A Comparative Study for Classification of Skin Cancer

Tri Cong Pham, Giang Son Tran, Thi Phuong Nghiem, Antoine Doucet, Chi Mai Luong, Van-Dung Hoang

A comparative study for classification of skin cancer

Tri Cong Pham[1]
School of Computer Science and Engineering, Thuyloi University
175 Tay Son, Dong Da, Hanoi, Vietnam
phtcong@tlu.edu.vn

Giang Son Tran
[1]*ICTLab, University of Science and Technology of Hanoi, VAST*[*]
tran-giang.son@usth.edu.vn

Thi Phuong Nghiem
[1]*ICTLab, University of Science and Technology of Hanoi, VAST*[*]
nghiem-thi.phuong@usth.edu.vn

Antoine Doucet
L3i Lab, University of La Rochelle
Av M. Crepeau, 17042 La Rochelle, France
antoine.doucet@univ-lr.fr

Chi Mai Luong[1]
Institute of Information Technology VAST[*]
lcmai@ioit.ac.vn

Van-Dung Hoang[**]
QuangBinh University
Dong Hoi, Quang Binh, Vietnam
dunghv@qbu.edu.vn

***Abstract*—Skin cancer is one of the most common types of cancer all over the world. It is easily treatable when it is detected in its beginning stage. Melanoma is the most dangerous form of skin cancer. Early detection of melanoma is important in reducing the mortality rate of skin cancer. Recently, machine learning has become an efficient method in classifying skin lesions as melanoma or benign. Main features for this task include color, texture and shape. A comparative study about color, texture and shape features of melanoma is useful for future research of skin cancer classification. Inspired by this fact, our study compares the classification results of 6 classifiers in combination with 7 feature extraction methods and 4 data preprocessing steps on the two largest datasets of skin cancer. Our findings reveal that a system consisting of Linear Normalization of the input image as data preprocessing step, HSV as feature extraction method and Balanced Random Forest as classifier yields best prediction results on the HAM10000 dataset with 81.46% AUC, 74.75% accuracy, 90.09% sensitivity and 72.84% specificity.**

***Index Terms*—Skin Cancer, Classification, Feature Extraction, Melanoma.**

I. Introduction

Melanoma is one of the most malignant, metastatic and dangerous types of skin cancer that causes a majority of deaths related to skin cancer. It was estimated that in 2018 there were about 91,270 new cases of skin cancer from melanoma with 9,320 deaths [1]. Geller et al., 2007 [2] indicated that melanoma is a curable disease if it is diagnosed early and correctly. Due to this, it is necessary to examine and observe melanoma closely when it is still at the early stage.

In order to detect skin cancer from melanoma, besides clinical tests, dermatologists often use their eyes to examine characteristics of skin lesions such as color, texture and shape to diagnose if the lesion is a benign or malignant tumor. Nowadays, advances in technologies allow the widely use of dermoscopy images in examining and diagnosing melanoma skin cancer. To support this task, many computer aided diagnosis (CAD) systems are designed to detect melanoma from dermoscopy images.

One important step of the CAD system for melanoma skin cancer is to classify if the melanoma skin lesion is benign or malignant. Due to this, many methods are proposed in the literature to detect malignant melanoma from skin lesions. Hamd et al., 2013 [3] proposed a method to predict skin cancer from symmetry and color matching for lesion pigments. In detail, the method detects and segments lesion edges to compute symmetrization for all images to isolate benign tumor. The suspicious images are then nominated into one of the three classes: Melanoma, Basal Cell Carcinoma (BCC), or Squamous Cell Carcinoma (SCC) tumor based on symmetrization and pigment-color matching score table. The experimental results of two matching procedures are compared to 40 pre-classified images where 80% of true classification is obtained for the first procedure and 92.5% is for second procedure.

Celebi at el. [4] presented a method to classify pigmented skin lesions from dermoscopy images using color, texture and shape features. For color and texture features, the image is divided into a set of regions presenting significant clinical properties of the lesions. The extracted feature data are fed into an optimization framework to rank the features for finding the optimal subset of features. For shape features, the method performs lesion border detection to separate lesion from background skin. The detected border is then used to extract shape features of the lesions. The method obtains a specificity of 92.34% and a sensitivity of 93.33% on a set of 564 tested images.

Barata et al., 2014 [5] introduced two systems for melanoma detection in dermoscopy images. The first system uses global features to classify skin lesions while the second system uses local features and the bag-of-features classifier to categorize skin lesions. The experimental results showed that color features obtained better performance than texture features

[*] Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam
[**] Corresponding: Van-Dung Hoang, dunghv@qbu.edu.vn

when used alone for the problem of skin lesion classification. Besides, both texture and color features achieve good classification results of 96% sensitivity and 80% specificity for global features and 100% sensitivity and 75% specificity for local features on 176 dermoscopy images.

Since there are many methods in the literature to perform melanoma classification of skin cancer, a comparative study of these methods will be helpful in summarizing and demonstrating the best methods of melanoma classification. In this paper, we perform a throughout experiment to classify melanoma using pattern recoginition techniques, including a set of 4 preprocessing methods, 7 feature extraction methods and 6 different classifiers on two largest public melanoma dataset and summarize our findings from this in-depth evaluation.

The rest of the paper is described as follows: we present the materials and methods we used in section II. An analysis of results and our discussion are explained in section III. Finally, the conclusion and future works are presented in section IV.

II. Materials and Methods

In this section, we describe the dataset and the methodology that we used for our experiments.

A. Dataset

We perform evaluations on two different datasets, organized by two challenges: ISIC 2016 [6] and HAM10000 [7]. Both datasets contain photos of Melanoma cases, divided into two classes as Melanoma and Benign.

The ISIC 2016 challenge dataset contains 900 images with labels, including 172 Melanoma images - 728 Benign images. The second dataset ISIC 2018 consists of 10,015 images in total, including 1,113 Melanoma images - 8,902 Benign images. The images are in high resolutions, can be approximately up to $1,800 \times 1,200$ pixels. Figure 1 shows several examples of melanoma cases and benign cases provided by the two given datasets.

We perform our evaluations, including training and testing, on each dataset separately to determine the effectiveness of the dataset size. For each dataset, we use 90% of the images as training data and the remaining 10% as testing data.

B. Methodology

Our classification evaluation procedure is illustrated on figure 2, including 4 main steps:

- Data Preprocessing
- Feature Extraction
- Melanoma Classification
- Result Analysis

In the upcoming sections, we will describe experiment methods in each step.

1) Data Preprocessing: Firstly, due to the variety of input image resolution, we perform proportional scale of each input image to the image with 600 pixels in width. We then propose to use three different preprocessing methods: Gaussian Blur, Normalization and combination of Gaussian Blur and Normalization.

Gaussian Blur (GB) is a common way to reduce noise since the dataset images are collected from several sources. We use standard 5×5 Gaussian blur filter with $\sigma = 1.1$.

Additionally, each dataset itself is collected from several sources, therefore its brightness and contrast should be normalized. We perform **linear normalization** (LN) to $[0, 255]$ range for each pixel provided in each input image as follows:

$$n(x, y) = 255 \times \frac{i(x, y) - min}{max - min}, \tag{1}$$

in which $i(x, y)$ denotes the original input signal and $n(x, y)$ denotes the corresponding normalized value.

In our experiments, we also use a combination of GB and LN to verify the effectiveness of this mixture toward Melanoma classification result.

2) Feature Extraction: Many machine learning tasks require a feature selection step, reducing the number of dimensions from the feature space. It is mainly done by removing redundant, noisy and unimportant features. This step brings several benefits: reducing feature extraction time, reducing complexity for the next classification step, improving prediction results, reducing training and testing time. Each extracted feature is a vector representing an input image. We propose to use below features for different aspects of an input image:

HSV (Hue-Saturation-Value) [8] represents the color features of the input image. We convert the input image (in RGB format) to HSV color space and calculate a 3D histogram for all channels (H, S, V), each divided into 8 bins. We then flatten this 3D histogram to achieve a color feature vector of 512 ($= 8 \times 8 \times 8$) dimensions.

LBP (Local Binary Pattern) [9] is a visual descriptor, representing textures of the input image. The input image is divided into 8×8 cells, each pixel in this cell is compared with its neighbour, providing a number for each pixel. We calculate the histogram of each cell, combine all together and perform normalization. The result is a 242-dimensional feature vector, representing textures of the input image.

HOG (Histogram of Oriented Gradients) [10] is another visual descriptor of an input image by counting gradient orientations of localized regions. The output of HOG is a shape feature vector composed of 65,520 dimensions.

SIFT (Scale-Invariant Feature Transform) [11] extracts keypoints of an input images, regardless of image transformation, scaling and rotation. The SIFT keypoints are then used for calculate the similarities of images.

Not only are the previous features assessed separately, we also evalute the effectiveness when using them in combination.

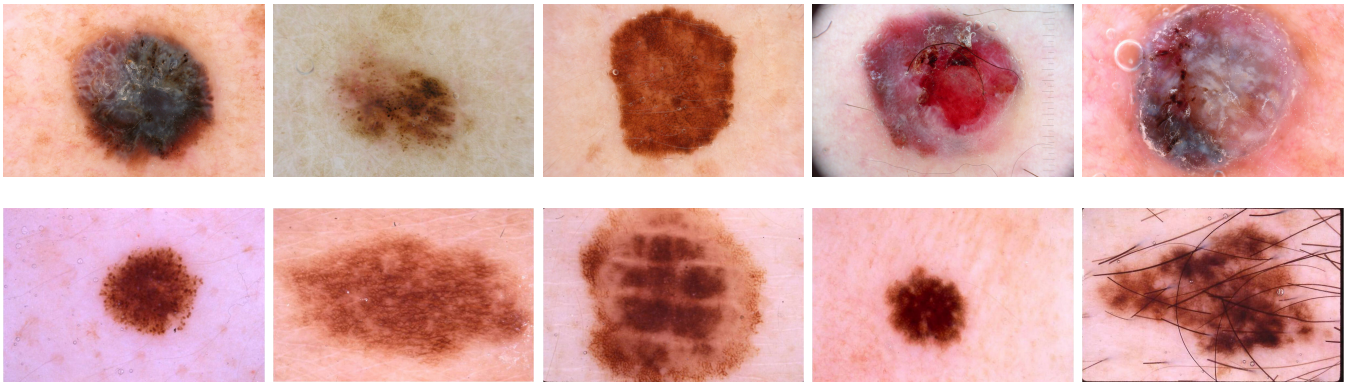

Fig. 1. Examples of images from the two datasets. First row: Melanoma cases. Second row: benign cases.

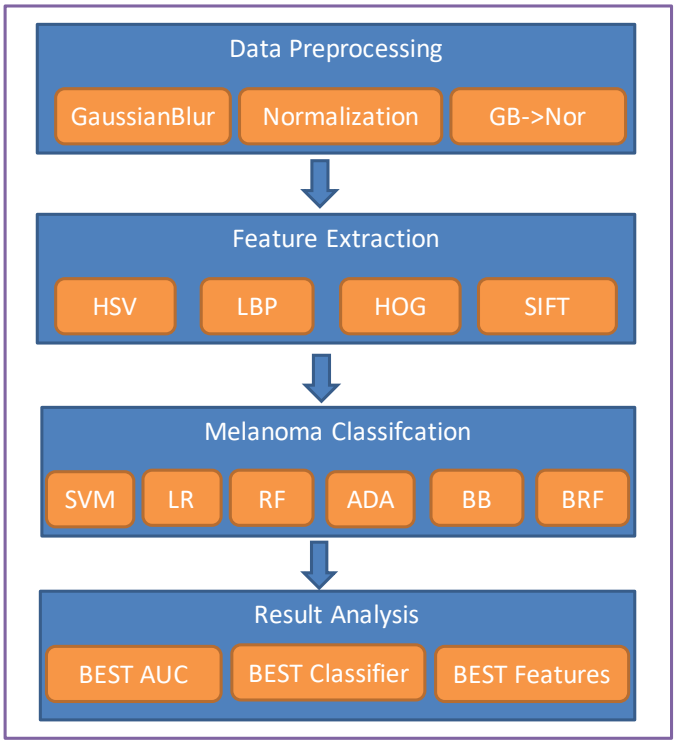

Fig. 2. Evaluation procedure.

Firstly we attempt to use the extracted features of HSV, LBP and HOG all together without transformation. Secondly, we evaluate their effectiveness of an additional linear normalization step after combination of them, e.g. HSV, LBP and HOG with normalization. Finally, an additional PCA (Principal Component Analysis) [12] is added after combining the 3 features, resulting in HSV, LBP and HOG with PCA, to evaluate the potency of dimension reduction.

To summarize, we use the following 7 feature extraction methods for evaluation of their effectiveness for Melanoma classification problem: HSV, LBP, HOG, SIFT, HSV+LBP+HOG, HSV+LBP+HOG with normalization, HSV+LBP+HOG with PCA.

3) Classification: The third step to solve the Melanoma classification problem is to perform classification, using the extracted features from previous step. This is a supervised learning task: for each image, its extracted features with their corresponding label (Melanoma or Benign) are fed to the classification model so that it can learn from the dataset. We evalute the following 6 models for the Melanoma classification problem due to their popularity in good performance on various datasets:

SVM (Support Vector Machine) is a discriminative classifier defined by a separating hyperplane [13]. The labelled training data are divided by an optimal hyperplane which could be used for categorizing new, untrained data.

LR (Logistic Regression) [14] is a common technique for solving binary classification problem. It uses a logistic function to model a binary dependent variable, then use this trained function to classify untrained data.

RF (Random Forest) [15] is a classification technique that leverage usage of multiple decision trees, each contains leaves representing class labels and branches representing conjunctions of features that lead to those labels. The trained decision trees are then used in a randomized fashion (therefore called Random Forest, in an attempt to overcome overfitting nature of decision trees) to classify an untrained data.

AdaBoost (Adaptive Boosting) [16] is used in our evaluation with Decision Trees in order to improve classification results. Adaboost gathers information at each stage of decision tree about the hardness of each training sample so that later trees can focus more on harder-to-classify examples.

BB (Balanced Bagging) [17] is another accuracy-improving algorithm that can be used with other classification methods. In our evaluation, we attempt to improve Decision Tree using BB method by constructing multiple decision trees and then combine their predictions to provide the final result.

BRF (Balanced Random Forest) [18] is an adaptive improvement of Random Forest to handle unbalanced datasets (i.e. the number of one class outweights the number of the other class). This approach tries to overcome the unbalanced

number of samples for each class by attempting to train the samples equally (for both classes) in the bootstraping stage.

4) Comparison Metrics: After the training process is finished, the model is then used for untrained images (test data) to evaluate. To compare effectiveness of each combination (preprocessing, feature extraction and classification), we use the following metrics:

Accuracy is measured as the ability to differentiate the Melanoma and benign cases correctly, defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Sensitivity is measured as the ability to determine the Melanoma cases correctly, defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

Specificity is measured as the ability to determine the benign cases correctly, defined as:

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

In these formulas, TP (true positive) represents the number of cases correctly identified as Melanoma; FP (false positive) represents the number of cases incorrectly identified as Melanoma; TN (true negative) represents the number of cases correctly identified as benign; and FN (false negative) represents the number of cases incorrectly identified as benign.

AUC (Area Under Curve) is a higher level metrics, combining the true positive rate (TPR, same as sensitivity) and false positive rate ($FPR = \frac{FP}{FP+TN}$, indicating how well a classification system distinguishes between the positive class and the negative class. When using only specificity and sensitivity separatedly, it is difficult to determine whether or not a classification method is overfit with positive samples (high Sensitivity) or overfit with negative samples (high specificity). Therefore, AUC is proposed in many classification system as the ultimate metric to measure their effectiveness, both for positive class and negative class. AUC is calculated as the area below a Receiver Operating Characteristic curve, composing of different combination of TPR and FPR when the classification threshold varies. Higher AUC value denotes the classification method is closer to a perfect prediction system.

Finally, we perform our experiments on a HP server, consisting of an Intel Xeon 2620 v3 (6 cores, 12 threads), 32GB of DDR4 memory and a NVIDIA GeForce GTX 1080.

III. Results and Discussion

In this section, we summarize our evaluation results for the previously described steps (preprocessing: original, GB, LN, GB + LN; feature extraction: HSV, LBP, HOG, SIFT, HSV+LBP+HOG, HSV+LBP+HOG with LN, HSV+LBP+HOG with PCA; classifier: SVM, LR, RF, AdaBoost, BB, BRF). For each combination of a dataset, a preprocessing method, a feature extraction method and a classifier, our experiment consists of a training step (using 90% the provided dataset) and a prediction step (using the remaining 10% data) and a prediction result is provided, consists of accuracy, sensitivity, specificity and AUC values. In total, we have 1344 metric results (2 datasets $\times$ 4 preprocessing methods $\times$ 7 feature extraction methods $\times$ 6 classifiers $\times$ 4 metrics). We mainly focus on the AUC metrics as they represent the ability of the whole experiment both for positive class and the negative class.

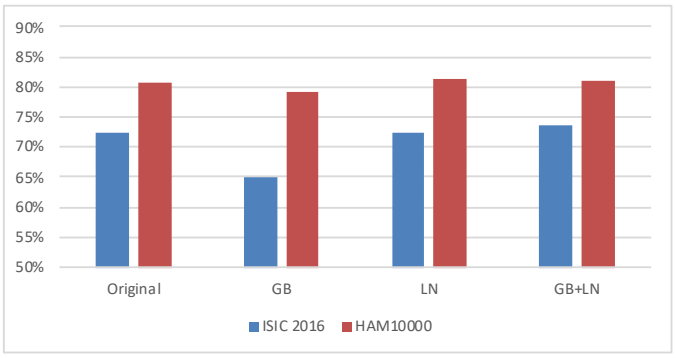

Fig. 3. Best AUC results of preprocessing methods on two datasets.

A. Dataset

Firstly we compare the effectiveness of classification results with regard to dataset used for training and testing. Our thorough evaluation obtains best AUC values of 81.46% and 73.37% on HAM10000 and ISIC 2016, respectively. The former result (81.46%) is achieved on HAM10000 using linear normalization as preprocessing, HSV as feature extraction and Balanced Random Forest as classifier. The later result (73.37%) is with ISIC 2016 using original image (no preprocessing), LBP as feature extraction and also with Balanced Random Forest as classifier. With this result, it can be inferred that bigger dataset can provide better classification accuracy.

B. Preprocess

In this section, we evaluate the effectiveness of preprocessing methods for each given dataset. Particularly, we evaluate the best results achievable using the original image, gaussian blur (GB) filter, linear normalization (LN) and a combination of GB and LN. Figure 3 illustrates our results in this regard.

It can be seen that the best prediction results are given by Linear Normalization on HAM10000 dataset and by Gaussian blur with Linear Normalization on ISIC 2016 dataset. Our best AUC values of 81.46% is achieved with Linear Normalization as data preprocessing on HAM10000 dataset, in combination with HSV (as feature extraction method) and Balanced Random Forest (as classifier).

The upcoming result analysis sections only consider HAM10000 since most of the AUC results on ISIC 2016 are significantly inferior than the HAM10000 counterpart.

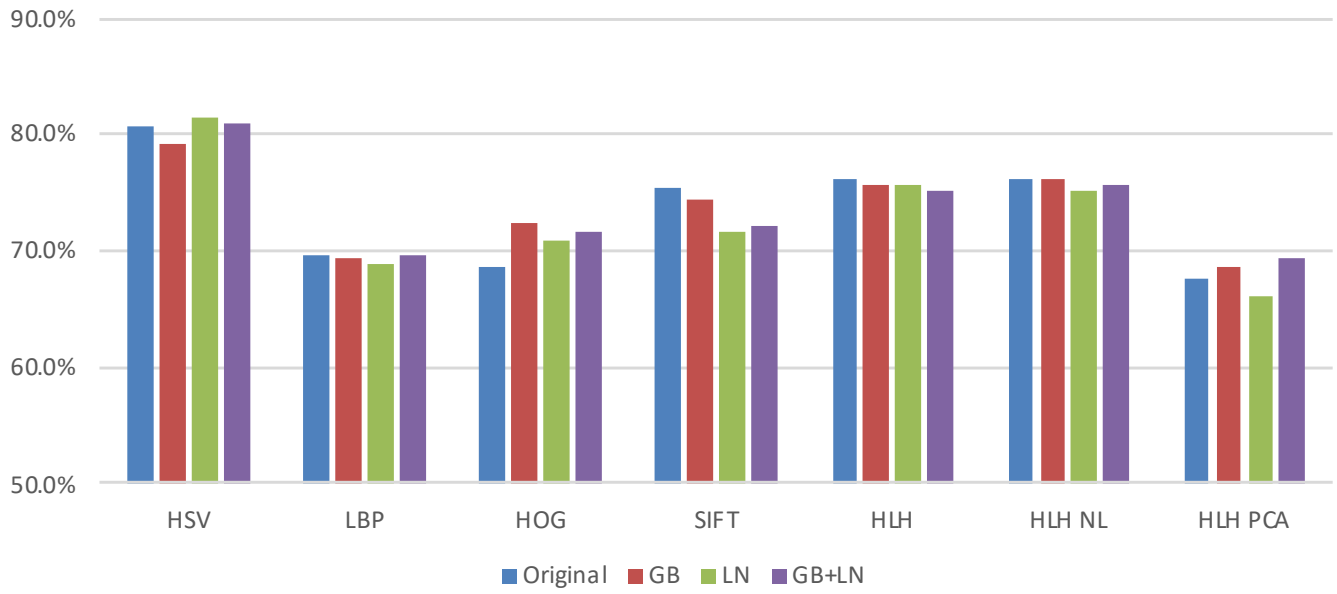

Fig. 4. Best AUC results of different feature extraction methods with previously discussed preprocessing methods.

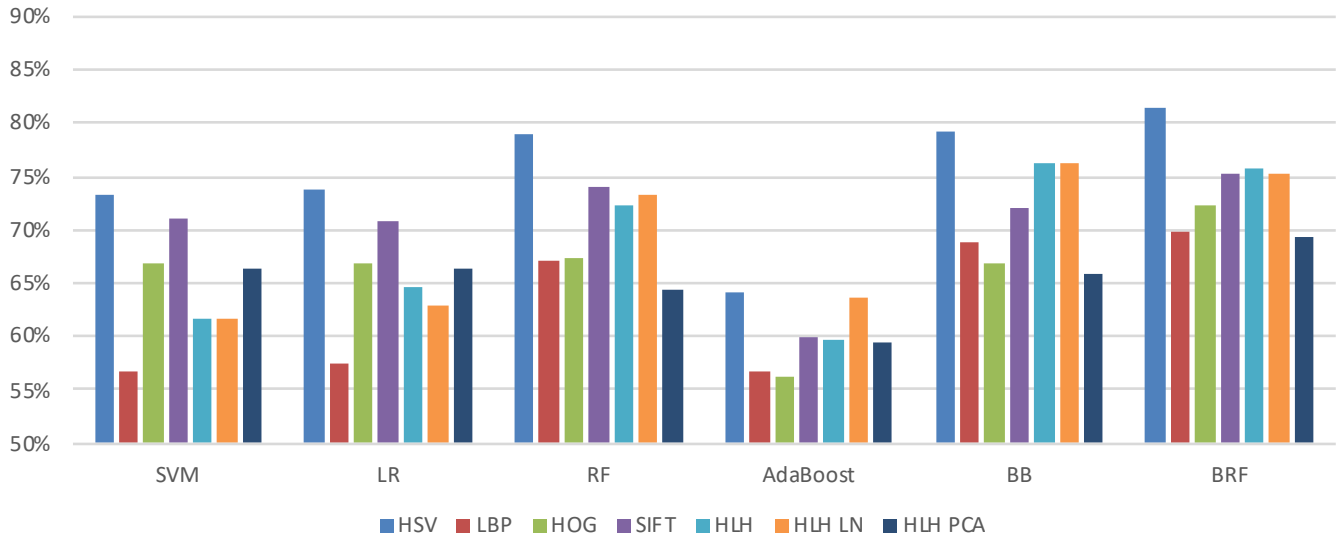

Fig. 5. Best AUC results of different classifiers with previously discussed feature extraction methods.

C. *Features*

In this section, we analyze the values of different features with regard to final prediction results. As previously discussed, the features being tested include HSV, LBP, HOG, SIFT, and combinations of HSV+LBP+HOG, HSV+LBP+HOG linearly normalized (HLH LN), HSV+LBP+HOG with principal component analysis (HLH PCA). Figure 4 summaries our findings in this regard.

An interesting result from this visualization is the superiority of HSV features (81.46% AUC) when compared with other feature extraction methods (65.99% - 76.22% AUC), even better than the combination of HSV+LBP+HOG. HSV alone represents color feature of the input image and is faster than other methods to calculate. Therefore, it can be concluded that color plays a very important part in identifying and classifying Melanoma from a candidate image. When being combined with other methods, the prediction result is reduced.

HSV+LBP+HOG with principal component analysis (HLH PCA), unexpectedly, does not compete well with other methods. Its prediction results are among the worst (from 65.99% to 69.31%), when compared with other methods (HSV, LBP, HOG, SIFT, HSV+LBP+HOG (HLH) and HSV+LBP+HOG

TABLE I
CLASSIFIER AUC PERFORMANCE AND ITS STANDARD DEVIATION WHEN USED WITH OTHER FEATURE EXTRACTION METHODS

Metrics	SVM	LR	RF	AdaBoost	BB	BRF
Mean	64.92%	65.70%	70.49%	60.24%	71.77%	73.51%
Std Dev	5.78%	5.34%	5.72%	2.69%	5.80%	4.88%

linearly normalized (HLH LN)).

Our experiment concludes that HSV is the best feature to use for Melanoma.

D. Classifier

Finally, we analyze the effectiveness of various classification methods to provide prediction results. The classifiers being used include Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), AdaBoost (Adaptive Boosting), Balanced Bagging (BB) and Balanced Random Forest (BRF). Figure 5 illustrates our findings.

In general, BRF achieves best AUC prediction score when compared with other feature extraction methods. Notably, when used with HSV, it achieves highest score (81.46% AUC) in all combinations.

Insisting on the variance of prediction results for each classifier, table I shows its average AUC values and the corresponding standard deviations when used with our previously discussed feature extraction methods. Among these methods, SVM's, LR's and RF's performances are unstable. Although they can reach the average AUC up to 70.49%, their stability when used with different features is not as low as BRF (4.88% STD). On the other side, Adaboost, being the most steady classifier, predicts with lowest AUC values among these classifiers. Therefore, we can conclude that BRF is the best classifier for our experiments.

Among all of our evaluations, the configuration providing best Melanoma prediction results in terms of AUC is as follows:

- Linear Normalization of the input image as data preprocessing step
- HSV as feature extraction method
- Balanced Random Forest as classifier

The prediction results of such a configuration is 81.46% AUC, 74.75% accuracy, 90.09% sensitivity and 72.84% specificity.

IV. CONCLUSION AND PERSPECTIVES

In this paper, we perform an in-depth evaluation for Melanoma classification using machine learning. In particular, we proposed a classification system consisting of a preprocessing step, a feature extraction method and a classifier. We experimented with 4 different methods for data preprocessing steps, 7 feature extraction methods and 6 classifiers. Our experiment results show that: using the HAM10000 dataset is better than the ISIC 2016 counterpart; Linear Normalizaion on HAM1000 provides better prediction results that other preprocessing techniques; HSV is the best model for feature extraction; and Balanced Random Forest is the best classifier.

This Melanoma classification work can be continued by applying more modern preprocessing step (such as data augmentation), feature extraction and classifier (such as convolutional neural network). We believe that using neural network can furtherly improve classification results.

REFERENCES

[1] R. Segal, K. Miller, and A. Jemal, "Cancer statistics, 2018," *CA Cancer J Clin*, vol. 68, pp. 7–30, 2018.

[2] A. C. Geller, S. M. Swetter, K. Brooks, M.-F. Demierre, and A. L. Yaroch, "Screening, early detection, and trends for melanoma: current status (2000-2006) and future directions," *Journal of the American Academy of Dermatology*, vol. 57, no. 4, pp. 555–572, 2007.

[3] M. H. Hamd, K. A. Essa, and A. Mustansirya, "Skin cancer prognosis based pigment processing," *International Journal of Image Processing*, vol. 7, no. 3, p. 227, 2013.

[4] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Computerized Medical imaging and graphics*, vol. 31, no. 6, pp. 362–373, 2007.

[5] C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques, "Two systems for the detection of melanomas in dermoscopy images using texture and color features," *IEEE Systems Journal*, vol. 8, no. 3, pp. 965–979, 2014.

[6] D. Gutman, N. C. F. Codella, M. E. Celebi, B. Helba, M. A. Marchetti, N. K. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)," *CoRR*, vol. abs/1605.01397, 2016. [Online]. Available: http://arxiv.org/abs/1605.01397

[7] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, p. 180161, 2018.

[8] J.-q. Ma, "Content-based image retrieval with hsv color space and texture features," in *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*. IEEE, 2009, pp. 61–63.

[9] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1657–1663, 2010.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] D. G. Lowe *et al.*, "Object recognition from local scale-invariant features." in *iccv*, vol. 99, no. 2, 1999, pp. 1150–1157.

[12] I. Jolliffe, *Principal component analysis*. Springer, 2011.

[13] B. Schölkopf, C. J. Burges, A. J. Smola *et al.*, *Advances in kernel methods: support vector learning*. MIT press, 1999.

[14] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft margins for adaboost," *Machine learning*, vol. 42, no. 3, pp. 287–320, 2001.

[17] S. Hido, H. Kashima, and Y. Takahashi, "Roughly balanced bagging for imbalanced data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 2, no. 5-6, pp. 412–426, 2009.

[18] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, "An empirical study of learning from imbalanced data using random forest," in *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, vol. 2. IEEE, 2007, pp. 310–317.