Improving binary skin cancer classification based on best model selection method combined with optimizing full connected layers of Deep CNN

Tri Cong Pham, van Dung Hoang, Cong Thanh Tran, Minh Sao Khue Luu, Duy Anh Mai, Antoine Doucet, Chi Mai Luong

Improving binary skin cancer classification based on best model selection method combined with optimizing full connected layers of Deep CNN

Authors: Tri Cong Pham; Van Dung Hoang; Cong Thanh Tran; Minh Sao Khue Luu; Duy Anh Mai; Antoine Doucet; Chi Mai Luong

***Abstract*— Melanoma is one of the most dangerous skin cancers, leading to high mortality rates. Early detection and resection are two important steps to reduce mortality. Recently, several studies have used artificial intelligence to solve the problem of binary skin cancer classification. However, the imbalance issue of the two sensitivity and specificity metrics affects the performance of models. Our research proposes an optimization of deep Convolutional Neural Network (CNN) combined with changing the best model selection for binary melanoma classification problem. Our research uses the latest and largest ISIC 2019 dataset including 17,302 skin lesion images for training and best model selection. The performance of the best models was compared based on the 10% data of ISIC 2019 dataset (test-10) and then compared with the performance of dermatologists on the same MClass-D dataset of 100 images. As a result, the first our proposed customized fully connected layers deep CNN solves the underfitting problem and avoids overfitting. Secondly, the proposed best model selection method helps to choose a better model than the traditional methods with Youden Index (YI) increased on both the test-10 dataset and MClass-D datasets. Finally, the proposed solution effectively outperformed 153 dermatologists out of a total of 157. This performance surpasses the current state-of-the-art solution by 17 dermatologists.**

Keywords—Skin cancer, melanoma, deep cnn, fully connected layers, best model selection.

I. Introduction

Skin cancer is one of the most common cancers and is easy to diagnose with rational tools. This cancer originates from the skin epithelium covering the outer surface of the body, including many layers of cells. Skin cancer is more common in whites, mainly in the elderly, more men than women. The disease usually occurs in open skin with a rate of 90% in the head and neck area. Approximately five million new cases are detected each year in the USA. Melanoma is the most serious type of skin cancer. It develops in the same skin cells that create moles. Because of this, melanoma is particularly dangerous. It can look like a harmless mole when it first develops. It accounts for only one percent of all skin cancer cases, estimates the American Cancer Society [1]. It is, however, responsible for the majority of deaths. Estimate that the number of new melanoma cases diagnosed in 2019 will increase by 7,7%. There are 192,310 cases of melanoma will be diagnosed in the USA in 2019 almost eight percent more than in 2018. An estimated 7,230 people (3,76%) will die of melanoma in 2019 in the USA [2]. If detected early, patients with melanoma, survival rate after five years is estimated at about 98%, this rate will be reduced to 64% when the disease has spread to the lymph nodes and only 23% when the disease has spread to distant organs.

The first step in the diagnosis of a malignant lesion by a dermatologist is the visual examination of the suspicious skin area. A correct diagnosis is important because of the similarities of some lesion types; moreover, the diagnostic accuracy correlates strongly with the professional experience of the physician [3]. Without additional technical support, professional dermatologists have a 65%-80% accuracy rate in melanoma diagnosis [4]. However, there are not enough experienced dermatologists all over the world. In suspicious cases, the visual inspection is supplemented with dermatoscopic images taken with a special high-resolution and magnifying camera. During the recording, the lighting is controlled and a filter is used to reduce reflections on the skin, thereby making deeper skin layers visible. The combination of visual inspection and dermatoscopic images ultimately results in an absolute melanoma detection accuracy of 75%-84% by dermatologists [5]. Artificial intelligence (AI) will gradually approach the medical imaging services to patients in many hospitals. It will be a powerful assistant for doctors by applying deep convolutional neural networks (DCNN) [6]. This technique will help to read medical images, classify and check them quickly and more accurately.

Many recent researches have used deep CNN for binary melanoma classification problems [7]–[12] but there are still challenges due to the limitation of data and data imbalance problems. These researches use CNN to classify melanoma and nevus and compare the performance of their algorithms with that of dermatologists. In 2017, Esteva et al was the first to compare the direct performance of the Deep CNN with that of 21 board-certified dermatologists on 111 (71 malignant, 40 benign) images dataset, achieved AUC of 91% [9] and lost at least one dermatologist. In 2018, the best-performing fusion algorithm of twenty-five teams of the 2016 International Skin Imaging Collaboration ISBI Challenge is compared with that of eight dermatologists on 100 images dataset. The best algorithm achieved greater AUC than dermatologists (86% vs 71%) and specificity of 70% at the sensitivity of 85,5% [13] respectively YI [14] of 55,5%. Tschandl et al's research compared the performance of Deep CNN with that of 95 dermatologists (including 62 board-certified dermatologists) on a test dataset of 2,072 cases in 2019. The CNN performance was AUC of 74.2% and sensitivity of 80.5% at specificity fixed at 51.3% and YI = 31,8%. It was better than that of humans with AUC of 69.5% and sensitivity of 77.6% at the same specificity, so YI = 28,9% [15]. In 2019, the deep CNN system proposed by Brinker et al outperformed 136 of 157 dermatologists from 12 university hospitals in Germany on 100 dermoscopic images of MClass-D (SOTA). At a mean sensitivity of 74.1%, it achieved higher specificity than the mean of dermatologists (86.5% vs 60%) with YI (60,6% vs 34,1%) [16]. Through these studies, we can see that CNN always has a higher YI than dermatologists.

This proves that the diagnosis of melanoma by CNN will be more effective. However, the problems are: 1) Sensitivity and Specificity measures are imbalanced; 2) Underfitting problem. Thus, in this study, we proposed a solution that uses CNN architecture in combination with two components: CNN and fully connected layers for binary melanoma classification problem by redesigning fully connected layers. Besides, instead of using the usual metrics as accuracy to choose the best model, research suggests some other metrics that are defined in section IV. In this study, we train the system with the ISIC 2019 dataset which is the latest and largest public image collection about skin cancer [17] with an epoch number of 150, then find the best models based on four metrics. The performance of the best models was compared based on the 10% data of the ISIC 2019 dataset (test-10) and then compared with the performance of dermatologists on the same MClass-D dataset of 100 images. The proposed best model selection method helps to choose a better model than the traditional method with YI increased and with the difference between SPE and SEN is the smallest on the test-10 dataset and MClass-D dataset.

II. MATERIALS

A. Materials

The data we used in this research is the ISIC 2019 [17] challenge train dataset, which is the latest and largest public image collection about skin cancer. It includes 25,331 dermoscopic images in 8 different categories. All melanoma diagnoses in the dataset were confirmed by histopathological evaluation of byopsies. As we only concerntrate on classifying melanoma and nevus for this research, we omit the unrelated images and keep total 17,302 melanoma and nevus images for training, validating and testing. Our final dataset consists of 4,503 malenoma images (minority class) and 12,799 nevus images (majority class). 80% of the data is used for training (namely train dataset), while 10% is used for validation (namely validation dataset) and the rest 10% is for testing (namely test-10 dataset). As shown in the Figure 1, the train dataset includes 3,603 melanoma images and 10,239 nevus images, whereas both the validation and test-10 dataset includes 450 melanoma images and 1,280 nevus images.

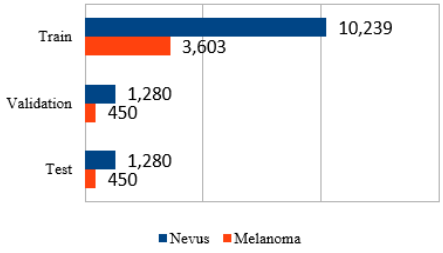

Fig. 1. Melanoma train, validation, and test datasets.

B. Dermatologist performances

To provide a new melanoma classification benchmark for comparing diagnostic performance between artificial intelligence algorithms and certified dermatologists, Titus J. Brinker et al. published an MClass-D dataset of 100 dermoscopic images including 80 nevi and 20 melanomas. Those images were sent to 157 dermatologists working in 12 university hospitals in Germany to record their professional experience and ask for their management decisions on whether to treat or reassure the patient. The MClass-D dataset used sensitivity (SEN), specificity (SPE) and area under the curve (AUC) to evaluate a dermatologist's performance. The Figure 2 summarizes the results of 157 dermatologists in the MClass-D dataset.

TABLE I. DIAGNOSTIC PERFORMANCES OF THE 157 DERMATOLOGISTS ON THE MCLASS-D DATASET.

Subset of dermatologists	**AUC**	**SEN**	**SPE**
All participants (n=157)	67.1	74.1	60.0
University hospital (n=151)	66.9	74.0	59.8
Private practice (resident) (n=6)	71.3	76.7	65.8
Position in hospital hierarchy			
Junior physicians (n=88)	66.5	74.8	58.2
Attendings (n=15)	66.4	72.7	60.0
Senior physicians (n=45)	67.7	73.0	62.3
Chief physicians (n=3)	71.3	73.3	69.2
Practical experience (pe)			
pe ≤ 2 years (n=46)	66.2	76.0	56.5
2 years < pe ≤ 4 years (n=37)	66.4	73.8	59.1
4 years < pe ≤ 12 years (n=32)	67.9	73.3	62.5
pe > 12 years (n=42)	67.9	73.0	62.8

In general, the majority of participants are working in hospitals, with a number of 151 (92.2%) out of the total dermatologists. Only 6 participants (3.8%) are dermatologic resident physicians working in a private office. There are two main groups of participants, junior physicians and board certified, accounting for 56.1% and 43.9% respectively. In terms of experience, almost half of dermatologists have more than four years of practical experience, in which 26% have been working in this field for more than 12 years. Highest AUC of 71.3% and SEN of 69.2% are achieved by the private practice group, whereas highest SPE is 69.2% by the chief physicians.

III. METHODS

A. Proposed binary skin cancer classification system

In this research, we proposed a binary skin cancer classification system which includes two main components: CNN and Custom fully connected layers as shown in Figure 2.

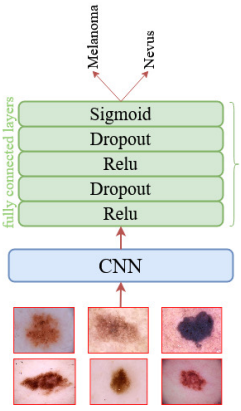

Fig. 2. Proposed optimizing architecture of Deep CNN for binary skin cancer classification.

CNN: In our system, popular CNN architecture is used for feature extraction. We investigated outstanding CNN architectures, such as InceptionV3 [18], ResNet50 [19] with

classic batch logic and loss function to determine network without underfitting problem in training process.

Fully connected layers (FC): In this study, we tested fully connected layers with a hidden layer and found that the system was underfitting, so we added a hidden layer to FC. As shown in Figure 1, FC consists of two hidden layers, the first hidden layer has 1,024 nodes, the second hidden layer is 512 nodes. Both hidden layers use the activation function, ReLU

Dropout: After each hidden layer, we use dropout to avoid overfitting problems [20] thereby improving the efficiency of the network. When using the dropout, it automatically deletes randomly selected units from the neural network during training at the set rate. In this study, both dropouts have a removal rate of 0.5.

Optimizer: The choice of the optimizer for network training is important, depending on the network design and the type of data used. In this research, we use the Adam optimizer to train the network. The parameter details used are: lr = 0.0001, beta_1 = 0.9, beta_2 = 0.999, decay = 0.0, epsilon = None and amsgrad = False.

Learning rate (lr) : lr is the most important parameter of the optimizer, for deep CNN, beside training the network with fixed lr, we can change lr after each epoch. We do not use fixed lr in this study, instead lr is changed after every step of each epoch using cyclical learning with base_lr = 0.0000001, max_lr = 0.0001, mode = triangular2 and step_size = 4 * steps_per_epoch. With this mode and step_size, after eight epochs, max_lr will be halved.

B. Best model selection method

Normally, when using deep learning for classification problems, network training is performed through many epochs. After each epoch, the network is trained on the entire dataset and create a new model. These three following methods are often used to select the best model among epochs' models, which are tested on the validation dataset: 1) last model, or model of the last epoch (last), 2) model with highest accuracy (acc), and 3) lowest loss model. The Keras library also allows customizing metrics to choose the best model, however, there could be two issues: 1) Metric construction requires high skills, and 2) How to know which metrics are good for solving the current problem.

In the medical binary classification problems including binary skin cancer classification, the performance of the system is not represented by the accuracy metric but instead the AUC, SEN and SPE indicators. This leads to the problem of how to select the best model based on the validation dataset and the models created after each epoch.

In this study, we propose methods for best model selection in the medical binary classification problem, which can be applied to binary skin cancer classification, as shown in Figure 3 below:

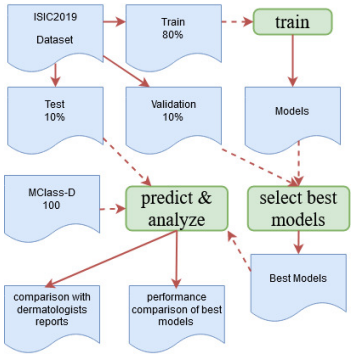

Fig. 3. Best model selection method base on custom metrics for binary skin cancer classification.

In this research, we train the neural network through 150 epochs. After the CNN complete the training at each epoch, a model is saved. After the whole training process, we receive 150 models that are evaluated on the validation dataset. Then all necessary indicators are collected for best model selection methods comparision. The indicators for selecting models are described in section IV.

IV. EXPERIMENTAL RESULTS

This study proposes an optimized Deep CNN system by redesigning fully connected layers for the binary skin cancer classification problem. Also, instead of using the usual metrics as acc or last model to choose the best model, we recommend methods using other metrics such as : 1) max of area under the curve (auc); 2) max of sensitivity (sen); 3) max of sensitivity + specificity, in other words, mean recall (sen+spe); 4) max of custom balanced accuracy (bacc) (formula 4).

We train our network with the ISIC 2019 dataset with 150 epochs as described in section III, then find best models based on the above four metrics. These best models are then evaluated by test-10 and their performance are compared with each other as well as with the two common methods, acc and last model.

Effectiveness measures: to evaluate the effectiveness of the binary skin cancer classification task, we rely on the 3 classical measures: area under the ROC curve (AUC), Sensitivity (SEN) and Specificity (SPE), all converted to p. Mathematically, ACC, SEN and SPE can be expressed with respect to true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) as equatation (1), (2) and (3). Also, we propose a metric called Custom Balanced Accuray (BACC) as in formula (4). Besides, YI (Youden index) [16] and Δ of the difference between SEN and SPE are calculated as in formula (5) and (6).

$$ACC = \frac{TP}{TP + TN} (1)$$

$$SEN = \frac{TP}{TP + FN} (2) \qquad SPE = \frac{TN}{TN + FP} (3)$$

$$BACC = \frac{\left(SEN + SPE - \left|\left(\frac{(SEN - SPE)}{2.0}\right)\right|\right)}{2.0} (4)$$

$$YI = SEN + SPE - 1 (5) \qquad \Delta = |SPE - SEN| (6)$$

In the binary classification problem using the CNN network with sigmoid function, the result returned by the network for each image input is a real number with a value in range [0 1]. Conversion of this value into negative and positive is done using a threshold, if prediction value is greater

than the threshold then the result is positive, otherwise negative. Normally the threshold is 0.5. However, with only one threshold of 0.5, the calculated sensitivity and specificity do not fully reflect the model's performance. In order to thoroughly analyze the system's performance, it is necessary to analyze the ROC curve and its critical thresholds. In this study, we use the threshold of 0.5 to compare the performance of best models on the test-10 dataset in Section IV.A, then evaluate the performance of these models on the MClass-D dataset with a threshold of 0.5 and analyzing the ROC curve in section IV.B.

A. Comparison of performances over the test-10 dataset

The Deep CNN network is trained on the training dataset of 13,842 photos. After training the network with 150 epochs, twelve best models of InceptionV3 and ResNet50 network architectures are selected based on the algorithm proposed in section III.B. Then, to evaluate their performance with the test-10 set of 1,730 skin lesion images (includes 450 melanoma images and 1,280 nevus images). At the threshold of 0.5, the AUC, SEN, and SPE performance of these models is shown in Figure 4.

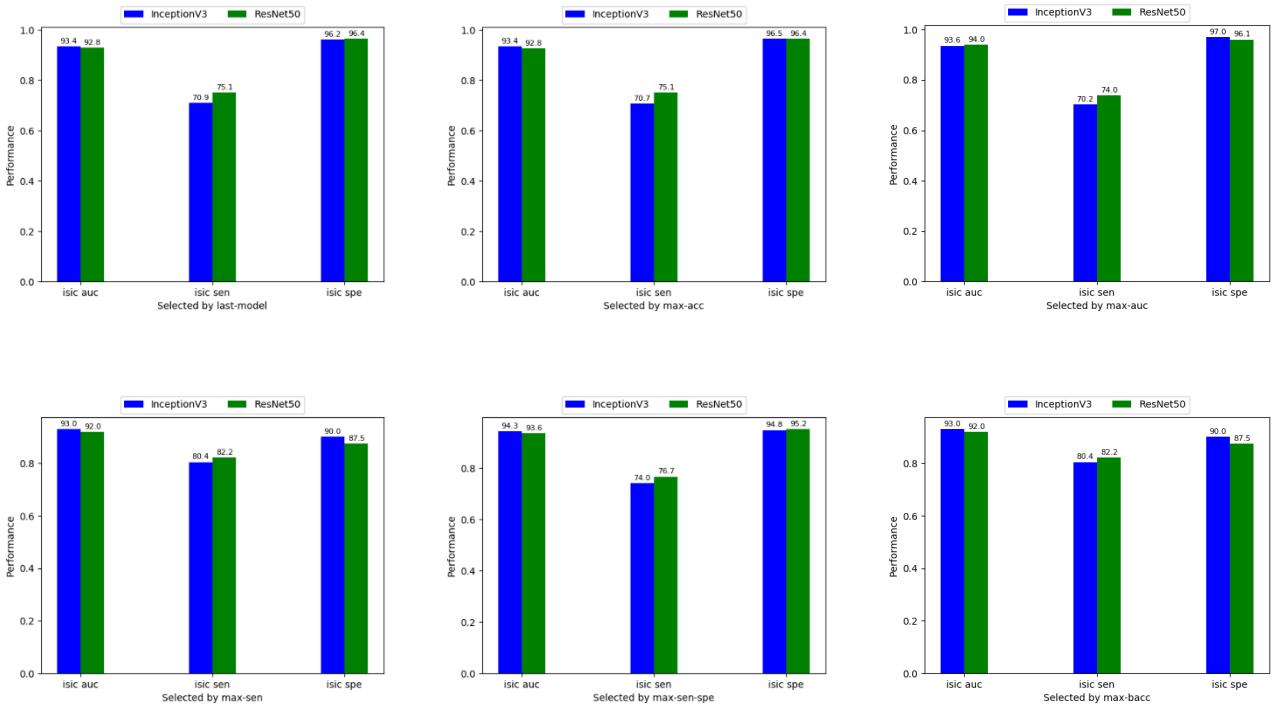

Fig. 4. Performances of twelve selected best models over the test-10 dataset using a prediction threshold of 0.5.

Overall, models built with InceptionV3 architecture performs better since most of the AUC values generated for InceptionV3 models are higher than the ResNet50's, except for the auc model. Moreover, the highest value of AUC among models built with IceptionV3 architecture (94.3%) was higher than the one with ResNet50 (94%). Among model selection methods, sen+spe model with InceptionV3 was the best with highest AUC value (94.3%). While our proposed sen and bacc models returned highest YI (0.704%) and lowest Δ (9.6%), the two popular methods last model and acc model had fairly low YI and received the highest Δ. In other words, the traditinal methods to select best model for a CNN in medical binary classification have a high proportion of misclassified images and do not preform well in balancing sensitivity and specificity. In summary, we can conclude from the test-10 set that, in terms of AUC, using InceptionV3 architecture with sen+spe best model selection method is the best solution for melanoma binary image classification.

B. Comparison with dermatologists over the MClass-D dataset

To compare the performance of the proposed solutions with 157 dermatologists, we uses the MClass-D dataset. Twelve best models are evaluated on this dataset with a threshold of 0.5 with results described in Table II. We also display analysis of the receiver operating characteristic curves of these best models in Figures 5 and 6 and compare our performance with SOTA and dermatologists' solutions at important thresholds shown in Table III.

TABLE II. PERFORMANCES OF TWELVE SELECTED BEST MODELS OVER THE MCLASS-D DATASET USING A PREDICTION THRESHOLD OF 0.5.

Methods	**AUC**	**SEN**	**SPE**	**YI**	**Δ**
InceptionV3					
last	87.0	45.0	97.5	0.425	52.5
acc	87.0	45.0	97.5	0.425	52.5
auc	86.1	50.0	96.2	0.462	46.2
sen	85.4	70.0	85.0	0.550	15.0
sen+spe	88.4	50.0	91.2	0.412	41.2
bacc	85.4	70.0	85.0	0.550	15.0
ResNet50					
last	80.0	60.0	95.0	0.550	35.0
acc	79.5	55.0	95.0	0.500	40.0
auc	81.9	55.0	93.8	0.488	38.8
sen	85.5	80.0	82.5	0.625	2.5
sen+spe	86.1	55.0	96.2	0.512	41.2

bacc	85.5	80.0	82.5	0.625	2.5

From the table, we can see that sen+spe is still the best model selection method with the highest AUC of 88.4% for InceptionV3 and 86.1% for ResNet50 architecture. Like the ISIC 2019 test results, sen and bacc are proved to handle imbalance sensitivity and specificity very well. Compare to the last model and acc model, sen and bacc have reduced the difference between sensitivity and specificity from 52.5% to 15 % with InceptionV3 architecture and from 35% and 40.0 to only 2.5% with ResNet50.

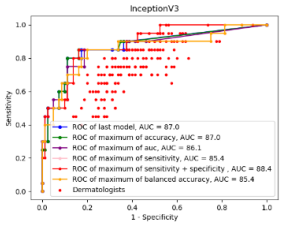

Fig. 5. The receiver operating characteristic curves of six InceptionV3's best models.

The Figure 5 displays the performance in terms of ROC of 6 InceptionV3 models. Generally, our proposed solutions have higher performance over most skin experts. The red line that represents the sen+spe model achieves the highest AUC (88.4%) and outperformes 153 over 157 dermatologists. Follows are last model and acc model (both with AUC of 87%), represented by the blue and green lines respectively.

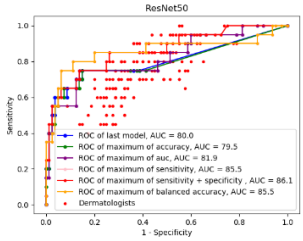

Fig. 6. The receiver operating characteristic curves of six ResNet50's best models.

Similarly, Figure 6 shows ROC curves of 6 models applied for the ResNet50 architecture. The sen+spe continues to be the best model with AUC reaches 86.1% and outperforms 140 out of 157 doctors. Meanwhile, last model and acc model had lowest AUC (90.0% and 79.5%).

TABLE III. PERFORMANCES BASE ON TEST-10 USING DIFFERENT THRESHOLDS OF SEN AND SPE.

Thres by	**InceptionV3**						**ResNet50**							
	last	**acc**	**auc**	**sen**	**sen+spe**	**bacc**	**last**	**acc**	**auc**	**sen**	**sen+spe**	**bacc**	**SOTA**	**Derm**
SEN (%)	**SPE**	**SPE**	**SPE**	**SPE**	**SPE**	**SPE**	**SPE**	**SPE**	**SPE**	**SPE**	**SPE**	**SPE**	**SPE**	**SPE**
90.0	63.7	65.0	62.5	66.2	62.5	66.2	0.0	0.0	41.2	58.8	50.0	58.8	-	
85.0	82.5	83.8	81.2	80.0	**83.8**	80.0	0.0	0.0	48.8	80.0	73.8	80.0	-	
76.7	88.8	88.8	81.2	82.5	86.2	82.5	0.0	0.0	53.8	88.8	75.0	88.8	-	65.8
74.1	88.8	88.8	88.8	83.8	86.2	83.8	85.0	85.0	86.2	93.8	86.2	93.8	86.5	60.0
SPE (%)	**SEN**	**SEN**	**SEN**	**SEN**	**SEN**	**SEN**	**SEN**	**SEN**	**SEN**	**SEN**	**SEN**	**SEN**	**SEN**	**SEN**
69.2	85.0	85.0	85.0	85.0	85.0	85.0	75.0	75.0	75.0	85.0	85.0	85.0	84.5	73.3
60.0	90.0	90.0	90.0	90.0	90.0	90.0	75.0	75.0	75.0	85.0	85.0	85.0	87.5	74.1

The Table III displays several thresholds and their corresponding sensitivity and specificity values on the test-10 set, in comparison with the SOTA and dermatologist's performance. At a mean sensitivity of 74.1%, our sen+spe model achieves specificity of 86.2%, which is almost equal to the SOTA's value and significantly surpasses dermatologists by 26.2%. More importantly, at the highest SEN achieved by private dermatologists of 76.7%, the sen+spe model has 20% higher performance in terms of SPE value. For a mean specificity of 69.2%, the sen+spe outperforms both the SOTA and dermatologists with the highest sensitivity of 85% (0.5% higher than the SOTA and 11.7% higher dermatologists). We also propose a more optimized threshold of 85% sensitivity and 83.8% specificity, at which the sensitivity and specificity are the most balanced but still high enough to assure the accuracy of the algorithm.

V. DISCUSSION

The results described in the previous section indicates that our solution has proved to perfom better both SOTA system and certified medical experts. First of all, our proposed best model selection method sen+spe achieves the highest AUC in all test sets and with both architecture InceptionV3 and ResNet50. To be precise, the maximum AUC of sen+spe in test-10 set is 94.3% and in MClass-D is 88.4%. These number means the model generates very little misclassified images. Secondly, the sen+spe method's performance outperforms total 153 of 157 dermatologists from different German university hospitals and surpasses the current best solution 12.5%. This result is visualized in Figure 5 and 6 and has been the greatest achievement so far. Finally, our CNN with customized fully connected layer provides a solution for underfitting and overfitting problems.

In summary, a customized CNN could apply for binary image classification in melanoma diagnosis to avoid underfitting, and its combination with the sen+spe metric to choose the best model significantly increase the result's quality. This provides excellent outcomes and even outperforms human qualification for this specific task.

VI. CONCLUSIONS

In this research, we propose a customized deep convolutional neural network architecture as well as analyze

a number of best model selection methods for the melanoma classification problem. Our major contributions are as follows.

1) The proposed best model selection method sen+spe outperforms two traditional method (last and acc) with the highest AUC on both test-10 and MClass-D set (94.3% and 88.4% respectively).

2) Our solution also outperforms 153 out of 157 dermatologists participated in the MClass-D dataset, which surpasses the current state-of-the-art solution 17 dermatologists.

3) The proposed InceptionV3 network with customized fully connected layers proved to solve underfitting issue and avoid overfitting.

The study provides a significant solution for the architecture designing and imbalance data issue in binary melanoma image classification. We also suggest that YI and the difference between sensitivity and specificity are also important metrics to evaluate deep neural networks' performance for medical image classification problems. With the limited resources and timeframe, we only conducted our experiment on one customized fully connected layer of two hidden layers. We believed that more intensive studies should be applied to develop and boarden this solution to other skin cancer typed as well as general medical image diagnosis prolems.

REFERENCES

[1] A. C. Society, "Cancer Facts & Figures 2018," 2018.

[2] "Skin Cancer Foundation." [Online]. Available: https://www.skincancer.org/skin-cancer-information/melanoma/.

[3] H. A. Haenssle *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.*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018, doi: 10.1093/annonc/mdy166.

[4] R. P. Braun, J.-H. Saurat, and L. E. French, "Dermoscopy of pigmented lesions: a valuable tool in the diagnosis of melanoma.," *Swiss Med. Wkly.*, vol. 134, no. 7–8, pp. 83–90, Feb. 2004, doi: 2004/07/smw-10318.

[5] T. J. Brinker *et al.*, "Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review," *J. Med. Internet Res.*, vol. 20, no. 10, p. e11936, Oct. 2018, doi: 10.2196/11936.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[7] T. J. Brinker *et al.*, "Deep neural networks are superior to dermatologists in melanoma image classification," *Eur. J. Cancer*, vol. 119, pp. 11–17, Sep. 2019, doi: 10.1016/j.ejca.2019.05.023.

[8] A. Hekler *et al.*, "Superior skin cancer classification by the combination of human and artificial intelligence," *Eur. J. Cancer*, vol. 120, pp. 114–121, Oct. 2019, doi: 10.1016/j.ejca.2019.07.019.

[9] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[10] T.-C. Pham, C.-M. Luong, M. Visani, and V.-D. Hoang, "Deep CNN and Data Augmentation for Skin Lesion Classification," 2018, pp. 573–582.

[11] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Exp. Dermatol.*, vol. 27, no. 11, pp. 1261–1267, Nov. 2018, doi: 10.1111/exd.13777.

[12] T. J. Brinker *et al.*, "Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark," *Eur. J. Cancer*, vol. 111, pp. 30–37, Apr. 2019, doi: 10.1016/j.ejca.2018.12.016.

[13] M. A. Marchetti *et al.*, "Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images," *J. Am. Acad. Dermatol.*, vol. 78, no. 2, pp. 270-277.e1, Feb. 2018, doi: 10.1016/j.jaad.2017.08.016.

[14] A. Kallner, "Formulas," in *Laboratory Statistics*, Elsevier, 2018, pp. 1–140.

[15] P. Tschandl *et al.*, "Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks," *JAMA Dermatology*, vol. 155, no. 1, p. 58, Jan. 2019, doi: 10.1001/jamadermatol.2018.4378.

[16] T. J. Brinker *et al.*, "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *Eur. J. Cancer*, vol. 113, pp. 47–54, May 2019, doi: 10.1016/j.ejca.2019.04.001.

[17] "ISIC 2019," 2019. [Online]. Available: https://challenge2019.isic-archive.com.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Dec. 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015.

[20] R. S. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res. 15*, p.) 1929-1958, 2014.