biomolecules

MDPI

Article

The Development of a Skin Cancer Classification System for Pigmented Skin Lesions Using Deep Learning

Shunichi Jinnai [1,*], Naoya Yamazaki [1], Yuichiro Hirano [2], Yohei Sugawara [2], Yuichiro Ohe [3] and Ryuji Hamamoto [4,5,*]

1 Department of Dermatologic Oncology, National Cancer Center Hospital, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan; nyamazak@ncc.go.jp
2 Preferred Networks, 1-6-1 Otemachi, Chiyoda-ku, Tokyo 100-0004, Japan; hirano@preferred.jp (Y.H.); suga@preferred.jp (Y.S.)
3 Department of Thoracic Oncology, National Cancer Center Hospital, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan; yohe@ncc.go.jp
4 Division of Molecular Modification and Cancer Biology, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan
5 Cancer Translational Research Team, RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
* Correspondence: sjinnai@ncc.go.jp (S.J.); rhamamot@ncc.go.jp (R.H.)

Received: 23 June 2020; Accepted: 28 July 2020; Published: 29 July 2020

Abstract: Recent studies have demonstrated the usefulness of convolutional neural networks (CNNs) to classify images of melanoma, with accuracies comparable to those achieved by dermatologists. However, the performance of a CNN trained with only clinical images of a pigmented skin lesion in a clinical image classification task, in competition with dermatologists, has not been reported to date. In this study, we extracted 5846 clinical images of pigmented skin lesions from 3551 patients. Pigmented skin lesions included malignant tumors (malignant melanoma and basal cell carcinoma) and benign tumors (nevus, seborrhoeic keratosis, senile lentigo, and hematoma/hemangioma). We created the test dataset by randomly selecting 666 patients out of them and picking one image per patient, and created the training dataset by giving bounding-box annotations to the rest of the images (4732 images, 2885 patients). Subsequently, we trained a faster, region-based CNN (FRCNN) with the training dataset and checked the performance of the model on the test dataset. In addition, ten board-certified dermatologists (BCDs) and ten dermatologic trainees (TRNs) took the same tests, and we compared their diagnostic accuracy with FRCNN. For six-class classification, the accuracy of FRCNN was 86.2%, and that of the BCDs and TRNs was 79.5% ($p = 0.0081$) and 75.1% ($p < 0.00001$), respectively. For two-class classification (benign or malignant), the accuracy, sensitivity, and specificity were 91.5%, 83.3%, and 94.5% by FRCNN; 86.6%, 86.3%, and 86.6% by BCD; and 85.3%, 83.5%, and 85.9% by TRN, respectively. False positive rates and positive predictive values were 5.5% and 84.7% by FRCNN, 13.4% and 70.5% by BCD, and 14.1% and 68.5% by TRN, respectively. We compared the classification performance of FRCNN with 20 dermatologists. As a result, the classification accuracy of FRCNN was better than that of the dermatologists. In the future, we plan to implement this system in society and have it used by the general public, in order to improve the prognosis of skin cancer.

Keywords: melanoma; skin cancer; artificial intelligence (AI); deep learning; neural network

1. Introduction

Skin cancer is the most common malignancy in Western countries, and melanoma specifically accounts for the majority of skin cancer-related deaths worldwide [1]. In recent years, many skin cancer classification systems using deep learning have been developed for classifying images of skin tumors, including malignant melanoma (MM) and other skin cancer [2]. There are reports that their accuracy was at the same level as or higher than that of dermatologists [3–5].

The targeted detection range of previous reports was from only malignant melanoma to the entire skin cancer. Image data used for machine learning were clinical images and dermoscopic images. Up to now, there has been no report of training a neural network using clinical image data of pigmented skin lesions and evaluating the accuracy of the system to classify skin cancer, such as MM and basal cell carcinoma (BCC). When developing a system, it is important to determine the appropriate endpoints according to the type of skin tumor to be targeted, as well as the method of imaging. When new patients come to a medical institution with skin lesions as the chief complaint, they are generally concerned not about whether they are malignant melanomas, but whether they are skin cancers. Therefore, there is a need to develop a system that can also detect other skin tumors that have a pigmented appearance similar to malignant melanoma. There are also erythematosus skin malignancies, such as mycosis fungoides [6], extramammary Paget's disease [7], and actinic keratosis [8], which is a premalignant tumor of squamous cell carcinoma. It is often difficult to distinguish these cancers from eczema. Since we are focusing on the detection of brown to black pigmented skin lesions, including MM, we have excluded these cancers in this study.

In recent years, with the progress of machine learning technology mainly on deep learning, the expectations of artificial intelligence has been increasing, and research on its medical application has been actively progressing [9–12]. In the present study, we used the faster, region-based convolutional neural network (Faster R-CNN, or FRCNN) algorithm, which is a result of merging region proposal network (RPN) and Fast R-CNN algorithms, into a single network [13,14]. The pioneering work of region-based target detection began with the region-based convolutional neural network (R-CNN), including three modules: regional proposal, vector transformation, and classification [15,16]. Spatial pyramid pooling (SPP)-net optimized the R-CNN and improved detection performance [16,17]. Fast R-CNN combines the essence of SPP-net and R-CNN, and introduces a multi-task loss function, which is what makes the training and testing of the whole network so functional [16,18]. FRCNN merges RPN and Fast R-CNN into a unified network by sharing the convolutional features with "attention" mechanisms, which greatly improves both the time and accuracy of target detection [13,16]. Indeed, FRCNN has shown higher detection performance in the biomedical filed than other state-of-the-art methods, such as support vector machines (SVMs), visual geometry Group-16 (VGG-16), single shot multibox detectors (SSDs), and you only look once (YOLO), in terms of time and accuracy [19–21]. In particular, FRCNN has achieved the best performance for diabetic foot ulcer (DFU) detection; the purpose of the DFU study was similar to our research goal [21]. Therefore, we ultimately chose the FRCNN architecture in this study. Moreover, in the medical science field, transductive learning models have widely been used in addition to supervised learning models [22,23]. Meanwhile, given that diagnosis is a medical practice and requires authorized training data by medical doctors, we chose supervised learning in the present study.

Importantly, many mobile phone applications that can detect skin cancers have been developed and put on the market [24–26]. In those applications, skin cancer detection is performed using smartphone camera images rather than the magnified images of dermoscopy, which is commonly used by dermatologists in medical institutions. Our goal is to develop a skin cancer detection system that can be easily used by people who are concerned about the possibility that the skin lesion is cancers. Therefore, in this study, we developed a neural network-based classification system using clinical images rather than dermoscopic images. We evaluated the accuracy of the system and asked dermatologists to take the same test, in order to compare the accuracy with the deep learning system we developed.

2. Materials and Methods

2.1. Patients and Skin Images

This study was approved by the Ethics Committee of the National Cancer Center, Tokyo, Japan (approval ID: 2016-496). All methods were performed in accordance with the Ethical Guidelines for Medical and Health Research Involving Human Subjects; with regard to the handling of data, we followed the Data Handling Guidelines for the Medical AI project. Of approximately 120,000 clinical images taken from 2001 to 2017 at the Department Dermatologic Oncology in the National Cancer Center Hospital, we extracted 5846 clinical images of brown to black pigmented skin lesions from 3551 patients. The clinical images were taken by digital cameras and stored as digital images. Additionally, we confirmed that all images were of sufficient quality that dermatologists could diagnose (Supplementary Table S1). The target diseases are malignant tumors (MM and BCC) and benign tumors (nevus, seborrheic keratosis (SK), senile lentigo (SL) and hematoma/hemangioma (H/H)). The breakdown of the extracted images was 1611 MM images (from 710 patients), 401 BCC images (from 270 patients), 2837 nevus images (from 1839 patients), 746 SK images (from 555 patients), 79 SL images (from 65 patients), and 172 H/H images (from 147 patients). All malignant tumors were biopsied and diagnosed histopathologically. Benign tumors were diagnosed clinically using dermoscopy, and those cases that were still difficult to differentiate were biopsied to make confirmed diagnosis. All of the images were taken with digital, single-lens reflex cameras, which had at least 4.95 million pixels, a macro lens, and macro ring flash. No dermoscopic images were included in this study. Out of the 3551 patients, we randomly selected 666 patients, and picked one image per patient for the test dataset. The remaining 4732 images from 2885 patients were used for training. The breakdown of the 666 images of the test dataset was 136 MM images (from 136 patients), 44 BCC images (from 44 patients), 349 nevus images (from 349 patients), 96 SK images (from 96 patients), 15 SL images (from 15 patients), and 26 H/H images (from 26 patients). The breakdown of the 4732 images of the training dataset was 1279 MM images (from 566 patients), 344 BCC images (from 222 patients), 2302 nevus images (from 1474 patients), 606 SK images (from 451 patients), 62 SL images (from 51 patients), and 139 H/H images (from 121 patients). We gave bounding-box annotations (where and what class each lesion is) to all the images, and a dermatologist (S.J.) confirmed their validity.

To reduce each dermatologist's burden, we randomly sampled 200 images from 666 images and created tests of 10 patterns, so that each image was selected at least three times (200 images $\times$ 10 sets = 2000 images; 2000 $\div$ 666 patients = 3). Thus, each test consisted of 200 images. The whole flow diagram is shown in Figure 1.

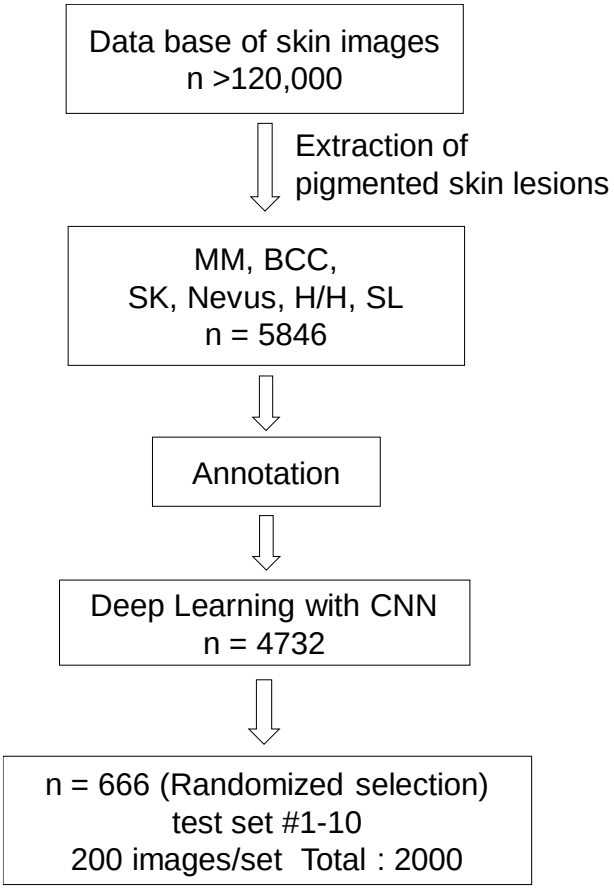

Figure 1. Flow diagram of this study: extracting the pictures of pigment lesions, annotation of lesions in images, deep learning with a convolutional neural network (CNN), and evaluation by the test dataset.

2.2. Training of a Deep Learning Model

With regard to the deep learning architecture, we placed the highest priority on accuracy and rapidity in choosing a model, because accurate and prompt classification is required in the medical field. As a result of various comparison, we finally selected the FRCNN; this model stably showed high classification accuracy, robustness, and rapidity [13,14,27–29]. Then, we trained an FRCNN model with the training dataset. We used Visual Geometry Group-16 (VGG-16) [30] as its backbone, and a Momentum stochastic gradient descent (SGD) [31] optimizer with learning rate of 1×10^{-3} and momentum of 0.9. We used weight decay of 5×10^{-4} and the batch size was 4. The model was trained for 100 epochs, and the learning rate was decreased by a factor of 10 after 40 and 80 epochs finished. Images of BCC, SL, and H/H were twice oversampled during training. Horizontal flip, random distort [32], 90 and 180 degree rotations, random cropping, and zoom were used for data augmentation. We used Chainer [33], ChainerCV [34], and Cupy [35] for the implementation of our network.

2.3. Test-Time Augmentation

During inference, we used test-time augmentation. Specifically, an input image underwent transformations of horizontal flip (two patterns); 72 degree rotations (five patterns); and 1×, 1.2×, or 1.4× zoom (three patterns), yielding 30 patterns of images in total. Predictions were made on all 30 images, and the predicted region with the highest confidence among all predictions was selected as the final prediction for that input image.

2.4. Model Validation and Verification

Our model (FRCNN), 10 board-certified dermatologists (BCDs), and 10 trainees (TRNs) were assessed using 10 patterns of tests, and we compared their performances. We compared the results in two patterns: a six-class classification (judge what class each sample is) and a two-class classification (judge whether each sample is benign or malignant). We calculated the accuracy for both six- and two-class classifications by the following formula: accuracy (%) = (total number of correct predictions)/(total number of all samples) × 100. For two-class classification, we also calculated sensitivity, specificity, false negative rates, false positive rates, and positive predictive values. The accuracy of two- and six-class classification was compared with the equivalent of each other using a paired *t*-test, and *p*-values < 0.05 were considered significant.

3. Results

3.1. Six-Class Classification of FRCNN, BCDs, and TRNs

The results (200 questions × 10 tests) of six-class classification of FRCNN, BCD, and TRN are shown in Table 1. The accuracy of the six-class classification of FRCNN was 86.2% (1724/2000), while those of BCD and TRN were 79.5% (1590/2000) and 75.1% (1502/2000), respectively. The accuracy of six-class classification of each examinee is shown in Table 2. Except for test #2, FRCNN had higher accuracy than the dermatologists. The standard deviation of the accuracy of six-class classification of FRCNN was 2.80%, and that of the dermatologists was 4.41%. The accuracy of six-class classification by FRCNN (86.2 ± 2.95%) was statistically higher than that of BCD (79.5 ± 5.27%, $p = 0.0081$) and TRN (75.1 ± 2.18%, $p < 0.00001$). The accuracy of six-class classification by BCD was not statistically higher than that of TRN ($p = 0.070$) (Figure 2).

Table 1. The results of six-class classification of the faster, region-based CNN (FRCNN); board-certified dermatologists (BCDs); and trainees (TRNs). Gray cells indicate correct answers.

FRCNN								
		Prediction						
		MM	BCC	Nevus	SK	H/H	SL	Total
True diagnosis	MM	327	9	48	21	0	3	408
	BCC	6	108	12	6	0	0	132
	Nevus	42	6	967	30	3	0	1048
	SK	21	9	36	223	0	0	289
	H/H	3	0	18	0	57	0	78
	SL	0	0	0	3	0	42	45
	Total	399	132	1081	283	60	45	2000
BCDs								
		Prediction						
		MM	BCC	Nevus	SK	H/H	SL	Total
True diagnosis	MM	340	12	22	26	3	5	408
	BCC	10	104	3	14	1	0	132
	Nevus	131	11	823	68	11	4	1048
	SK	18	24	17	225	0	5	289
	H/H	9	1	6	1	61	0	78
	SL	0	1	0	7	0	37	45
	Total	508	153	871	341	76	51	2000

Table 1. *Cont.*

		TRNs						
		Prediction						
		MM	BCC	Nevus	SK	H/H	SL	Total
True diagnosis	MM	327	15	42	12	8	4	408
	BCC	22	87	6	12	5	0	132
	Nevus	136	17	812	57	20	6	1048
	SK	26	17	37	191	1	17	289
	H/H	8	1	16	2	51	0	78
	SL	1	0	3	7	0	34	45
	Total	520	137	916	281	85	61	2000

MM: malignant melanoma; BCC: basal cell carcinoma; SK: seborrheic keratosis; H/H: hematoma/hemangioma; SL: senile lentigo.

Table 2. The accuracy of six-class classification for each examinee. The best accuracy for each test (test #1–10) is shown in gray.

TEST #	FRCNN	BCD	TRN
1	90.00%	84.00%	76.50%
2	82.50%	86.00%	72.00%
3	84.50%	83.50%	74.50%
4	90.00%	79.00%	74.50%
5	83.00%	78.00%	73.00%
6	86.50%	85.50%	75.00%
7	88.00%	70.50%	79.00%
8	86.50%	79.50%	75.00%
9	82.50%	73.50%	78.00%
10	88.50%	75.50%	73.50%

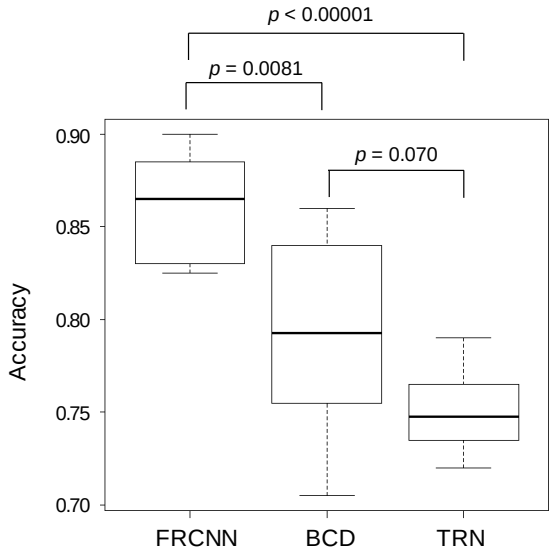

Figure 2. The accuracy of six-class classification by FRCNN, BCDs, and TRNs. In six-class classification, the accuracy of the FRCNN surpassed that of BCDs and TRNs.

3.2. Two-Class Classification of FRCNN, BCDs, and TRNs

The results of two-class classification (benign or malignant) of FRCNN, BCDs, and TRNs are shown in Table 3. Malignant tumors include MM and BCC, and benign tumor includes nevus, SK, SL, and H/H. The accuracy of two-class classification of the FRCNN was 91.5% (1829/2000), while those of BCDs and TRNs were 86.6% (1829/2000) and 85.3% (1705/2000), respectively. The accuracy of two-class classification by the FRCNN (91.5 ± 1.79%) was also statistically higher than that of BCDs (86.6 ± 4.01%, $p = 0.0083$) and TRNs (85.3 ± 2.18%, $p < 0.001$). The accuracy of two-class classification by BCD was not statistically higher than that of the TRNs ($p = 0.40$) (Figure 3).

Table 3. The results of two-class classification (benign or malignant) of the FRCNN, BCDs, and TRNs. Gray cells indicate correct answers.

FRCNN				
		Prediction		
		malignant	benign	Total
True diagnosis	malignant	450	90	540
	benign	81	1379	1460
	Total	531	1469	2000
BCDs				
		Prediction		
		malignant	benign	Total
True diagnosis	malignant	466	74	540
	benign	195	1265	1460
	Total	661	1339	2000
TRNs				
		Prediction		
		malignant	benign	Total
True diagnosis	malignant	451	89	540
	benign	206	1254	1460
	Total	657	1343	2000

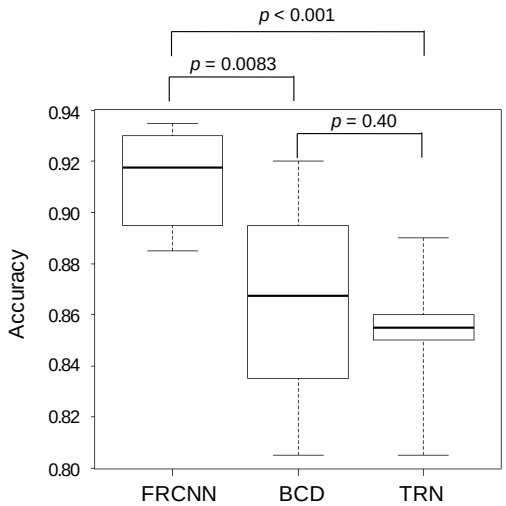

Figure 3. The accuracy of two-class classification (benign or malignant) by FRCNN, BCDs, and TRNs. The accuracy of the FRCNN surpassed that of the BCDs and TRNs.

3.3. Two-Class Classification of FRCNN, BCDs, and TRNs

The accuracy of six-class classification of each examiner is shown in Table 4. BCDs had the highest accuracy in test #2, and the BCDs and FRCNN had the same accuracy in test #6. In all the tests other than #2 and #6, FRCNN had the highest accuracy among all examiners. The standard deviation of the accuracy of two-class classifications of FRCNN was 1.69%, and those of BCDs and TRNs were 9.79% and 3.13%, respectively.

Table 4. The accuracy of two-class classification for each examinee. The best accuracy for each test (test #1–10) is shown in gray. The accuracy of the BCDs was the best in test #2. In test #6, the BCDs and FRCNN achieved the same accuracy.

TEST #	FRCNN	BCD	TRN
1	93.50%	89.50%	85.00%
2	88.50%	92.00%	86.00%
3	91.00%	89.00%	85.00%
4	93.50%	87.00%	80.50%
5	89.50%	84.50%	85.50%
6	91.50%	91.50%	85.50%
7	92.50%	83.50%	89.00%
8	92.00%	86.50%	86.50%
9	89.50%	81.50%	86.00%
10	93.00%	80.50%	83.50%

3.4. Summary of Classification Conducted by FRCNN, BCDs, and TRNs

We show the summary of the classification accuracy, sensitivity, specificity, false negative rates, false positive rates, and positive predictive values by FRCNN, BCDs, and TRNs in Table 5. FRCNN achieved highest accuracy and sensitivity. On the other hand, BCDs achieved the highest specificity. The false negative rates of all of them are almost the same, but the false positive rates of the dermatologists (BCDs: 13.4%; TRNs: 14.1%) were higher than that of the FRCNN (5.5%). The false positive rates of the dermatologists were higher than that of the FRCNN, and the positive predictive values of them were lower (BCDs: 70.5%, TRNs: 68.5%) than that of the FRCNN (84.7%).

Table 5. Summary of classification accuracy, sensitivity, specificity, false negative rates, false positive rates, and positive predictive values by the FRCNN, BCDs, and TRNs.

	FRCNN	BCDs	TRNs
Accuracy (six classes)	86.2	79.5	75.1
Accuracy (two classes)	91.5	86.6	85.3
Sensitivity	83.3	86.3	83.5
Specificity	94.5	86.6	85.9
False negative	16.7	13.7	16.5
False positive	5.5	13.4	14.1
Positive predictive value	84.7	70.5	68.5

4. Discussion

In this study, we developed a classification system by deep learning for brown to black pigmented skin lesions, as the target disease. Then, the same test dataset was used for examining 20 dermatologists, and the accuracy of them was compared with that of the FRCNN. The results showed that only one out of 20 dermatologists had higher accuracy than the FRCNN in six-class classification. The skin

tumor classification system using deep learning showed better results in both six- and two-class classification accuracy than BCDs and TRN dermatologists. Many similar tests have been reported in previous research [3,36,37], and it is considered that the machine learning algorithm has reached dermatologist-level accuracy in skin lesion classification [4,5,36]. In the present study, although the FRCNN and the dermatologists had similar results in terms of sensitivity, false positive rates were BCDs: 13.4%, TRNs: 14.1%, and FRCNN: 5.5%. It is likely that when the dermatologists were uncertain whether skin lesions were malignant or benign, they might tend to diagnose them as malignant. The dermatologists had higher false positive rates, and the positive predictive values were 70.5% by the BCDs and 68.5% by the TRNs, and lower than 84.7% by the FRCNN. False negative rates have been regarded as more important than false positive rates in such diagnostic systems for malignancy, but false positive rates must be carefully monitored. This is because false positive predictions give users unwanted anxiety. In addition, although the results of the dermatologists varied, the results of the FRCNN showed less variation. Brinker et al. reported that CNNs indicated a higher robustness of computer vision compared to human assessment for clinical image classification tasks [3]. This is due to the lack of concentration during work, which is unique to humans. It is considered that there may be differences in clinical ability depending on the years of experience of dermatologists.

We think that it is important to determine how to implement these results socially after system development and connect them to users' benefit. Depending on the concept of system development, the endpoint and the type of image data required for the development will change. For example, if the person who uses the system is a doctor, highly accurate system development closer to a confirmed diagnosis will be required. Training neural networks that can detect cancers from dermoscopic images will be also in need. However, for in-hospital use there is already a diagnostic method: biopsy. Biopsy is a method of taking a part of skin tissue and making a pathological diagnosis. Through a biopsy, it is possible to make an almost 100% diagnosis (confirmed diagnosis). Moreover, the procedure of biopsy takes only about 10 min. It is an advantage of dermatologists to be able to perform biopsy more easily than other department doctors, and it seems that there is no room for new diagnostic functions of any diagnostic imaging systems in medical institutions. On the other hand, when considering their use by the general public outside medical institutions, it is difficult to fully demonstrate their diagnostic performance. This is because the reproducibility of shooting conditions cannot be ensured, and the shooting equipment is different. Therefore, when using an imaging system outside medical institutions, it may be better to use the system to call attention to skin cancer rather than focus on improving diagnostic performance. Also, no one can say that the accuracy of the system needs to be improved when it is used outside the medical institution.

Mobile phone applications that can detect skin cancer and malignant melanoma have already been launched in countries around the world [24]. However, usage of such applications for the self-assessment of skin cancer has been problematic, due to the lack of evidence on the applications' diagnostic accuracy [38,39]. In addition to the problem of low accuracy, there is also a problem that they sometimes cannot recognize images well [25]. The reason is that the quality of images may be lower, and that there is more variability in terms of angles, distances, and the characteristics of the smartphone [40]. If the shooting conditions are bad, the accuracy is naturally low. This is an unavoidable task in terms of social implementation, in which the users are general public and the device used is a mobile phone camera. The main risk associated with the usage of mobile phone application software by general public is that malignant tumor may be incorrectly classified as low-risk, and its diagnosis and appropriate treatment are delayed. To solve these problems. and to improve the accuracy of the application over time, a large dataset is necessary to cover as many image-taking scenarios, as well as other information (i.e., ages, position of the primary lesion, the period time from first awareness to visit a dermatologist, etc.) as possible. However, it takes a lot of effort to create such a dataset. Udrea et al. have succeeded in improving accuracy by changing the learning method and training, with a large number of skin tumor images taken with a mobile phone [40]. We must be careful to make users fully aware that mobile phone application software is a system that also has the negative

aspects. In fact, SkinVision, an application for detecting skin cancers, also states that "assessment does not intend to provide an official medical diagnosis, nor replace visits to a doctor [40]."

We are also planning a future social implementation system of skin cancer classification to be used by the general public, with wearable devices, such as mobile phones. The original concept is to have early skin cancer detection, early treatment, and improved prognosis of skin cancer patients. In Japan, the incidence of skin cancer is lower than in Western countries, and its awareness is also low. The proportion of advanced stage cases of melanoma is higher than in Europe and the United States [41,42]. As a result, many patients tend to have poor outcomes. In recent years, the prognosis of melanoma has been improved by new drugs, such as immune checkpoint inhibitors and molecular-targeted therapy [43], but at the same time, the problem of rising medical costs has arisen [44]. In Japan, there is no official skin cancer screening, and there is no intervention that can be performed early for the entire Japanese population. Additionally, since melanoma is one of the rarer skin cancers for Japanese people, it is not well-recognized, and people tend not to see a dermatologist at the early stages [43]. The average period from first awareness to visit of Japanese melanoma patients was 69.5 months; the median was 24 months. In other countries, the median period is reported to be 2 months to 9.8 months, which is very different from the reports in Japan [45–48]. The rate of late-stage is high, due to the longer period from first awareness to visit. Because the stage of disease at the first visit is highly related to the prognosis of skin cancer [49], early detection of skin cancer is very important. If skin cancer is detected at an early stage, it will be easier to treat, and the prognosis will be much better [50]. We think that an intervention that shortens the period from awareness to visit is essential for improving the skin cancer prognosis. Some mobile phone application software that is on the market may have diagnosed skin cancers that were not diagnosed as skin cancer by dermatologists, which helps in the early detection and treatment of skin cancer [38]. In the future, we think that the intervention of skin image diagnostic application software, as described above, can solve various problems, such as improving the prognosis of skin cancer and reducing the treatment costs. Also, by reducing the waiting time for patients and unnecessary visits to outpatient clinics, and facilitating consultations, medical treatment will be efficient [40]. It would be of great value if such an image diagnosis system actually improved the prognosis after social implementation. Such application software has not appeared yet, and we hope we can create such an application in the future.

There are several limitations to this study. First, although all malignant tumors were biopsied and diagnosed histopathologically, benign tumors were confirmed as benign using biopsy, or for those not excised were deemed clinically benign. Second, the neural network was trained using clinical images of brown to black pigmented skin lesions from only our institution, and biases may exist in those data (e.g., portion of disease, type of camera). It will be necessary for future work to check whether the neural network generalizes well with images taken outside our institution. Third, in the present study, we showed only the ability of judging clinical images, but in routine medical care, human medical doctors make a definitive diagnosis by taking biopsies and other clinical information into consideration. Therefore, it is risky to judge that artificial intelligence (AI) is superior to human medical doctors based on this study. Further validation is essential; we need to make a careful judgment on how to implement our findings in society. In addition, this is only the first step, and there is no doubt that large-scale verification will be required as the next step, according to the suitable social implementation method. Lastly, although we used the FRCNN architecture in the present study, we need to carefully choose the best method for achieving our goal, because deep learning technologies have recently been progressing massively [51]. In particular, FRCNN has been reported to have difficulty identifying objects from low-resolution images, due to its weak capacity to identify local texture [52]. We plan to improve the algorithm appropriately, according to the direction of our social implementation.

5. Conclusions

We have developed a skin cancer classification system for brown to black pigmented skin lesions using deep learning. The accuracy of the system was better than that of dermatologists. It successfully

detected not only malignant melanoma, but also basal cell carcinoma. System development that fits the needs of society is important. We would like to seek the best method for the early detection of skin cancer and improvement of the prognosis.

Supplementary Materials: The following are available online at http://www.mdpi.com/2218-273X/10/8/1123/s1, Table S1: Information of the digital cameras, which took the skin images in this study.

Author Contributions: Conceptualization, S.J., and N.Y.; methodology, S.J., Y.H. and Y.S.; software, Y.H. and Y.S.; validation, S.J., Y.H. and Y.S.; data curation, S.J., Y.H. and Y.S.; writing—review and editing, S.J., Y.H., Y.S., and R.H.; supervision, R.H. and Y.O.; project administration, R.H.; funding acquisition, R.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Japan Science and Technology Agency (JST) Core Research for Evolutional Science and Technology (CREST) (Grant Number JPMJCR1689), and a Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research on Innovative Areas (Grant Number JP18H04908).

Acknowledgments: We would like to thank Kazuma Kobayashi for his helpful feedback. This work was partly supported by National Cancer Center Research and Development Fund (29-A-3).

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

References

1. Schadendorf, D.; van Akkooi, A.C.J.; Berking, C.; Griewank, K.G.; Gutzmer, R.; Hauschild, A.; Stang, A.; Roesch, A.; Ugurel, S. Melanoma. *Lancet* **2018**, *392*, 971–984. [CrossRef]
2. Nasr-Esfahani, E.; Samavi, S.; Karimi, N.; Soroushmehr, S.M.; Jafari, M.H.; Ward, K.; Najarian, K. Melanoma detection by analysis of clinical images using convolutional neural network. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2016**, *2016*, 1373–1376. [PubMed]
3. Brinker, T.J.; Hekler, A.; Enk, A.H.; Klode, J.; Hauschild, A.; Berking, C.; Schilling, B.; Haferkamp, S.; Schadendorf, D.; Frohling, S.; et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *Eur. J. Cancer* **2019**, *111*, 148–154. [CrossRef]
4. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
5. Fujisawa, Y.; Otomo, Y.; Ogata, Y.; Nakamura, Y.; Fujita, R.; Ishitsuka, Y.; Watanabe, R.; Okiyama, N.; Ohara, K.; Fujimoto, M. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *Br. J. Dermatol.* **2018**. [CrossRef]
6. Bilgic, S.A.; Cicek, D.; Demir, B. Dermoscopy in differential diagnosis of inflammatory dermatoses and mycosis fungoides. *Int. J. Dermatol.* **2020**, *59*, 843–850. [CrossRef]
7. Morris, C.R.; Hurst, E.A. Extramammary Paget's Disease: A Review of the Literature Part II: Treatment and Prognosis. *Derm. Surg.* **2020**, *46*, 305–311. [CrossRef]
8. Marques, E.; Chen, T.M. *Actinic Keratosis*; StatPearls Publishing: Treasure Island, FL, USA, 2020.
9. Asada, K.; Kobayashi, K.; Joutard, S.; Tubaki, M.; Takahashi, S.; Takasawa, K.; Komatsu, M.; Kaneko, S.; Sese, J.; Hamamoto, R. Uncovering Prognosis-Related Genes and Pathways by Multi-Omics Analysis in Lung Cancer. *Biomolecules* **2020**, *10*, 524. [CrossRef]
10. Hamamoto, R.; Komatsu, M.; Takasawa, K.; Asada, K.; Kaneko, S. Epigenetics Analysis and Integrated Analysis of Multiomics Data, Including Epigenetic Data, Using Artificial Intelligence in the Era of Precision Medicine. *Biomolecules* **2020**, *10*, 62. [CrossRef]
11. Yamada, M.; Saito, Y.; Imaoka, H.; Saiko, M.; Yamada, S.; Kondo, H.; Takamaru, H.; Sakamoto, T.; Sese, J.; Kuchiba, A.; et al. Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Sci. Rep.* **2019**, *9*, 14465. [CrossRef]
12. Yasutomi, S.; Arakaki, T.; Hamamoto, R. Shadow Detection for Ultrasound Images Using Unlabeled Data and Synthetic Shadows. *arXiv* **2019**, arXiv:1908.01439.
13. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
14. Liu, T.; Stathaki, T. Faster R-CNN for Robust Pedestrian Detection Using Semantic Segmentation Network. *Front. Neurorobot.* **2018**, *12*, 64. [CrossRef] [PubMed]

15. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 24–27.
16. Shao, F.; Wang, X.; Meng, F.; Zhu, J.; Wang, D.; Dai, J. Improved Faster R-CNN Traffic Sign Detection Based on a Second Region of Interest and Highly Possible Regions Proposal Network. *Sensors* **2019**, *19*, 2288. [CrossRef] [PubMed]
17. Huang, Y.Q.; Zheng, J.C.; Sun, S.D.; Yang, C.F.; Liu, J. Optimized YOLOv3 Algorithm and Its Application inTraffic Flow Detections. *Appl. Sci.* **2020**, *10*, 3079. [CrossRef]
18. Xiao, M.; Xiao, N.; Zeng, M.; Yuan, Q. Optimized Convolutional Neural Network-Based Object Recognition for Humanoid Robot. *J. Robot. Autom.* **2020**, *4*, 122–130.
19. Kuok, C.P.; Horng, M.H.; Liao, Y.M.; Chow, N.H.; Sun, Y.N. An effective and accurate identification system of Mycobacterium tuberculosis using convolution neural networks. *Microsc. Res. Tech.* **2019**, *82*, 709–719. [CrossRef]
20. Rosati, R.; Romeo, L.; Silvestri, S.; Marcheggiani, F.; Tiano, L.; Frontoni, E. Faster R-CNN approach for detection and quantification of DNA damage in comet assay images. *Comput. Biol. Med.* **2020**, *123*, 103912. [CrossRef]
21. Goyal, M.; Reeves, N.D.; Rajbhandari, S.; Yap, M.H. Robust Methods for Real-Time Diabetic Foot Ulcer Detection and Localization on Mobile Devices. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 1730–1741. [CrossRef]
22. Lee, N.; Caban, J.; Ebadollahi, S.; Laine, A. Interactive segmentation in multimodal med-ical imagery using a bayesian transductive learning approach. *Med. Imaging 2009 Comput.-Aided Diagn.* **2009**, *7260*, 72601W.
23. Wan, S.; Mak, M.W.; Kung, S.Y. Transductive Learning for Multi-Label Protein Subchloroplast Localization Prediction. *IEEE/Acm Trans. Comput. Biol. Bioinform.* **2017**, *14*, 212–224. [CrossRef]
24. Buechi, R.; Faes, L.; Bachmann, L.M.; Thiel, M.A.; Bodmer, N.S.; Schmid, M.K.; Job, O.; Lienhard, K.R. Evidence assessing the diagnostic performance of medical smartphone apps: A systematic review and exploratory meta-analysis. *Bmj Open* **2017**, *7*, e018280. [CrossRef]
25. Ngoo, A.; Finnane, A.; McMeniman, E.; Tan, J.M.; Janda, M.; Soyer, H.P. Efficacy of smartphone applications in high-risk pigmented lesions. *Australas J. Dermatol.* **2018**, *59*, e175–e182. [CrossRef]
26. Singh, N.; Gupta, S.K. Recent advancement in the early detection of melanoma using computerized tools: An image analysis perspective. *Ski. Res. Technol.* **2018**. [CrossRef]
27. Wu, W.; Yin, Y.; Wang, X.; Xu, D. Face Detection With Different Scales Based on Faster R-CNN. *IEEE Trans. Cybern* **2019**, *49*, 4017–4028. [CrossRef]
28. Li, H.; Weng, J.; Shi, Y.; Gu, W.; Mao, Y.; Wang, Y.; Liu, W.; Zhang, J. An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. *Sci. Rep.* **2018**, *8*, 6600. [CrossRef]
29. Ji, Y.; Zhang, S.; Xiao, W. Electrocardiogram Classification Based on Faster Regions with Convolutional Neural Network. *Sensors* **2019**, *19*, 2558. [CrossRef]
30. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
31. Qian, N. On the momentum term in gradient descent learning algorithms. *Neural Netw.* **1999**, *12*, 145–151. [CrossRef]
32. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
33. Tokui, S.; Oono, K.; Hido, S.; Clayton, J. Chainer: A next-generation open source framework for deep learning. In Proceedings of the Workshop on Machine Learning Systems (LearningSys) in the Twenty-Ninth Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7 December 2015.
34. Niitani, Y.; Ogawa, T.; Saito, S.; Saito, M. ChainerCV: A library for deep learning in computer vision. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1217–1220.
35. Nishino, R.; Loomis, S.H.C. CuPy: A numpy-compatible library for nvidia gpu calculations. In Proceedings of the Workshop on Machine Learning Systems (LearningSys) in the Thirty-first Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4 December 2017.

36. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A.; et al. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842. [CrossRef]
37. Cui, X.; Wei, R.; Gong, L.; Qi, R.; Zhao, Z.; Chen, H.; Song, K.; Abdulrahman, A.A.A.; Wang, Y.; Chen, J.Z.S.; et al. Assessing the effectiveness of artificial intelligence methods for melanoma: A retrospective review. *J. Am. Acad. Dermatol.* **2019**, *81*, 1176–1180. [CrossRef] [PubMed]
38. Kassianos, A.P.; Emery, J.D.; Murchie, P.; Walter, F.M. Smartphone applications for melanoma detection by community, patient and generalist clinician users: A review. *Br. J. Dermatol.* **2015**, *172*, 1507–1518. [CrossRef] [PubMed]
39. Wolf, J.A.; Moreau, J.F.; Akilov, O.; Patton, T.; English, J.C., 3rd; Ho, J.; Ferris, L.K. Diagnostic inaccuracy of smartphone applications for melanoma detection. *Jama Dermatol.* **2013**, *149*, 422–426. [CrossRef] [PubMed]
40. Udrea, A.; Mitra, G.D.; Costea, D.; Noels, E.C.; Wakkee, M.; Siegel, D.M.; de Carvalho, T.M.; Nijsten, T.E.C. Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *J. Eur. Acad. Dermatol. Venereol.* **2019**. [CrossRef] [PubMed]
41. Fujisawa, Y.; Yoshikawa, S.; Minagawa, A.; Takenouchi, T.; Yokota, K.; Uchi, H.; Noma, N.; Nakamura, Y.; Asai, J.; Kato, J.; et al. Classification of 3097 patients from the Japanese melanoma study database using the American joint committee on cancer eighth edition cancer staging system. *J. Dermatol. Sci.* **2019**, *94*, 284–289. [CrossRef]
42. American Cancer Society. Cancer Facts & Figures 2020. *Am. Cancer Soc. J.* **2020**. Available online: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf (accessed on 1 June 2020).
43. Fujisawa, Y.; Yoshikawa, S.; Minagawa, A.; Takenouchi, T.; Yokota, K.; Uchi, H.; Noma, N.; Nakamura, Y.; Asai, J.; Kato, J.; et al. Clinical and histopathological characteristics and survival analysis of 4594 Japanese patients with melanoma. *Cancer Med.* **2019**, *8*, 2146–2156. [CrossRef]
44. Gorry, C.; McCullagh, L.; Barry, M. Economic Evaluation of Systemic Treatments for Advanced Melanoma: A Systematic Review. *Value Health* **2020**, *23*, 52–60. [CrossRef]
45. Krige, J.E.; Isaacs, S.; Hudson, D.A.; King, H.S.; Strover, R.M.; Johnson, C.A. Delay in the diagnosis of cutaneous malignant melanoma. A prospective study in 250 patients. *Cancer* **1991**, *68*, 2064–2068. [CrossRef]
46. Richard, M.A.; Grob, J.J.; Avril, M.F.; Delaunay, M.; Gouvernet, J.; Wolkenstein, P.; Souteyrand, P.; Dreno, B.; Bonerandi, J.J.; Dalac, S.; et al. Delays in diagnosis and melanoma prognosis (I): The role of patients. *Int. J. Cancer* **2000**, *89*, 271–279. [CrossRef]
47. Tyler, I.; Rivers, J.K.; Shoveller, J.A.; Blum, A. Melanoma detection in British Columbia, Canada. *J. Am. Acad. Dermatol.* **2005**, *52*, 48–54. [CrossRef] [PubMed]
48. Fujisawa, Y. Japanese Melanoma Study: Annual Report 2017. *Jpn. Ski. Cancer Soc.* **2017**. Available online: http://www.skincancer.jp/report-skincancer_melanoma_2017.pdf (accessed on 31 December 2017).
49. Forbes, L.J.; Warburton, F.; Richards, M.A.; Ramirez, A.J. Risk factors for delay in symptomatic presentation: A survey of cancer patients. *Br. J. Cancer* **2014**, *111*, 581–588. [CrossRef]
50. Melanoma of the Skin 2019. In *Cancer Stat Facts*; National Cancer Institute: Bethesda, MD, USA, 2019.
51. Liu, W.; Wang, Z.; Liu, X.; Zeng, N.; Liu, Y.; Alsaadi, F.E. A survey of deep neural network architectures and their applications. *Neurocomputing* **2017**, *234*, 11–26. [CrossRef]
52. Cao, C.; Wang, B.; Zhang, W.; Zeng, X.; Yan, X.; Feng, Z.; Liu, Y.; Wu, Z. An improved faster R-CNN for small object detection. *IEEE Access* **2019**, *7*, 106838–106846. [CrossRef]