2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT) | 978-1-6654-1296-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/AINIT54228.2021.00038

CNN-based Diagnosis System on Skin Cancer using Ensemble Method Weighted by Cubic Precision

Pengcheng Jiang[1, *]
[1]*Waseda University*
Tokyo, Japan
[*]rexjiang@fuji.waseda.jp

***Abstract*— One of the most prevalent diseases, skin cancer, has been proven to be treatable at an early stage. Thus, techniques that allow individuals to identify skin cancer symptoms early are in great demand. This paper proposed an interactive skin lesion diagnosis system based on the ensemble of multiple sophisticated CNN models for image classification. The performance of ResNet50, ResNeXt50, ResNeXt101, EfficientNetB4, Mobile-NetV2, MobileNetV3, and MnasNet are investigated separately as ensemble components. Then, using various criteria, we constructed ensembles and compared the accuracy they achieved. Moreover, we designed a method to update the ensemble for new data and examined its performance. In addition, a few natural language processing (NLP) techniques were used to make our system more user-friendly. To integrate all the functionalities, we built a user interface with PyQt5. As a result, MobileNetV3 achieved 91.02% as the best accuracy among all single models; ensemble weighted by cubic precision achieved 92.84% accuracy as the highest one in this study; a notable improvement in accuracy demonstrated the effectiveness of the model updating approach, and a system with all of the desired features was successfully developed. These findings benefit in two aspects. For model performance, applying cubic precisions can increase ensemble learning classification accuracy. For the developed diagnosis system, it can aid in the early detection of the skin lesions for skin cancer treatment.**

Keywords- CNN, diagnosis system, medical image, ensemble learning, skin cancer, image classification, model update

I. INTRODUCTION

From the investigation done by the Skin Cancer Foundation, skin cancer is the most common cancer on the global scale where there are 9500 cases of it diagnosed every day, and an average of 2 people die from it every single hour in the United States [1]. Also, researchers show that melanoma is one of the most dangerous types of skin cancer for its extremely rapid spreading speed in the late stage, and the average age who gets a diagnosis with it is 65 [1-3]. American Academy of Dermatology researched the relationship between the survival rate of melanoma and gender and found that males are about twice likely to die by melanoma development than females at any age [4]. In addition to the protection from sun exposure, a regular examination of skin condition is also considered important since that melanoma is found to be greatly curable at early stages. As studied by Melanoma Research Alliance, the five-year survival rate for melanoma diagnosed at stage 0-2, stage 3, and stage 4 are 98.4%, 63.6%, and 22.5%, respectively [5]. Moreover, Jerant et al. conducted a more general study on the meaning of skin cancer screening. They provided evidence showing that early detection could largely increase the survival rate of melanoma [6]. Thus, early identification of suspicious lesions to skin cancer is significant regarding everyone's healthcare, which could be guaranteed by frequent skin screening. There are some existing applications developed for this purpose. For example, Michigan Medicine developed a mobile application named UMSkinCheck, which allows users to carry out full-body skin cancer screening and provides the support of tracking changes of skin lesions and detecting some that may lead to cancer [7]. SkinVision is another popular application for diagnosing skin cancer, which gives a notable high identification accuracy with the aid of machine learning techniques [8].

Since skin cancer screening is a visual exam of the skin, image analysis with either artificial or computer-aided methods is needed. As a computer-aided technique, machine learning is a powerful tool to extract features from images and analyze them for different tasks such as segmentation and classification. In the past decade, among plenty of machine learning algorithms, convolutional neural network (CNN) is found as one of the most effective and widely used algorithms for image processing [9]. As a branch, medical image analysis also achieved its success with CNNs. For instance, Moeskops et al. trained a single CNN to perform multiple segmentation tasks on brain MRI, breast MRI, and cardiac CTA, which achieved considerably low confusions between tasks [10]. Oh et al. proposed a patch-based CNN with limited data of chest X-ray for the diagnosis of COVID-19, which achieved 88.9% accuracy and 96.7% specificity that outperforms 70.7% and 89.7% obtained by the global approach [11]. Pham et al. used a deep CNN model for skin lesion classification, emphasizing the importance of data augmentation, and they achieved 89.0% accuracy as state of the art at the time [12]. CNN would not have such high performance and be so popular without the development of its architecture. Nowadays, most state-of-the-art CNN algorithms (e.g., ResNet, ResNeXt, EfficientNet, MobileNet, MnasNet) have abounding layers, a common feature of deep learning [13-17].

The remainder of this paper is laid out as follows. Section II provides a literature review of the medical diagnosis system, CNN, and ensemble learning. Section III illustrates our methodologies, including data pre-processing and system construction. Section IV presents our experimental results. Section V steps up to the discussion of concerns and future works. Finally, Section VI summarizes this study.

978-1-6654-1296-4/21/$31.00 ©2021 IEEE
DOI 10.1109/AINIT54228.2021.00038

II. Literature Review

Nowadays, various medical diagnosis systems are developed and widely used by people to get the initial detections of certain diseases [18]. Natural language processing (NLP) is one of the most powerful tools to develop such systems. For example, Doan built an NLP system to assist the quick diagnosis of Kawasaki disease so that the potential cardiac complication caused by the delayed diagnosis could be avoided [19]. Parthasarathy also developed an NLP-based tool to extract the electronic medical record data to diagnose serrated polyposis syndrome [20]. Besides the text information, the human voice is also found an analyzable data source based on which the disease detector could be built through audio-based machine learning. For instance, the University of Cambridge recently proposed an automatic approach to diagnose COVID-19 by training their model with a crowd-sourced dataset that includes the cough sounds from both infected and healthy people, which shows a notable accuracy for classification [21]. Although mentioned last, images would be the most common data type and are pervasively used to implement diagnosis systems where CNN plays an essential role in classifying diseases. For example, Liang et al. developed their own CNN to handle the classification task for malaria diagnosis [22]; Khan et al. introduced an approach combining feature extraction by deep CNNs and feature selection based on kurtosis controlled principal component to developing an automatic diagnosis system for skin lesion classification [23]. Alakwaa et al. implemented 3D-CNN with U-Net nodule detection to the lung cancer diagnosis [24].

An end-to-end learning method – CNN, extracts features from data without manual operations, which are necessary to many traditional extraction methods [25, 26]. In recent years, CNNs have had remarkable achievements in numerous applications such as NLP, audio analysis, image classification, and many others due to their convenience to implement and strong performance. Yih et al. developed a CNN-based framework for semantic parsing of single-relation questions, which achieved a 0.61 F1-score that outperforms its previous work, which achieved 0.54 [27]. Alayba proposed a model combining CNNs, Long Short-Term Memory (LSTM) networks and other NLP methods for sentiment analysis, and achieved an advantageous result of classification accuracy compared with using traditional learning algorithms such as SVM, Naïve Bayes or Logistic Regression [28]. Hershey et al. explored the ability of CNNs in audio classification with the data set consisting of 100 million YouTube videos and corresponding labels. Their notable results revealed that the state-of-the-art CNNs have more excellent performance on classifying audio than fully connected networks [29]. Image classification is one of the most well-known applications that benefited from the strengths of CNNs. For instance, Parkhi et al. verified the validity and the effectiveness of deep CNNs assisted with their embedding learning algorithm for face recognition. They achieved the validation accuracy of 97.3% on a dataset containing 2.6 million images of 1,595 people collected from YouTube [30]. Yu et al. proposed an efficient CNN architecture for hyperstructral image classification that could outperform other state-of-the-art CNNs [31]. In January 2021, Loey et al. proposed a hybrid transfer learning model to detect the face mask where the model is composed by CNNs for feature extraction and Support Vector Machine (SVM) for classification [32]. Impressively, they achieved 99.64%, 99.49%, and 100% accuracy for three different datasets, respectively. Therefore, CNN is currently one of the optimal choices for classification tasks.

Ensemble learning is an approach that includes multiple base learners in an ensemble which typically has a stronger generalization ability than any of these learners. Specifically, for classification tasks, we make the ensemble with several classifiers. The prediction result for a single input would be based on combining results from all the classifiers. Majority voting and weighted voting are two basic types of such combination. Majority voting means that we give the same right to each model in the ensemble, and the prediction receiving more than half of the votes from the candidates would be returned as the final prediction. In contrast, weighted voting means that different weights or rights are attached to different models so that the models with higher weight could affect more on the determination of prediction [33]. Kannojia proposed their hybrid classification model combining CNN and Extreme Learning Machine as an image classifier [34]. By applying ensemble learning with the majority voting rule, the accuracy of their model was increased to 99.33% compared with 99.24% by a single model and benchmarked on the MNIST dataset. Zhang et al. proposed their method with ensemble learning for the prediction of drug side effects where they applied the weighted voting rule to obtain the final prediction from multiple base predictors [35]. They proved that ensemble learning is effective in improving precision by carrying out experiments on three different datasets of drugs. More recently, Ali et al. developed a smart system based on boosting to predict heart disease where boosting is a type of weighted voting. They achieved 98.5% accuracy as a benchmark, indicating its better performance than other state-of-the-art methods [36]. Ensemble learning thus could be seen as an effective strategy to achieve stronger performance.

III. Methodology

A. Data Sources.

Two skin lesion datasets are used for the implementation of the ensemble in this paper. The first dataset is HAM10000, collected by Tschandl et al. through dermatoscopy and released for various machine learning purposes [37]. It consists of seven categories of skin lesions for 10015 images, which are ① actinic keratoses (AK), ② basal cell carcinoma

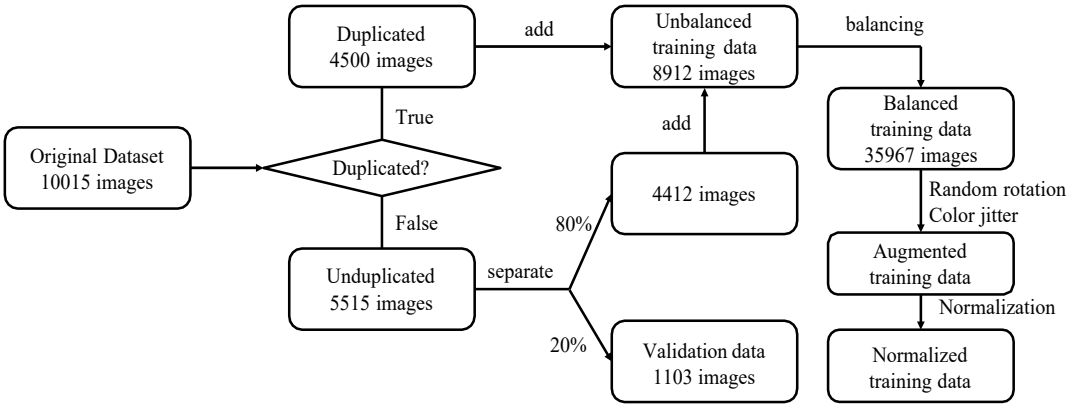


Fig. 1. Data pre-processing (HAM10000 as example)

(BCC), ③ benign keratosis-like lesions (BKL), ④dermato-fibroma (DF), ⑤ melanocytic nevi (NV), ⑥ vascular lesions (VSC) and ⑦ melanoma (MEL) for 327, 514, 1099, 115, 6705, 142, 1113 images, respectively. Another dataset we used for updating the ensemble is BCN20000, which consists of 19424 dermatoscopic images of skin lesions collected by Combalia et al. for the purpose of training classification models [38]. This dataset includes all categories in HAM10000 and has two extra ones: squamos cell carcinoma (SCC) and Unknown (UNK). It has 734, 2806, 1135, 121, 3995, 108, 2854 images correspondingly for ① to ⑦.

B. Data pre-processing.

In our study, we did similar data pre-processing to both datasets. Fig. 1 presents the process of data processing for HAM10000. Firstly, we checked the duplication of the images in the dataset with the information provided in the attached metadata. Next, we separated out 20% of the unduplicated data as the validation data and combined the remaining images as the raw training data without being processed. Since the dataset is not class balanced, we balanced the images for all categories by duplicating each of them by a certain rate. The balanced training data has 4455, 4790, 5055, 5350, 5822, 5160, 5335 images respectively to the categories mentioned above ① to ⑦. Then, we do the data augmentation to the balanced training data by random rotation with 20 degrees and color jitter with (brightness=0.1, contrast=0.1, hue=0.1). Finally, we normalized the data with computed mean and standardization to make it ready for training. The only difference in data processing for BCN20000 was that we removed the images of SCC and UNK to ensure the class consistency with HAM10000 and made the remained images as the original dataset to begin the processing.

C. CNN models.

Seven high performance CNN models were implemented in this paper, which are: ResNet50, ResNeXt50(32x4d), ResNeXt101(32x8d), EfficientNet-B4, MobileNetV2, MobileNetV3-Large and MnasNet. ResNet tackled the gradient vanishing problem by introducing the shortcut connections between layers that enable it to skip one or more layers so that it could at least stack a duplicated layer on the shallow layers by identity mapping to get better performance when gradient vanishing occurs [14]. ResNeXt was presented as an extension to ResNet with an approach called aggregated transformations, which can enhance accuracy without increasing parameter complexity while also reducing the number of hyper-parameters. It further reduces the computational overhead for network design [15]. EfficientNet implements a novel scaling method named compound model scaling that applies a fixed scaling coefficient set to scale each dimension uniformly and was demonstrated to outperform most state-of-the-art CNN models with this method [16]. MobileNet is a lightweight and efficient CNN model with a strong performance of accuracy, implementing depthwise separable convolution [17]. MobileNetV2 replaces the nonlinear activation (ReLU) with the linear bottleneck to avoid information loss in low-dimensional subspaces. It uses inverted residuals to balance the feature extraction by high-dimensional tensor and the calculation by low-dimensional tensor [39]. MnasNet refers to the inverted residual block in MobileNetV2, applies NAS to search a deep CNN with a novel objective function to optimize accuracy and latency and allow diversity between layers [18]. Based on MnasNet and MobileNetV2, MobileNetV3 extra introduces a new activation function, "h-swish", and a lightweight attention model based on squeeze and excitation, helping it outperform most state-of-the-art models [40].

D. Training.

We used *Pytorch* to do the training. A fully connected layer with 2048 input features and 7 output features was connected to the end of all CNN models for the training. With the pretrained models' parameters, we set the optimizer as Adam by a learning rate of 0.002. The criterion was set as cross entropy loss. To load the training data and validation data, we used the "DataLoader" function with batch size 16. Each model was trained by 20 epochs.

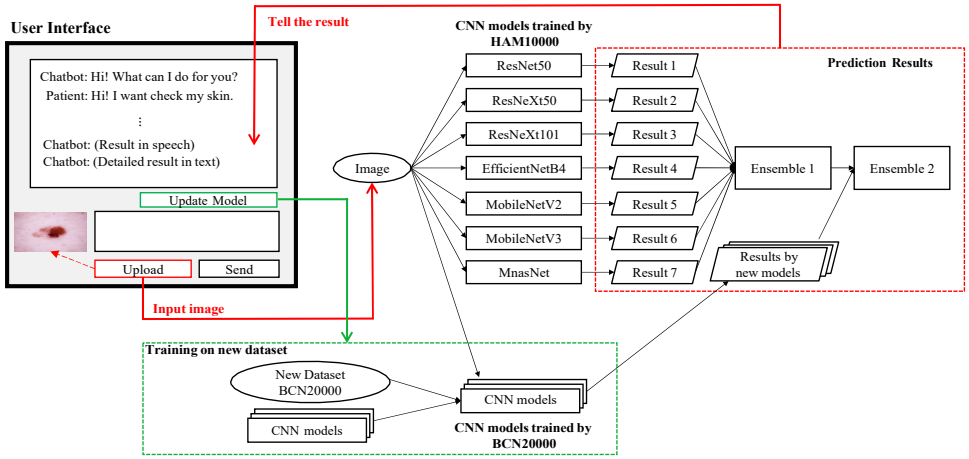


Fig. 2. Skin lesion diagnosis system design (red: upload image to receive diagnosis result, green: update the current model with a new dataset)

E. *System Design.*

We designed a system for skin lesion diagnosis. An overview of its architecture and functionalities is presented in Fig. 2. A UI built with *PyQt5* is provided for human-computer interaction. It has two text boxes where the upper one shows the chat history and the visualization of the diagnosis result, and the bottom one, together with the button "Send" allows the user to input and send the text. We implemented the tokenization function in *NLTK* and the "say" function in *pyttsx3* to enable the system to "understand" the text input from users and tell the diagnosis result in a speech to users, respectively. Button "Upload" allows the user to upload a skin image and starts the image classification. The image was firstly classified by seven models separately.

Then, the results by models would be analyzed by ensemble learning to get a more accurate prediction result. We attempted three different rules for "Ensemble 1" here: majority voting, weighted voting by precision, and weighted voting by cubic precision. To make the predictions with low precision have less impact on deciding the final result, and those with high precision could be considered, a cubic precision is performed in this study. The ensemble with the highest accuracy would be used to return the final result. We also created an "Update model" button for users to upload a new dataset and update the existing ensemble by combining it with some CNN models trained on the dataset.

IV. Results

A. *Performance Analysis for Models.*

For the implementation of different CNN models on the dataset HAM10000, two representative figures of training loss and validation loss during the training are shown in Fig. 3.

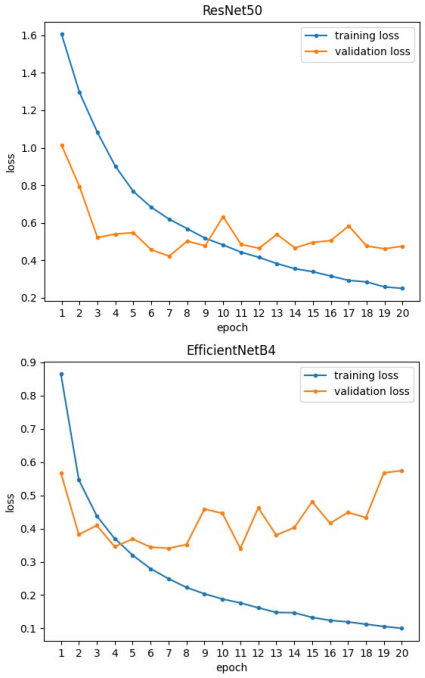


Fig. 3. Loss of models during training time

It shows that the validation loss for each of these CNN models would not change so much after a certain epoch. After epoch 7, when achieved the lowest value, the validation loss of ResNet50 did not converge any longer but vibrate to the end of training. Besides, by observing the loss curves for EfficientNetB4, we find a dramatic increment of validation loss after epoch 18 when training loss is lower than 0.11, indicating overfitting. As Efficient-NetB4 also achieved its lowest validation loss when training loss is around 0.2 on epoch 11, we believe that 20 epochs are also sufficient for other models.

We also noted down the training time and accuracy for each model, as shown in Table I. The result shows that

ResNet50, MobileNetV2, and MobileNetV3 took a relatively short time to train compared with others, and ResNeXt101 took the longest time, indicating the structural complexity of each model. For validation accuracy, Mobile-NetV3 won first place with 91.02%, EfficientNetB4 won second place with 90.57%, and MnasNet was ranked at the last place with 83.59% accuracy.

TABLE I. TRAINING TIME AND ACCURACY OF DIFFERENT MODELS ON HAM10000

	ResNet50	ResNeXt50	ResNeXt101	EfficientNetB4
Time (mins)	137.63	181.57	485.93	294.87
Accuracy	84.95%	88.94%	88.31%	90.57%
	MobileNetV2	MobileNetV3	MnasNet	
Time (mins)	128.33	128.63	300.15	
Accuracy	89.21%	91.02%	83.59%	

These models' performance of precision (Prec), sensitivity (SN), and specificity (SP) could be delved in Table II. By observation, we notice that different models have strengths in different skin lesion categories regarding the values of precision, sensitivity, and specificity. For instance, MobileNet-V2 and MobileNetV3 showed their advantage in predicting Vascular Skin Lesions with extremely high precision and specificity. ResNeXt101 and MobileNetV3 could more correctly diagnose Basal Cell Carcinoma than others with high precision. MnasNet showed its confidence in truly classifying the positives of Actinic Keratosis and Benign Keratosis, etc. For the average value, MobileNetV3, MobileNetV2, and EfficientNetB4 have the highest precision, sensitivity, and specificity, respectively.

TABLE II. EVALUATION PARAMETERS OF DIFFERENT MODELS ON HAM10000

		AK	BCC	BKL	DF	NV	VSC	MEL	Avg.
ResNet50	Prec	0.45	0.60	0.56	0.33	0.97	0.78	0.34	0.58
	SN	0.60	0.77	0.61	0.75	0.91	0.54	0.54	0.68
	SP	0.98	0.98	0.96	0.99	0.90	1.00	0.96	0.97
ResNeXt50	Prec	0.49	0.63	0.75	0.40	0.96	0.90	0.45	0.65
	SN	0.53	0.91	0.55	0.75	0.96	0.69	0.46	0.69
	SP	0.98	0.98	0.98	0.99	0.85	1.00	0.98	0.97
ResNeXt101	Prec	0.65	**0.81**	0.59	0.50	0.96	0.75	0.44	0.67
	SN	0.57	0.86	0.65	0.75	0.95	0.69	0.44	0.70
	SP	0.99	0.99	0.96	1.00	0.82	1.00	0.98	0.96
EfficientNetB4	Prec	0.76	0.79	0.69	0.46	0.97	0.89	0.46	0.72
	SN	0.53	0.86	0.72	0.75	0.97	0.62	0.46	0.70
	SP	1.00	0.99	0.97	0.99	0.86	1.00	0.98	**0.97**
MobileNetV2	Prec	0.75	0.62	0.69	0.63	0.97	**1.00**	0.42	0.72
	SN	0.70	0.89	0.65	0.63	0.95	0.69	0.54	**0.72**
	SP	0.99	0.98	0.97	1.00	0.87	1.00	0.97	0.97
MobileNetV3	Prec	0.68	**0.88**	0.69	0.70	0.96	**1.00**	0.56	**0.78**
	SN	0.63	0.80	0.78	0.88	0.97	0.54	0.44	0.72
	SP	0.99	1.00	0.97	1.00	0.84	1.00	0.99	0.97
MnasNet	Prec	**1.00**	0.52	0.93	0.44	0.90	0.88	0.27	0.71
	SN	0.10	0.71	0.16	0.50	0.96	0.54	0.41	0.48
	SP	1.00	0.98	1.00	1.00	0.55	1.00	0.95	0.92

B. Analysis of Ensemble Learning Method.

The ensemble of models based on different voting rules achieved notable results on HAM10000, and the accuracy for each is presented in Table III, where the first ensemble is based on the majority voting rule, and the others follow the weighted voting rule with different weights. As a result, the ensemble weighted by cubic precision (precision^3) achieved the highest accuracy, and the one weighted by precision had the lowest among all ensembles.

TABLE III. VALIDATION ACCURACY OF DIFFERENT ENSEMBLES ON HAM10000

	Ensemble (Majority)	Ensemble (Prec)	Ensemble (Prec^3)
Accuracy	92.20%	91.93%	**92.84%**

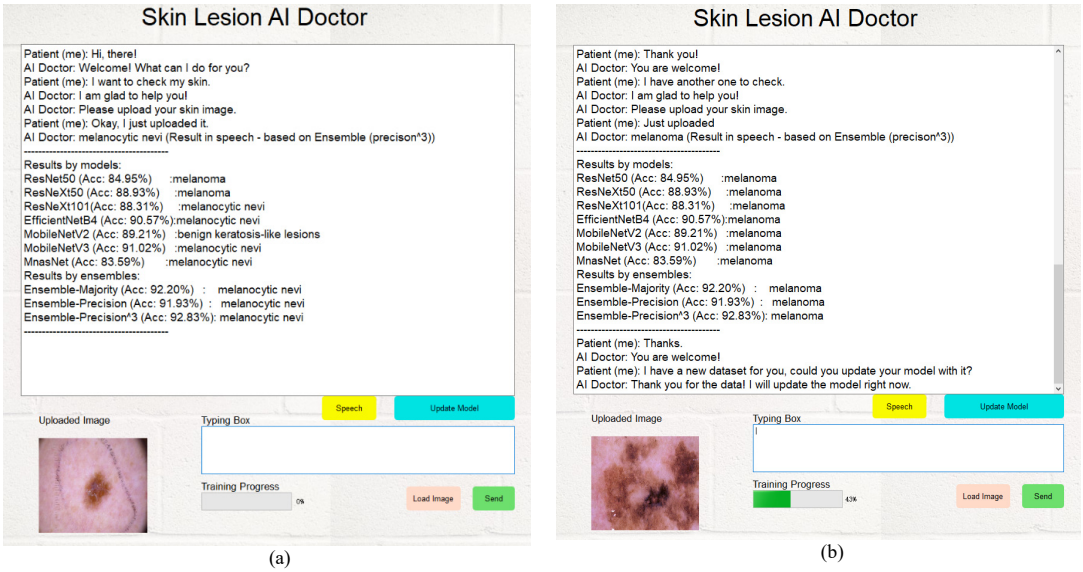
For further discussion, we evaluated all ensembles regarding their precision, sensitivity, and specificity for all lesion categories. The evaluation result is given in Table IV. As we observe, the ensemble with majority voting has the lowest precision regarding, and its average value is 0.80, while the ensemble weighted by cubic precision achieved the highest precision as 0.91. The sensitivity and specificity of the ensemble of majority voting are slightly higher than the other two ensembles. According to their specificity, both ensembles based on weighted voting perform better than those based on majority voting for lesions other than Melanocytic Nevi.

TABLE IV. EVALUATION PARAMETERS OF DIFFERENT ENSEMBLES ON HAM10000

		AK	BCC	BKL	DF	NV	VSC	MEL	Avg.
Ensemble (Majority)	Prec	0.74	0.81	0.79	0.70	0.96	1.00	0.59	**0.80**
	SN	0.57	0.97	0.72	0.88	0.98	0.62	0.59	0.76
	SP	0.99	0.99	0.98	1.00	0.85	1.00	0.98	0.97
Ensemble (Precision)	Prec	0.89	0.83	0.80	1.00	0.95	1.00	0.83	**0.90**
	SN	0.53	0.86	0.65	0.75	0.99	0.54	0.72	0.72
	SP	1.00	0.99	0.99	1.00	0.77	1.00	0.99	0.96
Ensemble (Precision^3)	Prec	1.00	0.91	0.78	1.00	0.95	1.00	0.74	**0.91**
	SN	0.60	0.89	0.67	0.75	0.99	0.69	0.61	0.74
	SP	1.00	1.00	0.98	1.00	0.78	1.00	0.99	0.97

C. Model Update with New data

We examined the validation accuracy of the ensemble of models previously trained on HAM10000 that weighted by cubic precision on the new dataset BCN20000. The results of the ensemble method and four different models are displayed in Table V, where the ensemble obtained the lowest accuracy and EfficientetB4 achieved the highest one.



(a) (b)

Fig. 4. Skin lesion diagnosis system – user interface

TABLE V. VALIDATION ACCURACY OF DIFFERENT MODELS ON BCN20000

	Ensemble (Old)	ResNeXt50	MobileNetV2	MobileNetV3	EfficientNetB4
Accuracy	**40.63%**	64.06%	65.38%	72.84%	**76.71%**

Then, MobileNetV3 and EfficientNetB4 were selected to update the ensemble. The details regarding precision, sensitivity, and specificity of models and ensembles are shown in Table VI. We can see that the average precision of the updated ensemble is higher than any of the old ensemble or the models used for the update. Moreover, the sensitivity and specificity of it are much higher than that of the old one.

The validation accuracy of the new ensemble on the new dataset was improved to 76.32% after the update.

TABLE VI. EVALUATION PARAMETERS OF MODELS AND ENSEMBLES ON BCN20000

		AK	BCC	BKL	DF	NV	VASC	MEL	Average
Ensemble (Old)	Prec	0.17	0.69	0.21	0.02	0.45	1.00	0.79	**0.48**
	SN	0.14	0.24	0.28	0.08	0.90	0.09	0.04	**0.25**
	SP	0.95	0.97	0.89	0.95	0.45	1.00	1.00	**0.89**
MobileNet-V3	Prec	0.47	0.79	0.58	0.55	0.77	0.61	0.80	0.65
	SN	0.67	0.77	0.62	0.71	0.82	0.86	0.63	0.72
	SP	0.95	0.94	0.95	0.99	0.88	1.00	0.95	0.95
EfficientNet-B4	Prec	0.68	0.78	0.65	0.47	0.86	0.86	0.73	**0.72**
	SN	0.54	0.88	0.70	0.58	0.73	0.82	0.79	0.72
	SP	0.98	0.92	0.96	0.99	0.94	1.00	0.90	0.96
Ensemble (Updated)	Prec	0.77	0.77	0.80	0.78	0.76	0.40	0.79	**0.72**
	SN	0.54	0.87	0.54	0.58	0.89	0.82	0.64	**0.70**
	SP	0.99	0.92	0.99	1.00	0.86	0.99	0.94	**0.95**

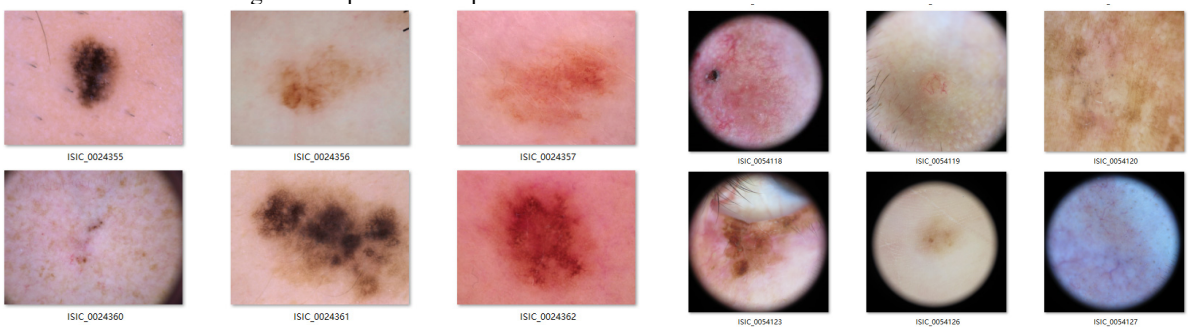
D. *Skin Cancer Diagnosis System*

The application we developed with PyQt5 is displayed in Fig. 4. Firstly, pictures in Fig. 4 present the interactive interface of our application built through NLP techniques. All of a user's messages sent to it by the "Send" button can be given the corresponding response. For example, when the user typed "I want to check my skin", the system tokenized the text and extracted "check" and "skin" as keywords, then responded appropriately by asking the user to upload a skin image. Figure 4(a) demonstrates how the user could use the "Load Image" option to upload a skin image to the system and get a Melanocytic Nevi diagnosis based on the ensemble weighted by cubic precision, which was a correct prediction. The diagnosis report was delivered in both voice and text formats, and the user could repeat the speech by pressing the "Speech" button. The validation accuracy in a bracket was attached to each model and ensemble in the text result to notify the user of some specifics. Some models in the ensemble identified the uploaded image as MEL or BKL but could not affect the final result. Figure 4(b) shows another example of diagnosis, in which the submitted image was categorized as Melanoma by all models and ensembles. Also, the fact that the user could upload fresh datasets and see the state of the backend training process via the "Training Progress" progress bar showed that the model update function was operating properly.

V. DISCUSSION

Concerns for precision. As shown in Table 5, the ensemble trained on HAM10000 failed to show its generalization ability on BCN20000 for its rather low accuracy. As a solution, we developed some CNN models on the new data and combined it with the old ensemble into a new ensemble based on cubic precision and achieved considerable accuracy. However, precision here was a prior knowledge tested on the new dataset, and it would be unknown if the source of new data is not given, which leads to a problem – Without knowing the data source,

it is unclear which precision we should apply to a new image. A feasible method would be to fit all of the CNN models in the ensemble with this new image and update their precisions dynamically, which would be the one focus of our future work [41].



(a) (b)

Fig. 6. Image samples in HAM10000 and BCN20000 (a: HAM10000, b: BCN20000)

Threats to Validity. This study's main threat to internal validity is the unbalanced image numbers for different skin lesion categories in the unprocessed training data, as revealed by Fig. 5. The distribution in BCN20000 is more balanced than HAM10000, and their total image numbers are close, which explains the precisions in the result. For example, EfficientNetB4 trained on BCN20000 has higher precision for melanoma than that trained on HAM10000. The unbalanced distribution also yielded a higher overall accuracy for HAM10000 since its validation data has a similar distribution, and still NV is the majority. In future work, we would combine the datasets and find more data for each category to balance and increase generalization ability.

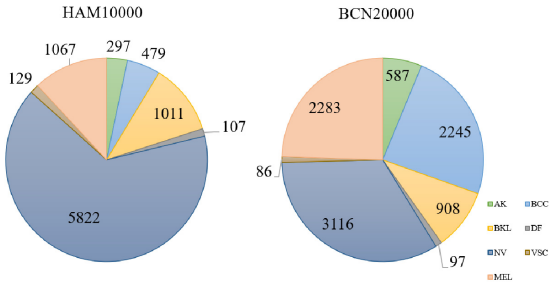


Fig. 5. Distribution of unbalanced training data for HAM1000 and BCN20000

Another threat to validity is the difference of collected images in different datasets. Fig. 6 presents some samples in both datasets we used in this paper. Skin images in HAM10000 almost have no shadow areas, while most images in BCN20000 are circled and filled with the black outside. This provides one more reason to explain the bad performance of the old ensemble on the new data. In the future, we would try to process more types of data augmentation with more data to improve the robustness of ensembles so that they could handle various interferences.

VI. Conclusion

As early detection of skin lesions is critical to skin cancer treatment, an efficient and accurate diagnosis system as a solution is highly demanded. Our study implemented seven high-performance CNN models, including ResNet50, ResNeXt50, ResNeXt101, EfficientNetB4, MobileNetV2, MobileNetV3 and MnasNet, to handle classification task with skin lesion images in HAM10000 dataset, which resulted in the highest accuracy 91.02% by MobileNetV3. To optimize the performance, we applied ensemble learning with these models. The first ensemble is based on the rule of majority voting and achieved 92.20% accuracy. We find that the precisions for different classes of lesions are different from these CNN models and build the second ensemble based on the precision-weighted voting, which achieved 91.93% accuracy. To minimize the impact of predictions with low precisions from models, we cubic the precision and introduced the third ensemble based on that, which achieved 92.83% accuracy. We also examined the classification ability of the trained ensemble on a different dataset – BCN20000, but only achieved 40.63% accuracy. Thus, we gave a corresponding solution: update the ensemble with the new dataset. The result shows that our ensemble could finally achieve 76.71% accuracy for the new dataset. As the integration of our works, we developed an application that allows users to communicate with it in text and upload their skin images to have a skin lesion diagnosis with the embedded ensemble model. The result of diagnosis would be given instantly in the forms of both speech and text in detail to the users. Additionally, this diagnosis application has the functionality of updating the existing ensemble with the new dataset. Users can upload their new dataset and operate the application through the UI interface. In general, our study provides an ensemble-algorithm-based diagnosis system, which can support skin cancer detection and screening in the early stage.

References

[1] *Skin Cancer Facts & Statistics*. (2021 January). Skin Cancer Foundation. https://www.skin-cancer.org/skin-cancer-information/skin-cancer-facts/

[2] *Melanoma Overview – A Dangerous Skin Cancer*. (2021 January). Skin Cancer Foundation. https://www.skincancer.org/skin-cancer-information/melanoma/

[3] *Key Statistics for Melanoma Skin Cancer*. (2021 January). American Cancer Society. https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html

[4] *Melanoma Strikes Men Harder*. (2017 February). American Academy of Dermatology Association. https://www.aad.org/public/diseases/skin-cancer/types/common/melanoma/men-50

[5] *Melanoma Survival Rates*. (2018 July). Melanoma Research Alliance. https://www.cure-melanoma.org/about-melanoma/melanoma-staging/melanoma-survival-rates/

[6] Jerant, A. F., Johnson, J. T., Sheridan, C. D., & Caffrey, T. J. (2000). Early detection and treatment of skin cancer. *American family physician, 62*(2), 357-368.

[7] *UMSkinCheck.* (2012 May). Michigan Medicine, University of Michigan. https://www.uofm-health.org/patient%20and%20visitor%20guide/my-skin-check-app

[8] *Explore SkinVision*. SkinVision. https://www.skinvision.com/getting-started/#explore_skin-vision

[9] Mohd Sanad Zaki Rizvi. (2020 February) *Learn Image Classification on 3 Datasets using Convolutional Neural Networks (CNN)*. Analytics Vidhya. https://www.analyticsvidhya.com-/blog/2020/02/learn-image-classification-cnn-convolutional-neural-networks-3-datasets/

[10] Moeskops, P., Wolterink, J. M., van der Velden, B. H., Gilhuijs, K. G., Leiner, T., Viergever, M. A., &Išgum, I. (2016, October). Deep learning for multi-task medical image segmentation in multiple modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 478-486). Springer, Cham.

[11] Oh, Y., Park, S., & Ye, J. C. (2020). Deep learning covid-19 features on cxr using limited training data sets. *IEEE transactions on medical imaging, 39*(8), 2688-2700.

[12] Pham, T. C., Luong, C. M., Visani, M., & Hoang, V. D. (2018, March). Deep CNN and data augmentation for skin lesion classification. In *Asian Conference on Intelligent Information and Database Systems* (pp. 573-582). Springer, Cham.

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[14] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500).

[15] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning* (pp. 6105-6114). PMLR.

[16] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[17] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., & Le, Q. V. (2019). Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2820-2828).

[18] Caballé-Cervigón, N., Castillo-Sequera, J. L., Gómez-Pulido, J. A., Gómez-Pulido, J. M., & Polo-Luque, M. L. (2020). Machine learning applied to diagnosis of human diseases: A systematic review. *Applied Sciences*, *10*(15), 5135.

[19] Doan, S., Maehara, C. K., Chaparro, J. D., Lu, S., Liu, R., Graham, A., ... & Pediatric Emergency Medicine Kawasaki Disease Research Group. (2016). Building a natural language processing tool to identify patients with high clinical suspicion for Kawasaki disease from emergency department notes. *Academic Emergency Medicine*, *23*(5), 628-636.

[20] Parthasarathy, G., Lopez, R., McMichael, J., & Burke, C. A. (2020). A natural language–based tool for diagnosis of serrated polyposis syndrome. *Gastrointestinal endoscopy, 92*(4), 886-890.

[21] Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., ... &Mascolo, C. (2020, August). Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3474-3484).

[22] Liang, Z., Powell, A., Ersoy, I., Poostchi, M., Silamut, K., Palaniappan, K., ... &Thoma, G. (2016, December). CNN-based image analysis for malaria diagnosis. In *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 493-496). IEEE.

[23] Khan, M. A., Javed, M. Y., Sharif, M., Saba, T., & Rehman, A. (2019, April). Multi-model deep neural network based features extraction and optimal selection approach for skin lesion classification. In *2019 international conference on computer and information sciences (ICCIS)* (pp. 1-7). IEEE.

[24] Alakwaa, W., Nassef, M., &Badr, A. (2017). Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). *Lung Cancer*, 8(8), 409.

[25] Lindeberg, T. (2012). Scale invariant feature transform.

[26] Ahonen, T., Hadid, A., &Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12), 2037-2041.

[27] Yih, W. T., He, X., & Meek, C. (2014, June). Semantic parsing for single-relation question answering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 643-648).

[28] Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2018, August). A combined CNN and LSTM model for arabic sentiment analysis. In International cross-domain conference for machine learning and knowledge extraction (pp. 179-191). Springer, Cham.

[29] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017, March). CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 131-135). IEEE.

[30] Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

[31] Yu, S., Jia, S., & Xu, C. (2017). Convolutional neural networks for hyperspectral image classification. *Neurocomputing*, 219, 88-98.

[32] Loey, M., Manogaran, G., Taha, M. H. N., & Khalifa, N. E. M. (2021). A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. *Measurement, 167*, 108288.

[33] [Online] Necati D., "Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Results", data science and databases, Developers, 2016. Available: https://www.toptal.com/machine-learning/ensemble-methods-machine-learning.

[34] Kannojia, S. P., & Jaiswal, G. (2018, February). Ensemble of hybrid CNN-ELM model for image classification. In *2018 5th international conference on signal processing and integrated networks (SPIN)* (pp. 538-541). IEEE.

[35] Zhang, W., Liu, F., Luo, L., & Zhang, J. (2015). Predicting drug side effects by multi-label learning and ensemble learning. *BMC bioinformatics*, 16(1), 1-11.

[36] Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, 63, 208-222.

[37] Tschandl, P., Rosendahl, C., &Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data, 5*(1), 1-9.

[38] Combalia, M., Codella, N. C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., ... &Malvehy, J. (2019). BCN20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.

[39] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).

[40] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1314-1324).

[41] Jason Brownlee. (2021, March). *How to Update Neural Network Models With More Data*. Machine Learning Mastery. https://machinelearning-mastery.com/update-neural-network-models-with-more-data/