Author Response

The authors are grateful for the reviewers' helpful comments that are valuable in improving this paper. We have revised the manuscript and used the MDPI editing service to enhance the writing quality. The modification for the reviewer's comments is shown in the following.

Revision made in accordance with comments by Reviewer No.2

1. In Table 1, there are several previous methods that focus on the same problem. What is the dataset of the former experiments and why not use the same dataset to compare the model performance between the proposed method and the former methods?

Answer: Thanks for this comment. For this issue, our concern can be summarized as the following points. First, the medical data is not easy to collect, due to privacy concerns. Second, the data sets of former methods are not public. Third, without permission from patients, it is not appropriate to experiment on collected data, from an ethical point of view. Therefore, to make the experiment more reliable, the experimental data were gathered from Kaohsiung Chang Gung Memorial Hospital based on the ethical approvals. For this comment, we have added the clarifications in Section 3.2.1 (P. 5) and Section 4.5 (P. 12).

2. The description about the model and the training is too simple and not detailed enough. Section 3 represents the method ideas, dataset preprocessing and framework of the CNN. But the detail about the model architecture and training is simple.

Answer: Thanks for this comment. For this comment, we added Tables 2 and 3 into Section 3.2.4 (P. 7). Also, the related description was added into Section 3.2.4 (P. 7). Please refer to the following statement for a quick review.

"Table 2 provides the architecture details of the considered CNN models, including the numbers of convolutions and pools, activation functions, and optimization functions. Based on these architectures, our primary intent was to approximate the nearly optimal settings for different CNN models, as depicted in Table 3. The detailed analysis for determining the best settings is shown in Section 4."

3. Section 4.3.1 shows that different data augmentation achieves different performance progress. But why different rotation leads to different results? And why the effect on different models are different, too?

Answer: Thanks for this comment.

For the first concern of "*why different rotation leads to different results?*", we are sorry that it is really our careless fault, which was caused by problem of performance instability. Actually, in CNNs, the filters are randomly generated for each convolution. Therefore, the models and prediction results are different. To reveal this point, we re-conducted the experiments by 20 testings for rotations of 15^0 and -15^0. The following Tables A and B show the best, worst and average accuracies of CNN models for these settings. From Tables A and B, we can know that there exists a gap between best and worst accuracies for each model. However, the best accuracy between rotations of 15^0 and -15^0 is pretty close for each model. Therefore, for this question, our response is the prediction difference between rotations 15^0 and -15^0 is not significant. Based on this discovery, we re-conducted Figure 9 in the manuscript (P. 10). Please refer to the following Figure A for a quick review.

Table A. Best, worst and average accuracies of CNN models for 20 testings under rotation 15^0.

	best	worst	average
VGG16	0.833333	0.680556	0.783333
VGG19	0.819444	0.694444	0.7625
RN50	0.861111	0.666667	0.759722
RN101	0.791667	0.5	0.60625
RN152	0.777778	0.444444	0.583333
DN121	0.861111	0.597222	0.757639
DN169	0.888889	0.694444	0.797222
DN201	0.902778	0.597222	0.809722
ENB0	0.902778	0.694444	0.796528
INv3	0.902778	0.694444	0.802778

Table B. Best, worst and average accuracies of CNN models for 20 testings under rotation -15^0.

	best	worst	average
VGG16	0.861111	0.722222	0.789583
VGG19	0.833333	0.694444	0.772222
RN50	0.833333	0.708333	0.776389
RN101	0.791667	0.486111	0.646528
RN152	0.777778	0.486111	0.590278
DN121	0.847222	0.638889	0.767361
DN169	0.861111	0.680556	0.814583
DN201	0.930556	0.5	0.79375
ENB0	0.861111	0.763889	0.808333
INv3	0.861111	0.708333	0.800694

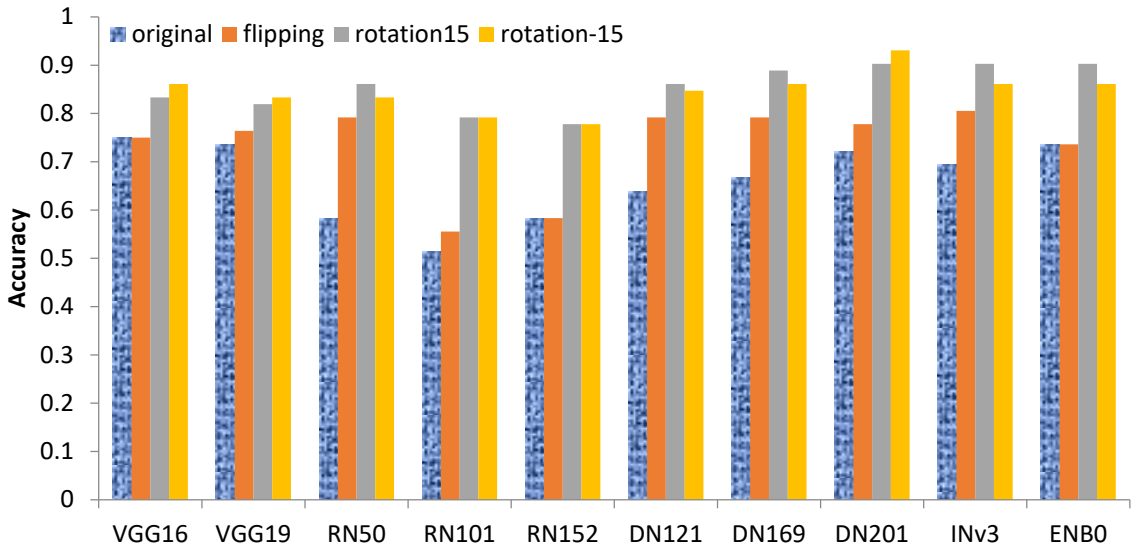

Figure A. Accuracies of different models with different data augmentations.

For the second concern of "*And why the effect on different models are different, too?*", the potential interpretation is that the model architectures were sensitive to data augmentation, considering the pre-trained model. In detail, the Residual idea works well for ResNets and DenseNets. This is also the main contribution to discerning the difference among AI Nets. The related explanation was added in Section 4.3.1 (P. 10).

4. Section 4 gives experiments on different training settings and gets the best training setting on average. However, the analysis points that different models fit different settings. So training all the model in same settings seems unfair and not reasonable. What about training the models in their best settings separately and get the best performance of different methods?

Answer: Thanks for this comment. Indeed, this is our careless neglect. For this issue, we re-conducted the overall comparisons with the best settings for each model. In the revision, Table 3 (P. 7) shows the best settings and the overall comparison result is shown in Table 8 (P. 12). Because the best setting has been adopted, the related figures (Figures 10-13) in the revision were updated thereupon. For a quick review, the summary of the related tables (Table C) and figures (Figures B-D) are shown in the followings.

- The new results in Table C (Table 8 in the revision) are slightly different from previous results, which are summarized as follows.

In Section 4.4, P. 12:

"First, the ranking of the five Nets was DenseNet, InceptionNet, EfficientNet, ResNet, and VGG, with average accuracies of 0.889, 0.889, 0.861, 0.852, and 0.806, respectively. Second, the top three

individual models were DN201, DN169, and RN101, in accordance with their accuracies. Third, overall, the best performances for each metric consistently occurred in two Nets; namely, DN201 and RN101. From this aspect, DN201 and RN101 were the two most reliable models. Fourth, from an average point of view, the top three individual models were DN201, RN101, and INv3, for which the averages of all metric results were 0.886, 0.882, and 0.879, respectively. This can be regarded as an echo of the third point that DN201 and RN101 had a balanced performance for all metrics."

Table C. Effectiveness of compared CNNs with transfer learning for different metrics.

Model \ Metric	Sensitivity	Specificity	Precision	F1-score	AUC	Kappa	Accuracy
VGG16	0.861	0.806	0.816	0.838	0.860	0.667	0.833
VGG19	0.833	0.722	0.750	0.789	0.870	0.556	0.778
RN50	0.889	0.833	0.842	0.865	0.910	0.722	0.861
RN101	0.889	*0.889	*0.889	0.889	*0.950	0.778	0.889
RN152	0.806	0.806	0.806	0.806	0.880	0.611	0.806
DN121	0.917	0.833	0.846	0.880	0.930	0.750	0.875
DN169	0.917	0.861	0.868	0.892	0.890	0.778	0.889
DN201	*0.944	0.861	0.872	*0.907	0.910	*0.806	*0.903
INv3	0.889	*0.889	*0.889	0.889	0.930	0.778	0.889
ENB0	*0.944	0.778	0.810	0.872	0.920	0.722	0.861

Note that, the * indicates the best performance in each metric (attribute).

- Second, the discussion for Figure B (Figure 11 in the revision) was modified as follows.

In Section 4.5, P. 12-13:

"The medical data in real applications is not easy to collect. To make the experiment more reliable, the experimental data were gathered from Kaohsiung Chang Gung Memorial Hospital, instead of crawling the Web. Therefore, due to the problem of insufficient data, overfitting occurred and heavily impacted the recognition quality. To cope with this problem, two operations were adopted: data augmentation and transfer learning. Here, an issue to clarify is the performance comparisons among learning by original data, learning by data augmentation, learning by transfer learning, and learning by fusing data augmentation and transfer learning. This comparison can be explained by Figure 11, summarizing Table 5, Figure 10, and Table 8. From this comparison, we can determine that, whatever the model, without the fusion of transfer learning and data augmentation, the best accuracy cannot be achieved."

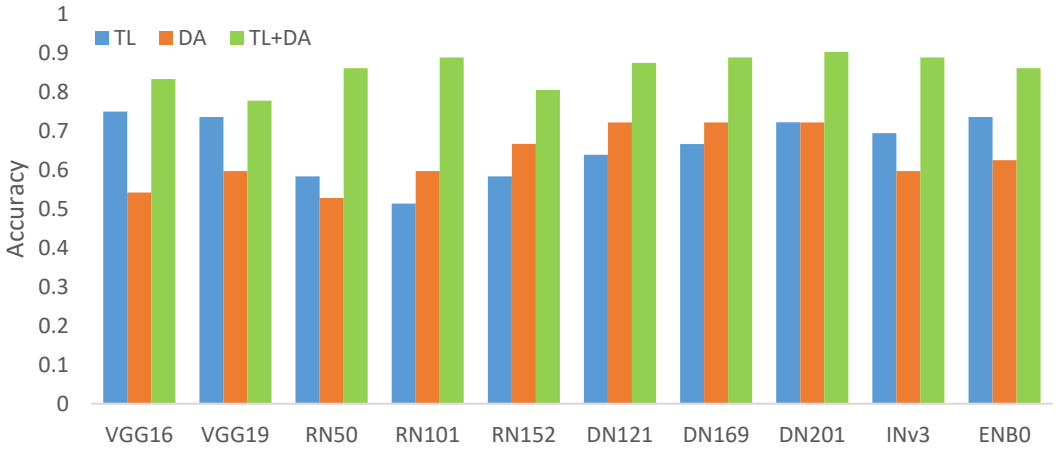

Figure B. Accuracies of compared CNNs by considering Transfer Learning (termed TL), Data Augmentation (termed DA), and fusion of TL and DA (termed TL+DA).

- Moreover, the description for Figure C (Figure 12 in the revision) was updated as follows. In Section 4.5, P. 13:

"Figure 12 reveals the AUC comparisons of models with accuracies larger than 0.9; namely, RN50, RN101, DN121, DN201, INv3, and ENB0. Basically, it can be divided into two observation spaces, by the line where the False-Positive Rate (FPR) equals 0.1. Before FPR 0.1, DN121 is better than the others. To the contrary, the performances of RN101 and DN201 are pretty close, but higher than those of the other models, after FPR 0.1. Overall, RN101, DN201, and INv3 are the highly considerable models, in terms of the AUC."

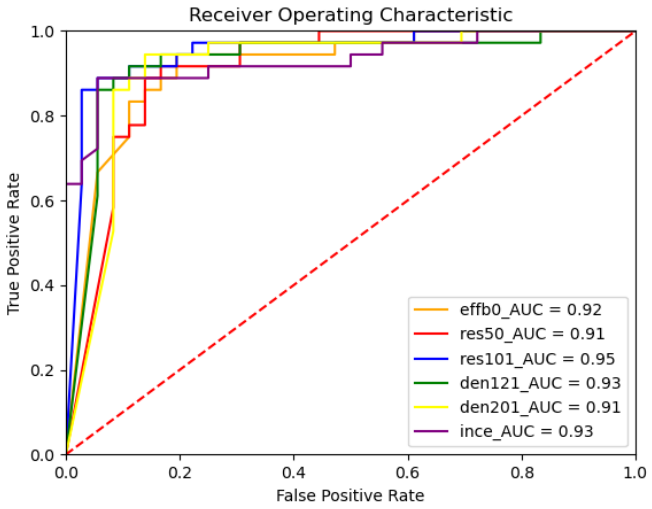

Figure C. AUCs of models with accuracies larger than 0.9.

- Finally, the updated description for Figure D (Figure 13 in the revision) is shown as follows. In Section 4.5, P. 13-14:

"Figure 13 depicts the validation result, showing that the top three models were DN201, DN169, and RN101, although their values were very close. This result is consistent with those given above. However, by considering the overall performances in Table 8, the best two models for recognizing scaphoid fractures are DN201 and RN101."

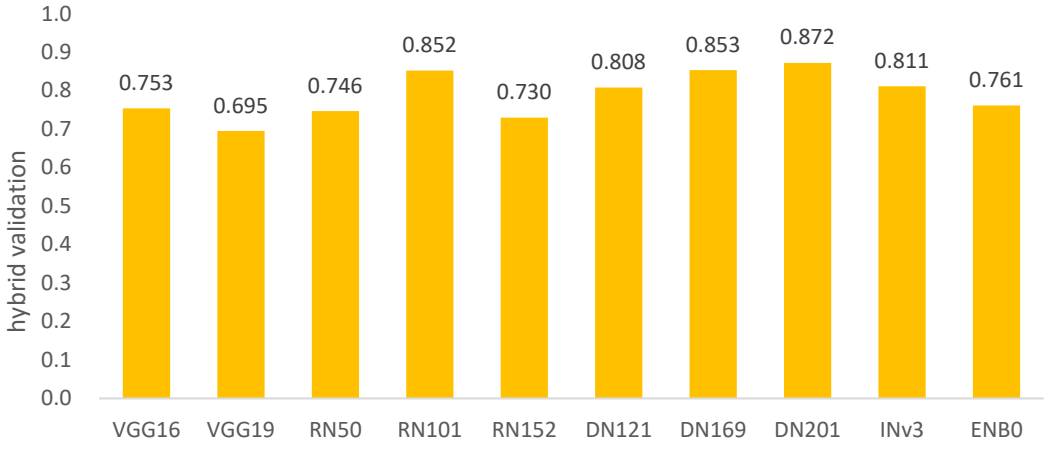

Figure D. Hybrid evaluations by averaging F1-score, AUC, and Kappa metrics.

5. The article points two paradigms of transfer-learning. However, the implementation of transfer-learning of the method is not detailed enough. Which paradigm does the method use and which performs better in the experiments?

Answer: Thanks for this comment. In this paper, the fine-tuning-based transfer learning is performed with VGG Nets, while the layer transfer-based learning was performed for the other Nets. The related description was added into Section 3.2.5 (P. 8).