A Genetic Programming Approach to Feature Construction for Ensemble Learning in Skin Cancer Detection

Qurrat Ul Ain, Harith Al-Sahaf, Bing Xue, Mengjie Zhang
School of Engineering and Computer Science
Victoria University of Wellington, P.O. Box 600, Wellington 6140, New Zealand
{Qurrat.Ul.Ain,Harith.Al-Sahaf,Bing.Xue,Mengjie.Zhang}@ecs.vuw.ac.nz

ABSTRACT

Ensembles of classifiers have proved to be more effective than a single classification algorithm in skin image classification problems. Generally, the ensembles are created using the whole set of original features. However, some original features can be redundant and may not provide useful information in building good ensemble classifiers. To deal with this, existing feature construction methods that usually generate new features for only a single classifier have been developed but they fit the training data too well, resulting in poor test performance. This study develops a new classification method that combines feature construction and ensemble learning using genetic programming (GP) to address the above limitations. The proposed method is evaluated on two benchmark real-world skin image datasets. The experimental results reveal that the proposed algorithm has significantly outperformed two existing GP approaches, two state-of-the-art convolutional neural network methods, and ten commonly used machine learning algorithms. The evolved individual that is considered as a set of constructed features helps identify prominent original features which can assist dermatologists in making a diagnosis.

CCS CONCEPTS

• **Computing methodologies** → **Genetic Programming**; *Ensemble methods*; Feature selection; • **Mathematics of computing** → *Dimensionality reduction*; • **Human-centered computing** → Information visualization.

KEYWORDS

Genetic Programming, ensemble classifiers, feature construction, melanoma detection, multi-class classification

1 INTRODUCTION

Skin cancer is a major public health problem, with over five million newly diagnosed cases every year in the United States [32]. In 2019, the global incidence of skin cancer was estimated to be over 104,350 cases, with almost 11,650 deaths [32]. Melanoma is the most serious form of skin cancer, which becomes life-threatening if not treated early [23]. Although the mortality is significant, when detected early melanoma survival exceeds 95% [23]. Since this cancer is visible on the skin, it is potentially detectable at a very early stage when it is curable. Because skin cancer treatment outcomes are substantially improved by early diagnosis, better diagnostic techniques using artificial intelligence and computer vision techniques are in great demand.

Automated skin cancer recognition from images is a very challenging task due to the presence of hair, gel, and reflection artifacts in the skin image, the huge intra-class variations with each cancer type, and the high degree of inter-class visual similarity between various types of skin cancers. These factors are the main obstacles in extracting useful information from skin lesion images, thereby stimulating the need to formulate methods that can capture informative features. The computer aided diagnostic (CAD) methods are expected to somehow mimic the medical properties such as the ABCD (Asymmetry, Border irregularity, Color variation and Dermoscopic structure) rule of dermoscopy [33], and the 7-point check-list method (Asymmetry, Streaks, Blue-whitish veil, Dots, Regression areas, Pigment network, and presence of six colors: red, white, light-brown, dark-brown, blue-gray, black) [5]. To incorporate these visual characteristics, the CAD systems utilize various texture, color, frequency, local, and global features to include as much information as possible. Using a single type of features may not help the classification algorithm to achieve good results. Moreover, such diagnostic systems are potentially useful, which not only diagnose a type of cancer quickly in real-time situations but also identify significant features effectively to help the dermatologist learn the critical visual patterns from these skin lesions.

The original set of features extracted from images may include redundant or irrelevant features, and may not contain enough information for accurately classifying these images. In such cases, feature selection (FS) and feature construction (FC) methods help pick important features and generate new high-level features from the original set of features to achieve improved performance [1, 34]. Genetic Programming (GP), a biologically inspired evolutionary algorithm, evolves models in successive generations to solve a specific problem by applying genetic operators such as crossover and mutation. GP keeps improving the evolved models iteratively by measuring their goodness using a fitness function. The evolved models can be considered as a classifier or a constructed feature, depending on the problem at hand. GP has been used successfully for FS and FC. However, generating new features specific to a single classifier may fit the training data too well, thereby producing poor results on the test data. An ensemble of classifiers combine the predictions of multiple classifiers and hence, each classifier contributes to produce more accurate results [15]. This study combines the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
GECCO '20, July 8–12, 2020, Cancun, Mexico
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7128-5/20/07…$15.00
https://doi.org/10.1145/3377930.3390228

benefits of feature construction and ensemble classification in a GP framework to construct informative features for the tasks of binary and multi-class skin cancer image classification.

In the recent years, convolutional neural networks (CNNs) have become popular in skin image analysis. Codella et al. [11] used the Caffe architecture to perform feature extraction. Esteva et al. [12] used a huge private dataset which consists of both clinical and dermoscopy images to train an Inception network from scratch, aiming at a performance close to a human expert. However, the deep learning approaches typically required thousands of images to effectively train a model, and due to a "black-box architecture", the models may not directly provide insights of prominent features. In addition, using a pre-trained CNN generally requires pre-processing a dataset to the same input configurations for which that CNN was originally designed for such as fixed-size images, and RGB or gray-scale images, which increases the computation time and decreases flexibility to apply to any size of image.

In order to deal with all these limitations, we are interested to combine feature construction and ensemble learning using a Genetic approach. Having multiple classifiers in an ensemble is expected to generate more generic and informative features suitable to multiple classifiers, thereby promoting generalization of the evolved ensemble classifier. This works aims at investigating the following objectives:

- Design a new feature construction method using GP to generate new features for an ensemble of classifiers.
- Assess the performance of the proposed classification method in comparison to bagging, boosting, random forests, other commonly used machine learning classification algorithms, and the existing deep learning and GP methods on two real-world skin cancer image datasets.
- Visualize the multiple constructed features and identify prominent image features.

2 BACKGROUND

2.1 Feature extraction Methods

To convert images into feature vectors, the following common feature extraction methods are reviewed, which are also used in the proposed method.

2.1.1 Local Binary Patterns (LBP). It is a dense image descriptor which has been successfully applied in computer vision applications for feature extraction [27]. LBP scans the whole image in a pixel-by-pixel fashion by using a sliding window of fixed radius. At each location of the window, it computes the value of the central pixel according to the intensity values of the neighboring pixels situated on the radius. These computed values are used to generate a LBP histogram (feature vector). LBP is divided into uniform and non-uniform patterns. There are 2^8 bins but only 58 of them are uniform. These uniform patterns have individual bins in the LBP histogram, while non-uniform patterns are binned together in one bin, making a total of 59 LBP features. Uniform patterns symbolize edges, corners and flat regions in an image, while non-uniform patterns cannot provide much textural information as shown in Figure 1. In skin lesions, uniform patterns allow detection of corners

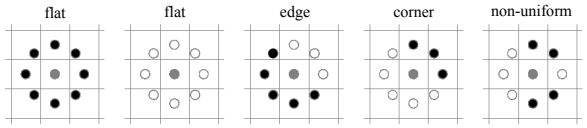


Figure 1: Examples of Uniform and non-uniform LBP patterns.

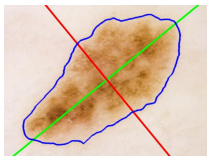


Figure 2: The greatest and shortest diameter for a dermoscopy image.

(lesion boundary), streaks (line ends) and dots (flat regions), which may help distinguish between different types of skin cancers.

2.1.2 Lesion Color Variation. Color is a significant component of the asymmetry, border, color, and diameter (ABCD) rule [33] as well as the 7-point checklist method [5]. These are the key medical properties which play a vital role in classifying a skin lesion. The number of colors present in a lesion generates huge variance in the RGB color space. This makes the features extracted from RGB color channels highly informative to discriminate between different classes of images. These global color features are extracted from the RGB color channel pixels in the segmented skin lesions. The mean (μ) and variance (σ^2) of each channel is represented, respectively, as μR, μG, μB and $\sigma^2 R$, $\sigma^2 G$, $\sigma^2 B$. To include complex non-uniform color distributions inside the lesion area, ratios of the mean values are also computed, i.e., $\frac{\mu_R}{\mu_G}$, $\frac{\mu_R}{\mu_B}$, $\frac{\mu_G}{\mu_B}$. To incorporate color variations of the lesion area with respect to the surrounding skin area, some other ratios are also calculated such as $\frac{\mu_R}{\overline{\mu}_R}$, $\frac{\mu_G}{\overline{\mu}_G}$, $\frac{\mu_B}{\overline{\mu}_B}$, where $\overline{\mu}$ shows the mean value of surrounding skin area. These 12 global color variation features are adapted from [31].

2.1.3 Lesion Geometrical Shape. Border shape and geometrical properties of a lesion demonstrate significant diagnostic information for detecting a type of cancer [14]. There are standard geometry features such as area, perimeter, greatest diameter, asymmetry index, circularity index, irregularity index A, and irregularity index B, which are adopted from [22]. Moreover, some other geometry features such as shortest diameter, irregularity index C, irregularity index D, and major and minor asymmetry indices are adopted from [14]. The details of these measures can be found in [22] and [14]. The greatest and shortest diameter of a lesion are shown in Figure 2 which seems important in capturing the shape of the lesion.

2.1.4 Wavelet Decomposition. The pyramid-structured wavelet analysis [10] captures both the local (detailed structure and internal texture) and global (overall properties) information of the lesion. We apply three-level pyramid-structured wavelet decomposition on red, blue, green, and luminance color channels of the skin images. The luminance is calculated as:

$$luminance = (0.3 \times R) + (0.59 \times G) + (0.11 \times B) \quad (1)$$

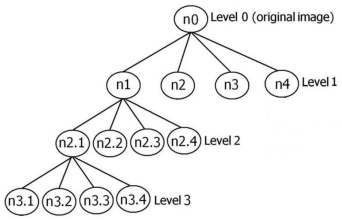


Figure 3: A schematic three-level wavelet tree with nodes in circle.

To extract informative features from the wavelet coefficients, eight statistical measures and ratios are used which include energy, mean, standard deviation, skewness, kurtosis, norm, entropy, and average-energy. The details and mathematical expressions of these measures can be found in [14]. Figure 3 shows a schematic representation of wavelet tree where circles represent nodes. There are 13 nodes in the wavelet tree (1 parent node which is the original image, and 4 nodes in each of the three subsequent levels ($4 \times 3 = 12$)). The eight measures computed on each tree node yield a total of 8×13 features, for each color channel. Hence, there are a total of 416 (= 8 measures $\times$ 13 nodes $\times$ 4 color channels) wavelet features extracted in this study.

2.2 Related Work

Earlier in 2012, Barata et al. [8] developed an automatic system for detection of pigment network in dermoscopy images. The system uses a set of three sequential steps: 1) pre-processing to remove hair and reflection artefacts, 2) detecting lines and pigment networks inside lesion area in skin images using directional Gabor filters, and 3) extracting features from the detected network to train an AdaBoost algorithm. The system provided good results for detection of pigment network, however, it does not provide information to further classify lesions into benign and malignant classes.

Ferris et al. [13] developed a computer-assisted diagnosis of dermoscopic images for classification of melanoma using random forest. This system computes 54 features from the lesion area. Fitness measures are sensitivity, specificity and area under the curve (AUC), showing trade-off between sensitivity and specificity. A study was conducted to compare the sensitivity and specificity of the classifier 30 dermatology clinicians. The classifier produced better sensitivity than dermatologists, however, produced lower specificity than dermatologists. Before extracting features, lesions are manually segmented which is time consuming and requires expert knowledge.

The use of ensembles of CNNs has been recently utilized for skin cancer image classification, which have shown promising results. Harangi et al. [17] used an ensemble of AlexNet, VGGNet, and GoogLeNet, and their results show that the ensemble-based approach outperformed all of its member CNNs. Valle et al. [35] explored ensembles of CNNs and transfer learning. Their results conclude that ensembles of models are a cost-effective alternative to the unstable sequential designs. Xie et al. [36] developed an artificial neural network (ANN) based ensemble model to identify tumors as benign or malignant. However, these deep learning approaches required thousands of images to effectively train a model, thereby need huge computing resources that most universities and research institutes cannot afford. Moreover, they remained unable to identify prominent features.

Identification of suitable data augmentation methods have gained immense importance recently, which can generally cope well with the limited size of datasets [29]. Transfer learning has gained attention, which has been explored with and without fine-tuning [25]. Moreover, other relevant criteria such as image size and selected architecture in CNNs has recently been studied [35]. Such methods require a lot of extra pre-processing work such as parameter tuning and identifying suitable data augmentation strategies.

Garnavi et al. [14] developed a CAD system to classify melanoma by employing various texture, border, and geometrical features. This diagnostic system by selecting an optimal feature set achieved an overall accuracy of 91.26%, with only 23 features. However, various types of feature are not combined in a suitable way which might limit the classification performance. Kawahara et al. [18] demonstrated how filters from a pre-trained CNN can be used to classify 10 classes of non-dermoscopy images in the Dermofit dataset [6]. However, they reported a standard overall classification accuracy of 81.80% for the highly imbalanced Dermofit dataset, which is not suitable as it may give biased results towards classes with more images.

Recently, Brinker et al. [9] proved that automated melanoma image classification using CNN achieved significantly better results than board-certified dermatologists. Barata et al. [7] used pre-trained DenseNet-161 architecture to perform a hierarchical diagnosis for three skin cancer classes. Additionally, they provided comparative studies on the importance of color normalization, lesion segmentation, and evaluation metrics. Their method required the same input configurations on which the architecture was originally trained. Generally, reducing the size of a skin image may distort aspect-ratio which may result in losing informative features.

GP has been widely explored for image analysis [4, 19, 30]. Ryan et al. [30] described a fully automated procedure to perform Stage-1 breast cancer detection using GP. It is a multi-stage method that mainly implements pre-processing, breast segmentation and feature extraction. Results revealed the ability of GP to produce human-readable solutions while being capable of examining the GP individuals. Al-Sahaf et al. [4] designed a novel GP-based image descriptor for multi-class texture image classification. Lensen et al. [19] showed that GP can automatically select regions of interest, extract informative features from these regions, and perform classification to achieve improved classification performance. Tran et al. [34] developed a feature selection and construction method using GP to improve classification performance on high-dimensional data. Ain et al. [2] proposed a binary classification method to effectively identify melanoma in images using GP. They have revealed the insights of the evolved GP individuals, which are not only automatically generated classification models but also are human-readable which may help a dermatologist learn the informative features to make diagnosis in real-world situations.

3 THE PROPOSED METHOD

This section presents our proposed algorithm, i.e., multiple feature construction with ensemble classification (MFCEC) using GP for skin cancer image classification.

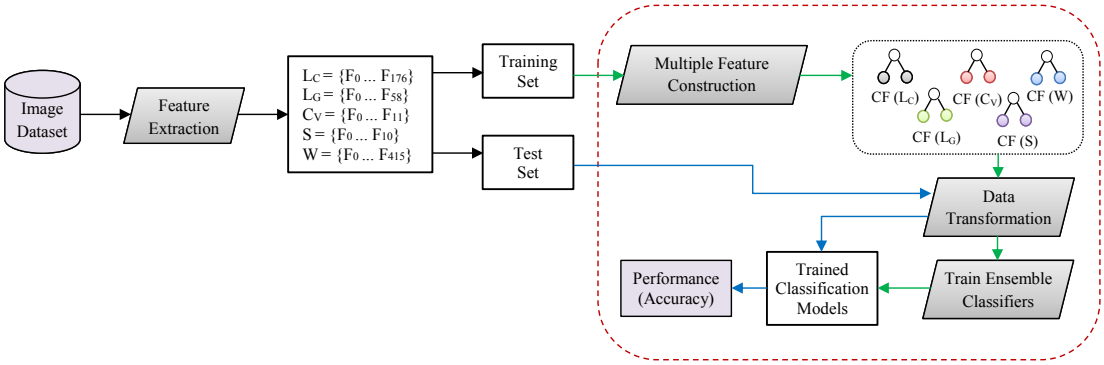


Figure 4: The workflow of the proposed algorithm.

3.1 Terminal Set and Function Set

The terminal set consists of five feature sets:

- LBP extracted from RGB color channel images (L_C),
- LBP extracted from gray image (L_G),
- global color variation features (C_V),
- lesion geometrical shape features (S), and
- frequency-based wavelet features (W).

The L_C feature vector consists of 177 (=59 × 3) LBP features extracted from red, green, and blue color channels concatenated together to make a single feature vector. L_G consists of 59 LBP features extracted from gray skin image. In C_V, S, and W feature vectors, there are 12, 11, and 416 features, respectively. These different types of features are employed in order to include sufficient information regarding texture, color, and domain-specific border shape properties of skin images which may help to achieve good performance. In this work, a GP individual consists of five trees where each tree evolves from a single set of features as its terminals.

The function set consists of seven operators: four arithmetic $\{+, -, \times, /\}$, two trigonometric $\{\sin, \cos\}$, and one conditional $\{if\}$ operator. Among the arithmetic operators, division is protected and the other three arithmetic operators have the standard arithmetic meaning. Protected division returns zero when a number is divided by zero. The $\{if\}$ operator takes four inputs; it returns the third input if the first input is greater than the second input, else it returns the fourth input.

3.2 Program Representation and Fitness Function

A GP individual is usually constructed as a single tree. However, it can be utilized to construct multiple trees in a single individual, which is called *multi-tree GP* [26]. In this work, using the multi-tree GP approach, GP constructs five trees in one individual during the evolutionary process. The five trees represent the five constructed features (CFs). Each CF is constructed from one and only one type of features as described in the Section 3.1. These CFs are utilized to transform the original training and test sets to new training and test sets. The transformed training set with five new CFs is provided as input to the ensemble classification algorithm, which is formed by Support Vector Machines (SVMs), Decision Trees (J48), and Random Forest (RF). These classifiers are trained on the training data during the evolutionary process. MFCEC uses the accuracy produced on the training data by the ensemble classifier as its fitness function, where each image is classified based on the majority voting. The balanced accuracy is used, which is defined as follows:

$$fitness = \frac{1}{m}\sum_{i=1}^{m}\frac{TP_i}{TP_i + FN_i} \tag{2}$$

where m is the number of classes. TP and FN are the true positives and false negatives, respectively. The ratio $\frac{TP_i}{TP_i+FN_i}$ shows the true positive rate of one class. The intuition behind using balanced accuracy is that the two medical datasets are highly imbalanced. To avoid bias towards the majority class, we used balanced accuracy instead of standard overall accuracy throughout in this work.

3.3 Crossover and Mutation

Since the proposed method requires to construct each tree from a single set of features, we have used *same-index-crossover/mutation* [20]. Through experiments, it has been observed that mixing different types of features to evolve a single tree does not provide good discriminative CFs. During the evolutionary process, GP evolves five trees (CFs) in a single individual namely CF(L_C), CF(L_G), CF(C_V), CF(S), and CF(W) as shown in Figure 4. The same-index-crossover/mutation ensures that CF(L_C) in one GP individual can only crossover/mutate with CF(L_C) of another GP individual, and it cannot crossover/mutate with CF(L_G), CF(C_V), CF(S), and CF(W). More details can be found in [2] and [20].

3.4 The Overall Algorithm

The overall structure of the proposed method is presented in Figure 4. First the images are transformed to feature vectors using the feature extraction methods described in Section 2. Note that the feature extraction is performed before the training and test data split because features are extracted image by image, so test images are not used for extracting features from training images, i.e., no bias produced. For each image, we get five feature vectors, namely (L_C), (L_G), (C_V), (S), and (W). The dataset is then divided into training and test sets. GP utilizes the training set to construct multiple features in one GP individual. Each tree in a GP individual is considered one CF. These constructed features are expected to have more discriminating ability between classes as compared to the original

Table 1: Real-world skin cancer image datasets.

Name	Classes	Instances	Image size	Optical Device
PH^2	Common Nevi	80	$763 \times 553 - 769 \times 577$	Dermatoscope
	Atypical Nevi	80	$764 \times 575 - 768 \times 576$	
	Melanomas	40	$764 \times 576 - 768 \times 576$	
Dermofit	Actinic Keratosis	45	$193 \times 221 - 777 \times 702$	Standard Camera (non-dermoscopy)
	Basal Cell Carcinoma	239	$189 \times 206 - 1341 \times 1130$	
	Melanocytic Nevus / Mole	331	$177 \times 189 - 857 \times 828$	
	Squamous Cell Carcinoma	88	$269 \times 273 - 1341 \times 1097$	
	Seborrhoeic Keratosis	257	$189 \times 229 - 1825 \times 1329$	
	Intraepithelial carcinoma	78	$565 \times 265 - 2176 \times 2549$	
	Pyogenic Granuloma	24	$292 \times 235 - 1870 \times 1834$	
	Haemangioma	96	$328 \times 193 - 914 \times 890$	
	Dermatofibroma	65	$436 \times 338 - 1498 \times 1492$	
	Melanoma	76	$367 \times 439 - 3055 \times 1630$	

Table 2: GP Parameter Settings.

Parameter	Value	Parameter	Value
Generations	50	Initial Population	Ramped half-and-half
Population Size	1024	Selection type	Tournament
Crossover Rate	0.80	Tournament size	7
Mutation Rate	0.19	Tree depth	2–6
Elitism	0.01		

sets of features. SVM, J48, and RF are selected as the three classifiers in the ensemble as they show the best performance among other settings of classifier selection in the ensemble. Since multiple classification algorithms (SVM, J48, and RF) are incorporated as an ensemble to use these CFs as input, the CFs constructed are generic to all the classification algorithms. The CFs are non-tailored to one specific classifier, rather generated regardless of which classifier is used to classify them. After finishing the evolutionary process, we get the three (SVM, J48, and RF) trained classification models. The original test set is transformed by utilizing the same CFs to a new test set. This new test set is used to evaluate these trained models to get the test performances. The highest accuracy produced among the three models is selected as the test accuracy.

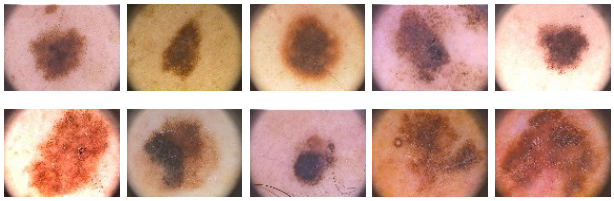
4 EXPERIMENT DESIGN

4.1 Benchmark Datasets

The dataset is divided using *10-fold cross validation* where nine folds are used for training and one for testing. The proposed method is examined on two real-world skin cancer image datasets: 1) PH^2 [24], and 2) Dermofit Image Library [6]. Details of these datasets are given in Table 1. The size of images in PH^2 dataset is almost same (≈768 ×570), however there is a huge variation in image sizes in Dermofit dataset ranging between 177×189 and 3055×1630. A specialized instrument called dermatoscope is used to capture the images in PH^2, whereas Dermofit has standard camera images. Among the three classes in PH^2, since atypical nevi refers to moles which are currently non-malignant but may develop melanoma later, this class is combined with common nevi which refers to moles, to form one class called "benign" to perform binary classification experiments. For Dermofit, the two classes: Melanocytic Nevus and Melanoma are used for binary classification experiments. For multi-class classification experiments, PH^2 has three classes (easy task) and Dermofit has ten classes (difficult task).

4.2 Benchmark Techniques

In this study, we compare the performance of our proposed method with ten commonly used machine learning algorithms: Naïve Bayes



(a) PH^2

(b) Dermofit

Figure 5: Image samples from the two benchmark datasets.

(NB), k-Nearest Neighbor (k-NN), Support Vector Machines (SVMs), Decision Trees (J48), Multi-layer Perceptron (MLP), Random Forest (RF), Bagging, AdaBoost, LogitBoost, and Random Committee. The number of k is set to 5 in k-NN. SVM uses a Radial Basis Function (RBF). In RF, the number of trees and the maximum depth of a tree are set to 10 and 5, respectively. In MLP, the momentum, learning rate, training epochs and the number of units in one hidden layer are 0.2, 0.1, 60, and 20, respectively. These settings are adopted from a previous study [3], where they have been empirically searched via experiments. All other settings are set to default as in the Waikato Environment for Knowledge Analysis (WEKA) package [16]. The ten classification algorithms are trained one time on the five sets of features, appended to make a single feature vector with 675 (= 177 L_C + 59 L_G + 12 C_V + 11 S + 416 W) features. The trained classifiers are then tested to obtain their test performance.

For GP implementation, the Evolutionary Computation in Java (ECJ) package is used [21]. We also compare MFCEC with the two existing GP approaches for skin cancer image classification:

- Embedded-GP [2] uses four types of features (L_C, L_G, C_V, and S) to evolve four trees in its GP individual. Since this is an embedded approach where GP also performs classification, each tree acts as a binary classifier. The best tree with highest accuracy on the training data is used to test the performance on the test data.
- Wrapper-GP [3] uses five types of features explained in Section 2 to evolve five trees in a single GP individual. These trees act as CFs to be classified by a machine learning algorithm such as decision tree. The trained model is applied on the test set to check the performance of this method.

In addition, we compare MFCEC with the state-of-the-art CNN methods recently developed for the PH^2 and Dermofit datasets:

- Patiño et al. [28] developed a lesion segmentation and classification method using morphological operations to estimate

Table 3: Results of binary classification on the two real-world skin cancer datasets (in terms of Sensitivity, Specificity, and balanced accuracy, where ↑ and + signs show the results of applying statistical significance tests).

Algorithm		PH^2			Dermofit		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Non-GP Methods	NB	60.00	94.38	77.19 ↑	97.32	96.67	**96.99** ↑
	SVM	25.00	99.38	62.19 ↑	27.68	100.0	63.84 ↑
	k-NN	57.50	90.63	74.06 ↑	76.07	98.79	87.43 ↑
	J48	57.50	87.50	72.50 ↑	93.57	97.27	95.42 ↑
	MLP	62.50 ± 4.59	95.82 ± 3.36	**78.66 ± 2.43** +	92.50 ± 2.50	98.79 ± 1.20	96.29 ± 1.53 +
Ensemble Methods	RF	55.00	98.13	76.56 ↑	61.79	99.70	80.74 ↑
	Bagging	80.71	90.35	**76.56** ↑	90.15	97.64	93.46 ↑
	AdaBoost	59.05	87.41	68.44 ↑	96.75	99.41	**98.30** ↑
	LogitBoost	70.12	87.82	70.00 ↑	97.78	99.12	97.82 ↑
	RandomCommittee	85.17	89.25	74.69 ↑	94.92	96.54	91.54 ↑
Embedded-GP [2]	–	73.65 ± 4.92	84.09 ± 5.10	78.87 ± 2.92 +	75.82 ± 3.08	73.32 ± 3.45	74.57 ± 1.86 +
Wrapper-GP [3]	NB	86.42 ± 1.16	93.12 ± 0.70	**89.77 ± 1.84** +	95.60 ± 0.49	97.22 ± 0.32	96.21 ± 1.09 +
	SVM	73.83 ± 1.27	99.13 ± 0.25	86.48 ± 2.35 +	95.18 ± 0.60	99.61 ± 0.11	**97.26 ± 1.25** +
	k-NN	30.00 ± 0.68	96.68 ± 0.28	63.34 ± 2.67 +	73.00 ± 0.45	99.42 ± 0.13	86.04 ± 2.52 +
	J48	77.08 ± 0.08	98.14 ± 0.04	87.61 ± 3.08 +	95.11 ± 0.54	99.14 ± 0.51	96.99 ± 0.70 +
MFCEC	–	98.46 ± 1.67	100.0 ± 0.00	**98.75 ± 1.64**	99.13 ± 1.79	100.0 ± 0.00	**99.86 ± 0.21**

asymmetry, border and color features of the lesions in the PH^2 dataset. The method incorporated SVM, logistic regression and a fully connected neural network where the neural network has shown the best performance achieving 86.5% on average for multi-class classification.

- Kawahara et al. [18] trained a logistic regression classifier with deep features extracted from a convolutional neural network, pre-trained on natural images, to classify ten classes of skin lesions in the Dermofit dataset. They reported a standard overall accuracy of 81.80%, whereas the balanced accuracy computed from the confusion matrix is 60.12%.

4.3 Parameter Settings

The parameters set for GP are listed in Table 2. GP keeps improving the performance of the ensemble classifier by either iterating over a maximum of 50 generations, or a perfect ensemble classifier is generated giving 100% accuracy on the training data.

5 EXPERIMENT RESULTS

The results are represented as the mean and standard deviation ($\bar{x} \pm s$) of the 30 GP runs, and are listed in Table 3. Since *10-fold cross validation* is used, the result of one GP run is the mean of the accuracies of the 10-folds. *Wilcoxon signed-rank test* (with a significance level of 5%) is applied to compare MFCEC to the other stochastic methods. *One-sample t-test* is applied to compare MFCEC to the other deterministic methods. For Wilcoxon signed-rank test, "+", "−" or "=" represents that MFCEC is significantly better, worse, or similar to the other algorithms. For one-sample t-test, ↑ or ↓ represents that MFCEC is significantly better or worse to the other algorithm.

5.1 Binary Classification

The binary classification results are presented in Table 3. Among the non-GP methods, MLP has shown the highest accuracy 78.75% on PH^2, whereas NB produced the best accuracy 96.99% on Dermofit. Among the four ensemble methods, Bagging outperformed the other three methods giving 76.56% accuracy on the dermoscopic (PH^2) dataset. AdaBoost showed the highest accuracy 98.30% on the standard camera (Dermofit) dataset. Although the Embedded-GP approach provided good accuracy on dermoscopic datasets outperforming all the non-GP and ensemble methods, they remain unable to achieve good results for standard camera images, where ensemble methods dominated all the non-GP and Embedded-GP methods. Similarly, among the non-GP, ensemble, Embedded-GP and Wrapper-GP methods, Wrapper-GP produced the best results on dermoscopic datasets, whereas ensemble AdaBoost method remain prominent on Dermofit images. However, MFCEC produced the best results among all the methods achieving 98.75% and 99.86% accuracies on dermoscopic and standard camera images, respectively. This shows that feature construction in ensemble learning has huge potential to solve complex real-world problems like melanoma detection. The main reason of dominance of MFCEC over Wrapper-GP is that MFCEC constructs features for an ensemble of classifiers which are expected to be more general as compared to features constructed for a single classifier in Wrapper-GP.

Table 4: Results of multi-class classification on the two real-world skin cancer datasets in terms of balanced accuracy.

Algorithm		PH^2	Dermofit
Non-GP Methods	NB	**71.00** ↑	45.92 ↑
	SVM	59.50 ↑	51.08 ↑
	k-NN	65.50 ↑	43.54 ↑
	J48	58.00 ↑	50.08 ↑
	MLP	67.50 ± 3.47 +	**64.92 ± 4.31** +
Ensemble Methods	RF	**71.50** ↑	47.92 ↑
	Bagging	**71.50** ↑	62.38 ↑
	AdaBoost	56.50 ↑	29.46 ↑
	LogitBoost	66.50 ↑	**62.62** ↑
	RandomCommittee	70.00 ↑	58.38 ↑
Wrapper-GP [3]	NB	80.31 ± 2.03 +	58.99 ± 1.25 +
	SVM	84.92 ± 2.31 +	53.05 ± 1.57 +
	k-NN	63.46 ± 2.55 +	47.46 ± 1.85 +
	J48	**85.82 ± 1.60** +	**74.05 ± 1.52** +
MFCEC	–	**98.03 ± 0.85**	**85.20 ± 1.20**

5.2 Multi-class Classification

The multi-class classification results are presented in Table 4. Among the five non-GP algorithms, RF achieved the best accuracy 71.50% on PH^2, whereas MLP achieved the best accuracy 64.92% on Dermofit. Similar to binary classification results, among the ensemble methods, bagging provided the best results for PH^2

whereas Boosting (LogitBoost) provided highest accuracy for Dermofit. Wrapper-GP with J48 outperformed all the non-GP and ensemble methods providing an increase in accuracy by around 14% and 7% on average on the PH^2 and Dermofit datasets, respectively. It is worthwhile to note here that PH^2 has 3 classes and Dermofit has 10 classes (more difficult). For Wrapper-GP, most of the single classifiers are performing well for a 3-class problem such as SVM and J48 producing 84.92% and 85.82% average accuracy, respectively, however, only J48 performed well enough for the complex 10-class problem reaching 74.05% average accuracy. MFCEC remained prominent among all the methods in multi-class classification as well achieving 98.03% and 85.20% on average on the PH^2 and Dermofit datasets, respectively.

From the results of the statistical tests presented in Table 4, clearly MFCEC outperformed all the non-GP, ensemble as well as the Wrapper-GP methods on the easy (PH^2) and difficult (Dermofit) datasets, which shows its effectiveness for these complex skin cancer image classification problems.

5.3 Comparison to the State-of-the-arts

For PH^2, the most recent state-of-the-art reported by Patino et al. [28] achieved 86.5% balanced accuracy using 10-fold cross validation. Since the experimental setup is the same as MFCEC, we can make a direct comparison. MFCEC outperformed this method by providing an increase of nearly 11% accuracy. To the best of our knowledge, the state-of-the-art result on Dermofit for this 10-class skin image classification problem is presented by CNNs [18]. The authors reported an overall accuracy of 81.80% using 5-fold cross validation which came out to be 60.12% balanced accuracy (as calculated from the confusion matrix provided in the study). Since comparison cannot be done directly (5-folds vs 10-folds), we have provided a general idea what accuracy has been achieved by the current state-of the art on Dermofit dataset.

6 FURTHER ANALYSIS

6.1 Overall analysis

The average of best-of-generation fitness value of the 30 independent GP runs using different seed values on the training data of the PH^2 dataset in multi-class classification experiments is depicted in Figure 6. The plot shows how the accuracies of individual classifiers (SVM, J48, and RF) progress with the increase in generations and how much each of them contribute to the ensemble classification curve. Since elitism is applied on the ensemble classification and not on the individual classifiers, the individual classifiers' accuracies show behaviours of increase and decrease during the evolutionary process. However, they ensure that the collective performance increases as the number of generations increase. The benefit of using ensemble of classifiers is evident from this plot which clearly illustrates that if one classifier cannot produce good results, the ensemble can still rely on other classifiers to maintain good performance. From this plot, we observe that RF and J48 are producing far better results individually than SVM. However, when there is a decrease in the performance of RF and J48 in the subsequent generation, SVM makes larger jumps to maintain or even improve the performance of the ensemble classifier. This behaviour is seen in the third and fourteenth generations.

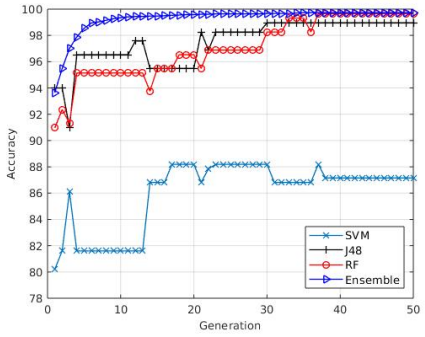


Figure 6: Graph between generation and accuracy values for SVM, J48, RF, and ensemble of these three classifiers.

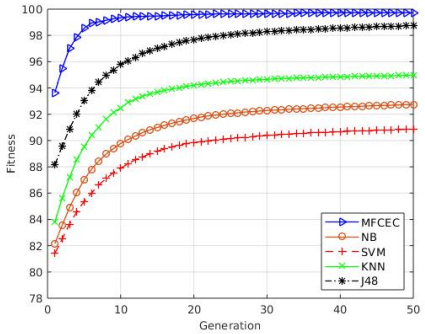


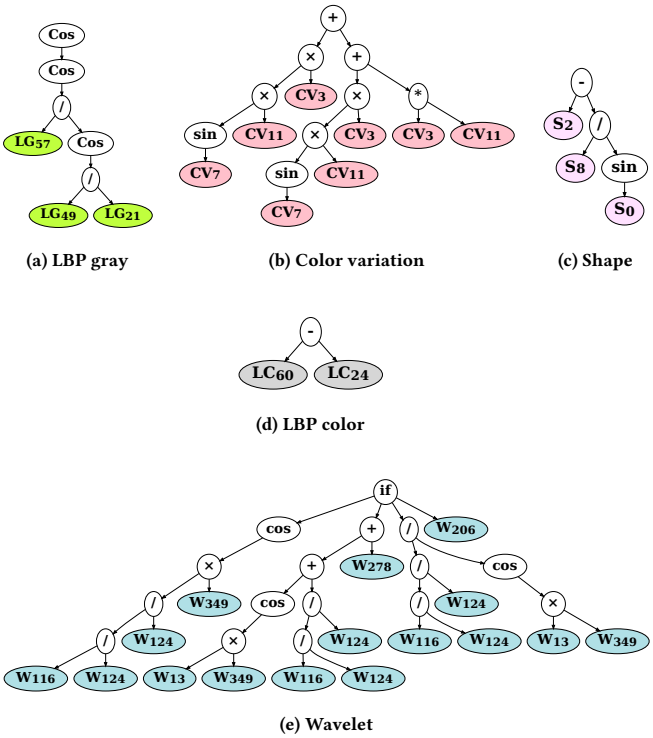
Figure 7: Graph between generation and accuracy values to compare MFCEC and the four existing Wrapper-GP [3] with NB, SVM, k-NN, and J48, respectively.

We also compare the evolutionary process of MFCEC with the previous Wrapper-GP [3] method as shown in Figure 7. The MFCEC curve shows an abrupt increase in the first five generations, being more powerful it achieves good performance in a very few earlier generations. However, the existing Wrapper-GP individual classifiers (NB, SVM, k-NN, and J48) start with lower average accuracy than MFCEC, thereby get the chance of making larger jumps as shown in first twenty generations. It is evident that MFCEC remained prominent and outperformed all the Wrapper-GP methods.

6.2 Analysis of an evolved GP program

GP has the ability to evolve models that can be interpretable. To analyse why MFCEC can achieve good performance, we show a good evolved GP individual in Figure 8. This individual is taken from the PH^2 experiments for the binary classification task. It has five trees evolved using the five types of features: a) L_G, b) C_V, c) S, d) L_C, and e) W. These CFs achieved 100.0% fitness produced by the ensemble classifier, where SVM produced 99.33%, J48 produced 99.83%, and RF produced 100% accuracy on the training data. Hence, selecting the highest performing RF model when applied to the test data, produced 100% accuracy on the test data. In Figure 8, colored nodes represent terminals (each color represents one type of features) and white nodes represent functions.

With the ability of FS and FC, GP plays a vital role in dimensionality reduction. From the evolved GP individual shown in Figure



(a) LBP gray (b) Color variation (c) Shape

(d) LBP color

(e) Wavelet

Figure 8: A good MFCEC evolved individual on the PH^2 dataset in the binary classification task.

8, GP has selected only 6 features among a total of 416 wavelet features, only 2 features among the 177 L_C features, only 3 features among the 59 L_G features, only 3 features among the 12 C_V features, and 3 features among the 11 S features. The wavelet texture-based features appearing in a tree of the GP individual shown in Figure 8(a) are listed in Table 5. The following conclusions can be derived from this table: 1) three out of six features belong to the nodes from the third level, which indicates our use of three-level wavelet decomposition as further decomposition may not obtain informative features for the purpose of classification, 2) texture features extracted from all the four color channels are selected to construct this informative CF, 3) the selected features are derived from both the low and the middle frequency channels as shown by the node column in Table 5, 4) among the eight statistical measures, norm, kurtosis, and entropy are prominent selected features. Moreover, the sub-trees "cos($W_{13} \times W_{349}$)" and "(W_{116} / W_{124}) / W_{124}" appear twice and thrice, which shows the potential of these subtrees getting selected multiple times to construct this informative wavelet-based CF.

Among the L_C and L_G features, the CFs shown in Figure 8(a) and (b) selected prominent LBP patterns corresponding to corners, edges and flat areas in these skin lesion images. Edges and corners identify various visual patterns such as streaks, blobs and pigment network inside a lesion area, whereas flat areas identify blue whitish veil and regions inside the blobs in the skin images. Therefore, these GP trees have selected prominent LBP patterns corresponding to significant visual characteristics of the skin lesions to build even more informative CFs. The C_V tree in Figure 8(b) is built from three features CV_3, CV_7, and CV_{11} which correspond to variance of red color channel (σR), ratio between mean of red and mean of blue color channels ($\frac{\mu_R}{\mu_B}$), and ratio between mean of blue color channel of lesion area and mean of blue color channel of skin area $\frac{\mu_B}{\mu_B}$. They are combined in simple arithmetic operators to produce a significant CF. In addition, the mathematical expression "$CV_3 \times (CV_{11} \times \sin(CV_7))$" appears twice which shows that this sub-tree captures significant information. In S tree (Figure 8(c)), S_0 and S_2, S_8 correspond to area of the lesion, greatest diameter, and the difference between greatest and shortest diameter of the lesion region, respectively. The lesion area, greatest and shortest diameter are vital in capturing the shape of the lesion. Here, rather selecting shortest diameter as individual feature, GP selected the difference of the greatest and shortest diameter, thereby incorporating important hand-crafted features effectively in evolving S tree. These border shape features can hugely assist the dermatologist in real-time situations by providing significant knowledge about the lesion geometrical properties and hence, making a diagnosis much easier.

Table 5: Wavelet features appearing in the GP individual shown in Figure 8(e).

Feature	Measure	Channel	Level	node
W_{13}	Norm	Green	0	–
W_{116}	Kurtosis	Red	3	3.4
W_{124}	Kurtosis	Red	3	3.1
W_{278}	Entropy	blue	2	2.4
W_{206}	Entropy	Green	3	3.3
W_{349}	Norm	Luminance	1	1.1

7 CONCLUSIONS

This study develops an ensemble classification method based on GP for feature construction to solve the complex task of skin cancer image classification. The method constructs new powerful features from the pre-extracted texture, color, frequency-based, local and global features. These new CFs when provided to an ensemble of classifiers in a GP framework result in generating good trained models. The results have revealed that the CFs constructed for ensemble of classifiers have more distinguishing ability between classes as compared to CFs constructed for a single classifier. The results are compared to the exiting GP approaches for skin cancer image classification, where the proposed method significantly outperformed all of them. In comparison to the state-of-the-art CNN methods for the two datasets, the proposed method has produced significantly better results. Moreover, the proposed method significantly outperformed the commonly used classification (NB, SVM, *k*-NN, J48, and MLP) and ensemble methods (RF, Bagging, AdaBoost, LogitBoost, and RandomCommittee). Since the CFs are interpretable, the insights of good evolved CFs identified important features selected from the original set of features. This information can be helpful to the dermatologist in making a diagnosis.

Although the proposed method has achieved very good results, its performance can be increased by generating more CFs and investigating suitable number of CFs. Selecting only prominent CFs, e.g. measuring their information gain, and providing those selected CFs to the ensemble classifiers may improve results and will be investigated in the future.

REFERENCES

[1] Soha Ahmed, Mengjie Zhang, Lifeng Peng, and Bing Xue. 2014. Multiple feature construction for effective biomarker identification and classification using genetic programming. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*. ACM, 249–256.

[2] Qurrat Ul Ain, Harith Al-Sahaf, Bing Xue, and Mengjie Zhang. 2018. A Multi-tree Genetic Programming Representation for Melanoma Detection Using Local and Global Features. In *Proceedings of the 31st Australasian Joint Conference on Artificial Intelligence. Lecture Notes in Computer Science*, Vol. 11320. Springer, 111–123.

[3] Qurrat Ul Ain, Bing Xue, Harith Al-Sahaf, and Mengjie Zhang. 2019. Genetic Programming for Multiple Feature Construction in Skin Cancer Image Classification. In *Proceedings of the 34th International Conference on Image and Vision Computing New Zealand*. IEEE, 1–6.

[4] Harith Al-Sahaf, Ausama Al-Sahaf, Bing Xue, Mark Johnston, and Mengjie Zhang. 2017. Automatically evolving rotation-invariant texture image descriptors by genetic programming. *IEEE Transactions on Evolutionary Computation* 21, 1 (2017), 83–101.

[5] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino. 1998. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of Dermatology* 134, 12 (1998), 1563–1570.

[6] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. 2013. A color and texture based hierarchical k-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*. Springer, 63–86.

[7] Catarina Barata and Jorge S Marques. 2019. Deep learning for skin cancer diagnosis with hierarchical architectures. In *Proceedings of the International Symposium on Biomedical Imaging*, Vol. 2. IEEE.

[8] Catarina Barata, Jorge S Marques, and Jorge Rozeira. 2012. A system for the detection of pigment network in dermoscopy images using directional filters. *IEEE Transactions on Biomedical Engineering* 59, 10 (2012), 2744–2754.

[9] Titus J Brinker, Achim Hekler, Alexander H Enk, Carola Berking, Sebastian Haferkamp, Axel Hauschild, Michael Weichenthal, Joachim Klode, Dirk Schadendorf, Tim Holland-Letz, et al. 2019. Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer* 119 (2019), 11–17.

[10] Tianhorng Chang and C-C Jay Kuo. 1993. Texture analysis and classification with tree-structured wavelet transform. *IEEE Transactions on Image Processing* 2, 4 (1993), 429–441.

[11] Noel Codella, Junjie Cai, Mani Abedini, Rahil Garnavi, Alan Halpern, and John R Smith. 2015. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In *Proceedings of the International Workshop on Machine Learning in Medical Imaging*. Springer, 118–126.

[12] A. Esteva, B. Kuprel, et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115–118.

[13] Laura K Ferris, Jan A Harkes, Benjamin Gilbert, Daniel G Winger, Kseniya Golubets, Oleg Akilov, and Mahadev Satyanarayanan. 2015. Computer-aided classification of melanocytic lesions using dermoscopic images. *Journal of the American Academy of Dermatology* 73, 5 (2015), 769–776.

[14] Rahil Garnavi, Mohammad Aldeen, and James Bailey. 2012. Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis. *IEEE Transactions on Information Technology in Biomedicine* 16, 6 (2012), 1239–1252.

[15] César Guerra-Salcedo and Darrell Whitley. 1999. Genetic approach to feature selection for ensemble creation. In *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 1*. 236–243.

[16] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations Newsletter* 11, 1 (2009), 10–18.

[17] Balazs Harangi, Agnes Baran, and Andras Hajdu. 2018. Classification of skin lesions using an ensemble of deep neural networks. In *Proceddings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2575–2578.

[18] Jeremy Kawahara, Aicha BenTaieb, and Ghassan Hamarneh. 2016. Deep features to classify skin lesions. In *Proceedings of the 13th International Symposium on Biomedical Imaging*. IEEE, 1397–1400.

[19] Andrew Lensen, Harith Al-Sahaf, Mengjie Zhang, and Bing Xue. 2016. Genetic Programming for Region Detection, Feature Extraction, Feature Construction and Classification in Image Data. In *Proceedings of the European Conference on Genetic Programming*. Springer, 51–67.

[20] Andrew Lensen, Bing Xue, and Mengjie Zhang. 2018. Generating Redundant Features with Unsupervised Multi-Tree Genetic Programming. In *Proceedings of the European Conference on Genetic Programming*. Springer, 84–100.

[21] Sean Luke. 2013. *Essentials of metaheuristics* (2nd ed.). Lulu. [Online] Available: http://cs.gmu.edu/~sean/book/metaheuristics/.

[22] Ilias Maglogiannis and Charalampos N Doukas. 2009. Overview of advanced computer vision systems for skin lesions characterization. *IEEE Transactions on Information Technology in Biomedicine* 13, 5 (2009), 721–733.

[23] Natalie H Matthews, Wen Qing Li, Abrar A Qureshi, Martin A Weinstock, and Eunyoung Cho. 2017. Epidemiology of melanoma. In *Cutaneous Melanoma: Etiology and Therapy [Internet]*. Codon Publications.

[24] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. 2013. PH^2 - A dermoscopic image database for research and benchmarking. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 5437–5440.

[25] Afonso Menegola, Michel Fornaciali, Ramon Pires, Flávia Vasques Bittencourt, Sandra Avila, and Eduardo Valle. 2017. Knowledge transfer for melanoma screening with deep learning. In *Proceedings of the 14th International Symposium on Biomedical Imaging*. IEEE, 297–300.

[26] D. P. Muni, N. R. Pal, and J. Das. 2006. Genetic programming for simultaneous feature selection and classifier design. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36, 1 (Feb 2006), 106–117. https://doi.org/10.1109/TSMCB.2005.854499

[27] Timo Ojala, Matti Pietikäinen, and David Harwood. 1996. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition* 29, 1 (1996), 51–59.

[28] Diego Patiño, Alberto M Ceballos-Arroyo, Jairo A Rodriguez-Rodriguez, German Sanchez-Torres, and John W Branch-Bedoya. 2020. Melanoma detection on dermoscopic images using superpixels segmentation and shape-based features. In *Proceedings of the 15th International Symposium on Medical Information Processing and Analysis*, Vol. 11330. International Society for Optics and Photonics, 1133018.

[29] Fábio Perez, Cristina Vasconcelos, Sandra Avila, and Eduardo Valle. 2018. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 303–311.

[30] Conor Ryan, Krzysztof Krawiec, Una-May O'Reilly, Jeannie Fitzgerald, and David Medernach. 2014. Building a stage 1 computer aided detector for breast cancer using genetic programming. In *Proceedings of the European Conference on Genetic Programming*. Springer, 162–173.

[31] TY Satheesha, D Satyanarayana, MN Giri Prasad, and Kashyap D Dhruve. 2017. Melanoma is Skin Deep: A 3D reconstruction technique for computerized dermoscopic skin lesion classification. *IEEE Journal of Translational Engineering in Health and Medicine* 5 (2017), 1–17.

[32] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. 2019. Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians* 69, 1 (2019), 7–34.

[33] W. Stolz, A. Riemann, A. B. Cognetta, L. Pillet, W. Abmayr, D. Holzel, P. Bilek, F. Nachbar, and M. Landthaler. 1994. ABCD rule of dermatoscopy: a new practical method for early recognition of malignant-melanoma. *European Journal of Dermatology* 4, 7 (1994), 521–527.

[34] Binh Tran, Bing Xue, and Mengjie Zhang. 2015. Genetic programming for feature construction and selection in classification on high-dimensional data. *Memetic Computing* 8, 1 (2015), 3–15.

[35] Eduardo Valle, Michel Fornaciali, Afonso Menegola, Julia Tavares, Flávia Vasques Bittencourt, Lin Tzy Li, and Sandra Avila. 2017. Data, depth, and design: Learning reliable models for melanoma screening. *arXiv preprint arXiv:1711.00441* (2017).

[36] Fengying Xie, Haidi Fan, Yang Li, Zhiguo Jiang, Rusong Meng, and Alan Bovik. 2017. Melanoma classification on dermoscopy images using a neural network ensemble model. *IEEE Transactions on Medical Imaging* 36, 3 (2017), 849–858.