applied sciences

MDPI

Article

The Usefulness of Gradient-Weighted CAM in Assisting Medical Diagnoses

Jong-Chih Chien [1], Jiann-Der Lee [2,3,4,*], Ching-Shu Hu [2] and Chieh-Tsai Wu [3]

[1] School of Informatics, Kainan University, Taoyuan 33857, Taiwan; jcchien@mail.knu.edu.tw
[2] Department of Electrical Engineering, Chang Gung University, Taoyuan 33302, Taiwan; sheaushing0905@hotmail.com
[3] Department of Neurosurgery, Chang Gung Memorial Hospital at Linkou, Taoyuan 33305, Taiwan; woodie2@adm.cgmh.org.tw
[4] Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City 24330, Taiwan
* Correspondence: jdlee@mail.cgu.edu.tw; Tel.: +886-3-211-8800 (ext. 5316)

Featured Application: Investigation into whether and how much AI-based heat-maps can assist radiologists when making diagnoses based on medical images.

Abstract: In modern medicine, medical imaging technologies such as computed tomography (CT), X-ray, ultrasound, magnetic resonance imaging (MRI), nuclear medicine, etc., have been proven to provide useful diagnostic information by displaying areas of a lesion or tumor not visible to the human eye, and may also help provide additional recessive information by using modern data analysis methods. These methods, including Artificial Intelligence (AI) technologies, are based on deep learning architectures, and have shown remarkable results in recent studies. However, the lack of explanatory ability of connection-based, instead of algorithm-based, deep learning technologies is one of the main reasons for the delay in the acceptance of these technologies in the mainstream medical field. One of the recent methods that may offer the explanatory ability for the CNN classes of deep learning neural networks is the gradient-weighted class activation mapping (Grad-CAM) method, which produces heat-maps that may offer explanations of the classification results. There are already many studies in the literature that compare the objective metrics of Grad-CAM-generated heat-maps against other methods. However, the subjective evaluation of AI-based classification/prediction results using medical images by qualified personnel could potentially contribute more to the acceptance of AI than objective metrics. The purpose of this paper is to investigate whether and how the Grad-CAM heat-maps can help physicians and radiologists in making diagnoses by presenting the results from AI-based classifications as well as their associated Grad-CAM-generated heat-maps to a qualified radiologist. The results of this study show that the radiologist considers Grad-CAM-generated heat-maps to be generally helpful toward diagnosis.

Keywords: CNN; Grad-CAM; assisted diagnosis

check for updates

Citation: Chien, J.-C.; Lee, J.-D.; Hu, C.-S.; Wu, C.-T. The Usefulness of Gradient-Weighted CAM in Assisting Medical Diagnoses. *Appl. Sci.* **2022**, *12*, 7748. https://doi.org/10.3390/app12157748

Academic Editor: Jan Egger

Received: 10 June 2022
Accepted: 17 July 2022
Published: 1 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In cases where physicians are required to make a clinical diagnosis based on medical images derived from technologies such as magnetic resonance imaging (MRI), or computed tomography (CT) scans, even a minor error in judgment may result in adverse effects or complications for the patient. Many applications of deep learning methodology in the detection and classification of abnormalities in pre-operative medical images have been proposed recently in the literature [1–3]. In many of these studies, methods based on the convolutional neural network (CNN) structure have shown great promise in their abilities in prediction and classification [4,5]. A CNN-based architecture is composed of convolution layers to help extract features and build feature maps from the input image(s), pooling layers to concentrate these features, and fully connected layers to classify or predict the

result using the features computed in the previous layers [6]. These modules of the CNN are shown below in Figure 1.

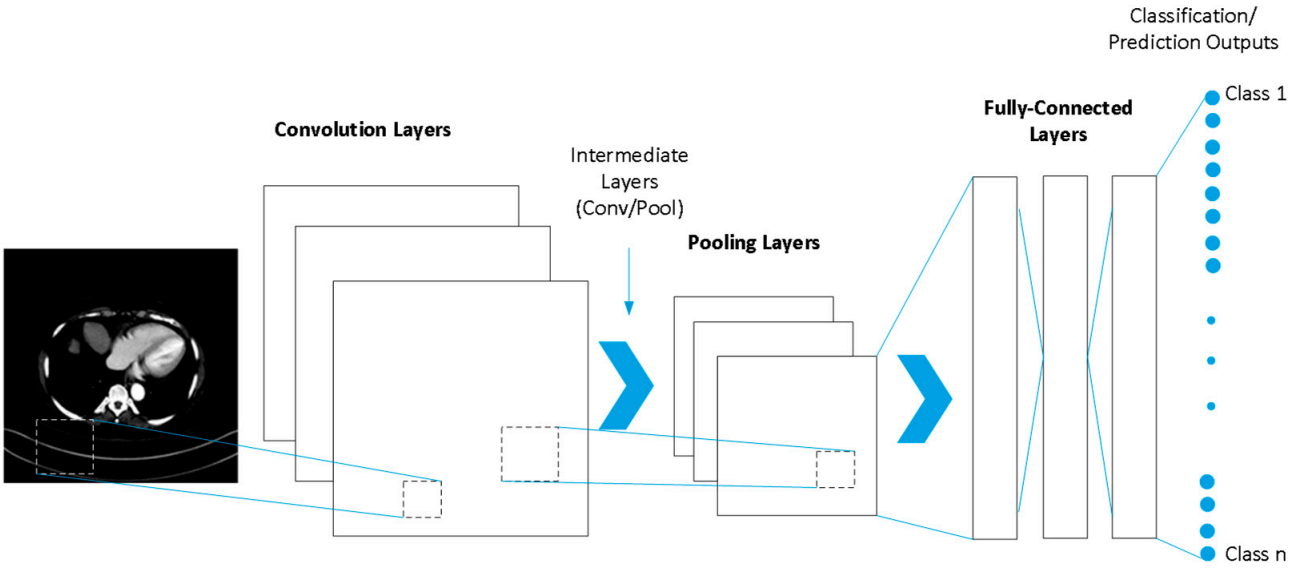


Figure 1. Components of a CNN deep learning network.

Variations of the basic CNN structure can be constructed by altering the numbers and sizes of these layers. Once constructed, the feature maps can then be extracted from the convolution layers for computational purposes and can also be visualized, if necessary. However, because most variations of the CNN architectures are composed of large numbers of these different layers, this produces incredibly high numbers of possible features due to variations in the combinations of different types of layers. Though these features contribute to the effectiveness of the CNN networks in terms of accuracy, they are disadvantageous in trying to explain the reason for the final results due to their large numbers. This lack of explainability is one of the main reasons for the lack of trust in medical results derived from AI-based systems [7]. If explanations for the AI's decisions could be presented to the physicians or radiologists during diagnosis, and are deemed to be helpful, then it may be possible to increase the rate of acceptance for AI-assisted diagnoses, and possibly reduce the number of false diagnoses by physicians. Several approaches have been developed in the literature in the hope of providing possible explanations for the results. Among these, the most promising approaches seek to provide some form of visual heat-maps with higher intensity values around the input image regions containing important features and information that the CNN network used in determining its results [8,9]. These promising methods include class-activation mapping (CAM) [10], saliency map [11], and a modified CAM called the gradient-weighted CAM (Grad-CAM) [12], and its variation, the Grad-CAM++ [13], which was proposed by a different author.

In terms of performance, comparisons between the methods mentioned above, in the applications for medical images, have already been made in the literature regarding the characteristics of the heat-maps they generate [14–16]. A quick summary of the content of these papers is shown below in Table 1.

Table 1. Summary of related papers.

Paper Reference	Summary
[14]	Compared the heat-maps generated using CAM, Grad-CAM, and Grad-CAM++ using 3 types of multiple sclerosis MRI images using the best-performing CNN models in terms of classification accuracy and concluded that Grad-CAM shows the best heat-map localizing ability.
[15]	Compared the heat-maps generated by Saliency Map, Grad-CAM, and Grad-CAM++ using chest X-ray images. Found that Grad-CAM generally localized pathologies better than the other methods, but slightly worse than the human benchmark.
[16]	Evaluated the performance of Grad-CAM on breast mammogram images, and showed that Grad-CAM has good localization capability after the pectoral muscle was removed from the images. Concluded that for improving diagnosis, classification accuracy as well as obtaining a reasonable heat-map is important.

These papers first found the best CNN-based classifiers for their chosen medical images, then generated the heat-maps using the above-mentioned methods, and compared the characteristics of the different heat-maps. These papers concluded that Grad-CAM shows the best performance in terms of localization, which is a desired property for heat-maps; i.e., a more localized heat-map is better at showing the separation of boundaries of the locations containing features that contributed the most toward the classification results, and thus may provide better discrimination and possible explanations for the decision(s) made by the CNN. It is important to note that these heat-maps are not the feature maps of the convolution layers; rather, they show the hierarchy of importance of locations within the feature maps that contributes to the final classification result.

The Grad-CAM method provides a visual form of explanation for the results of CNN models via the computation of heat-maps. It does so through the backpropagation of the result to the last convolution layer in the CNN model and weighting the gradient information to determine the importance of each neuron concerning the input image. It then generates a heat-map showing the importance of each region, as shown below in Figure 2. In Figure 2, w_k^c are the weights, where c represents the classification class and k the numbering of the feature map extracted from the CNN after the classification. If A represents the feature map, and Z is the total number of features in each A, then

$$w_k^c = Z \frac{\partial Y^c}{\partial A_{i,j}^k}, \text{ where } Y^c \text{ is gradient of score for class } c \quad (1)$$

Grad-CAM can be applied to the following families of CNN models: (1) CNNs with fully connected layers, (2) CNNs for structured outputs, and (3) CNNs with multimodal inputs or reinforcement learning tasks without any architectural changes or retraining. Combining Grad-CAM heat-maps with fine-graininess and suitability for visualization, it is possible to achieve the conditions for creating high-resolution visualizations in multiple categories and applying them to ready-made image classification. In the context of image classification models, the benefits are: (1) a deep understanding of their failure modes, (2) robustness to adversarial images, (3) superiority to previously mentioned CNN-explanatory methods in terms of localization, (4) greater faithfulness to the training results of the underlying model, and (5) assistance in achieving goals by identifying data-set biases. There have been many interesting investigations into the applicability of Grad-CAM in various CNN-based applications. In [17], a CNN-based AI structure was used to classify patients' races based on medical images alone, and the investigation not only examined the classification accuracy but also the underlying mechanism for the results of the classifications using Grad-CAM. The results presented in this paper found that the CNN-based AI could distinguish between races based on the medical images alone with an accuracy

around the 90th percentile, which is better than most doctors. The underlying mechanisms for such determinations were not as expected, based on the results of the heat-maps from Grad-CAM. This interesting investigation shows that Grad-CAM may provide explanations for decisions made by CNN-based AI methods that are beyond human expectations.

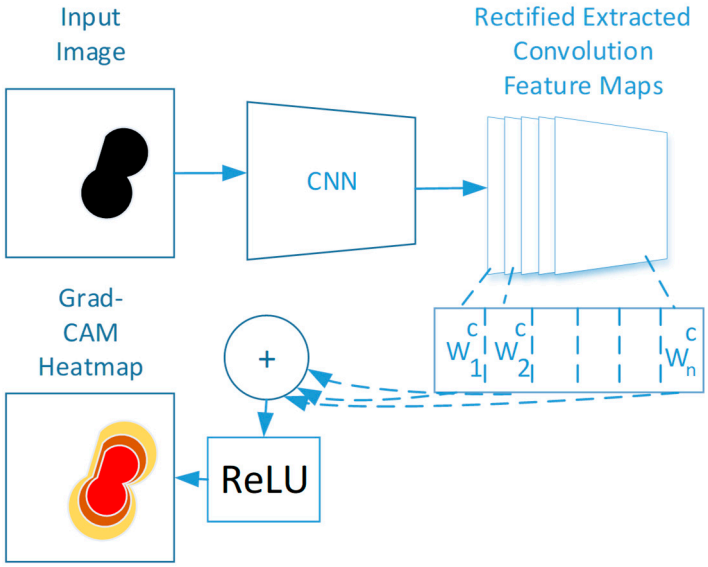


Figure 2. Flowchart and components of Grad-CAM.

Because the comparisons of the effectiveness of Grad-CAM against other CNN-explanatory methods have already been made in the literature, as mentioned before, it is not the purpose of this paper to reinvestigate this aspect nor to compare the characteristics of Grad-CAM with the other methods. Instead, this investigation seeks to probe the explanatory power of Grad-CAM in physician-centered diagnoses. The purpose is to determine whether and how much it can help physicians in making the correct diagnoses or avoiding false diagnoses. The organization of this paper is as follows: the methodology of this investigation will be presented in the following section, followed by the experimental results and discussions, and the conclusions/discussions at the end of this paper.

2. Method

The medical images for this study required pre-label metadata for training purposes, so the images were collected from a trustworthy source. For this purpose, about 1700 pre-labeled CT images were retrieved from the National Institute of Health (NIH) medical image database, the DeepLesion [18]. Examples of these images are shown below in Figure 3.

Figure 3. *Cont.*

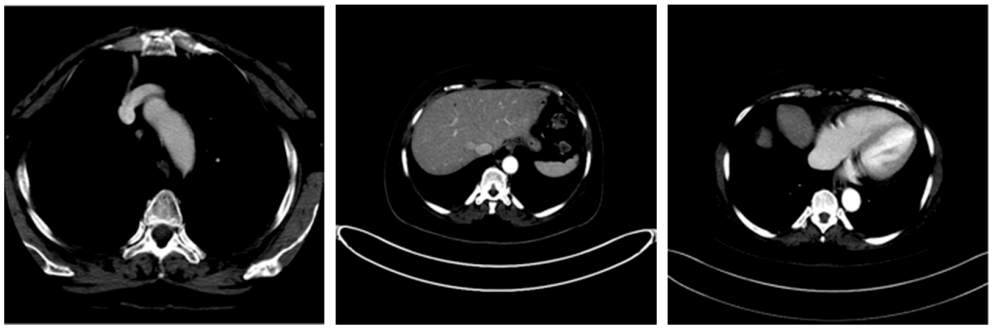


Figure 3. Example of pre-labeled CT images from DeepLesion.

Of these images, 80% were used for training a generic CNN network built using the python language, coded in the free tier of the Google co-laboratory environment [19], and the rest for testing. The training set was trained until a reasonably acceptable level of accuracy was reached, and then the testing set was fed into the trained CNN network. Finally, the Grad-CAM method was applied to each image in the testing data to compute the heat-maps of the images in the training set from their feature maps. A generic CNN model was used for this investigation. The structure of the CNN network created is shown below in Figure 4.

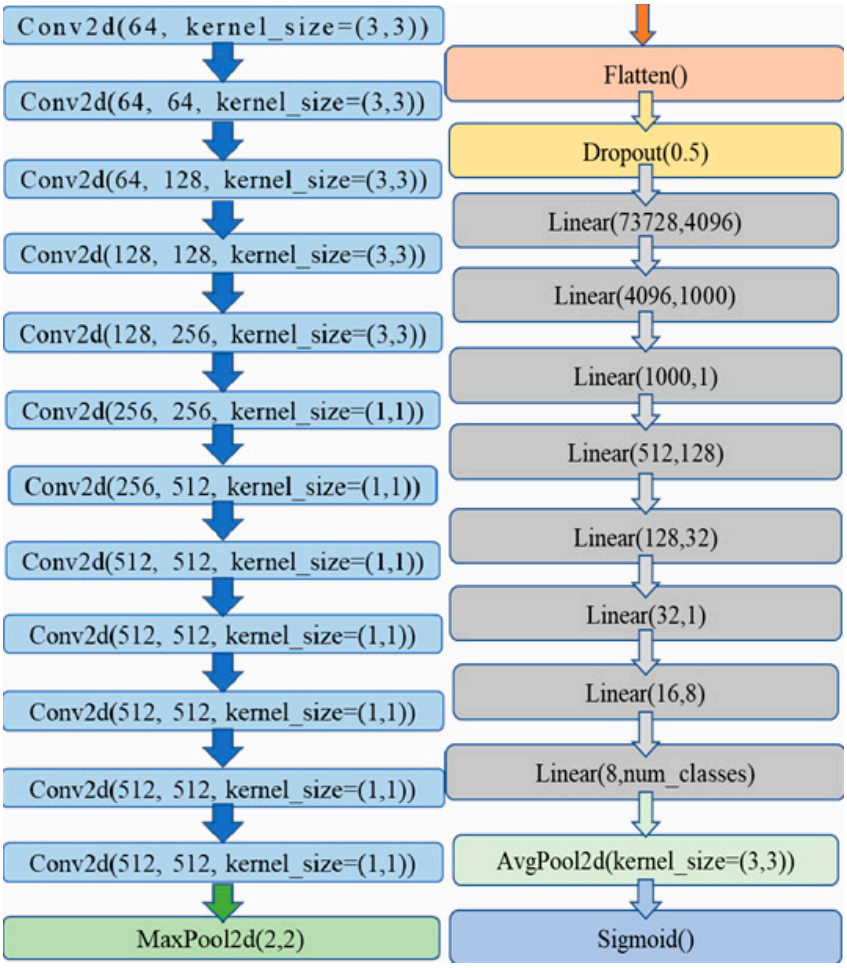


Figure 4. The generic CNN model was used for this investigation.

As seen from the above diagram, the blocks in building our generic CNN model include:

(1) Conv2d: This block constitutes a 2D convolutional layer that creates a convolutional kernel that varies with the layer input and helps to produce the output tensor. The kernel is a convolutional matrix or mask, which can be used for enhancing desired features on its input by convolving it with the input matrix, which can be an image or output from the previous layer.
(2) MaxPool2D: This block is the pooling layer, which retains only the largest value in each 2D sub-block of its input, respectively.
(3) Flatten: This block takes the multidimensional input from the MaxPool2D layer and flattens the whole thing into a one-dimensional output; this is commonly used in the transition from the convolutional layer to the fully connected layer.
(4) Dropout: This block is used in machine learning to help solve the overfitting problem when training. It prevents overfitting by randomly turning off the connections between the input and hidden neurons.
(5) Linear: This block is the module used to perform linear calculations, i.e., $y = A^T x + b$. It is part of the fully connected layer.
(6) AvgPool2d: This block is the pooling layer, which retains only the average values in each 2D sub-block of its input, respectively.
(7) Sigmoid: This block is the activation function, which can be mathematically described. It approximates the recognizable S-shaped curve, which is often used for logistic regression in basic neural network implementation.

In the training phase, our model uses Binary Cross Entropy as the loss function. It is a commonly used binary classification in machine learning. Its equation is expressed as follows:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \quad (2)$$

Finally, to determine the effectiveness of the explanations that can be provided by the heat-maps generated using Grad-CAM, a few heat-maps from false positive and false negative classification results were required. So, a highly accurate classification/prediction result would actually not be helpful for this investigation. Some of the heat-maps of the misclassified (based on the labeled meta-data) images along with the heat-maps of some the correctly identified images were presented to a qualified radiologist in order to assess whether the heat-maps help in reaching a correct diagnosis. The results are presented in the following section.

3. Results

For this experiment, a binary classification was used to predict whether the CT image contains a tumor or not. The threshold to determine whether the generic CNN was sufficiently trained was set arbitrarily at 95% accuracy for the testing set. So, the CT images in the training set were sampled, resampled, and trained until the 20% testing data set was able to reach a classification accuracy of 95.89%, with only 14 false classifications. The following figure, Figure 5, shows example outputs from the first run, where "Yes/No" indicates whether the tumor was detected, and the actual values of "1/0" are DeepLesion labels, where "1" indicates the image contains tumor", and "0" indicates "has no tumor".

The following table, Table 2, shows the statistics from the first and second runs.

After reaching an accuracy of 95.89%, the training was stopped, and the Grad-CAM method was applied for each of the test images, including those that were falsely classified. The following figure, Figure 6, shows some of the CT images with their associated heat-maps overlaid. In the legends of the figures, "Actual" shows the label for DeepLesion, and "Predicted" shows the output of the CNN classifier.

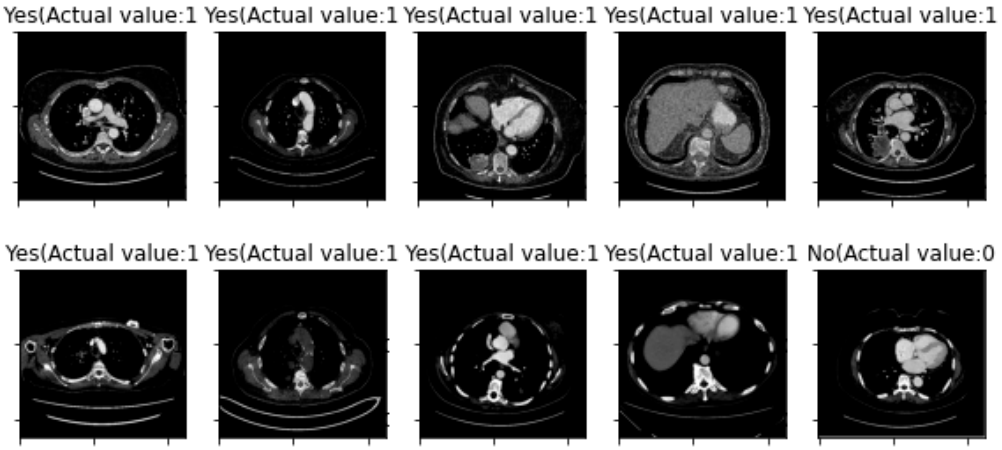


Figure 5. Example of outputs from the first run.

Table 2. Classification statistics for the first and second runs.

First Run:		**Second Run:**	
Training Sample Error Rate	5%	Training Sample Error Rate	5%
Testing Sample Error Rate	14.586%	Testing Sample Error Rate	8.091%
Accuracy	85.414%	Accuracy	91.909%

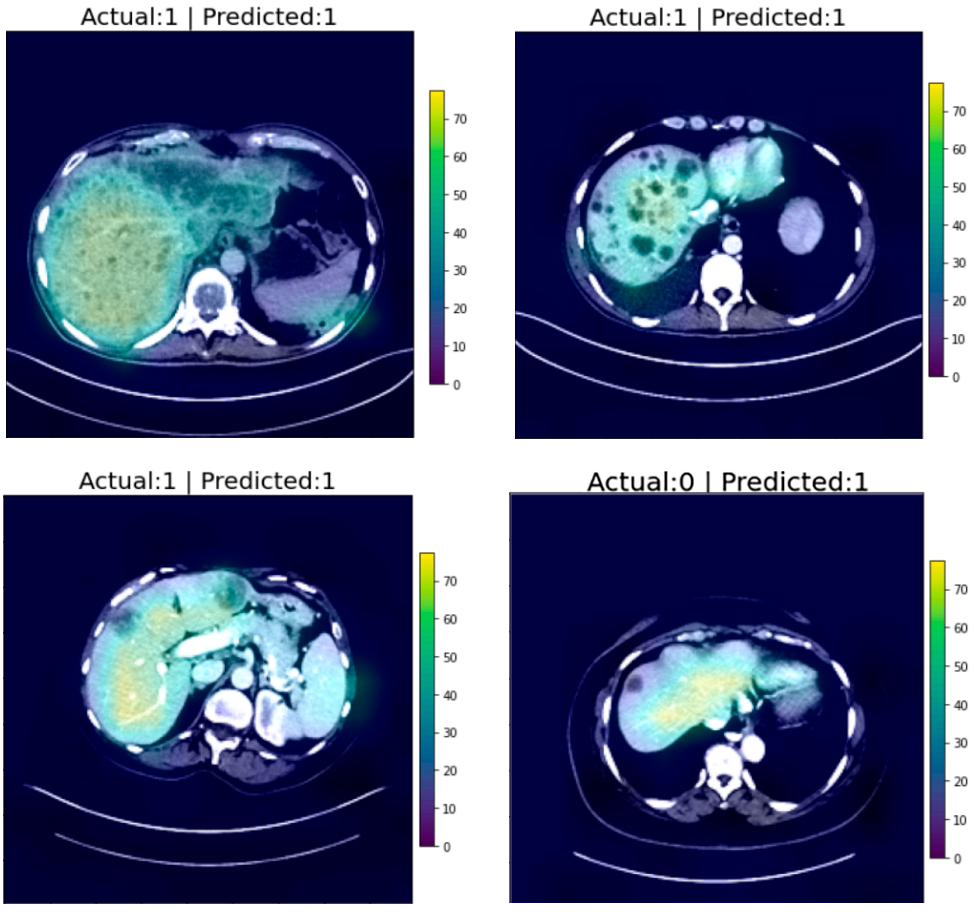


Figure 6. CT images with Grad-CAM-generated heat-maps overlaid.

A qualified clinical radiologist from the Center of Acute and Critical Imaging in the Chang Gung Memorial Hospital at Linkou, located in Taoyuan, Taiwan, was consulted for this investigation. He was shown the original CT images in the test set and those same images with their associated heat-maps overlaid. Unfortunately, because of time and other constraints, responses for only 16 images of all the images that were presented to the clinician were received. The following table, Table 3, lists the DeepLesion labels plus the diagnoses from the clinician. The first column is simply an arbitrary number assigned to the image; the second column contains the translucent version of the heat-map overlaying the original image. The third and fourth columns show the DeepLesion labels and the CNN classification results on whether the image contains tumor(s). The fifth column contains the trained radiologist's diagnosis based on each image. The last column shows the subjective evaluation of the radiologist of the helpfulness of the Grad-CAM-generated heat-map in reaching the diagnosis.

Table 3. Grad-CAM heat-map-overlaid CT images with respective labels and diagnoses.

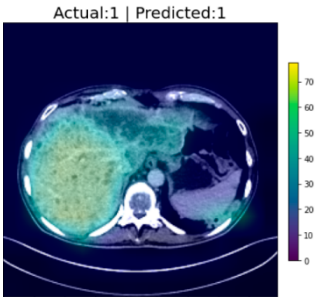
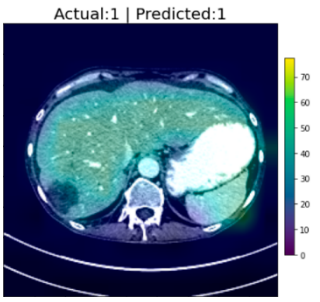
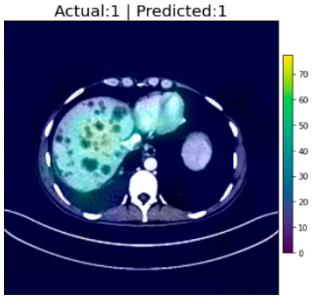
Image Number	Overlaid Image	Deep Lesion Label	CNN Result	Clinical Radiologist's Diagnosis	Grad-CAM Helpfulness
1		Has Tumor	Has Tumor	Has Tumor	Some
2		Has Tumor	Has Tumor	Has Tumor	Yes
3		Has Tumor	Has Tumor	Has Tumor	Yes

Table 3. *Cont.*

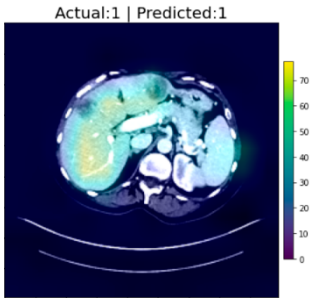
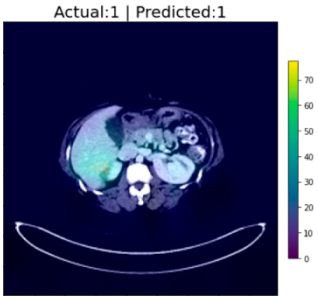
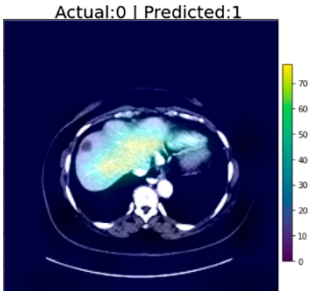
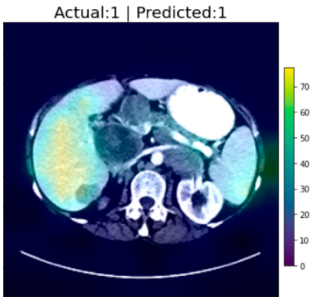
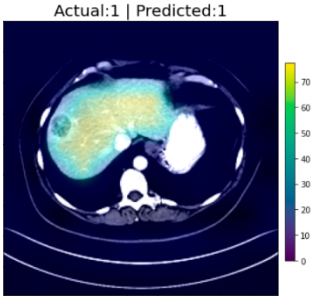
Image Number	Overlaid Image	Deep Lesion Label	CNN Result	Clinical Radiologist's Diagnosis	Grad-CAM Helpfulness
4		Has Tumor	Has Tumor	Has Tumor	Yes
5		Has Tumor	Has Tumor	Has Tumor	Yes
6		No Tumor	Has Tumor	Has Tumor	Yes
7		Has Tumor	Has Tumor	Has Tumor	Yes
8		Has Tumor	Has Tumor	Has Tumor	Yes

Table 3. *Cont.*

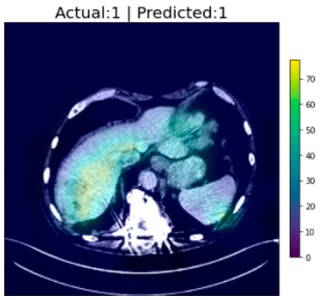
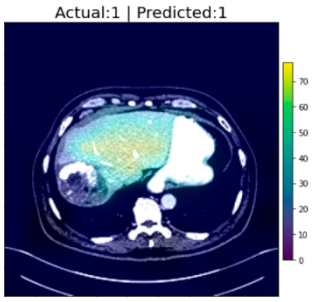
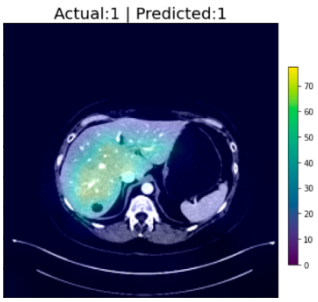
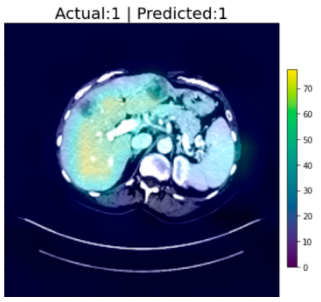
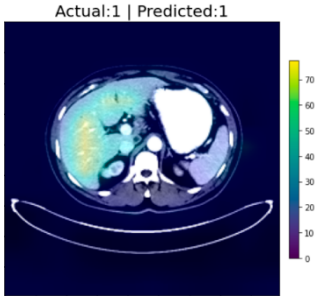
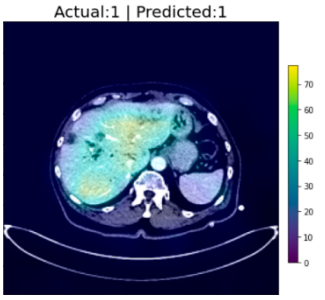
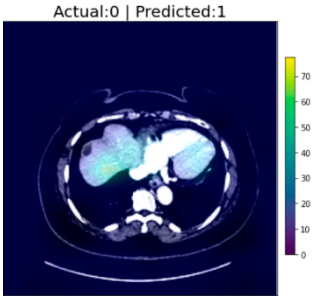
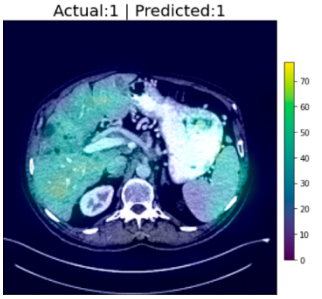
Image Number	Overlaid Image	Deep Lesion Label	CNN Result	Clinical Radiologist's Diagnosis	Grad-CAM Helpfulness
9		Has Tumor	Has Tumor	Has Tumor	Some
10		Has Tumor	Has Tumor	Has Tumor	Yes
11		Has Tumor	Has Tumor	No Tumor (Cyst)	Some
12		Has Tumor	Has Tumor	Has Tumor	Yes
13		Has Tumor	Has Tumor	No Tumor	Yes

Table 3. *Cont.*

Image Number	Overlaid Image	Deep Lesion Label	CNN Result	Clinical Radiologist's Diagnosis	Grad-CAM Helpfulness
14		Has Tumor	Has Tumor	No Tumor (Inflammation of the Biliary Tract)	Some
15		No Tumor	Has Tumor	No Tumor (Cyst)	Some
16		Has Tumor	Has Tumor	Has Tumor	Yes

Note: DeepLesion did not label cysts.

As an interesting sidenote: During the testing phase of the CNN classifier, an investigator noted that it appears that tumors in certain body regions appeared to have a higher rate of classification accuracy than images of other regions. Although it was not in the original plan, an impromptu experiment was conducted to test this hypothesis. From the 1700 original samples, those that contain tumors in specific regions of the body, i.e., the liver and lungs, were grouped together. The reason we chose these two regions is that patients residing in the Asia-Pacific region are more susceptible to cancers in these regions [20], and symptoms of cancers in these regions also tend to be ignored by patients in their early stages because these symptoms can be insignificant. If these types of cancer can be detected at an early stage, then treatments with higher rates of success can be prescribed.

The two groups, liver cancer, and lung cancer were trained and tested separately. The final results appear to justify the earlier suspicions: the classification of tumors in the lung region achieved 98.126% accuracy, while the classification of those in the liver region achieved an astounding 100% accuracy. However, because their sample sizes are small, the results are not truly representative. It is, however, an interesting observation even though the results are not useful for the Grad-CAM experiment due to the small sample sizes. The following figures, Figures 7 and 8, show the sample outputs of these experiments, where "Yes/No" indicates whether a tumor was detected, and the values of "1/0" are DeepLesion labels, where "1" indicates the image contains tumor", and "0" indicates "has no tumor".

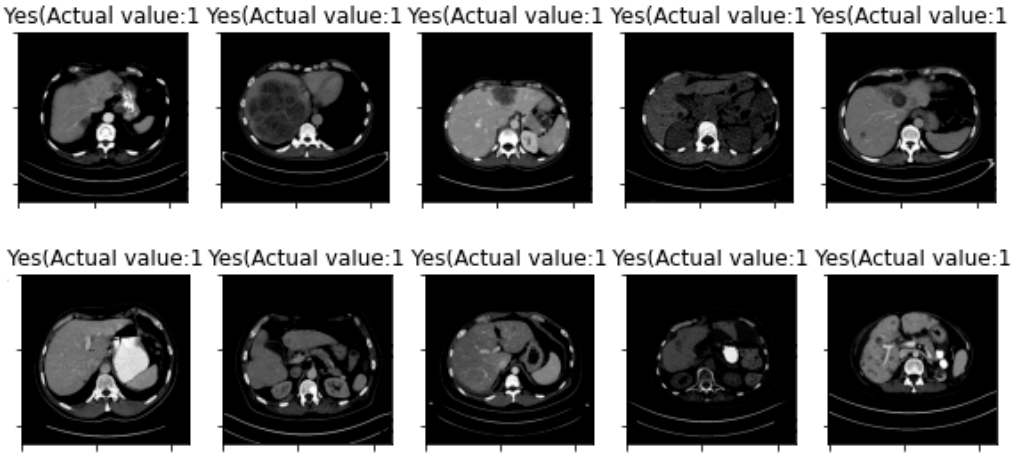


Figure 7. Examples of classification of CT images containing lung tumors.

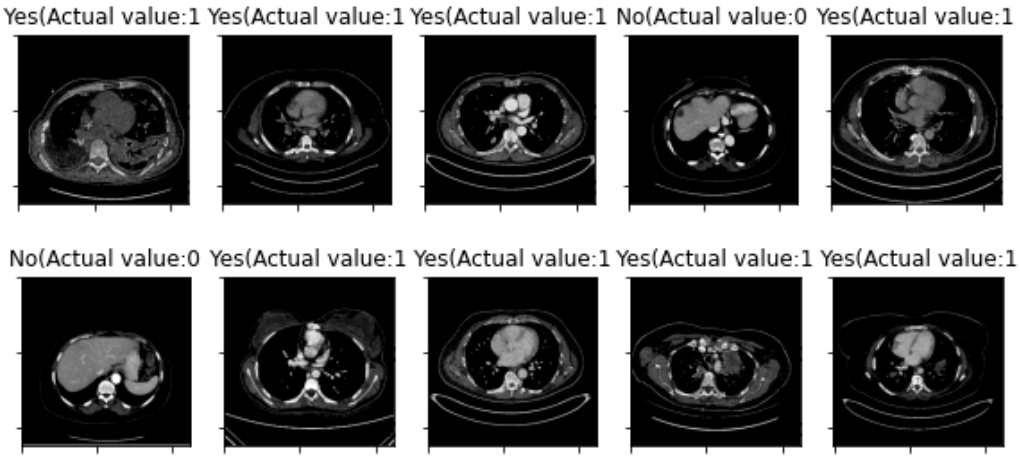


Figure 8. Examples of the classification of CT images containing liver tumors.

4. Discussion

For these 16 images, the radiologist's diagnoses agreed with the DeepLesion labels 12 times and agreed with the CNN results 12 times, though CNN the results differ from the labels in two cases. In these two cases, the radiologist agreed with the DeepLesion labels once and with the CNN results once. In the case where the radiologist agreed with the CNN result, he found the heat-map to be helpful. In the case where the agreement was with the DeepLesion label, the heat-map was considered less helpful. This observation may imply that presenting the heat-maps in addition to the AI-based diagnoses may help physicians in reducing false diagnoses and may help increase the rate of acceptance for AI-based results in the medical community. In addition, for all 16 images, the helpfulness to the radiologist of the Grad-CAM-generated heat-maps ranged from somewhat helpful to really helpful. In no case was a Grad-CAM heat-map considered not helpful at all. From the information contained in the table above, of the clinician's 16 diagnoses, three of them differ from the DeepLesion labels, i.e., images 6, 11 and 14. In the case of image number 6, where the DeepLesion label indicated no tumor, the clinician diagnosed the opposite, which was in agreement with the CNN result. In this case, the radiologist found that the heat-map was helpful. In the case of image number 11, the clinician diagnosed that what appeared to be tumor(s) is actually cyst(s). The Grad-CAM-generated heat-map for image number 11 was deemed only somewhat helpful. Finally, in the case of image number 14, where the DeepLesion label also indicates that the image does contain a tumor, the clinician diagnosed that there is no tumor, but an inflammation of the biliary tract; in this case, the Grad-CAM heat-map was again deemed somewhat helpful. Based on these three cases, and assuming

that the clinician's diagnoses are correct, it is probable that the Grad-CAM heat-maps can be more helpful in the case of a false negative diagnosis than a false positive diagnosis. Though there appears to be no literature discussing the incorrect labeling of DeepLesion, there are studies indicating that the labels in DeepLesion are incomplete [21].

There is a single case among the 16 images, image 15, where the CNN classifier misclassified. The image contained no tumor, as DeepLesion labeled and was verified by the clinician, but the classifier generated a false positive. The clinician further clarified that the image did contain cyst(s), but not a tumor, which may be the reason for the false classification. The false prediction may be an indication that the classifier was not trained sufficiently to differentiate between cysts and tumors, as the DeepLesion database did not label cysts. However, the Grad-CAM heat-map was somewhat helpful by pointing out the region(s) where the classifier thought tumors may exist.

The results of these impromptu experiments appeared to show that the classification accuracies for lung and liver tumors for CNN deep learning network models may be higher than average. This is an interesting result that may be worth pursuing in a separate investigation.

5. Conclusions

Based on the results from this investigation, it may be possible to claim that Grad-CAM-generated heat-maps of a sufficiently trained CNN classifier/predictor can range from somewhat helpful to truly helpful in the hands of a trained radiologist. The heat-maps may be more helpful in correcting a false negative diagnosis than a false positive diagnosis. However, because only 16 images were used in the final stage of this investigation, these claims cannot be presented as a firm result. For future investigations, a qualified radiologist should examine and diagnose more images in order to establish these claims more firmly. As there are differences between the clinical radiologist's diagnoses and the DeepLesion labels, more radiologists should be consulted in cases where the diagnoses are different from the labels. Overall, this investigation shows that the methods such as Grad-CAM, which attempt to provide explanations for the results of deep learning classifications, are heading in the right direction in reducing misdiagnoses when using medical images. Additionally, from other studies presented in the literature, such as [17], methods such as Grad-CAM may help provide explanations for AI-assisted decisions that are beyond humans' current understandings. In conclusion, the application of methods such as Grad-CAM in assisted diagnosis appears to exhibit great potential in reducing false diagnoses.

Author Contributions: Conceptualization, J.-D.L. and J.-C.C.; methodology, J.-D.L. and J.-C.C.; software, J.-D.L. and C.-S.H.; validation, J.-D.L., J.-C.C., C.-S.H. and C.-T.W.; formal analysis, J.-D.L. and J.-C.C.; investigation, J.-D.L., J.-C.C., C.-S.H. and C.-T.W.; resources, C.-S.H. and C.-T.W.; writing—original draft preparation, J.-D.L. and J.-C.C.; writing—review and editing, J.-D.L. and J.-C.C.; project administration, J.-D.L., J.-C.C. and C.-T.W.; funding acquisition, J.-D.L. and J.-C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by Ministry of Science and Technology (MOST), Taiwan, Republic of China, under Grants MOST110-2221-E-182-035 and MOST111-2221-E-182-018.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study was obtained from the DeepLesion (https://nihcc.app.box.com/v/DeepLesion, accessed on 5 January 2022) provided by the National Institute of Health.

Acknowledgments: The authors would like to thank the clinical radiology at Chang Gung Memorial Hospital for his time and help in this investigation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cheng, J.Z.; Ni, D.; Chou, Y.H.; Qin, J.; Tiu, C.; Chang, Y.; Huang, C.; Shen, D.; Chen, C. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Sci. Rep.* **2016**, *6*, 24454. [CrossRef]
2. Azour, F.; Boukerche, A. Design Guidelines for Mammogram-Based Computer-Aided Systems Using Deep Learning Techniques. *IEEE Access* **2022**, *10*, 21701–21726. [CrossRef]
3. Hu, Q.; Whitney, H.M.; Giger, M.L. A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI. *Sci. Rep.* **2020**, *10*, 10536. [CrossRef] [PubMed]
4. Fonollà, R.; van der Zander, Q.E.W.; Schreuder, R.M.; Masclee, A.A.M.; Schoon, E.J.; van der Sommen, F.; de With, P.H.N. A CNN CADx System for Multimodal Classification of Colorectal Polyps Combining WL, BLI, and LCI Modalities. *Appl. Sci.* **2020**, *10*, 5040. [CrossRef]
5. Khan, M.B.; Islam, M.T.; Ahmad, M. A CNN-based CADx Model for Pneumonia Detection from Chest Radiographs with Web Application. In Proceedings of the 2021 International Conference on Science & Contemporary Technologies (ICSCT), Dhaka, Bangladesh, 5–7 August 2021; pp. 1–5.
6. Available online: https://towardsdatascience.com/convolutional-neural-network-feature-map-and-filter-visualization-f75012a5a49c (accessed on 10 February 2022).
7. Shaban-Nejad, A.; Michalowski, M.; Brownstein, J.S.; Buckeridge, D.L. Guest Editorial Explainable AI: Towards Fairness, Accountability, Transparency and Trust in Healthcare. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2374–2375. [CrossRef]
8. Joshi, G.; Walambe, R.; Kotecha, K. A Review on Explainability in Multimodal Deep Neural Nets. *IEEE Access* **2021**, *9*, 59800–59821. [CrossRef]
9. Singh, A.; Sengupta, S.; Lakshminarayanan, V. Explainable Deep Learning Models in Medical Image Analysis. *J. Imaging* **2020**, *6*, 52. [CrossRef] [PubMed]
10. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
11. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Proceedings of the Workshop at International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
12. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
13. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
14. Zhang, Y.; Hong, D.; McClement, D.; Oladosu, O.; Pridham, G.; Slaney, G. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *J. Neurosci. Methods* **2021**, *353*, 109098. [CrossRef] [PubMed]
15. Saporta, A.; Gui, X.; Agrawal, A.; Pareek, A.; Truong, S.Q.H.; Nguyen, C.D.T.; Ngo, V.; Seekins, J.; Blankenberg, F.G.; Ng, A.Y.; et al. Benchmarking saliency methods for chest X-ray interpretation. *medRxiv* **2021**. [CrossRef]
16. Lizzi, F.; Scapicchio, C.; Laruina, F.; Retico, A.; Fantacci, M.E. Convolutional Neural Networks for Breast Density Classification: Performance and Explanation Insights. *Appl. Sci.* **2022**, *12*, 148. [CrossRef]
17. Gichoyo, J.W.; Banerjee, I.; Bhimireddy, A.R.; Burns, J.L.; Celi, L.A.; Chen, L.; Correa, R.; Dullerud, N.; Ghassemi, M.; Huang, S.; et al. AI recognition of patient race in medical imaging: A modelling study. *Lancet Digit. Health* **2022**, *4*, e406–e414. [CrossRef]
18. Available online: https://nihcc.app.box.com/v/DeepLesion (accessed on 5 January 2022).
19. Available online: https://colab.research.google.com (accessed on 15 March 2022).
20. Kimman, M.; Norman, R.; Jan, S.; Kingston, D.; Woodward, M. The burden of cancer in member countries of the Association of Southeast Asian Nations (ASEAN). *Asian Pac. J. Cancer Prev.* **2012**, *13*, 411–420. [CrossRef]
21. Yan, K.; Wang, X.; Lu, L.; Summers, R.M. DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J. Med. Imaging* **2018**, *5*, 036501. [CrossRef]