Explaining Massive-Training Artificial Neural Networks in Medical Image Analysis Task Through Visualizing Functions Within the Models

Ze Jin, Maolin Pang, Yuqiao Yang, Fahad Parvez Mahdi, Tianyi Qu, Ren Sasage, and Kenji Suzuki(✉)

Biomedical Artificial Intelligence Research Unit, Institute of Innovative Research, Tokyo Institute of Technology, Kanagawa, Japan
{jin.z.ab,suzuki.k.di}@m.titech.ac.jp

Abstract. In this study, we proposed a novel explainable artificial intelligence (XAI) technique to explain massive-training artificial neural networks (MTANNs). Firstly, we optimized the structure of an MTANN to find a compact model that performs equivalently well to the original one. This enables to "condense" functions in a smaller number of hidden units in the network by removing "redundant" units. Then, we applied an unsupervised hierarchical clustering algorithm to the function maps in the hidden layers with the single-linkage method. From the clustering and visualization results, we were able to group the hidden units into those with similar functions together and reveal the behaviors and functions of the trained MTANN models. We applied this XAI technique to explain the MTANN model trained to segment liver tumors in CT. The original MTANN model with 80 hidden units (F1 = 0.6894, Dice = 0.7142) was optimized to the one with nine hidden units (F1 = 0.6918, Dice = 0.7005) with almost equivalent performance. The nine hidden units were clustered into three groups, and we found the following three functions: 1) enhancing liver area, 2) suppressing non-tumor area, and 3) suppressing the liver boundary and false enhancement. The results shed light on the "black-box" problem with deep learning (DL) models; and we demonstrated that our proposed XAI technique was able to make MTANN models "transparent".

Keywords: Deep Learning · Explainable AI (XAI) · Visualizing Functions · Liver Tumor Segmentation · Unsupervised Hierarchical Clustering

1 Introduction

Artificial intelligence (AI) research has evolved rapidly, and unprecedented breakthroughs have been made in many fields. Applications of AI products can be witnessed in our daily life, such as autonomous driving, computer-aided diagnosis, automatic voice customer service, etc. The development of AI is undoubtedly a revolution in the course of human history.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 713–722, 2023.
https://doi.org/10.1007/978-3-031-43895-0_67

The most effective and commonly used AI model is the one based on deep neural networks [1]. However, with continuous research being held in methodologies, DL models are becoming more and more complicated. Researchers found that the deeper and more complex DL models are, the better the performance they could achieve for the tasks that traditional AI algorithms could not work well. The complexity of DL models reduces interpretability and transparency substantially; therefore, the current DL models are "black-box" [2]. It is difficult to find how the model works in a way that humans can understand. Because of that, what researchers can do is only to prepare enough data and spend time training a model to obtain a high performance. Therefore, researchers or users can hardly find the reason why a DL model made a wrong decision.

XAI is an old area in AI research, but was named relatively recently [3, 4], focusing now on the explainability of DL models. The final goal of XAI is to develop methods for revealing a basis for the decision made by a DL model and how the decision was made by the model to let users understand and trust the decision and model. Many XAI methods have been proposed to explain a trained DL model (i.e., post-hoc methods). Representative XAI methods include class activation mapping (CAM) [5], grad-CAM, layer-wise relevance propagation (LRP) [6], DL important features (DeepLIFT) [7], local interpretable model-agnostic explanations (LIME) [8], and SHapley additive explanations (SHAP) [9]. These XAI methods offer post-hoc explanations that indicate which areas in a given input image the trained model focuses on and identify which areas in the image have a positive or negative impact on the model decision. In other words, those XAI methods are "instance-based" and limited to the visual explanation of model's attentions in a given input image (i.e., an instance). However, they do not offer explanations of the learned functions of the network.

In this study, we developed and presented an original XAI approach that can reveal the learned functions of groups of neurons in a neural network, which we call "functional explanations" and define as explanations of the model behavior by a combination of functions, as opposed to the visualization of a pattern to which a neuron responds. To our knowledge, there is no XAI method that offers functional explanations. Thus, our method is a post-hoc method that offers both instance-based and model-based functional explanations**.** We applied our XAI method to an MTANN model to emphasize the explainability and trustability of the MTANN, so that users can trust the MTANN.

2 Method

2.1 MTANN Deep Learning

In the field of image processing, supervised nonlinear filters and edge enhancers based on an artificial neural network (ANN) [10] have been investigated for the reduction of the quantum noise in angiograms and supervised semantic segmentation of the left ventricles in angiography [11], which are called neural filters and neural edge enhancers, respectively. By extending the neural filter and edge enhancer, massive-training artificial neural networks (MTANNs) have been developed to reduce false positives in the computerized detection of lung nodules in computed tomography (CT) [12]. The MTANNs have also shown promising performance in pattern recognition and classification tasks [13, 14].

An MTANN is a deep learning model consisting of linear-output artificial neural network regression model that directly operates on pixels in an input image, as shown in Fig. 1. A large number of patches are extracted from input images; and corresponding pixels at the same positions in desired output images, named as teaching images, are extracted for the MTANN to learn. This patch-based training leads to the fact that the MTANN can be trained with only a small number of input and teaching images.

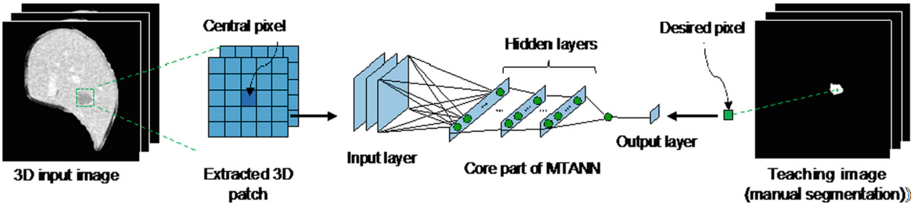


Fig. 1. Illustration of the structure of MTANN, extracting a patch from an input image and a desired pixel from a teaching image.

2.2 Sensitivity-Based Structure Optimization

The numbers of hidden layers and their units in an MTANN model are adjustable hyper-parameters. A relatively large structure is used to ensure that the model performs well on a specific task. A trained large model, however, may contain redundant units, and functions of neurons for the task would be "distributed and diluted" in many neurons in the model. This makes the analysis of the functions of neurons very difficult [15].

To address this issue, we applied our sensitivity-based structure optimization algorithm [16] to a trained large MTANN model to "consolidate" the diluted functions of neurons in the MTANN model. With this algorithm, redundant hidden units of the model are gradually removed; and a compact model with equivalent performance is obtained. The algorithm is described as the following steps:

Algorithm 1: Structure optimization for the MTANN.

```
Require: $D = \{(x_i, y_i) | 1 \le i \le N\}$: The training data
Require: $H = \{h_i | 1 \le i \le n\}$: The numbers of units in each hidden layer
  $t \leftarrow 0$ (Initialize timestamp)
  Initialize the weights in the model $NN_i$
  While $\exists h_i > 1 (h_i \in H)$ do
    $t \leftarrow t+1$
    Train $NN_t$ on $D$ until the loss value converges
    $l_t \leftarrow$ the loss value of $NN_t$ on $D$
    $m_t \leftarrow$ other necessary evaluation metrics of $NN_t$ on $D$ (like PSNR, dice coefficient, etc., which depend on the task)
      $l_{max} \leftarrow 1.0$ (Initialize the maximum loss value after removing a hidden unit from $NN_t$, and the loss value is supposed to be between 0 and 1)
      $i_{max} \leftarrow 0$ (Initialize the index of the hidden layer where the hidden unit belongs)
      $j_{max} \leftarrow 0$ (Initialize the index of the hidden layer until in the $i_{max}$-th hidden layer)
      for $i$ in {1...n} do (Go through each hidden layer)
        if $h_i = 1$ do (This layer has only one unit which cannot be removed)
          Skip to the next iteration
        for $j$ in {1...$h_i$} do (Go through each hidden unit in the $i$-th hidden layer)
          Remove the $j$-th hidden unit in the $i$-th hidden layer from $NN_t$ temporarily
          $l_0 \leftarrow$ the loss value of $NN_t$ on $D$
          if $l_0 < l_{max}$ do
            $l_{max} \leftarrow l_0$
            $i_{max} \leftarrow i$
            $j_{max} \leftarrow j$
          Put the $j$-th hidden unit in the $i$-th hidden layer back to $NN_t$
          $NN_{t+1} \leftarrow NN_t$ (Copy current model's weights and structure)
          Remove the $j$-th hidden unit in the $i$-th hidden layer from $NN_{t+1}$ permanently
return $\{(NN_i, l_i, m_i) | 1 \le i \le t\}$
```

With the proposed optimization algorithm, the hidden units of MTANN could be gradually removed until the performance drops greatly when any of the rest unit is deleted.

2.3 Calculation of Weighted Function Maps

After applying the structure optimization algorithm, every hidden unit in the compact model is expected to have an essential function for the target task. To understand the functions of the hidden units, function maps were obtained by performing the MTANN convolution of a hidden unit over a given input image. For better discrimination between enhancement and suppression, the function maps were normalized and then multiplied by the sign of the weight between the hidden units and the output unit. Weighted function maps were finally generated by shifting the range of the function map by 0.5. Namely, for a given hidden unit, in the weighted function map, a pixel value >0.5 means enhancement of patterns in the input image, whereas a pixel value <0.5 means suppression.

2.4 Unsupervised Hierarchical Clustering

To group similar functions of the hidden units of the MTANN, we applied an unsupervised hierarchical clustering algorithm [17] to the weighted functional visualization maps. With this algorithm, the hidden units were automatically divided into several groups based on the following distance function between the weighted function maps of the hidden units:

$$distance(x, y) = \alpha(1 - SSIM\,(x, y)) + NRMSE(x, y) \tag{1}$$

where SSIM is the structural similarity index, and NMRSE is the normalized root mean square error. With the unsupervised hierarchical clustering algorithm, we visualize the function maps of the hidden units group by group to explain the behavior of each group of the hidden units.

3 Experiments

3.1 Dynamic Contrast-Enhanced Liver CT

Our XAI technique was applied to explain the MTANN model's decision in a liver tumor segmentation task [20]. Dynamic contrast-enhanced liver CT scans consisting of 42 patients with 194 liver tumors in the portal venous phase from the LiTS database [21] were used in this study. Each slice of the CT volumes in the dataset has a matrix size of 512 × 512 pixels, with in-plane pixel sizes of 0.60–1.00 mm and thicknesses of 0.20–0.70 mm. The dataset consists of the original hepatic CT image with the liver mask and the "gold-standard" liver tumor region manually segmented by a radiologist, as illustrated in Fig. 2.

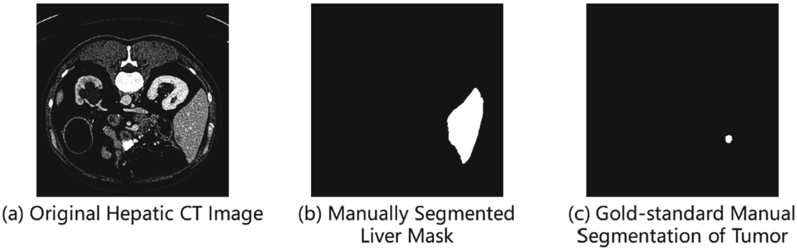

(a) Original Hepatic CT Image
(b) Manually Segmented Liver Mask
(c) Gold-standard Manual Segmentation of Tumor

Fig. 2. An example from the dynamic contrast-enhanced liver CT dataset.

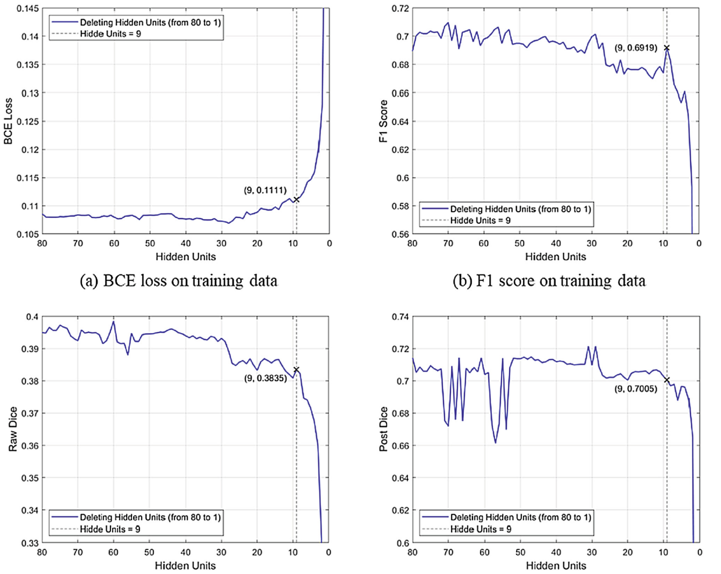
Firstly, to have the same physical scale on spatial coordinates, bicubic interpolation was applied on the original hepatic CT images together with the corresponding liver mask and "gold-standard" tumor segmentation to obtain isotropic images with a voxel size of 0.60 × 0.60 × 0.60 mm^3. Then, to unify the image size into the same size, the isotropic image was cropped to obtain the liver region volume of interest (VOI) with an in-plane matrix size of 512 × 512. An anisotropic diffusion filter was applied to reduce the quantum noise, which could substantially reduce the noise while major structures such as tumors and vessels maintained [22]. Finally, a Z-score normalization was applied to unify complex histograms of tumors in different cases. The final pre-processed CT images were used as the input images.

In addition, since most liver tumors' shape is ellipsoidal, the liver tumors can also be enhanced by the Hessian-based method and utilized in the model to improve the performance [23, 24]. Hence, the model consisted of these two input channels: segmented liver CT image and its Hessian-enhanced image. Also, the patches were extracted from input images from both channels: a 5 × 5 × 5 sized patch in the same spatial position was extracted to form a training patch with a size of 2 × 5 × 5 × 5 pixels.

Seven cases and 24 cases in the dynamic contrast-enhanced CT scans dataset were used for training and testing, respectively. 10,000 patches were randomly selected from the liver mask region in each case, summing up to a total of 70,000 training samples for training. The number of input units in the MTANN model with one hidden layer was 250. The structure optimization process started with 80 hidden units in the hidden layer. The binary cross-entropy (BCE) loss function was used to train the model. The MTANN model classified the input patches into tumor or non-tumor classes, and the output pixels represented the probability of being a tumor class. During the structure optimization process, the F1 score on the training patches and the Dice coefficient on the training images were also calculated as the reference to select a suitable compact model that performed equivalently to the original large model.

As observed in the four evaluation metric curves in Fig. 3, as the number of hidden units was reduced from 80 to 9, the performance of the model fluctuated up and down, and after it was reduced below 9, the performance of the model dropped dramatically. Therefore, we chose a number of hidden units of 9 as the optimized structure.

Then, we applied the unsupervised hierarchical clustering algorithm to the weighted function maps from the optimized compact model with 9 hidden units. Figure 4 shows that the 9 hidden units are clearly divided into 3 different groups. We denote hidden units 3, 4, and 7 as group A, hidden units 2, 6, 1, and 8 as group B, and hidden units 0 and 5 as group C. The hidden units in the same group should have a similar function, and the function maps from each group should show the function of the group.



(a) BCE loss on training data

(b) F1 score on training data

(c) Post dice coefficient (between post-processed output and teaching images) on training data

(d) Raw dice coefficient (between raw output and teaching images) on training data

Fig. 3. Performance change of an MTANN segmentation scheme (in terms of BCE loss, F1 score, raw dice, and post dice) in the structure optimization process.

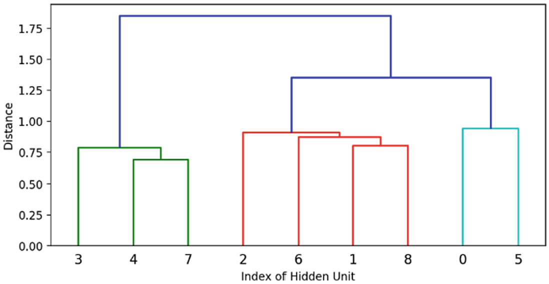
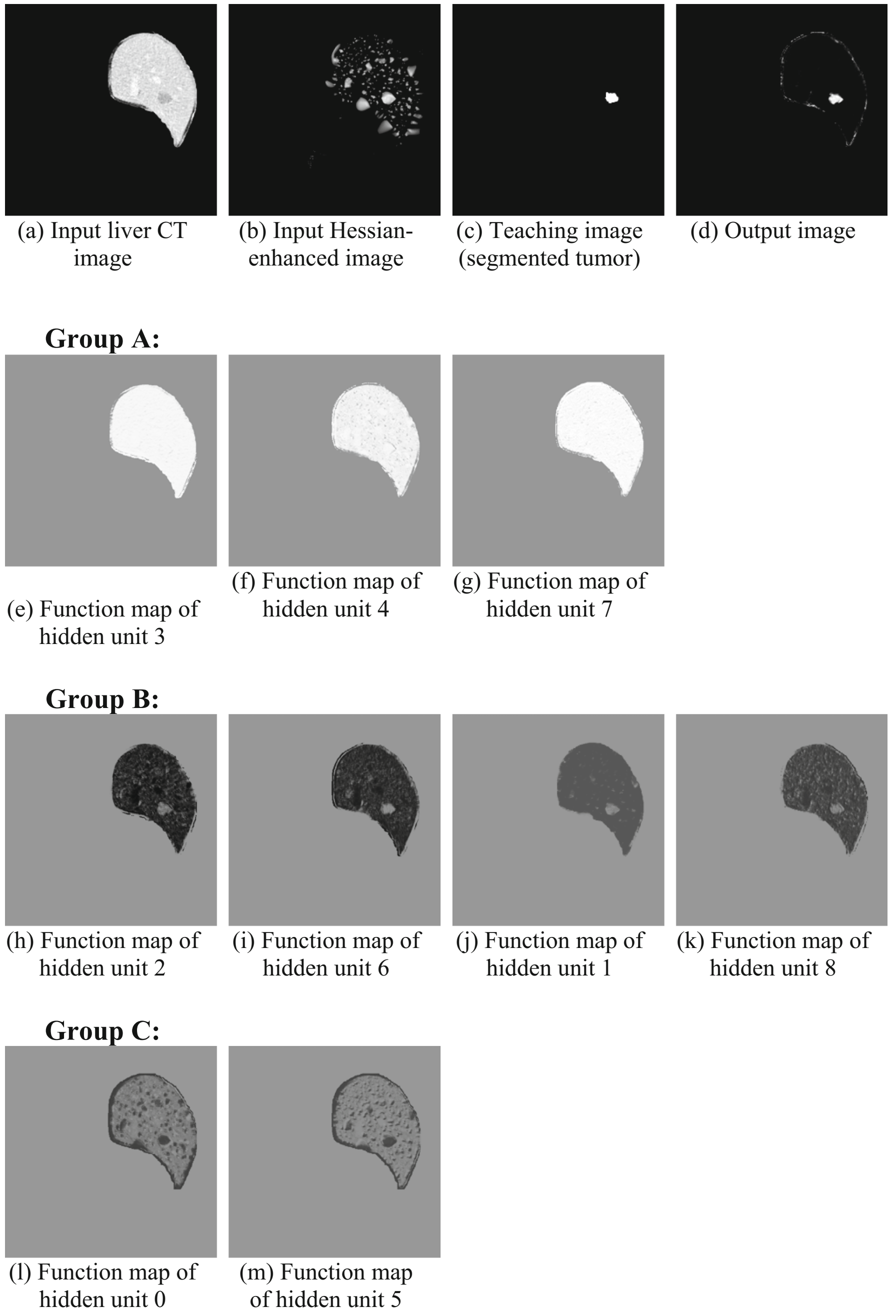


Fig. 4. Result of the unsupervised hierarchical clustering process for the function visualization maps for 9 hidden units.

As illustrated in Fig. 5, the low-intensity areas in the function maps of hidden units 0 and 5 in group C match the high-intensity areas in the Hessian-enhanced input image, which means they suppress the high-intensity areas. Likewise, group A enhances the liver area, and group B suppresses the non-tumor area. We also understood that groups A and B worked together to enhance the tumor area, and group C suppressed the liver's boundary as well as reduced the false enhancements inside the liver. Thus, our XAI method was able to reveal the learned functions of groups of neurons in the neural network, which we call "functional explanations" and define as the explanations of the



(a) Input liver CT image

(b) Input Hessian-enhanced image

(c) Teaching image (segmented tumor)

(d) Output image

(e) Function map of hidden unit 3

(f) Function map of hidden unit 4

(g) Function map of hidden unit 7

(h) Function map of hidden unit 2

(i) Function map of hidden unit 6

(j) Function map of hidden unit 1

(k) Function map of hidden unit 8

(l) Function map of hidden unit 0

(m) Function map of hidden unit 5

Fig. 5. Functional visualization maps for the 9 hidden units in groups A, B, and C obtained by using our XAI method, and the comparison with the input, teaching, and output images.

model behavior by a combination of functions. Our method is a post-hoc method that offers both instance-based and model-based functional explanations.

4 Conclusion

In this study, we proposed a novel XAI approach to explain the functions and behavior of an MTANN model for semantic segmentation of liver tumors in CT. Our structure optimization algorithm refined the structure and made every hidden unit in the model have a clear, meaningful function by removing redundant hidden units and "condensing" the functions into fewer hidden units, which solved the issue of unstable XAI results with conventional XAI methods. The unsupervised hierarchical clustering algorithm in our XAI approach grouped the hidden units with a similar function into one group so as to explain their functions by group. Through the experiments, we successfully proved that the MTANN model was explainable by functions.

Acknowledgment. This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

1. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
2. Castelvecchi, D.: Can we open the black box of AI? Nat. News **538**(7623), 20 (2016)
3. Gunning, D., Aha, D.: DARPA's explainable artificial intelligence (XAI) program. AI Mag. **40**(2), 44–58 (2019)
4. Adabi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
5. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
6. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**, 1–46 (2015)
7. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings International Conference Machine Learning, pp. 3145–3153 (2017)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
9. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
10. Suzuki, K., Horiba, I., Sugie, N.: Neural edge enhancer for supervised edge enhancement from noisy images. IEEE Trans. Pattern Anal. Mach. Intell. **25**(12), 1582–1596 (2003)
11. Suzuki, K., Horiba, I., Sugie, N., et al.: Neural filter with selection of input features and its application to image quality improvement of medical image sequences. IEICE Trans. Inf. Syst. **85**(10), 1710–1718 (2002)
12. Suzuki, K., et al.: Extraction of left ventricular contours from left ventriculograms by means of a neural edge detector. IEEE Trans. Med. Imaging **23**(3), 330–339 (2004)

13. Suzuki, K., Li, F., Sone, S., Doi, K.: Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network. IEEE Trans. Med. Imaging **24**(9), 1138–1150 (2009)
14. Suzuki, K., Rockey, D.C., Dachman, A.H.: CT colonography: advanced computer-aided detection scheme utilizing MTANNs for detection of 'missed' polyps in a multicenter clinical trial. Med. Phys **37**(1), 12–21 (2010)
15. Weigend, A.: On overfitting and the effective number of hidden units. In: Proceedings of the 1993 Connectionist Models Summer School, vol. 1 (1994)
16. Suzuki, K., Horiba, I., Sugie, N.: A simple neural network pruning algorithm with application to filter synthesis. Neural Process. Lett **13**(1), 43–53 (2001). https://doi.org/10.1023/A:1009639214138
17. Bar-Joseph, Z., Gifford, D.K., Jaakkola, T.S.: Fast optimal leaf ordering for hierarchical clustering. Bioinformatics **17**(1), 22–29 (2001)
18. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. **36**, 105–139 (1999). https://doi.org/10.1023/A:1007515423169
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. **13**(4), 600–612 (2004)
20. Sato, M., Jin, Z., Suzuki, K.: Semantic segmentation of liver tumor in contrast-enhanced hepatic CT by using deep learning with hessian-based enhancer with small training dataset size. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 34–37 (2021)
21. Simpson, A.L., Antonelli, M., Bakas, S., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. ArXiv Prepr. ArXiv190209063 (2019)
22. Huynh, H.T., Le-Trong, N., Bao, P.T., Oto, A., Suzuki, K.: Fully automated MR liver volumetry using watershed segmentation coupled with active contouring. Int. J. Comput. Assist. Radiol. Surg. **12**(2), 235–243 (2017). https://doi.org/10.1007/s11548-016-1498-9
23. Sato, Y., et al.: Tissue classification based on 3D local intensity structures for volume rendering. IEEE Trans. Vis. Comput. Graph. **6**(2), 160–180 (2000)
24. Jin, Z., Arimura, H., Kakeda, S., Yamashita, F., Sasaki, M., Korogi, Y.: An ellipsoid convex enhancement filter for detection of asymptomatic intracranial aneurysm candidates in CAD frameworks. Med. Phys. **43**(2), 951–960 (2016)