by the systems. There remains a lack of specific rules for existing medical decision-making systems. Many studies have applied interpretable machine learning to the field of medical treatment to provide physicians with a trustworthy medical support system. However, in many studies, the physicians still need to spend a lot of time comprehending the provided results, bringing about an additional burden for the physicians. To solve the problem, this study used the XGBoost classifier to predict the risk of dying from liver cancer, and interpreted the features in the dataset that affect the death rate according to the model's results, helping physicians to better understand the way of operating machine learning. In our evaluation, the classifier has a good performance in prediction, and the accuracy reaches 92.68%. Regarding interpretability, the interpretation of individual risks shows each person's potential risk of dying when they suffer from liver cancer, while visualization is used to comprehend the feature with the higher influence. The features and values in the dataset that cause the risks to increase or decrease have been observed. However, this study has not focused on the analysis of the correlation between features. The research results show that some features will affect the prediction results, and the features may affect each other. In the future, it is necessary to discuss the correlation between features in a dataset, and increase the number of datasets, to ensure that the system can generate more accurate results.

## References

1. Yan, Q., et al.: Application and progress of the detection technologies in hepatocellular carcinoma. Genes Dis. 2 (2022)

2. Sood, S.K., Rawat, K.S., et al.: A visual review of artificial intelligence and industry 4.0 in healthcare. Comput. Electr. Eng. **101**, 107948–107962 (2022)
3. Saba, T., Sameh, A., Khan, F., Shad, S.A., Sharif, M.: Lung nodule detection based on ensemble of hand crafted and deep features. J. Med. Syst. **43**(12), 1–12 (2019). https://doi.org/10.1007/s10916-019-1455-6
4. Li, J., et al.: Predicting breast cancer 5-year survival using machine learning: a systematic review. PLoS One **16**(4), e0250370 (2021)
5. Oza, P., et al.: A Bottom-up review of image analysis methods for suspicious region detection in mammograms. J. Imaging **7**(9), 1–40 (2021)
6. Moingeon, P., Kuenemann, M., et al.: Artificial intelligence-enhanced drug design and development: toward a computational precision medicine. Drug Discov. Today **27**(1), 215–222 (2022)
7. Maweu, B., et al.: CEFEs: A CNN explainable framework for ECG signals. Artif. Intell. Med. **115**, 102059–102074 (2021)
8. Qi, Z., Li, F.: Embedding deep networks into visual explanations. Artif. Intell. **292**, 1–27 (2017)
9. Yudistira, N., et al.: Learning where to look for COVID-19 growth: multivariate analysis of COVID-19 cases over time using explainable convolution-LSTM. Appl. Soft Comput. **109**, 107469–107487 (2021)
10. van der Velden, B.H.M., et al.: Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med. Image Anal. **79**, 102470–102490 (2022)
11. Holzinger, A., et al.: Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Inf. Fusion **79**, 263–278 (2022)
12. Kazmaier, J., van Vuuren, J.H.: The power of ensemble learning in sentiment analysis. Expert Syst. Appl. **187**, 115819–115834 (2022)
13. Painuli, D., Bhardwaj, S., et al.: Recent advancement in cancer diagnosis using machine learning and deep learning techniques: a comprehensive review. Comput. Biol. Med. **146**, 105580–105609 (2022)
14. Barredo Arrieta, A., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)
15. Gu, D., Su, K., et al.: A case-based ensemble learning system for explainable breast cancer recurrence prediction. Artif. Intell. Med. **107**, 101858–101866 (2020)
16. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**(3), 647–665 (2013). https://doi.org/10.1007/s10115-013-0679-x
17. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777. Curran Associates Inc., Long Beach (2017)
18. Chen, T., et al.: Xgboost: extreme gradient boosting. R Package Version 0.4-2 **1**(4), 1–4 (2015)
19. Santos, M.S., et al.: A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J. Biomed. Inform. **58**, 49–59 (2015)