Combination of Deep Learning and Ambiguity Rejection for Improving Image-Based Disease Diagnosis

Thanh-An Pham[1] and Van-Dung Hoang[2](✉)

[1] Ho Chi Minh University of Banking, Ho Chi Minh City, Vietnam
anpt@buh.edu.vn

[2] Ho Chi Minh City University of Technology and Education, Ho Chi Minh City, Vietnam
dunghv@hcmute.edu.vn

Abstract. Artificial intelligent (AI) based medical image recognition plays important task to assist in many disease diagnosis systems. In medical diagnosis, the incorrect decision is very serious. The healthcare diagnosis guides the treatment plan, and it is significant impact on the patient's health outcomes. An incorrect diagnosis can lead to delays in treatment or even the wrong treatment being administered, which results in serious harm to the patient. In this article, we propose an approach to reject ambiguity samples in the classification results, which improve the accuracy of the medical image- based diseases diagnose. In this study, we also experimented using some well-known deep learning models such as MobileNet (lightweight architecture) and DenseNet (more complex and dense connected architecture). Additionally, we combine with some solutions to address the problem of the data imbalance such as focal loss and data augmentation techniques. In the classification stage, there are still significant misclassification results. Therefore, we present the solution for ambiguity rejection of uncertain samples. Experimental results show that the accuracy increases significantly after removing uncertain samples. The high removal rate of uncertain samples also affects to the diagnosing quality. This approach eliminates uncertain samples, which utilizes for improving the diagnosing quality from results of deep learning classification around 10% recall and 70% coverage rate, respectively.

Keywords: Ambiguity rejection · Classification · Feature extraction · Medical image processing

1 Introduction

Skin cancer is a prevalent and dangerous disease that requires high accurate diagnosis for effective treatment. Melanoma, a type of skin cancer, has become increasingly common in recent decades and affects people of all ages. Although melanoma accounts for only 1% of skin cancers, it causes the majority of skin cancer deaths. Early prediction of skin cancer are crucial for effective treatment and cauterization. Advanced technology, particularly in the field of artificial intelligence, has led to the development of practical

N. T. Nguyen et al. (Eds.): ACIIDS 2023, LNAI 13995, pp. 147–160, 2023.
https://doi.org/10.1007/978-981-99-5834-4_12

applications for medical and healthcare. Deep learning (DL) has been widely applied in various fields, which includes medical diagnosis and healthcare, robotics and automation, and intelligent assistance systems and so on. The high performance with handling variety tasks become a popular choice for solving specific problems. DLis particularly useful for image processing tasks, such as medical image analysis and diagnosis, due to its ability to learn and extract features in high performance. DL techniques have been shown to produce better results compared to traditional shallow learning approaches. It is abilited to handle large datasets with many trainable parameters. However, a major challenge in training DL models is small dataset, data imbalance. This problem is leaded to biased classification models, with high performance on majority categories and low performance on minority categories. For example, in the ISIC 2018 dataset, the NV category is large samples, while other categories are a little samples. This problem leads to the NV categories dominating the model during training, and low performance on other categories. To address this issue, some techniques such as data augmentation and focal loss approach are used to improve performance. Augmented data techniques is a common technique to balance the dataset by artificially increasing the number of samples in under-fitting categories. However, this technique leads to overfitting or making noisy samples into the dataset. Therefore, in this study, we only focus on reject uncertain samples, which may lead incorrect diagnosis, for improving accuracy of decision with high rate of sample coverage and reject accuracy.

2 Related Works

These are some of the popular and well-known DL models in image classification and pattern recognition. Each of them has strength points and characteristics that make suitable for different datasets and application fields. The GoogleNet approach [1] is known as a deep architecture with multiple layers, MobileNet [2] is designed to be lightweight and efficient for mobile devices, ResNet [3] and DenseNet [4] are ability to train very deep neural networks and overcome the vanishing gradient problem, while EfficientNet [5] has shown to be highly accurate and efficient for various image recognition tasks. These are selected models, which have greatly improved the flexibility and accuracy of image recognition systems [6, 7]. Generally, it selects the appropriate model for specific dataset with expected that the system achieves higher accuracy without the need for manual tuning or hand craft selection. This is particularly useful in applications where the dataset is changing or evolving, the classified system should adapt to new data. Overall, DL models are more accessible and effective for a wider range of applications in image recognition and beyond. In industrial aspects, DLs-based methods have been widely used in many applications such as video surveillance system [8]. These approaches aim to find the optimal configuration of hyperparameters for the DL model, such as learning rate, batch size, number of filters, etc. The search method randomly selects a combination of hyperparameters and evaluates the performance of the model. The grid search method searches for the best combination of hyperparameters within a predefined range. The Bayesian optimization algorithm uses prior knowledge to guide the search for the best hyperparameters [9–11]. These methods have been shown to be effective in finding optimal hyperparameters for DL models [12, 13]. There are various approaches to improving the performance of DL models for image recognition tasks. These include using state-of-the-art models, selecting models automatically based on the data, optimizing the structure and hyperparameters of the models, and data augmentation to address the problem of imbalanced data [14, 15]. The selection approach depends on the specific problem and available resources, and combination of different approaches is necessary to reach higher accuracy.

In the field of medical images-based cancer disease diagnosis, dermoscopy is a skin surface imaging microscopic technique technology. Numerous studies have demonstrated that DL models produce high diagnostic performance when compared to standard imaging, dermatologists [16]. The paper [17] analysis methods and experimental results on the ISIC Challenge 2018. They presented a two-stage method to segment lesion regions from medical images based optimized training method and applied some parts for post-processing. The lesion images were acquired with a variety of dermatoscope types, from all anatomic sites, or historical sample of patients presented for skin cancer screening, from several different institutions. Each lesion image contains exactly one main lesion. Inspired by synthetic minority oversampling technique [18]. This method focuses the minority category samples before performing up sampling, which supports for better consideration of the uneven distribution of the samples. In another approach, MC-SMOTE method [19] combines of over-sampling the minority categories and under-sampling the majority categories, which achieves higher classifier performance than just using under-sampling the majority categories. This method uniformly increases minority categories samples by utilizing k-mean method, e.g., wind turbine fault detection for applied to practical application.

Other recent developments in the field of pattern recognition and classification based on the use of attention mechanisms in DL models [20]. In this approach, it allows the classification models to focus on the most informative parts of input images rather than processing on the entire image as equally importance. Nowadays, attention mechanisms have been shown that its outperformers accuracy than the DL models based on convolutional network in various tasks such as pattern classification, object recognition, image captioning, and so on.

In other approach, some research works report methods for eliminating uncertain samples [21–23]. These solutions are the inspiration for proposed solutions in the problem of diagnosing diseases, which improve the accuracy of medical image classification. This approach is integrated reject option that enables the network to reject input samples that are difficult to classify with high confidence. The authors argue that this can lead to better performance in real-world applications where the cost of misclassification is high. The reject option is implemented using a binary decision tree that operates on the output of the network. The decision tree takes as input the predicted class probabilities and other features such as the maximum and minimum probabilities and decides whether to reject the input sample or classify it with one of the predefined classes. The methods achieved state-of-the-art performance on several benchmark datasets and performs particularly well on imbalanced datasets.

3 Proposed Methodology

3.1 Overview Approach

This method aims to improve the performance of a DL by optimizing its architecture and ambiguity rejection. The general processing architecture, illustrated in Fig. 1, includes three major stages that should be investigated and customized: feature extraction, fully connected network for the classifier, and ambiguity rejection.

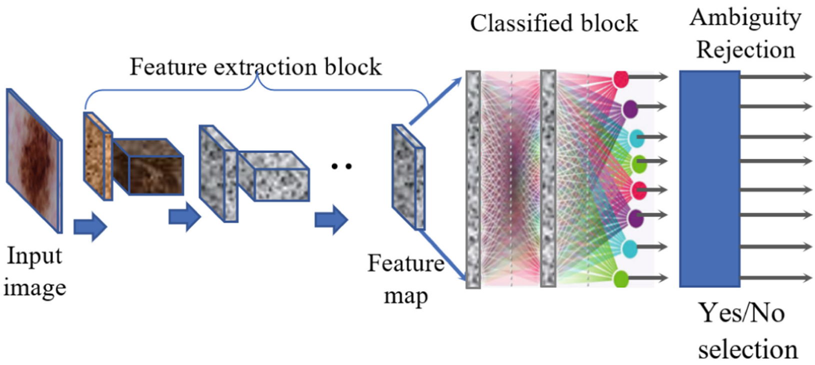


Fig. 1. General training flowchart of a DCNN based classification architecture.

3.2 Feature Extraction and Classification

In the first stage of feature extraction, the DL model is adjusting the training parameters, and refining the loss formulation. The approach has been evaluated empirically using various convolutional neural network (CNN) backbones for feature extraction tasks on different criteria. Our research does not focus on designing new deep learning architectures. Instead, we use the popular CNN model and customize fully connected layers for multiple category classification. There are many approaches to solve the feature extraction stage, such as using state-of-the-art backbone architectures with their pretrained parameters or initially constructing CNN architectures for selected searching the best model. The output feature maps are used as input for the classification stage. Experimental results prove the stability and efficiency on some predefined DCNN backbones, such as DenseNet and MobileNet family.

In this paper, two popular outstanding CNN architectures of DenseNets [4], MobileNets [2] were investigated. Among that, the family MobileNet architectures are known as lightweight model, which is efficiently model for limited resources. Two versions of MobileNet and MobileNetV3Large models were explored the performance ratio. The transfer learning was applied from a pretrained ImageNet model to ISIC2018 dataset for finetuning network hyperparameters. In contrast, DenseNets are more accurate and efficient, which are two versions of DenseNet121 and DenseNet201. The DenseNet is transferred learning from the pretrained model using ImageNet, without including the

Table 1. The list of backbones and their parameters

Backbone name	Number of layers	Number of parameters
MobileNet	90	3.757.255
MobileNetV3Large	280	4.885.895
DenseNet121	431	7.565.895
DenseNet201	711	19.309.127

last top layer, and the feature map is taken from its last layer named "ReLU". These architectures with trainable parameters are illustrated in Table 1.

In the classification stage, there are various approaches, such as using fully connected neural network (FCNN), support vector machines (SVM), or other machine learning approaches, which are appropriately applied. In this study, the FCNN for multiple classification, which takes the input feature maps from the feature extraction stage to classify. To avoid overfitting problems, we add some special layers to this neural network architecture, such as dropout layers. The optimal architecture was estimated using the trial-and-error method. Finally, the architecture consists of two dense connected layers with 1,024 nodes and 512 nodes following by activated layer. The activation function results to dropout layer with the ratio of 50% probabilities. The final output layer with c nodes following softmax activation function.

3.3 Imbalanced Data Processing

As mentioned above, to address imbalanced data issue, we investigated several solutions, such as data augmentation (AU) method and focusing on hard samples using focal loss (FL) [24] approach. The AU technique is also explored in this study. Augmentation processing involves applying image processing techniques such as geometric and artificial color transformations to augment data samples of minority categories and to concentrate on misclassification samples. This technique helps to address the problem of data imbalance. The method is suitable for multi-skin disease classification and effectively addresses issues of underfitting and overfitting, which is happened due to the imbalance of samples between the major categories and minor categories. Some image processing techniques are applied such as color normalization and geometrical transformations, which applied to the training dataset. We used color processing and affined transformations such as rotation, flip, skews, zoom, and crop. The augmented data was generated with random parameters within a predefined period, and each new sample was created and fixed for all methods. That means our approach is different to the image data generator, such as Tensorflow and PyTorch libraries, which generate new data from the original dataset for each epoch. In the data generator processing, training data is different each time a trained model, different methods. The image data generator is used to avoid overfitting, but it is difficult to show compared results of different methods because generated training dataset is different each time. The data augmentation method was used to balance the dataset between all categories with the expectation of improving the correct rates. The main problem with this approach is that it produces a huge training dataset from the original one, which requires high hardware requirements and significantly increases computational time. The details of the parameters used to generate the dataset are presented in Table 2.

Table 2. The details of parameters for data augmented processing.

Transformation	Random value
Rotation	[−10, 10]
Flip	Left-right, up-down
Contrast	[0.7, 1]
Tx	[−10, 10]
Ty	[−10, 10]
Zx	[0.8, 1]
Zy	[0.8, 1]
Shear	[−5, 5]

In this paper, we also investigate the weighting mechanism by FL [24] that affects the efficiency of the model for different categories of data. This approach deals the problem of data imbalance without data augmentation processing. Different to data augmentation, the loss functions (LFs) applied for multiple classification, but it may less effective because the performance metrics for this problem are composed of indicators such as one versus all accuracy, sensitivity/recall, and specificity. The training task aims to optimize the model's parameters to achieve the lowest loss cost across all datasets, thereby increasing classification performance. However, this approach leads to a seesaw problem where majority categories are more influential than minority categories, resulting in lower weighting towards performance scores.

3.4 Ambiguity Rejection

Normally, a multi-class classification model can be defined as a set of probabilities P $= \{p_1, p_2, .., p_m\}$ where each p_i denotes predicted probability of classifier of the m categories, p_i is the predicted probability of the i^{th} category and the output of the classifier is defined as a function $f(x) = argmax(p_i)$, with $i \in$ {1,2,..,m}.When we use a per-class confidence thresholds ambiguity rejection module to reject confusion region, the function $f(x)$ is adjusted as the Formula below.

$$f(x) = \begin{cases} reject, if\ p_i \le \delta_i, \forall i \in \{1, 2, .., m\} \\ argmax(p_i), i \in \{1, 2, .., m\}\ otherwise \end{cases} \tag{1}$$

where $\delta = \{\delta_1, \delta_2, .., \delta_m\}$ denotes confidence thresholds, of which δ_i is the threshold of i^{th} category (c_i). δ set is usually obtained from a training sample so that the correctly classified accuracy on test dataset is greater than or equal to the pre-set select accuracy e.g., 95%.

In this study, we use the validation dataset to determine the threshold δ_i of the class c_i. More specifically, from the validation dataset, by using classifier, we calculate a set of probabilities P of each sample in this dataset. For a given class c_i, we determine the potential thresholds ($\delta_{possible}$), which are the unique values of the list probability p_i. The most importance question is how to choose the best threshold for the class c_i. For a given threshold $\delta_i \in \delta_{possible}$ of the class c_i, we determine rejected samples by the Eq. 1. For example, we have n rejected samples and there are k samples that are failures (corrected classified by our model). The probability of having more than k failures is ProbFailure(k,n). A given δ_i is acceptable when ProbFailure(k,n) is greater than 1-β, β is a given significance level. For each acceptable δ, we calculate select accuracry and coverage respectively, and threshold of the class c_i is the one with the highest select accuracry and coverage. In this research, ProbFailure(k,n) is estimate by using Binomial Cumlative Distribution function in Eq. 2 as the following formula.

$$binom.cdf\,(k, n, p) = \sum\nolimits_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \tag{2}$$

where n denotes the number of rejected samples, k denotes the number of failures in n rejected samples, and p denotes the probability that a given rejected sample is failure. A given rejected sample is failure as random, so $p = 0.5$.

4 Experimental Results and Analysis

4.1 Materials and Preprocessing

In this study, the ISIC2018 [25, 26] skin cancer dataset is used to experiment and evaluate the solution. Due to this dataset is still used for a competition then the ground truth labels of testing images are not available. Therefore, the experiment and comparison are based on the training and validation datasets. The original dataset for training contains 10,015 samples and 193 samples for evaluation. The dataset consists of 7 categories, which include Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic keratosis (AKIEC), Benign keratosis (BKL), Dermatofibroma (DF), and Vascular lesion (VASC). The image samples are uniformed 450×600 resolution. For evaluation, the original validation dataset is used as the validation1 dataset. The original training dataset is split into 80% for training and 20% for evaluation as the validation2. Details about the dataset used in this experiment is presented in Table 3.

Table 3. Details of the experimental dataset

	MEL	NV	BCC	AKIEC	BKL	DF	VASC	Total
Training	890	5364	411	262	879	92	114	8012
Validation1	21	123	15	8	22	1	3	193
Validation2	223	1341	103	65	220	23	28	2003
Augmentation	5340	5364	5343	5240	5274	5336	5358	37255

4.2 Evaluation Metrics

To evaluate the performance of the studied methods on the task of feature extraction and classification, we assessed using popular effectiveness measures such as Recall (REC), Accuracy (ACC), Precision (PRE), Specificity (SPE), and F1. Notice that the accuracy metric of multiple classification is different to that of binary classification problem. The accuracy is estimated based on the one versus all retained classes. For each category, the samples are treated as positive samples and other retained classes are treated as negative samples in the binary classification problem. So, the accuracy score criterion differs between binary and multiple classification. However, some other metrics are the same as in binary classification. The effectiveness measured metrics are computed as follows:

$$ACC_i = (TP_i + TN_i)Ns \tag{3}$$

$$ACC = \frac{1}{Ns}\sum_{i=1}^{c} n_i * ACC_i \tag{4}$$

$$\mathrm{Re}call = TP/(TP + FN) \tag{5}$$

$$PRE = TP/(TP + FP) \tag{6}$$

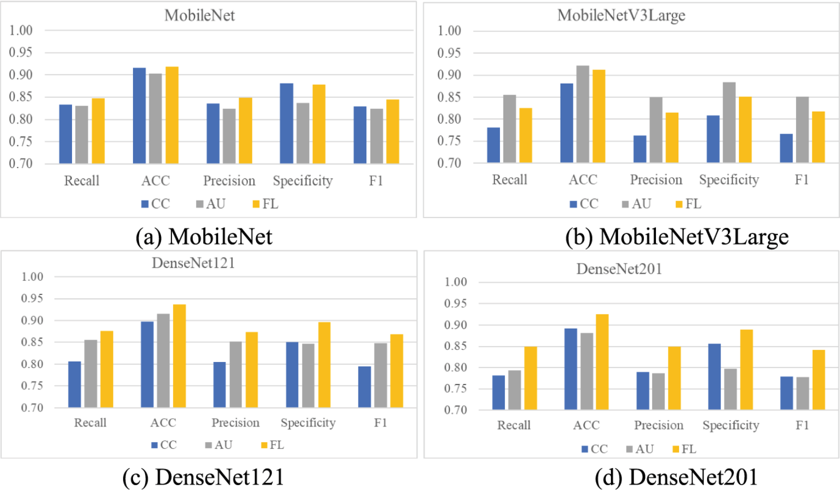
$$SPE = TN/(TN + FP) \tag{7}$$

$$F_1 = TP/[TP + \frac{1}{2}(FP + FN)] \tag{8}$$

where Ns is the total number of samples in dataset, $Ns = TP_i + FP_i + FN_i + TN_i$ where TP_i and FP_i are the number of true positive and false positive samples belonging to the category i^{th}, respectively; FN_{i} and $\mathrm{TN_i}$ are the number of false negative and true negative samples belonging to the category i^{th}, respectively. The number of samples of the class i^{th} is n_i. In that approach the accuracy of each class c^{th} is calculated by $\mathrm{TP_c}$/total instances of the class c^{th}. However, this performance measurement is same with Recall ratio. Therefore, we used the above formulation for estimating the accuracy rate.

4.3 Evaluation Results and Analysis

In this study, we experimentalize and analyze feature extraction and classification task using the category cross entropy and FL, AU method and then ambiguity sample rejection for improving high confident disease diagnosis. In amount of solutions for data imbalance treatment, the AU requires higher computational cost for model training due to that it generates more significant new samples for balancing training dataset. We also customized two kinds of feature extraction backbones, such as MobileNet, DenseNet family. These kinds of backbones are representative for different approaches. The MobileNet backbone represents for a small and compact architecture. It is suitable for applying to limited resource computing systems. The DenseNet backbone represents for the dense connected network with a heaving trainable parameter. In general, MobileNets are lightweight architectures, which consist of several million of trainable parameters. However, they achieve high accuracy with different applications. The MobileNet models are efficient mechanisms based on the depth-wise separable convolutions. The DenseNet architecture with dense connection layers through dense blocks. The network layers relate to matching feature-map sizes directly with each other. Each layer obtains additional inputs from all preceding layers and passes on its feature maps to all subsequent layers. The experimental results on the evaluated dataset show that DenseNet121 + FL method reach outperformer on validation dataset1 at 88.08% recall and 94.18 accuracy rate, as depicted in Table 5 of appendix section. Meanwhile, DensseNet201 and FL method reach the best result on validation dataset2 with all criteria. So, the DensNet network family is more stable results comparing to other methods (Fig. 2). Meanwhile, CC method get the lowest with 76.68% Recall at 88.77% accuracy. In overall, the DenseNet family and FL response the best results on ISIC2018 dataset, as illustrated in Fig. 3.



(a) MobileNet (b) MobileNetV3Large

(c) DenseNet121 (d) DenseNet201

Fig. 2. Experimental results on both evaluation datasets

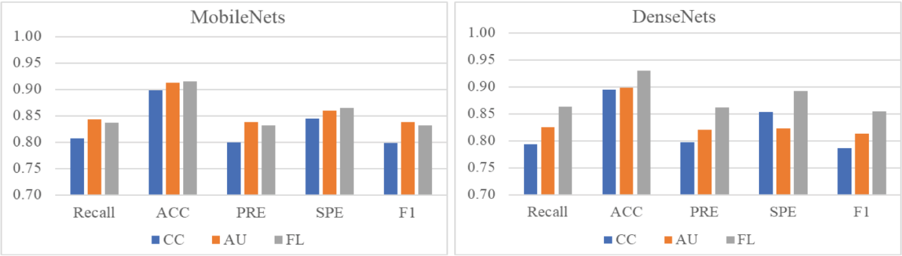


Fig. 3. Average evaluated result of MobileNets and DenseNets family on both validation sets.

In ambiguity reject stage, we adjust δ set so that select accuracy is high, around 95.0% corresponds to error rate at 5%, to ensure acceptable error rate in real-world applications and to compare performances of methods. The validation dataset1 and validation dataset2 are used determine the threshold $\delta_{possible}$ with expected to reach accepted select_recall rate with highest coverage ratio of the class of each category. Some experimental results are shown in Table 4. Ambiguity rejection with $\delta = 0.1$, selected recall ratio was reached about 96.25% at 75% correct coverage ratio with DenseNet121 + AU. Meanwhile, MobileNetV3Large + CC archives 93.19% select recall at 66.24 correct coverage ratio only, as depicted in Table 4 (a). Experimental results also illustrated that the determined coefficient of delta = 0.3, the DenseNet201 + FL achieved the highest precision with 94.91% select_recall at 81.06% correct coverage ratio, while the MobileNet + CC achieved the lowest accuracy with 91.69% select_recall at 80.93% correct coverage ratio, as illustrated in Table 4 (b). According to experimental result shows that CC loss function archives the lowest recall ratio in both situation classification and ambiguity rejection.

5 Conclusions

In this article, we presented a new approach for improving medical image-based diseases diagnosing by applying DL classification and rejecting ambiguous samples. Our approach concentrates on balancing of the influence coefficient ratio of each category to the other ones instead of focusing hard samples of LF method or augmenting image data with expected higher precision ratio. The CNN architecture was also customized fully connected layers and transformed for ISIC dataset. Applying ambiguity rejection stage to removing uncertain samples support for significantly improves accuracy. The solution was able to improve the diagnosing quality from results of classification stage, e.g. recall rate is improved from 85.63% to 96.25% at 75% coverage rate with DenseNet121 + AU, Experimental results demonstrated that this solution utilizes for archiving higher accuracy, but it also gaps a problem of eliminating uncertainty samples, which is not fully coverage ratio in disease diagnosis.

Table 4. Experimental results of ambiguity rejection on both evaluation datasets

(a) Thresh_func is b_cdf and delta=0.1

MobileNet- based ambiguity rejection (delta=0.1)

	base_recall	select_recall	coverage	reject_acc
CC	0.8331	0.9405	0.7481	0.4790
AU	0.8311	0.9544	0.7435	0.4930
FL	0.8478	0.9513	0.7498	0.4583

MobileNetV3Large-based ambiguity rejection (delta=0.1)

	base_recall	select_recall	coverage	reject_acc
CC	0.7806	0.9319	0.6624	0.5088
AU	0.8546	0.9551	0.7772	0.4864
FL	0.8255	0.9375	0.7523	0.5166

DenseNet121- based ambiguity rejection (delta=0.1)

	base_recall	select_recall	coverage	reject_acc
CC	0.8059	0.9535	0.6572	0.4719
AU	0.8563	0.9625	0.7500	0.4586
FL	0.8765	0.9544	0.7950	0.4286

DenseNet201- based ambiguity rejection (delta=0.1)

	base_recall	select_recall	coverage	reject_acc
CC	0.7822	0.9524	0.6546	0.5426
AU	0.7932	0.9496	0.6588	0.4952
FL	0.8498	0.9567	0.7841	0.5165

(b) Thresh_func is b_cdf and delta=0.3

MobileNet- based ambiguity rejection (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.8331	0.9169	0.8093	0.5186
AU	0.8311	0.9446	0.7719	0.5139
FL	0.8478	0.9353	0.8208	0.5457

MobileNetV3Large-based ambiguity rejection (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.7806	0.9273	0.6875	0.5340
AU	0.8546	0.9354	0.8236	0.5156
FL	0.8255	0.9188	0.7988	0.5528

DenseNet121- based ambiguity rejection (delta=0.3)

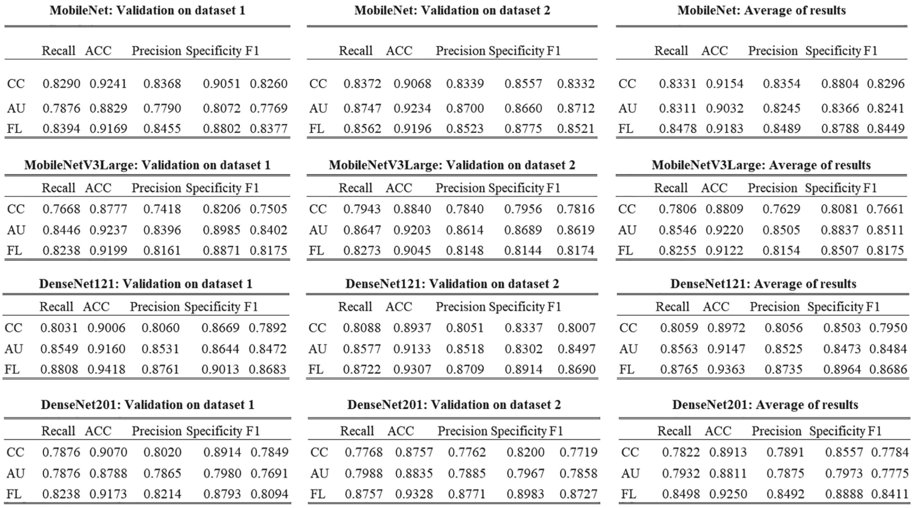
	base_recall	select_recall	coverage	reject_acc
CC	0.8059	0.9389	0.7006	0.5015
AU	0.8563	0.9334	0.8193	0.4928
FL	0.8765	0.9300	0.8942	0.6092

DenseNet201- based ambiguity rejection (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.7822	0.9315	0.7036	0.5725
AU	0.7932	0.9343	0.6899	0.5120
FL	0.8498	0.9491	0.8106	0.5479

Appendix

Table 5. Detail of classified results on validation dataset1, validation dataset2. The average result is formed (result1 on dataset1 + result1 on dataset1)/2.



MobileNet: Validation on dataset 1

	Recall	ACC	Precision	Specificity	F1
CC	0.8290	0.9241	0.8368	0.9051	0.8260
AU	0.7876	0.8829	0.7790	0.8072	0.7769
FL	0.8394	0.9169	0.8455	0.8802	0.8377

MobileNet: Validation on dataset 2

	Recall	ACC	Precision	Specificity	F1
CC	0.8372	0.9068	0.8339	0.8557	0.8332
AU	0.8747	0.9234	0.8700	0.8660	0.8712
FL	0.8562	0.9196	0.8523	0.8775	0.8521

MobileNet: Average of results

	Recall	ACC	Precision	Specificity	F1
CC	0.8331	0.9154	0.8354	0.8804	0.8296
AU	0.8311	0.9032	0.8245	0.8366	0.8241
FL	0.8478	0.9183	0.8489	0.8788	0.8449

MobileNetV3Large: Validation on dataset 1

	Recall	ACC	Precision	Specificity	F1
CC	0.7668	0.8777	0.7418	0.8206	0.7505
AU	0.8446	0.9237	0.8396	0.8985	0.8402
FL	0.8238	0.9199	0.8161	0.8871	0.8175

MobileNetV3Large: Validation on dataset 2

	Recall	ACC	Precision	Specificity	F1
CC	0.7943	0.8840	0.7840	0.7956	0.7816
AU	0.8647	0.9203	0.8614	0.8689	0.8619
FL	0.8273	0.9045	0.8148	0.8144	0.8174

MobileNetV3Large: Average of results

	Recall	ACC	Precision	Specificity	F1
CC	0.7806	0.8809	0.7629	0.8081	0.7661
AU	0.8546	0.9220	0.8505	0.8837	0.8511
FL	0.8255	0.9122	0.8154	0.8507	0.8175

DenseNet121: Validation on dataset 1

	Recall	ACC	Precision	Specificity	F1
CC	0.8031	0.9006	0.8060	0.8669	0.7892
AU	0.8549	0.9160	0.8531	0.8644	0.8472
FL	0.8808	0.9418	0.8761	0.9013	0.8683

DenseNet121: Validation on dataset 2

	Recall	ACC	Precision	Specificity	F1
CC	0.8088	0.8937	0.8051	0.8337	0.8007
AU	0.8577	0.9133	0.8518	0.8302	0.8497
FL	0.8722	0.9307	0.8709	0.8914	0.8690

DenseNet121: Average of results

	Recall	ACC	Precision	Specificity	F1
CC	0.8059	0.8972	0.8056	0.8503	0.7950
AU	0.8563	0.9147	0.8525	0.8473	0.8484
FL	0.8765	0.9363	0.8735	0.8964	0.8686

DenseNet201: Validation on dataset 1

	Recall	ACC	Precision	Specificity	F1
CC	0.7876	0.9070	0.8020	0.8914	0.7849
AU	0.7876	0.8788	0.7865	0.7980	0.7691
FL	0.8238	0.9173	0.8214	0.8793	0.8094

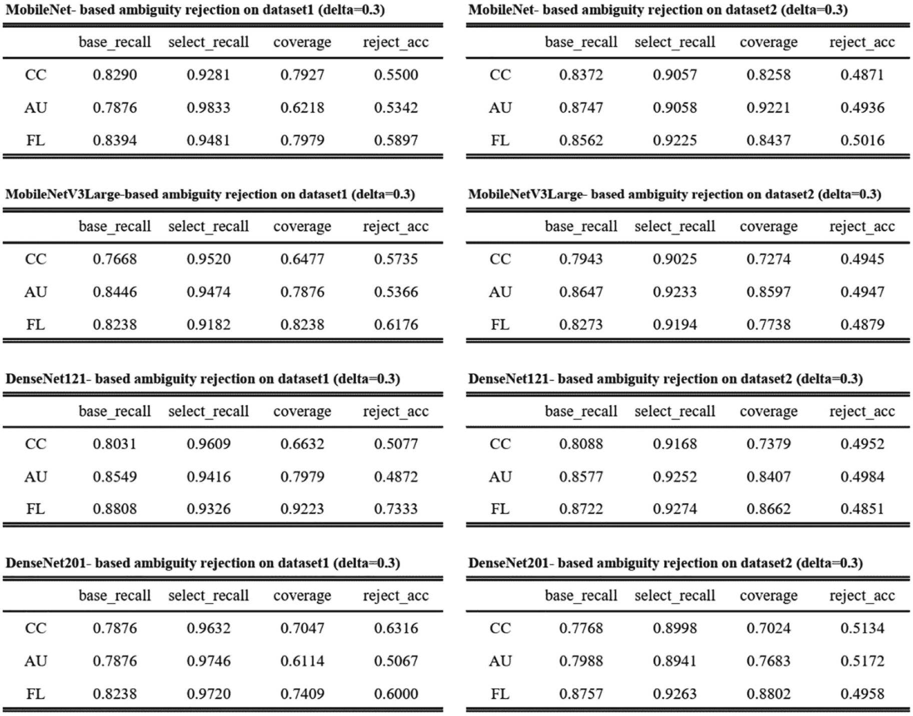
DenseNet201: Validation on dataset 2

	Recall	ACC	Precision	Specificity	F1
CC	0.7768	0.8757	0.7762	0.8200	0.7719
AU	0.7988	0.8835	0.7885	0.7967	0.7858
FL	0.8757	0.9328	0.8771	0.8983	0.8727

DenseNet201: Average of results

	Recall	ACC	Precision	Specificity	F1
CC	0.7822	0.8913	0.7891	0.8557	0.7784
AU	0.7932	0.8811	0.7875	0.7973	0.7775
FL	0.8498	0.9250	0.8492	0.8888	0.8411

Table 6. Detail of ambiguity rejection results on validation dataset1 and validation dataset2



MobileNet- based ambiguity rejection on dataset1 (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.8290	0.9281	0.7927	0.5500
AU	0.7876	0.9833	0.6218	0.5342
FL	0.8394	0.9481	0.7979	0.5897

MobileNet- based ambiguity rejection on dataset2 (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.8372	0.9057	0.8258	0.4871
AU	0.8747	0.9058	0.9221	0.4936
FL	0.8562	0.9225	0.8437	0.5016

MobileNetV3Large-based ambiguity rejection on dataset1 (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.7668	0.9520	0.6477	0.5735
AU	0.8446	0.9474	0.7876	0.5366
FL	0.8238	0.9182	0.8238	0.6176

MobileNetV3Large- based ambiguity rejection on dataset2 (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.7943	0.9025	0.7274	0.4945
AU	0.8647	0.9233	0.8597	0.4947
FL	0.8273	0.9194	0.7738	0.4879

DenseNet121- based ambiguity rejection on dataset1 (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.8031	0.9609	0.6632	0.5077
AU	0.8549	0.9416	0.7979	0.4872
FL	0.8808	0.9326	0.9223	0.7333

DenseNet121- based ambiguity rejection on dataset2 (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.8088	0.9168	0.7379	0.4952
AU	0.8577	0.9252	0.8407	0.4984
FL	0.8722	0.9274	0.8662	0.4851

DenseNet201- based ambiguity rejection on dataset1 (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.7876	0.9632	0.7047	0.6316
AU	0.7876	0.9746	0.6114	0.5067
FL	0.8238	0.9720	0.7409	0.6000

DenseNet201- based ambiguity rejection on dataset2 (delta=0.3)

	base_recall	select_recall	coverage	reject_acc
CC	0.7768	0.8998	0.7024	0.5134
AU	0.7988	0.8941	0.7683	0.5172
FL	0.8757	0.9263	0.8802	0.4958

References

1. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
2. Howard, A.G., et al.: Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv preprint arXiv:1704.04861 (2017)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
4. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
5. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019)
6. Li, L., Talwalkar, A.: Random Search and Reproducibility for Neural Architecture Search. arXiv preprint arXiv:1902.07638 (2019)
7. Bertrand, H., Ardon, R., Perrot, M., Bloch, I.: Hyperparameter optimization of deep neural networks: combining hyperband with Bayesian model selection. Conférence sur l'Apprentissage Automatique (2017)
8. Hoang, V.-T., Huang, D.-S., Jo, K.-H.: 3-D facial landmarks detection for intelligent video systems. IEEE Trans. Industr. Inf. **17**, 578–586 (2020)

9. Kotthoff, L., Thornton, C., Hoos, H.H., Hutter, F., Leyton-Brown, K.: Auto-WEKA 2.0: automatic model selection and hyperparameter optimization in WEKA. J. Machine Learning Res. **18**, 826–830 (2017)
10. Tran, D.-P., Nguyen, G.-N., Hoang, V.-D.: Hyperparameter optimization for improving recognition efficiency of an adaptive learning system. IEEE Access **8**, 160569–160580 (2020)
11. Dikov, G., van der Smagt, P., Bayer, J.: Bayesian Learning of Neural Network Architectures. arXiv preprint arXiv:1901.04436 (2019)
12. Huang, C., Lucey, S., Ramanan, D.: Learning policies for adaptive tracking with deep feature cascades. Proceedings of the IEEE International Conference on Computer Vision, pp. 105–114 (2017)
13. Long, M., Cao, Y., Cao, Z., Wang, J., Jordan, M.I.: Transferable representation learning with deep adaptation networks. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 3071–3085 (2018)
14. Le, N.Q.K., Huynh, T.-T., Yapp, E.K.Y., Yeh, H.-Y.: Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles. Comput. Methods Programs Biomed. **177**, 81–88 (2019)
15. Snoek, J., et al.: Scalable bayesian optimization using deep neural networks. International Conference on Machine Learning, pp. 2171–2180 (2015)
16. Carli, P., et al.: Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. Br. J. Dermatol. **148**, 981–984 (2003)
17. Qian, C., et al.: A Two-Stage Method for Skin Lesion Analysis. arXiv preprint arXiv:1809.03917 (2018)
18. Yi, H., Jiang, Q., Yan, X., Wang, B.: Imbalanced classification based on minority clustering synthetic minority oversampling technique with wind turbine fault detection application. IEEE Trans. Industr. Inf. **17**, 5867–5875 (2020)
19. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artificial Intell. Res. **16**, 321–357 (2002)
20. Niu, Z., Zhong, G., Yu, H.: A review on the attention mechanism of deep learning. Neurocomputing **452**, 48–62 (2021)
21. Geifman, Y., El-Yaniv, R.: Selectivenet: A deep neural network with an integrated reject option. International Conference on Machine Learning, pp. 2151–2159 (2019)
22. Franc, V., Prusa, D., Voracek, V.: Optimal strategies for reject option classifiers. J. Mach. Learn. Res. **24**, 1–49 (2023)
23. Kashani Motlagh, N., Davis, J., Anderson, T., Gwinnup, J.: Learning when to say "i don't know. Advances in Visual Computing: 17th International Symposium, ISVC 2022, San Diego, CA, USA, October 3–5, 2022, Proceedings, Part I, pp. 196–210 (2022)
24. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2980–2988 (2017)
25. Codella, N., et al.: Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
26. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**, 1–9 (2018)