*Article*

# Ensemble Learning of Multiple Models Using Deep Learning for Multiclass Classification of Ultrasound Images of Hepatic Masses

**Norio Nakata[1,*], and Tsuyoshi Siina [2]**

[1] Division of Artificial Intelligence in Medicine; Department of Radiology, The Jikei University, School of Medicine, 3-25-8, Nishi-Shinbashi, Minato-ku, Tokyo 1058461, Japan; nakata@jikei.ac.jp
[2] Shibaura Institute of Technology, Graduate School of Science and Engineering, 3-7-5 Toyosu Koto-ku, Tokyo 135-8548, Japan; shiinat@shibaura-it.ac.jp
[*] Correspondence: nakata@jikei.ac.jp; +81-3-3433-1111.

**Abstract:** Ultrasound (US) is commonly used for the diagnosis of liver masses. Ensemble learning has been widely used for image classification, but its methods have not been fully optimized. This study was performed to investigate the usefulness of ensemble learning and compare a number of ensemble learning techniques using multiple convolutional neural network (CNN)-trained models for image classification of liver masses in US images. The US imaging data set was classified into four categories: benign liver tumor (BLT, 6320 images), liver cyst (LCY, 2320 images), metastatic liver cancer (MLC, 9720 images), and primary liver cancer (PLC, 7840 images). In this study, 250 test images were randomly selected for each class, for a total of 1000 images, and the remaining images were used for training. Sixteen different CNNs were used for to train and test the US images. All four types of ensemble learning—soft voting (SV), stacking (ST), weighted average voting (WAV), and weighted hard voting (WHV)—showed greater accuracy than the single CNN. All four types also showed significantly better deep learning performance than ResNeXt101 alone. For image classification of liver masses using US images, ensemble learning improved the performance of deep learning over a single CNN.

**Keywords:** ensemble learning; deep learning; convolutional neural network; liver; ultrasonography; artificial intelligence

## 1. Introduction

Hepatocellular carcinoma (HCC) is the most common primary liver cancer, the sixth most common cancer, and the second leading cause of cancer death, resulting in almost 800 000 deaths worldwide annually [1–3]. Survival of patients with HCC is mainly influenced by the disease stage at the time of diagnosis [4]. Ultrasound (US) is a simple, noninvasive, safe, portable, relatively inexpensive, and easily accessible imaging modality with no risk of radiation exposure, making it useful as a diagnostic and monitoring tool in medicine [5,6]. Conventional US is the first choice for surveillance of HCC [4,7,8]. The reported sensitivity of US alone for HCC diagnosis ranges from 60% to 90%, with excellent specificity of greater than 90% [9, 10]. Ultrasonography is often used to screen the liver, especially for HCC, due to its low cost, accessibility, and lack of X-ray exposure [11].

There have been a number of reports of the application of artificial intelligence (AI) in US, including examination of the thyroid, breast, abdomen, and pelvic area, and in obstetrics and gynecology, such as monitoring of the fetus, heart, and vascular system. AI-based image classification of liver masses has been applied to computed tomography (CT) [12,13]. In addition, AI-based image classification of colorectal cancer liver metastases has been studied in magnetic resonance imaging (MRI) [14], and in the differential diagnosis of masses in US [15, 16].

Ensemble learning, which uses multiple learning models, has recently been applied for image classification using deep learning. Ensemble learning can improve the performance of AI by due to the use of multiple models in conjunction with each other. Conventional ensemble learning related to diagnostic imaging includes studies using voting and stacking methods [17–28]. Ensemble learning has been reported involving the selection of the top three of nine trained convolutional neural network (CNN) models [17, 18], and one study reported ensemble learning from all of 15 trained CNNs [19]. However, the bases for determining which ensemble learning methods should be used and in what combinations remain unclear.

This study was performed to compare and investigate the usefulness of some ensemble learning techniques using multiple CNN-trained models for multiclass image classification of liver masses in US images.

## 2. Materials and Methods

### *2.1. Data Set*

The study was performed using 26 200 US B-mode, grayscale still images of liver masses collected as part of a research project funded by the Japan Agency for Medical Research and Development (AMED) from 2018 to 2021, entitled "National Database Construction of Digital Ultrasound Images and Artificial Intelligence-Assisted Ultrasound Diagnostic System Development." The data for this project were collected by US experts from several institutions selected by the Japanese Society of Ultrasound Medicine. The images used in this project were classified into four categories: benign liver tumor (BLT, 6320 images), liver cyst (LCY, 2320 images), metastatic liver cancer (MLC, 9720 images), and primary liver cancer (PLC, 7840 images). In this study, 250 test images were randomly selected for each class, for a total of 1000 images, and the remaining images (6970 BLT images, 2310 LCY images, 9470 MLC images, and 7590 PLC images) were used in training. The study was approved by the Ethics Committee of Jikei University Hospital.

### *2.2 Hardware and Software*

The hardware consisted of a desktop computer built for AI with an Intel core i9 CPU, 128 GB of random access memory (RAM), and NVIDIA RTX A6000 graphics processing unit (GPU), running Ubuntu Linux version 20.04 (https://jp.ubuntu.com/). The programming languages used was Python version 3.8 (https://www.python.org/). TensorFlow version 2.8 (https://www.tensorflow.org/), Keras version 2.8 (https://keras.io/), a machine learning (ML) library, and scikit-learn 0.24 (https://scikit-learn.org) were used to create the programs.

### *2.3 Training and Testing with 16 CNNs*

Sixteen different CNN models were used for training and testing US images (Table 1). EfficientNetB0–B6 [29] were pretrained by Noisy Student [30], and the other nine models were fine-tuned by replacing only all coupling layers or by replacing part of the convolutional layer with all coupling layers based on the models pretrained by ImageNet. Sixteen types of training were performed using the models and the abovementioned training images. The resolution of all input images was set to 256 × 256 for comparison them and to ensure reproducibility of training results. The batch sizes were all set to 32, the number of epochs to 50, and the random number seed was fixed to an arbitrary seed value. Width shift, height shift, and horizontal flip were used as data augmentation. The floating-point number for width shift and height shift was set to 0.1. We used k-partition cross-validation (k = 10) to validate the predictive performance of the ML models. The model fitting callback was the early stopping function of Keras, which stops training before overtraining occurs, and training was stopped if there was no improvement during 10 epochs of val_loss values. We also reduced the learning rate by 0.1 if there was no improvement for three epochs using ReduceLROnPlateau in Keras. For each of the 16 CNN training models created in training, we performed a test on a test image using the same model. Precision, sensitivity (recall), specificity, and f1 score were calculated for each class as indices for comparing the accuracy of the test results for each model and as evaluation indices for the two types of ensemble learning described below, and the macro average of each of these indicators was calculated. The area under the curve (AUC) of the receiver operating characteristic (ROC) curves for each of the four classes and the macro average were also calculated for each model. The accuracy of the results was calculated and ranked from the one with the highest value.

### *2.4 Ensemble Learning*

Ensemble learning was performed using the soft voting (SV), weighted average voting (WAV), weighted hard voting (WHV), and stacking (ST) methods. In general, stacking methods are ensemble ML algorithms that can learn how to optimally combine predictions from several base models using meta-learning algorithms (Figure 1). The ST method used in this study combines a CNN as the base model with LightGBM version 3.2.1 (https://lightgbm.readthedocs.io/), a type of gradient boosting method, as the metamodel. For each of these four types of ensemble learning, we calculated the 2–16 best methods one by one based on the order of increasing accuracy of each algorithm. For LightGBM, k = 5 was set using k-partition cross-validation. The number of gradient boosting iterations (num_boost_round) was set to 1000, and the number of times that training was terminated (early_stopping) was set to 50 if the score did not improve a certain number of consecutive times in the evaluation data, even if the specified training count was not reached. The multiclass log loss (multi_logloss) was used as the metric for the LightGBM evaluation function. As the hyperparameters need to be optimized when performing the LightGBM calculation, we used Optuna version 2.10.0 (https://www.preferred.jp/ja/projects/optuna/), an open source software framework for automating the optimization of hyperparameters, for all three of the abovementioned algorithms. We also fixed the random number seeds in the program to ensure reproducibility when using Optuna with LightGBM. Finally, of the 15 SV, WA, WV, and ST methods, we selected the one with the highest accuracy for each of the classes. The precision, sensitivity (recall), specificity, and f1 score were calculated for each class as well as the macro average for each of these indices. The ROC curves for the four classes and the macro average, as well as the AUC of the ROC for the macro average, were calculated for each model.

### *2.5 Statistical Analysis*

The test results of 16 single learning models and ensemble learning were examined for statistical significance. Among the 16 independent learning models, we selected the top three models with the highest accuracy from 1 to 3. For ensemble learning, all 15 SV methods and all 15 ST methods were used. For these two methods, i.e., single learning and ensemble learning, we determined the correctness of the known 4-class classification of test results and the correctness from the predicted result class for each method, and also determined the true positives and negatives between the two methods. The total numbers of false positives, false negatives, and true negatives were tabulated. Finally, the McNemar test [31] was used to determine the significance of the total number of 90 total results for single learning and ensemble learning. In addition, to examine the statistical significance of the differences between SV and ST, a total of 15 McNemar tests were conducted to test for significant differences between the ensemble learning models using the same individual learning models (top 2–16). Significant difference tests were performed.

### *2.6 Heat map image*

Heat map images for each class were created for the model with the highest accuracy among the 16 individual training models. Heat map images allow visualization of the part of the image examined by the ML algorithm by focusing on the features extracted by the last convolutional layer of the CNN. Gradient-weighted class activation mapping (Grad-CAM) (http://gradcam.cloudcv.org/) code was used to create heat map images for each class.

doi:10.20944/preprints202211.0574.v1

## 3. Results

### 3.1. Evaluation of Test Results for 16 CNN models

The evaluation metrics for each of the 16 CNNs are shown in Tables 2 and 3. Ranked in order of increasing accuracy, the top 1 to 3 (values of accuracy in parentheses) were ResNext101 (0.719), Xception (0.715), and InceptionResNetV2 (0. 7). With the exception of EfficientNetB0 (0.694) ranked 6th, the CNNs ranked 4–9 were ResNet-based CNNs, and those ranked 11–16, of which EfficientNetB4 (0.617) had the lowest accuracy, were EfficientNet-based CNNs. The ranking based on the highest f1 score from 1 to 3 was consistent with the accuracy ranking. All 16 CNNs were ranked high in specificity, precision, and sensitivity. There was no significant difference between the test results of the top-ranking CNN model, ResNeXt101, and those of the models ranked 2nd to 9th, while a clear significant difference was seen compared to the 10th to 16th ranked models ($P < 0.05$) (Table 3).

### 3.2. Evaluation of Test Results for Ensemble Learning

A comparison of the accuracy values for ensemble learning is shown in Figure 2. Among SV methods, SV16 had the highest accuracy of 0.776. Among ST methods, ST9 had the highest accuracy of 0.776. Among WHV methods, WHV13 had the highest accuracy of 0.779. Among WAV methods, WAV7 had the highest accuracy of 0.783, which was the highest accuracy value for all methods (Table 4). In addition, statistical analysis of all test results among the four ensemble learning methods for the same number of models used in the 15 methods from the 2nd to 16th ranking methods showed that SV2 had significantly better test results than WHV2 ($P < 0.01$). There were no significant differences in test results for the ensemble learning method among the other SVs, STs, WAVs, and WHVs.

The respective metrics for SV16, ST9, WAV7, and WHV13 are shown in Table 5. In addition to accuracy (0.783), WAV7 had the highest precision (0.790), sensitivity (0.783), specificity (0.928), f1 score (0.786), and macro-AUC (0.935) of the ROC curve were the highest values among all methods used in this study.

The significance of differences between the 2nd to 16th ranking test results of these four ensemble training methods and the test results for ResNeXt101 with the top-ranking accuracy is shown in Table 5. For SV and ST, the test results were significantly higher ($P < 0.05$) than those of ResNeXt101 for those ranking 4th to 16th. For WAV, the 3rd to 16th ranking showed significantly higher test results with ResNeXt101 ($P < 0.05$), and for WHV, the 5th to 16th ranking showed significantly better test results with ResNeXt101 ($P < 0.05$).

The average computation times required to output the results from the test data of each of the 15 ensemble learning methods from the 2nd to 16th ranking were 4.26 s for SV, 18 min 7.38 s for S, 13 min 22.23 s for WAV, and 40 min 55.73 s for WHV.

In addition, we compared the test results between the same ensemble learning methods with the highest accuracy (SV16, ST9, WAV7, and WHV13) and the other methods of the same ensemble learning method (Table 6). There were no significant differences between SV16 and SV2–16 or ST4–16, while ST9 and ST2 and 3 were significantly different ($P < 0.05$). WAV7 was significantly different from WAV2–4 ($P < 0.05$), but not from WAV5–16. The results for WHV13 were significantly different from those for WHV2–4 ($P < 0.05$) but not significantly different from those for WHV5–16 ($P < 0.05$).

### 3.3. Details of Image Classification of Ultrasound Images of Liver Masses

Table 7 shows the image classification results of four different liver masses for ResNeXt101, the model with the highest accuracy among the 16 CNNs examined, and WAV7, the model with the highest accuracy among ensemble learning models examined in this study. All of the indices (precision, sensitivity, specificity, f1 score, AUC) for both ResNeXt101 and WAV7 were highest for LCY, followed by BLT and PLC, and lowest for MLC. In comparison of the values of precision, sensitivity, and specificity, only the liver cyst of WAV7 had 100% precision and specificity, while all the others, both ResNeXt101 and WAV7, had 100% precision and specificity, but all others, both ResNeXt101 and WAV7, had high specificity, precision, and sensitivity, in that order. Four classes of heat maps were created using ResNeXt101, the model with the highest accuracy among the 16 CNNs, and representative examples are shown in Figure 3.

### *3.4. Figures, Tables, and Schemes*

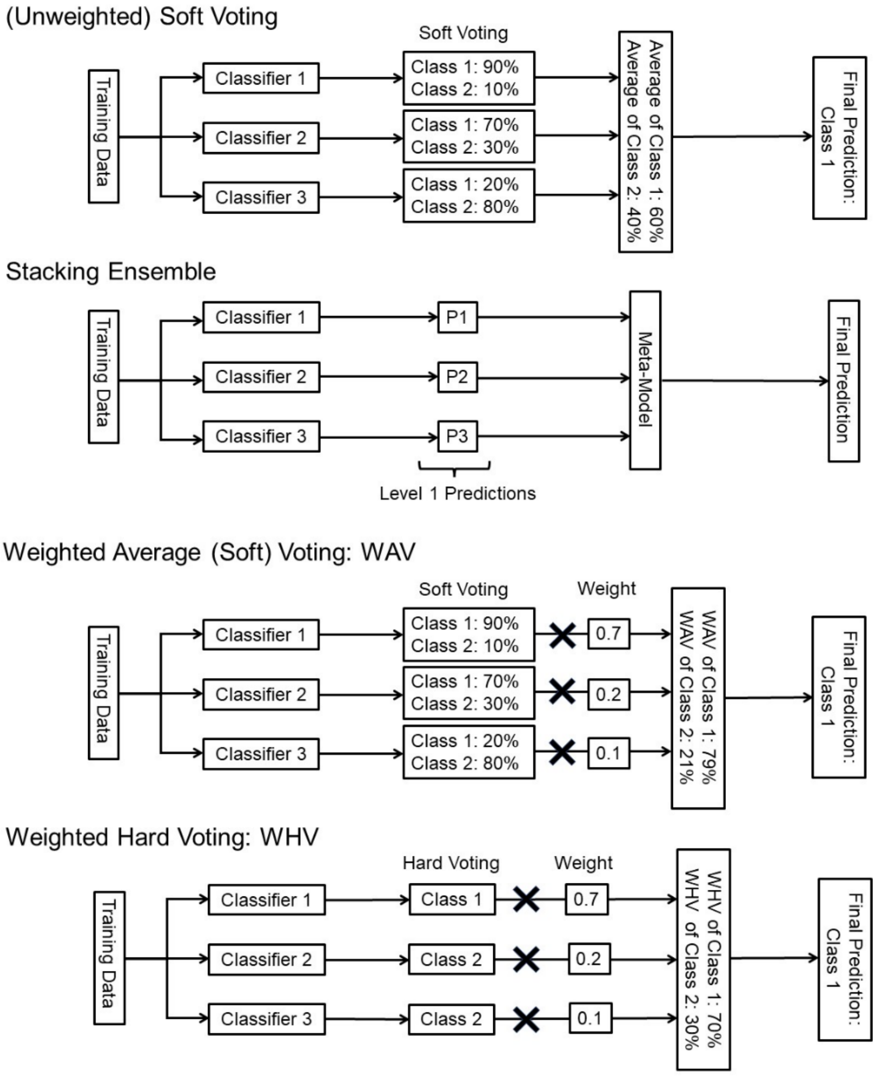

**Figure 1.** Examples of Four Ensemble Learning Methods.

This figure shows examples of three different classifiers and two class classifications. In this study, 16 different classifiers and 4 class classifications were used.

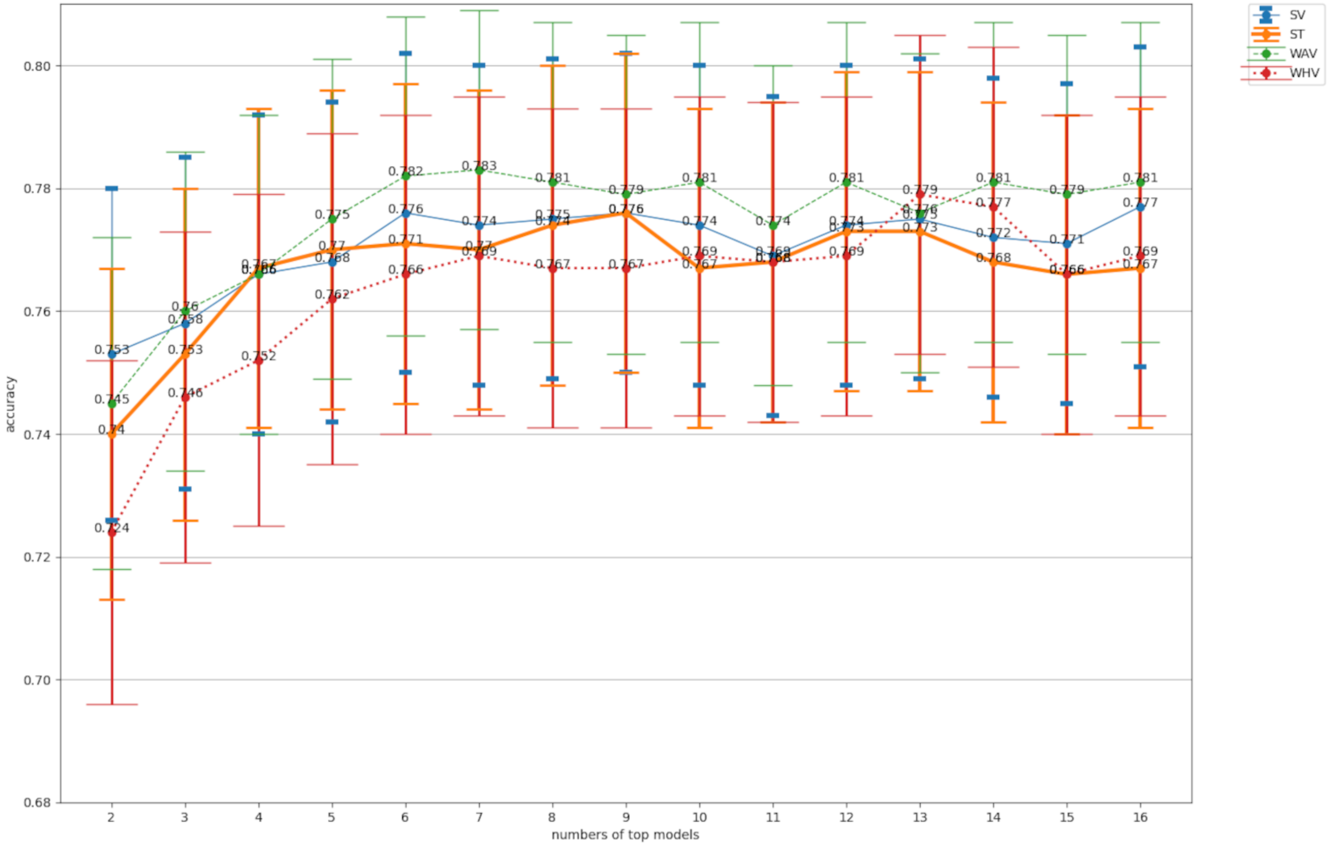

**Figure 2.** Comparison of Accuracy for Ensemble Learning.

Numbers of top models representing the number of classifiers used for ensemble training from those ranked 2–16 in accuracy.

SV, soft voting; ST, stacking; WAV, weighted average voting; WHV, weighted hard voting.

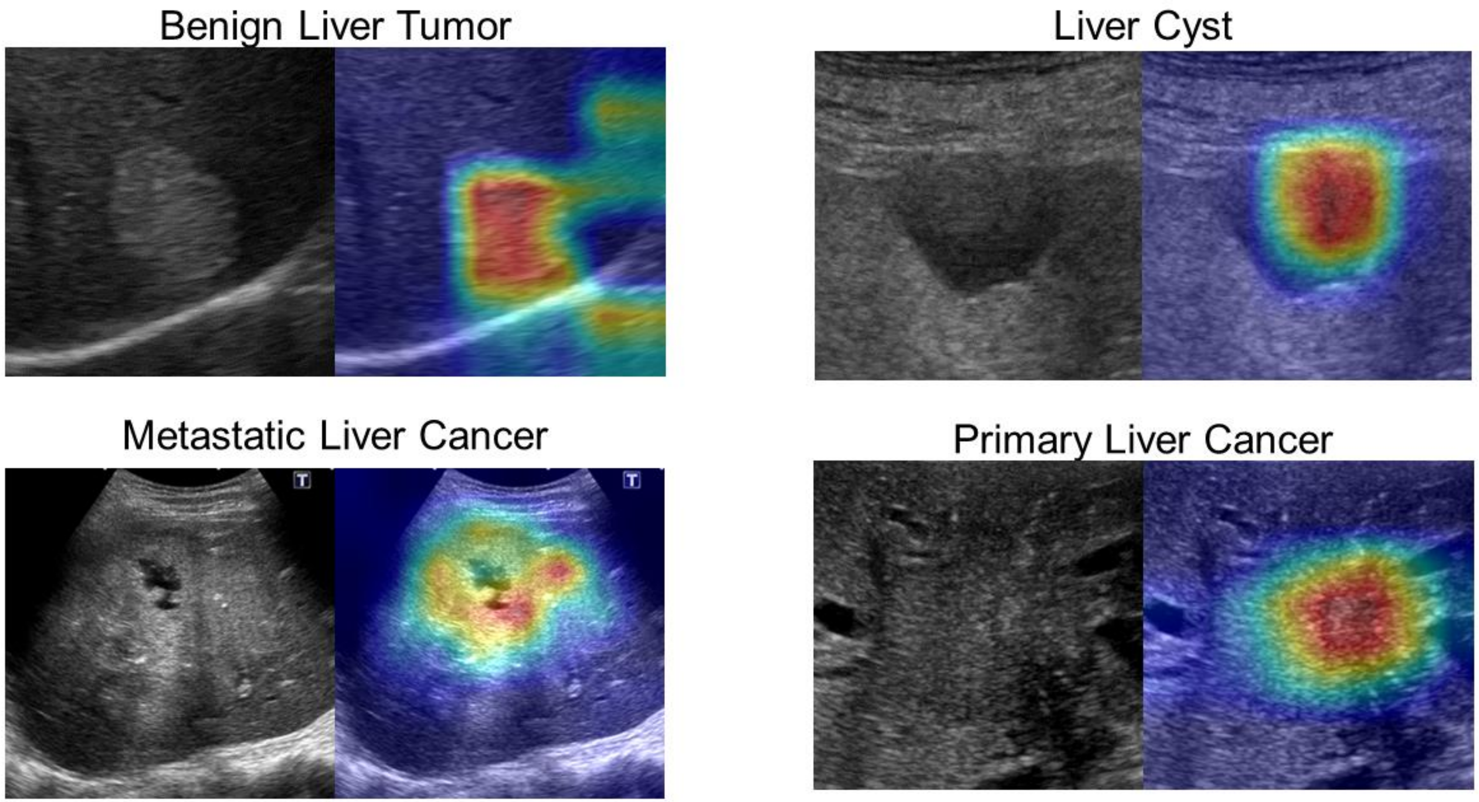

**Figure 3.** Ultrasound and Heat Map Images of the Four Classes.

A 256 × 256 ultrasound B-mode image (grayscale image on the left of each class) and a corresponding heat map image (color mapped image on the right side of each class) for each of the four classes (benign liver tumor, liver cyst, metastatic liver cancer, primary liver cancer).

doi:10.20944/preprints202211.0574.v1

**Table 1.** Sixteen Different CNN models

| Model | Total parameters | Trainable parameters | FLOPs |
|---|---|---|---|
| Xception | 22 963 756 | 9 970 324 | 11.9 G |
| InceptionV3 | 23 905 060 | 13 227 844 | 7.73 G |
| InceptionResNetV2 | 55 914 724 | 5 737 060 | 17.7 G |
| ResNet50 | 25 689 988 | 25 636 868 | 10.1 G |
| ResNet101 | 44 760 452 | 44 655 108 | 19.9 G |
| ResNeXt50 | 25 150 413 | 25 082 183 | 11.1 G |
| ResNeXt101 | 44 368 845 | 4 353 028 | 20.9 G |
| SeResNeXt50 | 27 681 396 | 27 613 172 | 11.1 G |
| SeResNeXt101 | 49 146 548 | 49 008 692 | 20.9 G |
| EfficientNetB0 | 5 365 408 | 2 445 236 | 1.05 G |
| EfficientNetB1 | 7 891 076 | 2 677 204 | 1.54 G |
| EfficientNetB2 | 9 215 478 | 3 090 844 | 1.78 G |
| EfficientNetB3 | 12 361 516 | 3 531 108 | 2.59 G |
| EfficientNetB4 | 19 513 948 | 4 491 508 | 4.03 G |
| EfficientNetB5 | 30 615 796 | 5 558 404 | 6.3 G |
| EfficientNetB6 | 43 324 556 | 6 731 796 | 8.96 G |

EfficientNetB0–B6 are models pretrained by Noisy Student, and the other nine models were fine-tuned models based on models pretrained by ImageNet, replacing only all coupling layers, or replacing a part of the convolution layer and all coupling layers

**Table 2.** Evaluation Metrics for Each of the 16 CNNs

| Model | Precision*, ** | Sensitivity*, ** | Specificity*, ** | Accuracy** | Ranking of accuracy |
|---|---|---|---|---|---|
| ResNeXt101 | 0.731 (0.778, 0.686) | 0.719 (0.772, 0.666) | 0.906 (0.925, 0.888) | 0.719 (0.747, 0.691) | 1 |
| Xception | 0.728 (0.777, 0.679) | 0.715 (0.769, 0.661) | 0.905 (0.924, 0.886) | 0.715 (0.743, 0.687) | 2 |
| InceptionResNetV2 | 0.720 (0.766, 0.675) | 0.700 (0.756, 0.644) | 0.900 (0.919, 0.881) | 0.700 (0.728, 0.672) | 3 |
| SeResNeXt50 | 0.709 (0.758, 0.660) | 0.699 (0.752, 0.646) | 0.900 (0.919, 0.880) | 0.699 (0.727, 0.671) | 4 |
| ResNeXt50 | 0.710 (0.758, 0.663) | 0.695 (0.750, 0.640) | 0.898 (0.917, 0.879) | 0.695 (0.724, 0.666) | 5 |
| EfficientNetB0 | 0.715 (0.762, 0.669) | 0.694 (0.750, 0.638) | 0.898 (0.917, 0.879) | 0.694 (0.723, 0.665) | 6 |
| SeResNeXt101 | 0.698 (0.746, 0.650) | 0.690 (0.744, 0.636) | 0.897 (0.916, 0.877) | 0.690 (0.719, 0.661) | 7 |
| ResNet101 | 0.698 (0.750, 0.647) | 0.686 (0.739, 0.633) | 0.895 (0.915, 0.875) | 0.686 (0.715, 0.657) | 8 |
| ResNet50 | 0.697 (0.748, 0.645) | 0.684 (0.737, 0.631) | 0.895 (0.914, 0.875) | 0.684 (0.713, 0.655) | 9 |
| InceptionV3 | 0.705 (0.754, 0.656) | 0.676 (0.733, 0.619) | 0.892 (0.912, 0.872) | 0.676 (0.705, 0.647) | 10 |
| EfficientNetB2 | 0.674 (0.725, 0.621) | 0.664 (0.719, 0.607) | 0.888 (0.909, 0.867) | 0.664 (0.692, 0.634) | 11 |
| EfficientNetB5 | 0.680 (0.730, 0.630) | 0.662 (0.718, 0.606) | 0.887 (0.908, 0.867) | 0.662 (0.691, 0.633) | 12 |
| EfficientNetB3 | 0.681 (0.733, 0.630) | 0.661 (0.718, 0.604) | 0.886 (0.908, 0.866) | 0.661 (0.690, 0.632) | 13 |
| EfficientNetB6 | 0.673 (0.726, 0.620) | 0.660 (0.716, 0.604) | 0.887 (0.908, 0.866) | 0.660 (0.689, 0.631) | 14 |
| EfficientNetB1 | 0.647 (0.700, 0.595) | 0.628 (0.687, 0.569) | 0.876 (0.898, 0.854) | 0.628 (0.658, 0.598) | 15 |
| EfficientNetB4 | 0.638 (0.689, 0.588) | 0.617 (0.675, 0.559) | 0.872 (0.894, 0.851) | 0.617 (0.647, 0.587) | 16 |

*Macro average.

**95% confidence interval.

**Table 3.** Evaluation Metrics for Each of the 16 CNNs (Table 2 continued) and Significant Differences Between ResNeXt101 (Top Ranking) and the Other 15 Models

| Model | f1 score* | ROC Macro AUC | Ranking of accuracy | *P*-value ** |
|---|---|---|---|---|
| ResNeXt101 | 0.724 | 0.902 | 1 | – |
| Xception | 0.720 | 0.900 | 2 | 0.882 |
| InceptionResNetV2 | 0.707 | 0.880 | 3 | 0.377 |
| SeResNeXt50 | 0.699 | 0.874 | 4 | 0.341 |
| ResNeXt50 | 0.701 | 0.903 | 5 | 0.238 |
| EfficientNetB0 | 0.701 | 0.866 | 6 | 0.224 |
| SeResNeXt101 | 0.692 | 0.870 | 7 | 0.148 |
| ResNet101 | 0.685 | 0.887 | 8 | 0.107 |
| ResNet50 | 0.682 | 0.888 | 9 | 0.081 |
| InceptionV3 | 0.684 | 0.859 | 10 | 0.037 |
| EfficientNetB2 | 0.666 | 0.868 | 11 | 0.006 |
| EfficientNetB5 | 0.668 | 0.890 | 12 | 0.005 |
| EfficientNetB3 | 0.667 | 0.867 | 13 | 0.006 |
| EfficientNetB6 | 0.664 | 0.877 | 14 | 0.004 |
| EfficientNetB1 | 0.635 | 0.852 | 15 | < 0.001 |
| EfficientNetB4 | 0.624 | 0.856 | 16 | < 0.001 |

*Macro average.

***P*-value of the McNemar's test for comparison of ResNeXt101 with the other 15 models.

doi:10.20944/preprints202211.0574.v1

**Table 4.** Top Evaluation Metrics for Each of the Four Types of Ensemble Learning

| Model | Precision*,** | Sensitivity*,** | Specificity*,** | f1 score* | ROC Macro AUC | Accuracy** |
|---|---|---|---|---|---|---|
| SV16 | 0.783 (0.824, 0.733) | 0.777(0.822, 0.724) | 0.926 (0.941, 0.906) | 0.779 | 0.94 | 0.777 (0.802, 0.750) |
| ST9 | 0.783 (0.822, 0.734) | 0.776 (0.821, 0.723) | 0.925 (0.940, 0.906) | 0.778 | 0.924 | 0.776 (0.801, 0.749) |
| WAV7 | 0.790 (0.831, 0.749) | 0.783 (0.832, 0.734) | 0.928 (0.943, 0.912) | 0.786 | 0.935 | 0.783 (0.809, 0.757) |
| WHV13 | 0.784 (0.829, 0.740) | 0.779 (0.827, 0.731) | 0.926 (0.943, 0.910) | 0.781 | – | 0.779 (0.805, 0.753) |

*Macro average.

**95% confidence interval.

SV, soft voting; ST, stacking; WAV, weighted average voting; WHV, weighted hard voting.

**Table 5.** *P*-values of McNemar Test Between the Top-Ranking Model (ResNeXt101) and Top Four Ensemble Methods (SV, ST, WAV, and WHV)

| *P*-value | SV2 | SV3 | SV4 | SV5 | SV6 | SV7 | SV8 | SV9 | SV10 | SV11 | SV12 | SV13 | SV14 | SV15 | SV16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ResNeXt101 | *0.092* | *0.054* | 0.016 | 0.013 | 0.003 | 0.005 | 0.003 | 0.003 | 0.004 | 0.0099 | 0.004 | 0.004 | 0.006 | 0.007 | 0.002 |
| *P*-value | ST2 | ST3 | ST4 | ST5 | ST6 | ST7 | ST8 | ST9 | ST10 | ST11 | ST12 | ST13 | ST14 | ST15 | ST16 |
| ResNeXt101 | *0.300* | *0.088* | 0.012 | 0.007 | 0.006 | 0.008 | 0.004 | 0.003 | 0.013 | 0.01 | 0.004 | 0.005 | 0.011 | 0.014 | 0.012 |
| *P*-value | WAV2 | WAV3 | WAV4 | WAV5 | WAV6 | WAV7 | WAV8 | WAV9 | WAV10 | WAV11 | WAV12 | WAV13 | WAV14 | WAV15 | WAV16 |
| ResNeXt101 | *0.192* | 0.037 | 0.014 | 0.003 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.004 | 0.001 | 0.003 | 0.001 | 0.001 | 0.001 |
| *P*-value | WHV2 | WHV3 | WHV4 | WHV5 | WHV6 | WHV7 | WHV8 | WHV9 | WHV10 | WHV11 | WHV12 | WHV13 | WHV14 | WHV15 | WHV16 |
| ResNeXt101 | *0.836* | *0.183* | *0.096* | 0.027 | 0.017 | 0.0095 | 0.013 | 0.013 | 0.009 | 0.0103 | 0.009 | 0.002 | 0.002 | 0.017 | 0.0095 |

SV, soft voting; ST, stacking; WAV, weighted average voting; WHV, weighted hard voting.

**Table 6.** *P*-values of McNemar Test Between the Same Ensemble Learning Model With the Highest Accuracy (SV16, ST9, WAV7, and WHV13) and Other Models of the Same Ensemble Learning Method

| *P*-value | SV2 | SV3 | SV4 | SV5 | SV6 | SV7 | SV8 | SV9 | SV10 | SV11 | SV12 | SV13 | SV14 | SV15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SV16 | *0.058* | *0.110* | *0.347* | *0.412* | *1.000* | *0.820* | *0.905* | *1.000* | *0.804* | *0.322* | *0.761* | *0.868* | *0.424* | *0.180* |
| *P*-value | ST2 | ST3 | ST4 | ST5 | ST6 | ST7 | ST8 | ST10 | ST11 | ST12 | ST13 | ST14 | ST15 | ST16 |
| ST9 | 0.003 | 0.007 | *0.200* | *0.418* | *0.500* | *0.263* | *0.774* | *0.281* | *0.341* | *0.798* | *0.788* | *0.322* | *0.253* | *0.272* |
| *P*-value | WAV2 | WAV3 | WAV4 | WAV5 | WAV6 | WAV8 | WAV9 | WAV 10 | WAV 11 | WAV 12 | WAV 13 | WAV 14 | WAV 15 | WAV 16 |
| WAV7 | < 0.001 | 0.022 | 0.036 | *0.280* | *1.000* | *0.815* | *0.672* | *0.885* | *0.336* | *0.897* | *0.477* | *0.899* | *0.749* | *1.000* |
| *P*-value | WAV2 | WAV3 | WAV4 | WAV5 | WAV6 | WAV8 | WAV9 | WAV 10 | WAV11 | WAV12 | WAV13 | WAV 14 | WAV 15 | WAV 16 |
| WAV7 | < 0.001 | 0.004 | 0.022 | *0.132* | *0.223* | *0.332* | *0.207* | *0.219* | *0.289* | *0.215* | *0.174* | *0.883* | *0.136* | *0.220* |

SV, soft voting; ST, stacking; WAV, weighted average voting; WHV, weighted hard voting.

doi:10.20944/preprints202211.0574.v1

**Table 7.** Metrics of Four Different Liver Masses for ResNeXt101, the Model with the Highest Accuracy Among 16 CNNs, and WAV7

| ResNeXt101 | Precision* | Sensitivity* | Specificity* | f1 score | ROC AUC |
|---|---|---|---|---|---|
| BLT | 0.745 (0.797, 0.685) | 0.688 (0.742, 0.628) | 0.921 (0.900, 0.939) | 0.715 | 0.915 |
| LCY | 0.991 (1.000, 0.979) | 0.904 (0.941, 0.867) | 0.997 (1.000, 0.994) | 0.946 | 0.994 |
| MLC | 0.567 (0.623, 0.510) | 0.664 (0.723, 0.605) | 0.831 (0.858, 0.804) | 0.611 | 0.843 |
| PLC | 0.625 (0.685, 0.565) | 0.620 (0.680, 0.560) | 0.876 (0.900, 0.852) | 0.622 | 0.854 |
| WAV7 | | | | | |
| BLT | 0.775 (0.826, 0.723) | 0.784 (0.835, 0.733) | 0.924 (0.943, 0.905) | 0.779 | 0.944 |
| LCY | 1.000 (1.000, 1.000) | 0.920 (0.954, 0.886) | 1.000 (1.000, 1.000) | 0.958 | 0.999 |
| MLC | 0.664 (0.720, 0.609) | 0.736 (0.791, 0.681) | 0.876 (0.900, 0.852) | 0.698 | 0.891 |
| PLC | 0.721 (0.778, 0.664) | 0.692 (0.749, 0.635) | 0.911 (0.931, 0.890) | 0.706 | 0.903 |

*95% confidence interval.

WAV, weighted average voting.

## 4. Discussion

In this study, we used accuracy as a metric to compare multiple models and rank the test results of 16 different CNN models [32]. In comparison of the accuracy of multiple training models, the random seed must be fixed to ensure reproducibility of the test results. Therefore, we fixed the random seed at multiple locations in the program as appropriate. However, as this study focused on comparison of multiple models and the usefulness of ensemble learning, fixing the random seed did not necessarily mean that the test results of the training model would be optimal. It should be considered that fixing the random seed tends to make each of the listed evaluation metrics somewhat lower.

The type of model chosen to achieve high accuracy may also depend on the type and resolution of the target medical images. Therefore, a simple comparison with other reports regarding the usefulness of ensemble learning and subjects that differ from each other is limited. In this study, 16 CNN models were selected from Inception, ResNet, and EfficientNet systems. Comparison of the model with the top accuracy (RexNeXt101) with the other 15 types showed no significant difference in the confusion matrix evaluation with EfficientNetB0, which is ranked 6th, and the 2nd–9th ranking models, which included all other ResNet systems. The top10 to 16, which included models of the efficientnet system up to EfficientNetB2-6, showed significant differences from the RexNeXt101 confusion matrix evaluation. This result was contrary to the prediction that the EfficientNet system would be expected to perform better, even though Noisy Student rather than ImageNet was used as the learning model. Further studies are required to examine the usefulness of EfficientNet by increasing or decreasing the number of images.

In the comparison of SV, ST, WAV, and WHV, WAV showed the best accuracy, similar to the results of previous studies. Although the gradient boosting method was used as a meta-model, ST did not perform as well as SV and WAV, as in a previous study on ensemble learning using medical images [14]. To systematize the number of models used in ensemble learning, we also calculated 15 different ensemble learning models from the 2nd to 16th ranking models, and compared them to the model with the highest accuracy for four different ensemble learning models. While there was no significant difference in the SV model, the WAV and WHV models clearly showed lower performance from rank 2 to 4, and the S model clearly showed lower performance in ranks 2 and 3. These observations suggested that care is required when performing ensemble learning with a small number of participants. Therefore, although there was no clear difference in the performance of 9 of 16 CNNs, those ranked 4–16 in SV and ST, 3–16 in WAV, and 5–16 in WHV were clearly better than ResNeXt101, a single CNN with top1 accuracy in the confusion matrix evaluation. The results demonstrated the usefulness of ensemble learning, and that the accuracy of WAV tended to be higher than that of ResNeXt101, which was the model with the highest accuracy. However, ensemble learning has the disadvantage that it is a complex and time-consuming procedure, and its potential for commercialization is likely to be limited. In this study, the average computation time with SV was 4.26 s, while WAV took 13 min. Advances in hardware have made faster computation speeds possible, thus allowing shorter processing times. The results of this study also showed that ensemble learning clearly improves accuracy compared to single learning models, and it is expected that ensemble learning will be widely adopted in future with continuing trends toward improved computation speed.

## 5. Conclusions

In image classification of liver masses using US B-mode images, ensemble learning was shown to improve accuracy and deep learning performance over a single CNN.

**Author Contributions:** Conceptualization, N.N. and T.S.; methodology, N.N.; software, N.N.; validation, N.N.; formal analysis, N.N.; investigation, N.N.; resources, N.N.; data curation, N.N., T.S.; writing—original draft preparation, N.N.; writing—review and editing, N.N., T.S.;

visualization, N.N.; supervision, N.N.; project administration, N.N.; funding acquisition, N.N., T.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Japan Agency for Medical Research and Development (AMED), Research on the construction of a national database of digital ultrasound images and the development of an artificial intelligence-assisted ultrasound diagnosis system, adopted in the ICT Infrastructure Construction and Artificial Intelligence Implementation Research Project for Clinical Research (grant number FY2019).

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Jikei University Hospital (protocol code: 30-281(9302) on Nov 12, 2018).

**Informed Consent Statement:** This study was approved by the institutional review board and the requirement for written informed consent was waived due to its retrospective design.

**Data Availability Statement:** The models are available upon request via this GitHub link: https://github.com/imedix2021/ame_liver_ensemble. The datasets used are uploaded at the following URL: https://drive.google.com/drive/folders/1aeq6AazaesStWGfAEokRhsg-rKziBMVx. For permission to use the dataset, please contact the author. The original hepatic US data set used in this study was selected from US images of hepatic mass lesions from the national project "the construction of big database of US digital image: toward the diagnostic aid with artificial intelligence as part of the ICT Infrastructure Establishment and Implementation of Artificial Intelligence for Clinical and Medical Research fund (https://www.amed.go.jp/en/program/list/14/02/002.html), which is an effort to develop ICT infrastructure that utilizes the results of medical data analysis in a fully integrated manner by the government of Japan. The contact information for this database is rinshoict@amed.go.jp.

**Acknowledgments:** We would like to thank Sho Takahashi for advice regarding statistical analysis.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Rawla P.; Sunkara T.; Muralidharan P.; Raj JP. Update in global trends and aetiology of hepatocellular carcinoma. *Contemp Oncol* **2018**, *22*, 141–150. doi: 10.5114/wo.2018.78941. Epub 2018 Sep 30. PMID: 30455585; PMCID: PMC6238087.
2. Tang A.; Hallouch O.; Chernyak V.; Kamaya A.; Sirlin CB. Epidemiology of hepatocellular carcinoma: target population for surveillance and diagnosis. *Abdom Radiol* **2018**, *43*, 13–25. doi: 10.1007/s00261-017-1209-1. PMID: 28647765.
3. Global Burden of Disease Liver Cancer Collaboration. 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 Oncol* **2017**, *3*, 1683–1691. doi:10.1001/jamaoncol.2017.3055.
4. Kee K-M, Lu S-N. Diagnostic efficacy of ultrasound in hepatocellular carcinoma diagnosis. *Expert Rev Gastroenterol Hepatol* **2017**, *11*, 277–279. doi: 10.1080/17474124.2017.1292126. Epub 2017 Feb 15. PMID: 28162003.
5. Bierig SM, Jones A. Accuracy and cost comparison of ultrasound versus alternative imaging modalities, including CT, MR, PET, and angiography. *J Diagn Med Sonogr* **2009**, *25*, 138–144.
6. Terkawi AS, Karakitsos D, Elbarbary M, Blaivas M, Durieux ME. Ultrasound for the anesthesiologists: present and future. *ScientificWorldJournal* **2013**, *20*, 2013:683685. doi: 10.1155/2013/683685. PMID: 24348179; PMCID: PMC3856172.
7. Wang F, Numata K, Nihonmatsu H, Okada M, Maeda S. Application of new ultrasound techniques for focal liver lesions. *J Med Ultrason* **2020**, *47*, 215–237.
8. Ahn JC, Lee Y-T, Agopian VG, Zhu Y, You S, Tseng H-R, et al. Hepatocellular carcinoma surveillance: current practice and future directions. *Hepatoma Res* **2022**, *8*. doi:10.20517/2394-5079.2021.131.
9. Miller ZA, Lee KS. Screening for hepatocellular carcinoma in high-risk populations. *Clin Imaging* **2016**, *40*, 311–314.
10. Cassinotto C, Aubé C, Dohan A. Diagnosis of hepatocellular carcinoma: an update on international guidelines. *Diagn Interv Imaging* **2017**, *98*, 379–391.
11. Jiang H-Y, Chen J, Xia C-C, Cao L-K, Duan T, Song B. Noninvasive imaging of hepatocellular carcinoma: From diagnosis to prognosis. *World J Gastroenterol* **2018**, 24, 2348–2362.
12. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology* **2018;**286. 887–896. doi: 10.1148/radiol.2017170706. Epub 2017 Oct 23. PMID: 29059036.
13. Azer SA. Deep learning with convolutional neural networks for identification of liver masses and hepatocellular carcinoma: A systematic review. *World J Gastrointest Oncol* **2019**;11, 1218–1230. doi: 10.4251/wjgo.v11.i12.1218. PMID: 31908726; PMCID: PMC6937442.
14. Rompianesi G, Pegoraro F, Ceresa CD, Montalti R, Troisi RI. Artificial intelligence in the diagnosis and management of colorectal cancer liver metastases. *World J Gastroenterol* **2022**, 28, 108-122. doi: 10.3748/wjg.v28.i1.108. PMID: 35125822; PMCID: PMC8793013.
15. Tiyarattanachai T, Apiparakoon T, Marukatat S, Sukcharoen S, Geratikornsupuk N, Anukulkarnkusol N, et al. Development and validation of artificial intelligence to detect and diagnose liver lesions from ultrasound images. *PLoS One* **2021**, *16*, e0252882.
16. Hu HT, Wang W, Chen LD, Ruan SM, Chen SL, Li X, Lu MD, Xie XY, Kuang M. Artificial intelligence assists identifying malignant versus benign liver lesions using contrast-enhanced ultrasound. *J Gastroenterol Hepatol* **2021**, *36*, 2875-2883. doi: 10.1111/jgh.15522. Epub 2021 May 5. PMID: 33880797; PMCID: PMC8518504.
17. Kang J, Ullah Z, Gwak J. MRI-based brain tumor classification using ensemble of deep features and machine learning classifiers. *Sensors (Basel)* **2021**, *21*, 2222. doi: 10.3390/s21062222. PMID: 33810176; PMCID: PMC8004778.
18. Moon WK, Lee YW, Ke HH, Lee SH, Huang CS, Chang RF. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Comput Methods Programs Biomed* **2020**, *190*, 105361. doi: 10.1016/j.cmpb.2020.105361. Epub 2020 Jan 25. PMID: 32007839.
19. Gifani P, Shalbaf A, Vafaeezadeh M. Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans. *Int J Comput Assist Radiol Surg* **2021**, *16*, 115-123. doi: 10.1007/s11548-020-02286-w. Epub 2020 Nov 16. PMID: 33191476; PMCID: PMC7667011.
20. Assiri AS, Nazir S, Velastin SA. Breast tumor classification using an ensemble machine learning method. *J Imaging* **2020**, 6, 39. doi: 10.3390/jimaging6060039. PMID: 34460585; PMCID: PMC8321060.
21. Wang Z, Dong J, Zhang J. Multi-model ensemble deep learning method to diagnose COVID-19 using chest computed tomography images. *J Shanghai Jiaotong Univ Sci* **2022**, 27, 70-80. doi: 10.1007/s12204-021-2392-3. Epub 2021 Dec 26. PMID: 34975263; PMCID: PMC8710815.
22. Wei X, Gao M, Yu R, Liu Z, Gu Q, Liu X, Zheng Z, Zheng X, Zhu J, Zhang S. Ensemble deep learning model for multicenter classification of thyroid nodules on ultrasound images. *Med Sci Monit* **2020**, 26:e926096. doi: 10.12659/MSM.926096. PMID: 32555130; PMCID: PMC7325553.
23. Guo P, Xue Z, Mtema Z, Yeates K, Ginsburg O, Demarco M, Long LR, Schiffman M, Antani S. Ensemble deep learning for cervix image selection toward improving reliability in automated cervical precancer screening. *Diagnostics (Basel)* **2020** , 10, 451. doi: 10.3390/diagnostics10070451. PMID: 32635269; PMCID: PMC7400120.
24. El Asnaoui K. Design ensemble deep learning model for pneumonia disease classification. *Int J Multimed Inf Retr* **2021**, 10, 55-68. doi: 10.1007/s13735-021-00204-7. Epub 2021 Feb 20. PMID: 33643764; PMCID: PMC7896551.

25. Zhou T, Lu H, Yang Z, Qiu S, Huo B, Dong Y. The ensemble deep learning model for novel COVID-19 on CT images. *Appl Soft Comput* **2021**, 106885. doi: 10.1016/j.asoc.2020.106885. Epub 2020 Nov 6. PMID: 33192206; PMCID: PMC7647900.

26. Heisler M, Karst S, Lo J, Mammo Z, Yu T, Warner S, Maberley D, Beg MF, Navajas EV, Sarunic MV. Ensemble deep learning for diabetic retinopathy detection using optical coherence tomography angiography. *Transl Vis Sci Technol* **2020**, 9, 20. doi: 10.1167/tvst.9.2.20. PMID: 32818081; PMCID: PMC7396168.

27. He M, Wang X, Zhao Y. A calibrated deep learning ensemble for abnormality detection in musculoskeletal radiographs. *Sci Rep* **2021**, 11, 9097. doi: 10.1038/s41598-021-88578-w. PMID: 33907257; PMCID: PMC8079683.

28. Mouhafid M, Salah M, Yue C, Xia K. Deep ensemble learning-based models for diagnosis of COVID-19 from chest CT images. *Healthcare (Basel)* **2022**,10, 166. doi: 10.3390/healthcare10010166. PMID: 35052328; PMCID: PMC8776223.

29. Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning. *In Proceedings of Machine Learning Research* **2019**, 6105-6114.

30. Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* **2020**. 10687-10698.

31. McNemar, Q. (June 18, 1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika.* **1947**, 12, 153–157. doi:10.1007/BF02295996

32. Raschka, S., Mirjalili, V. *Python machine learning: machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd: Birmingham, U.K., 2019; pp. 317.