applied sciences

MDPI

Article

Evaluating the Overall Accuracy of Additional Learning and Automatic Classification System for CT Images

Hiroyuki Sugimori

Faculty of Health Sciences, Hokkaido University, Sapporo 060-0812, Japan; sugimori@hs.hokudai.ac.jp; Tel.: +81-11-706-3410

Received: 28 December 2018; Accepted: 14 February 2019; Published: 17 February 2019

Featured Application: This article describes the evaluation of the automatic classification system for computed tomography (CT) images using a deep learning technique. Additional learning for automatic training will help to create various classification models in the medical fields. The results in this study will be useful for creating new classification models.

Abstract: A large number of images that are usually registered images in a training dataset are required for creating classification models because training of images using a convolutional neural network is done using supervised learning. It takes a significant amount of time and effort to create a registered dataset because recently computed tomography (CT) and magnetic resonance imaging devices produce hundreds of images per examination. This study aims to evaluate the overall accuracy of the additional learning and automatic classification systems for CT images. The study involved 700 patients, who were subjected to contrast or non-contrast CT examination of brain, neck, chest, abdomen, or pelvis. The images were divided into 500 images per class. The 10-class dataset was prepared with 10 datasets including with 5000–50,000 images. The overall accuracy was calculated using a confusion matrix for evaluating the created models. The highest overall reference accuracy was 0.9033 when the model was trained with a dataset containing 50,000 images. The additional learning for manual training was effective when datasets with a large number of images were used. The additional learning for automatic training requires models with an inherent higher accuracy for the classification.

Keywords: deep learning; medical image classification; additional learning; CT image; automatic training; GoogLeNet

1. Introduction

Deep learning techniques [1–3], including deep convolutional neural networks (CNNs), are being employed widely in the field of image processing to conduct image classification [4–6], object detection [7,8], and image segmentation [9–12] tasks. Recently, many studies [4–17] have investigated the applications of deep learning techniques in medical imaging, which now serve as an expansion to this field.

Image diagnosis using computed tomography (CT) and magnetic resonance imaging (MRI) is currently becoming indispensable in the medical field. Although a large number of CT and MRI images are being generated from daily medical examinations, these images are referred to as a follow-up for only a few specific patients. There are many existing models [4–7,13] for the classification of medical images; however, these models are not usually updated since they are created only when needed. Thus, it is not possible to improve such models because they lack procedures and feasibility to retrain

the additional medical images. Additionally, creating models requires a large number of images that usually are registered images in a training dataset because training images for CNN are processed using supervised learning algorithms. Herein, we focus on additional learning and automatic learning for CT images because a current CT scanner has the ability to generate a large number of images per examination. Although there is an existing report [13] on the classification of CT images including contrast enhancement data, there are no reports on the classification of medical images based on the evaluation of the additional learning and automatic image learning system. This study aims to evaluate the overall accuracy of the additional learning and the automatic classification systems for CT images.

2. Materials and Methods

2.1. Subjects and CT Images

The study included 700 patients (male: 371, female: 329; mean age ± standard deviation (SD): 59.2 ± 19.5 years), who were subjected to either a contrast or non-contrast CT examination of the brain, neck, chest, abdomen, or pelvis in January, 2016. This study was approved by the ethics committee of the Hokkaido University Hospital. The CT images were obtained on a 320-detector-row CT scanner (Aquilion ONE; Canon Medical Systems, Otawara, Japan), an 80-detector-row CT scanner (Aquilion PRIME; Canon Medical Systems, Otawara, Japan), and a 64-detector-row Light Speed VCT (GE Medical Systems, Milwaukee, WI, USA).

2.2. Datasets

The dataset of CT images for creating models for classification was divided in 10 classes for brain, neck, chest, abdomen, and pelvis with contrast-enhanced (CE) and non-contrast-enhanced examination, defined as plain (P). The number of images in each class was 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000 images from the earliest date that they were acquired from the 700 patients; the names of the corresponding datasets were defined as 5 K, 10 K, 15 K, 20 K, 25 K, 30 K, 35 K, 40 K, 45 K, and 50 K, respectively, where the letter K represents one thousand; e.g., the 5 K dataset includes a total of 5000 images, 500 images each of the 10 classes in that dataset. For the validation dataset, another three different datasets (A, B, and C) of 1000 images for each class were prepared (a total of 30,000 images), which were exclusive from the above datasets. The names and details of each dataset are listed in Table 1.

Table 1. Names of datasets and the number of images in each label.

Class Name	Dataset										Validation Dataset		
	5 K	10 K	15 K	20 K	25 K	30 K	35 K	40 K	45 K	50 K	A	B	C
Brain (P)	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	1000	1000	1000
Brain (CE)	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	1000	1000	1000
Neck (P)	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	1000	1000	1000
Neck (CE)	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	1000	1000	1000
Chest (P)	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	1000	1000	1000
Chest (CE)	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	1000	1000	1000
Abdomen (P)	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	1000	1000	1000
Abdomen (CE)	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	1000	1000	1000
Pelvis (P)	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	1000	1000	1000
Pelvis (CE)	500	1000	1500	2000	2500	3000	3500	4000	4500	5000	1000	1000	1000
Total number of images	5000	10,000	15,000	20,000	25,000	30,000	35,000	40,000	45,000	50,000	10,000	10,000	10,000

P: plain, CE: contrast enhanced.

The image range of each class was defined as follows. Brain: slice from the anterior tip of the parietal bone to the foramen magnum; neck: slice from the foramen magnum to the pulmonary apex; chest: slice from the pulmonary apex to the diaphragm; abdomen: slice from the diaphragm to the top of an iliac crest; pelvis: slice from the top of an iliac crest to the distal end of the ischium (Figure 1). The range of each class was the same as that of a previous report [13] for the classification of CT images.

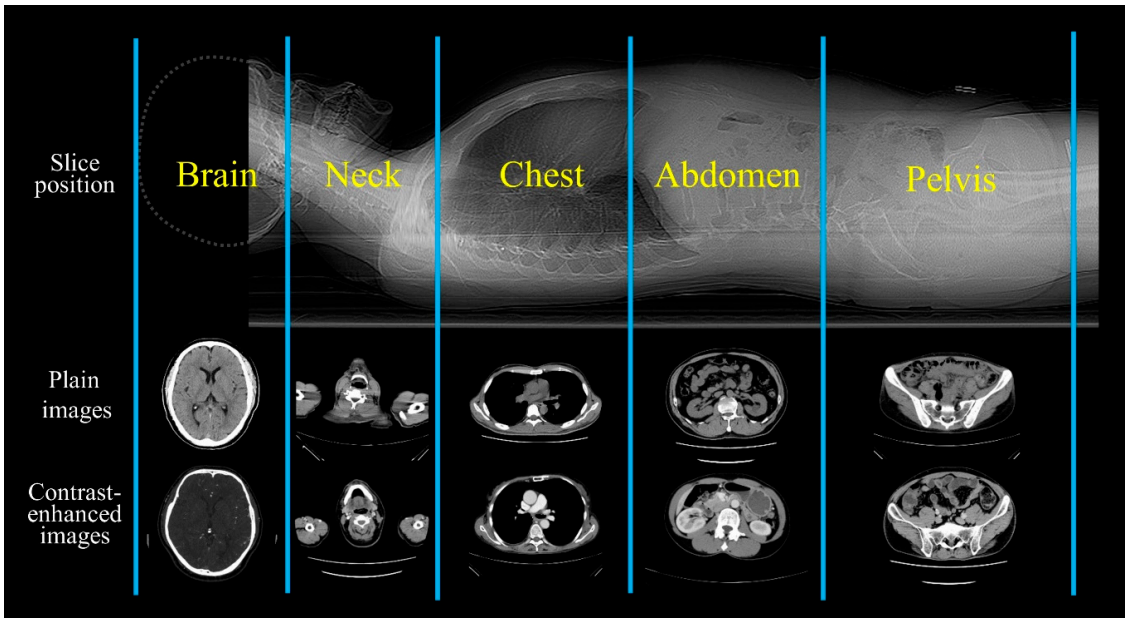

Figure 1. Slice position and sample CT images of the 10 classes.

CE examination involved the intravascular injection of contrast media before examination. The timing of the scan from injection was not considered. Exclusion criteria of CT images for the datasets were images with excessive magnification, images with the reconstruction kernel of bone or lung, images with nothing above the anterior tip of the parietal bone, and images with only arms or legs.

2.3. Preprocessing of Images for Creating the Models

The CT images were retrieved from the picture archiving and communication system. To convert the images for use by the training database, the CT images were converted from digital imaging and communications in medicine (DICOM) format to joint photographic experts group (JPEG) format using a dedicated DICOM software (XTREK view, J-MAC SYSTEM Inc., Sapporo, Japan). The window width and level of DICOM image were used to preset values in the DICOM tag. The DICOM images were converted to JPEG images with a size of 512 × 512 pixels. JPEG files were sorted into folders according to the class that each image belonged to.

2.4. Manual Training of the Images for Creating the Models

The outline of the training performed for creating the models is shown in Figure 2. The authoring software for deep learning was performed via in-house MATLAB (The Mathworks Inc., Natick, MA, USA) software, and a deep learning optimized machine with two GTX1080 Ti GPUs with 11.34 TFlops of single precision, 484 GB/s of memory bandwidth, and 11 GB of memory per board were used. Herein, GoogLeNet [3] with 22 layers was used as the CNN architecture (Figure 3). The hyper-parameters of the training models are as follows: Maximum training epochs were 10 and an initial learning rate was 0.0001. The learning rate was fixed throughout the training. The overall accuracy was calculated using the confusion matrix in the software. The results were evaluated using the validation datasets. Each dataset for training was sorted by a radiological technologist with 17 years of experience. Datasets were divided into 500 images per class for every 5000 images and 1000 images per class for every 10,000 images to create a model for each dataset. Additional learning processes, which were repeated up to the 50 K dataset, were performed to evaluate the accuracy after additional effects.

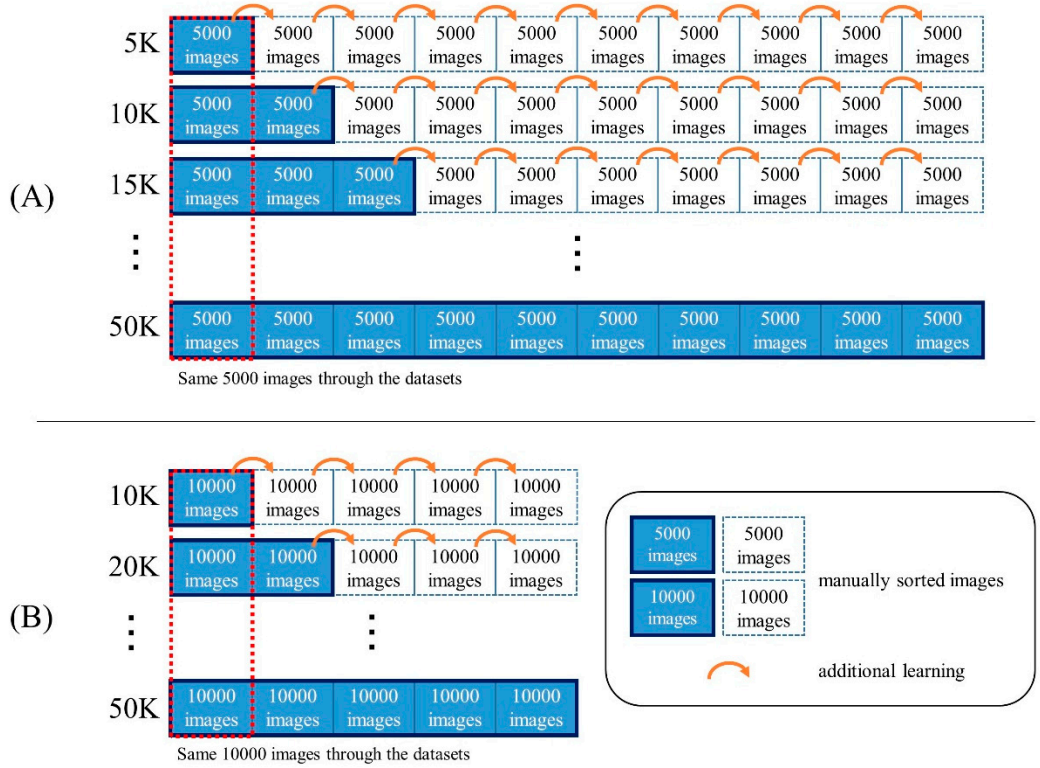

Figure 2. Outline of the training performed for creating the models. Additional learning for every (**A**) 5000 images (**B**) 10,000 images.

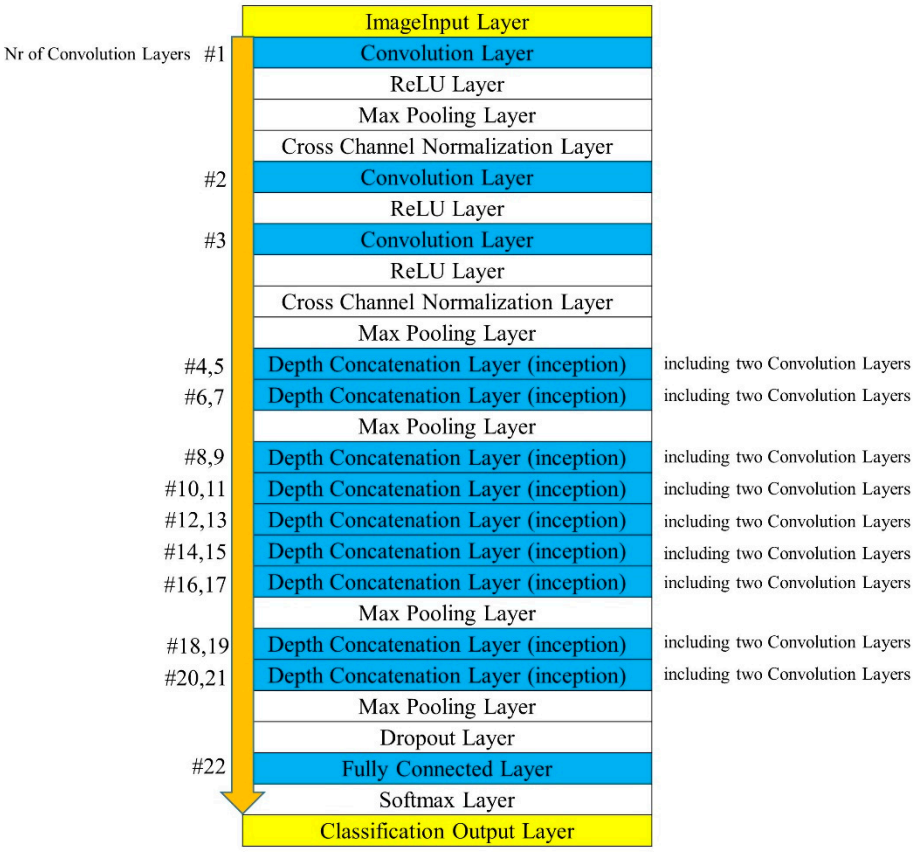

Figure 3. The CNN architecture, which has 22 convolutional layers.

2.5. Automatic Training for Creating Models

The outline of the training for creating the models is shown in Figure 3. The authoring software, machine, CNN architecture, and hyper-parameters of training models were the same as those discussed in Section 2.4. The automatic training system was developed with MATLAB software because supervised learning usually requires images that were classified by humans. Differing from manual training, the following functions were added to the software. (i) The created models with each dataset were used to automatically classify new images into the classes to which they should belong. (ii) The classified JPEG files were sorted into each folder according to their image classes. (iii) The classified images were used for the training to create new models. (iv) The automatic classification and creation of a model was repeated up to the 50 K dataset (Figure 4). The new images provided were divided into 500 images per class for every 5000 images and 1000 images per class for every 10,000 images.

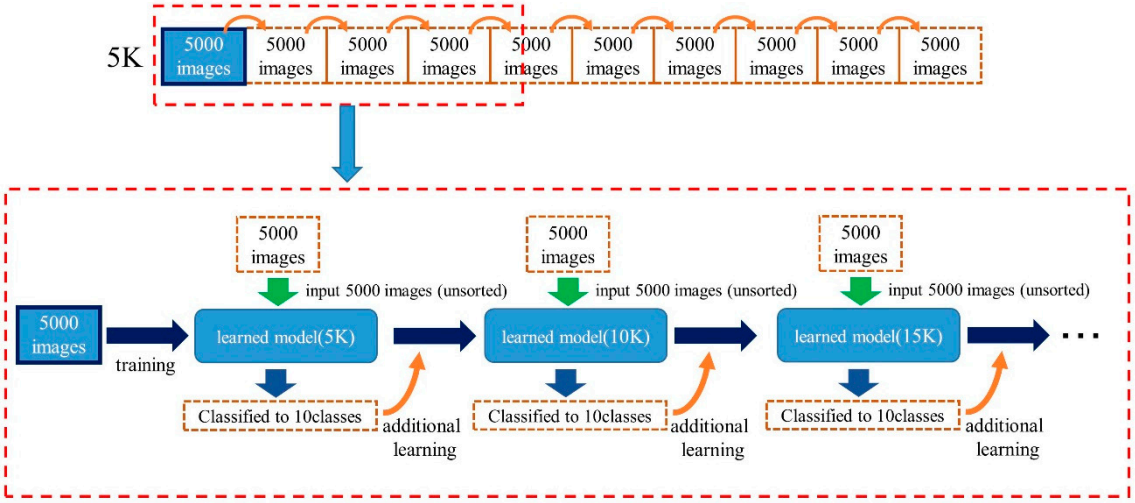

Figure 4. Details of the additional learning process for automatic training (Example case using 5 K datasets).

2.6. Evaluation of the Created Models

The confusion matrix obtained using each dataset, shown in Figure 5, is an indicator of the performance of the created models. The training performed with 10 image classes is shown as a 10×10 table and all performances were based on numbers obtained by applying the classifier to the validation dataset. The overall accuracy was obtained as a ratio of the number of correctly classified images in all validation images to the total number of images. The overall accuracies in each dataset were calculated as reference accuracy. Accuracies of the manual and automatic training were calculated for each dataset. Furthermore, the overall accuracies were evaluated three times with each validation dataset and presented at mean regardless of the dataset.

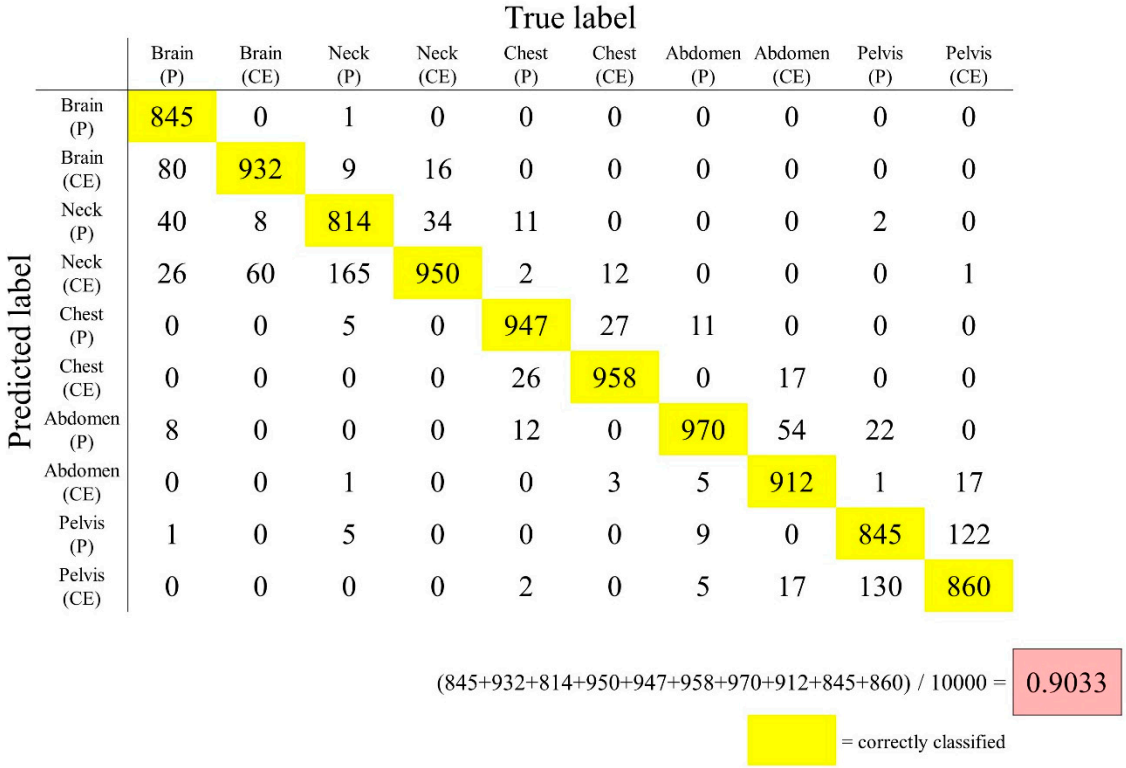

Figure 5. Confusion matrix for evaluating the overall accuracy, which was calculated using the validation dataset A with 50 K dataset.

3. Results and Discussions

3.1. Reference Accuracy

Table 2 shows the overall accuracy for each dataset. With an increase in the size of image datasets, the overall accuracy became higher. The highest overall accuracy for the datasets used was 0.9033 and the model was trained using the 50 K dataset.

Table 2. Overall accuracy for each dataset.

Dataset Type	Group	Dataset									
		5 K	10 K	15 K	20 K	25 K	30 K	35 K	40 K	45 K	50 K
Validation dataset	A	0.6028	0.6532	0.7293	0.7914	0.8334	0.8369	0.8615	0.8947	0.8986	0.9033
	B	0.4833	0.5352	0.5713	0.6166	0.6789	0.7056	0.7693	0.7877	0.7884	0.8422
	C	0.4927	0.5101	0.5899	0.613	0.7121	0.7397	0.8208	0.8472	0.8633	0.8974
	mean	0.5263	0.5662	0.6302	0.6737	0.7415	0.7607	0.8172	0.8432	0.8501	0.881

3.2. Manual Training

Figure 6 shows the relation between datasets and the overall accuracy of the created model for manual training. For the additional learning of every 5000 images, the overall accuracy when additional learning started from 5 K to 20 K increased continuously up to 25 K. However, after exceeding the 30 K dataset, the overall accuracy fluctuated. For the additional learning of every 10,000 images, the overall accuracy increased continuously up to 40 K. However, the overall accuracy of the dataset of 40 K slightly declined compared to that of 30 K.

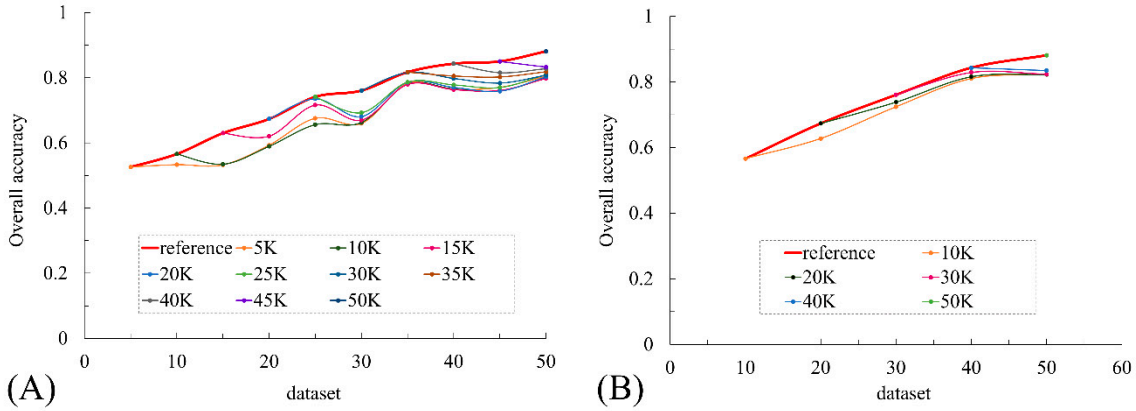

Figure 6. Relation between datasets and overall accuracy for manual training. (**A**) Additional learning for every 5000 images, (**B**) additional learning for every 10,000 images.

3.3. Automatic Training

Figure 7 shows the relation between datasets and the overall accuracy of the created model for automatic training. For the additional learning of every 5000 images, there was little increase in the overall accuracy when the additional learning started from 5 K to 20 K. There was a gradual decrease in the overall accuracy when the additional learning started from 25 K to 35 K and over 40 K dataset. There were no subsequent data when the additional learning started from 5 K to 20 K because some created models could not classify new images up to 10 classes because they had incomplete classification models. For the additional learning for every 10,000 images, there was little increase in the overall accuracy. However, when the additional learning started from 40 K, the overall accuracy was maintained at a high value.

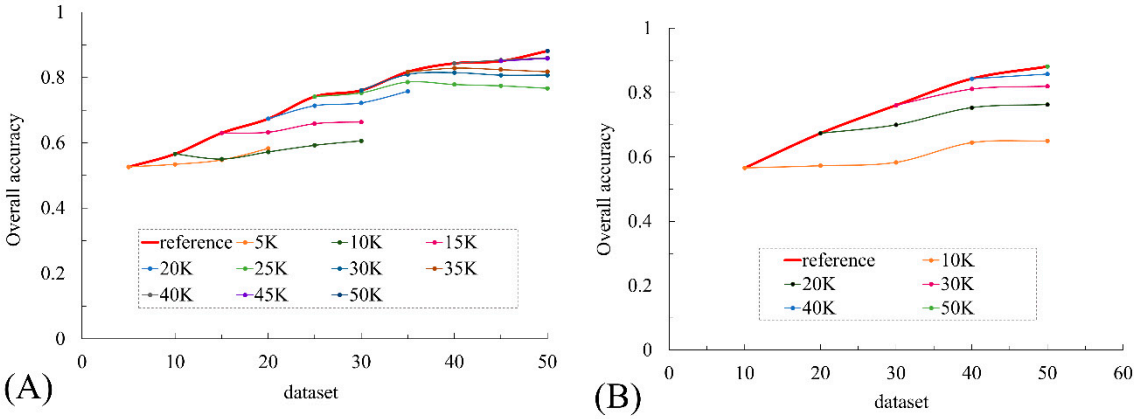

Figure 7. Relation between datasets and the overall accuracy for automatic training. (**A**) Additional learning for every 5000 images, (**B**) additional learning for every 10,000 images.

This study evaluated the overall accuracy of the additional learning and automatic classification system for CT images. From the viewpoint of additional learning, there was a significant improvement of the overall accuracy for the manual training. However, the additional dataset to be added should be prepared with a large number of images because the training for every 5000 images might be affected by specific feature amount. One of the reasons for the fluctuating accuracy, as shown in Figure 6A, might be insufficient feature information in the dataset. For the additional learning of every 5000 images, the number of images for the additional training was small perhaps because, as shown by a previous report [13], the number of CT images affected the accuracy of training the dataset. If the additional dataset included specific patients' data (for instance, the patient who suffered serious traffic

accident), the feature amount through the training may be changed dramatically. Therefore, additional images with a variety of features should be prepared by using a high enough number of images for additional learning. On the contrary, automatic training showed no improvement in the overall accuracy, one reason being that the inherent accuracy is not affected by the created models. As the reference overall accuracy, the datasets between 5 K and 20 K were under 0.8 of the overall accuracy. Inaccurate classification affected the models created for automatic training. As a result, there was no further improvement in the overall accuracy. However, when additional learning started from the 40 K and larger datasets, the reference accuracy around 0.9 maintained the overall accuracy at this value. This means that automatic training with a model of higher inherent accuracy might be effective in performing accurate classifications.

The limitations of this study are as follows. First, the hyper-parameters of the training models used are fixed parameters. Although a previous study [13] showed that the hyper-parameters and CNN architecture affected the overall accuracy, the CNN architecture of GoogLeNet is suitable for performing classification in many fields owing to its high accuracy; thus, we used fixed parameters. Second, the process of training accuracy and loss were not showed in this study because the ability to generalize was most important for the intended application [18] in the training process; thus, we only focused on the overall accuracy. However, the overfitting would hardly cause problems during training in this study because GoogLeNet adopted the inception module [19] and global average pooling [20] for preventing overfitting. Third, the additional image data was fixed at 500 images per class. Actual human CT images are often taken from a specific region such as from the lung or liver. The number of images in each class was unstable and imbalanced, as observed during the daily routine examinations. Therefore, the standard of the additional images was required to be set to the number of images and not patients because the additional learning needs to be evaluated in the same situation. In the future, we plan to investigate the effects of an imbalanced number of images when creating an additional model. As for the images, the CT images were converted from DICOM to JPEG images in this study. The CT images have Hounsfield Units (HUs, CT-specific numbers); by definition, water is zero HU and air is −1000 HU. A previous study [21] showed the strong correlation between HUs and grayscales though the JPEG images have no information of absolute values. We supposed the classification of the slice position might not be affected in this study.

4. Conclusions

Herein, we evaluated the overall accuracy of the additional learning and the automatic classification system for CT images. It was found that additional learning for manual training was effective when a large number of images were used. The additional learning for automatic training requires models with the inherent higher accuracy for the classification.

Author Contributions: H.S. proposed the idea, contributed to data acquisition, performed manual classification, data analysis, algorithm construction, wrote the article, and edited the paper.

Funding: This study was supported in part by Grants-in-Aid for Regional R&D Proposal-Based Program from Northern Advancement Center for Science & Technology of Hokkaido Japan.

Acknowledgments: The author thanks the laboratory students Kazuya Sasaki for his help.

Conflicts of Interest: The author declares no conflict of interest.

References

1. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.

3. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
4. Lakhani, P.; Sundaram, B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* **2017**, *284*, 574–582. [CrossRef] [PubMed]
5. Qayyum, A.; Anwar, S.M.; Awais, M.; Majid, M. Medical image retrieval using deep convolutional neural network. *Neurocomputing* **2017**, *266*, 8–20. [CrossRef]
6. Gao, X.W.; Hui, R.; Tian, Z. Classification of CT brain images based on deep learning networks. *Comput. Methods Programs Biomed.* **2017**, *138*, 49–56. [CrossRef] [PubMed]
7. Masood, A.; Sheng, B.; Li, P.; Hou, X.; Wei, X.; Qin, J.; Feng, D. Computer-Assisted Decision Support System in Pulmonary Cancer detection and stage classification on CT images. *J. Biomed. Inform.* **2018**, *79*, 117–128. [CrossRef] [PubMed]
8. Zhao, X.; Liu, L.; Qi, S.; Teng, Y.; Li, J.; Qian, W. Agile convolutional neural network for pulmonary nodule classification using CT images. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 585–595. [CrossRef] [PubMed]
9. Wachinger, C.; Reuter, M.; Klein, T. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *Neuroimage* **2018**, *170*, 434–445. [CrossRef] [PubMed]
10. Akkus, Z.; Galimzianova, A.; Hoogi, A.; Rubin, D.L.; Erickson, B.J. Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. *J. Digit. Imaging* **2017**, *30*, 449–459. [CrossRef] [PubMed]
11. Ren, X.; Xiang, L.; Nie, D.; Shao, Y.; Zhang, H.; Shen, D.; Wang, Q. Interleaved 3D-CNNs for joint segmentation of small-volume structures in head and neck CT images. *Med. Phys.* **2018**, *45*, 2063–2075. [CrossRef] [PubMed]
12. Avendi, M.R.; Kheradvar, A.; Jafarkhani, H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Med. Image Anal.* **2016**, *30*, 108–119. [CrossRef] [PubMed]
13. Sugimori, H. Classification of Computed Tomography Images in Different Slice Positions Using Deep Learning. *J. Healthc. Eng.* **2018**, *2018*, 9. [CrossRef] [PubMed]
14. Kim, K.H.; Choi, S.H.; Park, S.-H. Improving Arterial Spin Labeling by Using Deep Learning. *Radiology* **2017**, *287*, 658–666. [CrossRef] [PubMed]
15. Liu, F.; Jang, H.; Kijowski, R.; Bradshaw, T.; McMillan, A.B. Deep Learning MR Imaging–based Attenuation Correction for PET/MR Imaging. *Radiology* **2017**, *286*, 676–684. [CrossRef] [PubMed]
16. Yasaka, K.; Akai, H.; Kunimatsu, A.; Abe, O.; Kiryu, S. Liver Fibrosis: Deep Convolutional Neural Network for Staging by Using Gadoxetic Acid–enhanced Hepatobiliary Phase MR Images. *Radiology* **2017**, *287*, 146–155. [CrossRef] [PubMed]
17. Chen, M.C.; Ball, R.L.; Yang, L.; Moradzadeh, N.; Chapman, B.E.; Larson, D.B.; Langlotz, C.P.; Amrhein, T.J.; Lungren, M.P. Deep Learning to Classify Radiology Free-Text Reports. *Radiology* **2017**, *286*, 845–852. [CrossRef] [PubMed]
18. Zheng, Q.; Yang, M.; Yang, J.; Zhang, Q.; Zhang, X. Improvement of Generalization Ability of Deep CNN via Implicit Regularization in Two-Stage Training Process. *IEEE Access* **2018**, *6*, 15844–15869. [CrossRef]
19. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
20. Lin, M.; Chen, Q.; Yan, S. Network in Network. *arXiv*, 2013; arXiv:1312.4400.
21. Kamaruddin, N.; Rajion, Z.A.; Yusof, A.; Aziz, M.E. Relationship between Hounsfield unit in CT scan and gray scale in CBCT. *AIP Conf. Proc.* **2016**, *1791*, 020005. [CrossRef]