Genetic Programming for Multiple Feature Construction in Skin Cancer Image Classification

Qurrat Ul Ain, Bing Xue, Harith Al-Sahaf, and Mengjie Zhang
School of Engineering and Computer Science
Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand
Email: {qurrat.ul.ain, bing.xue, harith.al-sahaf, mengjie.zhang}@ecs.vuw.ac.nz

***Abstract*—Skin cancer is a common cancer worldwide, with melanoma being the most deadly form which is treatable when diagnosed at an early stage. This study develops a novel classification approach using multi-tree genetic programming (GP), which not only targets melanoma detection but is also capable of distinguishing between ten different classes of skin cancer effectively from lesion images. Selecting a suitable feature extraction method and the way different types of features are combined are important aspects to achieve performance gains. Existing approaches remain unable to effectively design a way to combine various features. Moreover, they have not used multi-channel multi-resolution spatial/frequency information for effective feature construction. In this work, wavelet-based texture features from multiple color channels are employed which preserve all the local, global, color and texture information concurrently. Local Binary Pattern, lesion color variation, and geometrical border shape features are also extracted from various color channels. The performance of the proposed method is evaluated using two skin image datasets and compared with an existing multi-tree GP method, ten single-tree GP methods, and six commonly used classification algorithms. The results reveal the goodness of the proposed method which significantly outperformed all these classification methods and demonstrate the potential to help dermatologist in making a diagnosis in real-time situations.**

I. Introduction

The incidence of melanoma which is the deadliest type of skin cancer, has increased rapidly over the past 30 years [1]. Skin cancer is highly curable when detected in its earliest stages, with a 5-year survival rate of 92% [1]. Recent developments in computer aided diagnostic (CAD) systems facilitate earlier diagnosis of various skin cancers. Dermatologists study several important visual characteristics for making a diagnosis based on dermoscopy criteria; the Asymmetry, Border deformity, variation in Color, and lesion Diameter (ABCD) rule [2]. These are the basic medical properties that help dermatologists accurately diagnose different kinds of skin cancers.

Genetic programming (GP) is a nature-inspired method that genetically breeds a population of computer programs (models or trees) to solve a particular task [3]. GP applies genetic operators such as reproduction, crossover, and mutation, iteratively to transform a current generation of programs into a new generation of programs [3]. The automatically evolved program possess a tree-like framework with terminal nodes and internal nodes. Features appear at terminal nodes while functions appear at internal nodes. GP utilizes its inbuilt feature selection ability by having the most prominent features as its terminals. The evolved program can be considered as a newly constructed feature (CF) developed from the selected features at the terminals, which often has more distinguishing ability between classes compared to the original features, which highly impacts on achieving good performance. GP has not only been used for classification, but has also been studied widely for feature selection and feature construction [4].

Traditionally, a GP individual consists of only one tree. However, GP can be designed to generate multiple trees (CFs) in an individual which is termed as multi-tree GP (MTGP) [5]. In the literature, MTGP has been explored for multi-class classification [6], multiple feature construction [7], and automatically evolving image descriptors [8].

To classify skin cancer images, various discriminative features, which include local and global as well as color and texture properties, must be provided to the classification algorithm to achieve good performance. A study on the human visual system indicates that the spatial/frequency representation contains both local and global information which inspired researchers to develop multi-scale texture models for image classification [9]. The multi-scale characteristics of wavelets make them a useful texture analysis technique to construct informative features [10]. This motivated us to construct wavelet-based features in this study.

We have utilized MTGP to construct multiple features each based on a single type of features, and give these CFs to a classification algorithm. Some existing approaches [10], [11] extract different types of features from lesion images. These methods evaluate the potential of these features to detect melanoma in a binary image classification problem using machine learning algorithms, but remain unable to combine the features effectively to achieve performance gains. Moreover, they limit the use of their method to binary classification (melanoma detection) only; not testing on the difficult multi-class skin image classification task. Moreover, most of the existing approaches are developed for only single image modality (images captured from one device), this work focuses on developing a robust classification method which can perform well across multiple image modalities.

Different from the existing approaches, our proposed method uses a wrapper approach to construct informative features, which are then given to a classification algorithm (such as decision trees) for classification. A wrapper approach includes a learning algorithm in the CF evaluation whereas a filter approach is independent of any classification algorithm

978-1-7281-4187-9/19/$31.0 ©2019 IEEE

[12]. An embedded approach combines feature construction and classifier learning into a single process. The feature construction ability of the proposed method generates highly informative CFs necessary to achieve performance gains.

Goals: This work develops a new MTGP method for skin cancer binary and multi-class image classification using a wrapper approach. Different from the previous methods, the proposed method aims at utilizing GP to construct features based on different types of texture, color, border shape and geometrical information features for skin images taken from different optical instruments (specialized dermatosocope and standard camera). Each GP individual consists of multiple trees each of which is a CF using only one type of features and all the CFs are collectively used for classification. By doing so, the proposed method is expected to automatically construct informative features, using the best type of image features. This work addresses the following research questions:

- Which type of the features are significant in providing good performance across different datasets?
- Can addition of the new wavelet-based features improve the performance of binary and multi-class classification?
- Can MTGP approach provide better discriminating ability as compared to single-tree GP wrapper and embedded approaches across different datasets?
- Can the proposed GP method outperform the other non-GP classification algorithms and the existing GP skin image classification methods?

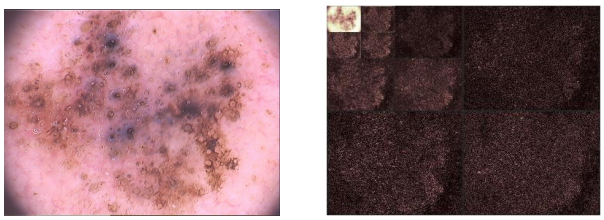
II. Background

Feature extraction is used to extract the image features, similar to those visually detected by dermatologists, that can accurately characterize a type of skin cancer [10]. In this work, we capture texture information from images using three-level pyramid-structured wavelet decomposition [9], local information using Local Binary Pattern image descriptor [13], global information using lesion color variation [11], and border shape features [10], [14]. These different types of features are incorporated to: 1) provide necessary discriminative information to GP for effective feature construction, 2) analyze which type of features are more prominent to classify which type of images (dermoscopy and standard camera).

1) Wavelet-based Features: Texture analysis helps identify the visual characteristics of a lesion which constitutes the basis of clinical diagnosis (e.g., ABCD rule of dermoscopy) [10]. The pyramid-structured wavelet analysis [9] captures both the local (detailed structure and internal texture) and global (overall properties) information of the lesion. We have applied three-level pyramid-structured wavelet decomposition on red, green, blue, and luminance color channels of the skin images, where luminance is represented by Eq. 1.

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

Various statistical measures are used to extract informative features from the wavelet coefficients such as energy, mean, standard deviation, skewness, kurtosis, norm, entropy, and average-energy, details can be found in [10]. Fig. 1(a) shows a skin lesion image and Fig. 1(b) shows the pyramid-structured wavelet decomposition applied on this image. To the best of our knowledge, this is the first time four color channels with three-level of pyramid-structured wavelet decomposition has been reported.



(a) (b)

Fig. 1. Three-level pyramid-structured wavelet decomposition.

2) Local Binary Patterns (LBP): LBP is an image descriptor widely used for feature extraction, developed by Ojala et al. [13]. It scans an image in a pixel-by-pixel fashion, using a sliding window of fixed radius. The central pixel value is calculated based on the intensity of surrounding pixels values lying on the radius. It generates a histogram (feature vector), from the computed values. The size of feature vector can be reduced from 2^p bins to $p\,(p-1)+3$ bins using only uniform LBP patterns and putting all non-uniform patterns in one bin. For LBP, a window size of 3×3 pixels and a radius of 1 pixel ($\mathrm{LBP}_{8,1}$) is used. In skin images, LBP features allow detection of corners (lesion boundary), streaks (line ends) and blobs (flat regions) which may add to performance gains.

3) Lesion Color Variation: Color, being a significant component of the ABCD rule [2], plays a vital role in classifying skin lesions. Variation in color triggers high variance in the RGB color space. Hence, features extracted from RGB color channels may have good discriminating ability between classes. To incorporate global color features, the pixels in the segmented skin lesion of RGB color channels are used. The mean (μ) and variance (σ^2) of each channel is calculated and represented as μR , μG , μB and $\sigma^2 R$, $\sigma^2 G$, $\sigma^2 B$. To capture complex non-uniform color distributions within the skin lesion region, mean ratios of the mean values are calculated, i.e., $\frac{\mu_R}{\mu_G}$, $\frac{\mu_R}{\mu_B}$, $\frac{\mu_G}{\mu_B}$. Variations in color of the skin lesion with respect to the surrounding skin is also considered. These features are calculated as $\frac{\mu_R}{\overline{\mu}_R}$, $\frac{\mu_G}{\overline{\mu}_G}$, $\frac{\mu_B}{\overline{\mu}_B}$, where $\overline{\mu}$ represents the mean value of surrounding skin. These features are adopted from [11].

4) Geometry-based Features: Border shape and geometrical properties of a lesion provide significant diagnostic information. We used some standard geometry features such as area, perimeter, greatest diameter, circularity index, irregularity index A, irregularity index B, and asymmetry index adopted from [14], and shortest diameter, irregularity index C, irregularity index D, and major and minor asymmetry indices adopted from [10].

III. The Proposed Method

The proposed method, 5-tree GP wrapper (WGP-5), for skin image classification is described in this section. The overall structure is presented in Fig. 2. Each image in the dataset is given to different feature extraction methods discussed in Section II to get five feature vectors, namely Wavelet, $\mathrm{LBP}_{\mathrm{Gray}}$,

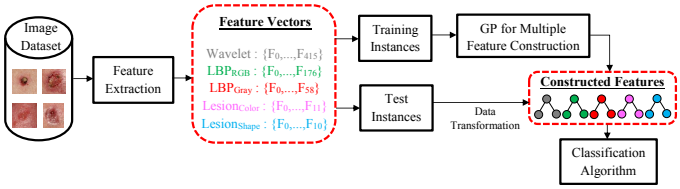


Fig. 2. The flowchart of the proposed method.

LBP_{RGB} $\text{Lesion}_{\text{Color}}$, and $\text{Lesion}_{\text{Shape}}$. These images, i.e., feature vectors, of the whole dataset are divided into training and test sets. The training set is given to GP to evolve five trees each based on a single type of features in one GP individual. Using these five trees (CFs), the training and test sets are transformed to new training and test sets. Then a classification algorithm (such as decision tree) uses the transformed training set to evolve a classification model. The learnt classification model is applied on the transformed test set to obtain the test classification performance.

A. Representation

A GP individual consists of five trees. The five types of features (Wavelet, LBP_{gray}, LBP_{RGB}, $\text{Lesion}_{\text{color}}$, and $\text{Lesion}_{\text{shape}}$) are fed into multi-tree GP method where it is ensured that during the evolutionary process, each tree can select from only one type of features. In other words, an individual in our multi-tree GP method consists of five CFs; one evolves using Wavelet features, second using LBP_{Gray} features, third using LBP_{RGB} features, fourth using $\text{Lesion}_{\text{Color}}$ features, and fifth using $\text{Lesion}_{\text{Shape}}$ features, as shown in Fig. 4.

B. Terminal Set and Function Set

The terminal set consists of five types of features, extracted from the feature extraction methods discussed in Section II.

1) Wavelet: 416 wavelet-based texture features extracted from RGB and luminance color channels of the images.
2) LBP_{RGB}: 59 LBP features extracted from each of the RGB channels and concatenated to make 177 (= 59 LBP features $\times$ 3 channels) features.
3) LBP_{gray}: 59 LBP features extracted from gray images.
4) $\text{Lesion}_{\text{color}}$: Color variation across the lesion area and skin region is calculated by 12 $\text{Lesion}_{\text{color}}$ features.
5) $\text{Lesion}_{\text{shape}}$: Border shape information of the lesion region is included by extracting 11 $\text{Lesion}_{\text{shape}}$ features.

The value of the ith feature for the above five feature types is indicated as W_i, C_i, G_i, L_i, and S_i, respectively, as shown by the GP individual in Fig. 4. The function set consists of seven operators; four arithmetic $\{+, -, \times, /\}$, two trigonometric $\{sin, cos\}$, and one conditional $\{if\}$ operator. Among the arithmetic operators, the first three operators have the original arithmetic meaning, however, division is protected that returns 0 when divided by 0. The *if* operator takes four inputs and returns the third input if the first input is greater than the second input; else, it returns the fourth input.

C. Crossover and Mutation

The genetic operators, such as crossover and mutation, are utilized accordingly to fit the requirements of the proposed method, which we call *same-index-crossover/mutation*. According to our initial experiments in this study, using different types of features to evolve a single tree results in poor feature construction because it ruins the effectiveness of the original features selected by the tree. Hence, in order to retain only one type of features in a single tree, we use *same-index-crossover/mutation* [15]. For illustration, the tree evolved from *wavelet* features in Parent-1 can only crossover/mutate with the tree evolved from the same *wavelet* features in Parent-2, and it cannot crossover/mutate with the other four trees.

D. Fitness Function

The fitness function is the balanced classification accuracy defined as

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

where m shows the number of classes, TP refers to true positive, FN refers to false negative, and the ratio $\frac{TP_i}{TP_i+FN_i}$ represents the true positive rate of class i. When there is a class imbalance problem (different number of instances in different classes), using standard overall accuracy, which is defined as the ratio between correctly classified instances and the total number of instances, may produce results biased towards the majority class. Hence, it is more suitable to use balanced accuracy to give equal weights to all the classes in a dataset. Using this fitness function (Eq. (2)) will help the GP individual to achieve overall good performance on the different classes in a dataset.

E. Classification

After completing the GP evolutionary process, we get the best GP individual with five constructed features on the training data. These constructed features are used to transform the original training and the original test data to a new training and a new test data. The transformed training data is used to train a classification method (such as decision tree). Then the transformed test data is given to the trained classification model to achieve the performance on the test data.

IV. Experiment Design

For carrying out the experiments, the datasets are split by *10-fold cross validation* such that nine folds are used for training and the remaining one fold for testing. Stratified random sampling is used to split the data to 10 folds. The number of GP runs is 30 and the results are reported in terms of the mean and standard deviation of the fitness values. For evolving an individual having five trees on the training data (9 folds), the fitness given in Eq. 2 is used, which is the accuracy of the wrapper classification algorithm. These five CFs are then used to transform the test data (1-fold). This procedure is repeated 10 times to get the result for 10-*fold cross validation* where each fold is considered for testing only one time. Hence, the above procedure is repeated 30 times, using 30 different seed values, to get 30 pairs of training and test accuracies. The implementation of WGP-5 is done using the Evolutionary Computing Java-based (ECJ) package [16].

(a) (b)

Fig. 3. Samples of (a) PH^2 dataset, and (b) Dermofit dataset.

TABLE I
GP PARAMETER SETTINGS.

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

A. Datasets

1) PH^2 *dataset:* This dataset [17] consists of 200 dermoscopy images with a size of roughly 768 $\times$ 560 pixels captured using a specialized device to capture skin images called dermatoscope. The images belong to three classes: melanoma, common nevus, and atypical nevus. In dermatology, melanoma and common nevus refers to malignant and non-malignant lesions, respectively. Atypical nevus is a currently non-malignant lesion, but can develop tumor cells later. For binary classification experiments, 40 melanoma are considered as "*malignant*" class, and 80 common nevus and 80 atypical nevus are together considered as "*benign*" class. Samples of this dataset are shown in Fig. 3(a).

2) Dermofit dataset: The Dermofit Image Library [18] consists of 1300 lesion images taken from a standard camera, under standardized conditions. The lesions belong to ten classes, and each image is provided with a gold standard diagnosis. For binary classification, we have used two classes; 1) Melanocytic Nevus as "*benign*", and 2) Malignant Melanoma as "*malignant*". Samples of this dataset are shown in Fig. 3(b).

B. GP Parameters

The GP parameters adopted in the proposed method are shown in Table I. The evolutionary process keeps evolving until a maximum number of 50 generations is reached or a perfect individual with 100% accuracy is found.

C. Classification Methods for Comparison

1) Non-GP Methods: To check the effectiveness of WGP-5 on the test set, six classification methods are used: Naïve Bayes (NB), Support Vector Machines (SVMs), k-Nearest Neighbor (k-NN) where $k = 1$ (the closest neighbor), Decision Trees (J48), Random Forest (RF), and Multilayer Perceptron (MLP). For implementation of these methods, the commonly used Waikato Environment for Knowledge Analysis (WEKA) package is used [19]. Similar to the previous methods [7], [20], we adopt a Radial basis Function kernel which has shown better performance compared to the default linear kernel in WEKA. For MLP, the momentum, learning rate, training epochs and the number of units in a single hidden layer are adopted from the previous methods [7], [20]. To make a valid comparison in our experiments, the non-GP methods are given all the five types of features.

2) GP Methods:

- **WGP-4**: This is similar to WGP-5 with a wrapper approach. However, it constructs four CFs from four types of features defined in Section III-B(2-5).
- **EGP-4** [20]: This is similar to WGP-4 with same four types of features used to evolve four trees in a GP individual. However, it is an embedded approach where GP is used as a classification method as well. Each tree in a GP individual is considered a binary classifier. The best tree (classifier) among the four trees with the highest accuracy on the training data is used to test the performance on the test data.
- **WGP-1**: This is the traditional GP with a single tree in a GP individual. Only one type of features such as $\mathrm{LBP}_{\mathrm{Gray}}$ is given to GP to evolve a single CF in a GP individual which will be given to J48 for classification.
- **EGP-1**: This is the traditional GP approach which evolves a single tree in its individual to perform binary classification. In one set of experiments, the terminal set consists of a single type of features to evolve a single tree in a GP individual.

V. RESULTS AND DISCUSSIONS

The results of our experiments are presented in Tables II and III. The values of these results represent the mean and standard deviation among the 30 GP runs, where value of one GP run is computed as the mean of applying *10-fold cross validation* to the datasets. The deterministic methods are run once, hence, their results are represented as the mean and standard deviation of applying *10-fold cross validation* to the datasets. In order to correctly identify the significance of our proposed method, the results are investigated using *Wilcoxon signed-rank test* with a significance level of 5%. This test is applied on the test results to check which method has better ability to correctly classify the lesion images. Three symbols "+", "$-$" and "=" are used which represents that the proposed method significantly outperforms, is significantly worse, and performs similarly in comparison to the corresponding method.

A. Binary Classification

The results for the task of binary classification for the two datasets are presented in Table II. Vertically, the table consists of seven blocks where the first and second blocks give the results WGP-4 and WGP-5, respectively. The third block gives the results of the existing embedded method (EGP-4) [20]. The fourth block shows results of the other non-GP classification methods, the fifth and the sixth block show results of the single-tree GP methods (WGP-1 and EGP-1).

Among the four classification algorithms in the WGP-4 and our proposed WGP-5 methods (Table II), it has been observed that in case of WGP-4, NB and J48 achieved the highest performances with $85.70 \pm 2.65\%$ and $84.18 \pm 4.11\%$ on PH^2 and Dermofit datasets, respectively. However, there is a significant increase in performance when using the proposed WGP-5 method where the CF generated with wavelet features help improve the distinguishing ability of the classifier. In our method WGP-5, NB and SVM achieved the highest

TABLE II
RESULTS OF MULTI-TREE GP METHOD FOR **BINARY CLASSIFICATION**: ACCURACY (%) ON THE TEST SET OF BOTH DATASETS.

Classification Algorithm		PH^2	Dermofit
WGP-4	NB	**85.70 ± 2.65** +	80.45 ± 2.18 +
	SVM	81.52 ± 3.58 +	80.33 ± 2.71 +
	k-NN	61.26 ± 4.05 +	69.27 ± 2.89 +
	J48	85.18 ± 3.72 +	**84.18 ± 4.11** +
WGP-5	NB	**89.77 ± 1.84**	96.21 ± 1.09
	SVM	86.48 ± 2.35	**97.26 ± 1.25**
	k-NN	63.34 ± 2.67	86.04 ± 2.52
	J48	87.61 ± 3.08	96.99 ± 0.70
EGP-4	–	78.87 ± 2.92 +	74.57 ± 1.86 +
Non-GP Methods	NB	77.19 ± 9.06 +	96.99 ± 3.13 =
	SVM	62.19 ± 7.96 +	63.84 ± 8.26 +
	k-NN	74.06 ± 10.83 +	87.43 ± 5.76 +
	J48	72.50 ± 10.99 +	95.42 ± 3.87 +
	RF	75.00 ± 8.73 +	93.44 ± 3.94 +
	MLP	78.75 ± 10.81 +	95.64 ± 4.63 +
WGP-1	LBP_{Gray}	60.19 ± 4.73 +	53.88 ± 3.44 +
	LBP_{RGB}	65.70 ± 6.25 +	53.80 ± 3.36 +
	$Lesion_{Color}$	61.81 ± 4.56 +	65.79 ± 5.90 +
	$Lesion_{Shape}$	61.65 ± 4.28 +	64.88 ± 3.69 +
	Wavelet	67.75 ± 4.25 +	96.94 ± 1.33 =
EGP-1	LBP_{Gray}	65.96 ± 3.96 +	59.91 ± 3.57 +
	LBP_{RGB}	73.87 ± 2.34 +	63.26 ± 3.19 +
	$Lesion_{Color}$	65.70 ± 3.61 +	74.13 ± 2.67 +
	$Lesion_{Shape}$	49.89 ± 5.34 +	61.74 ± 7.06 +
	Wavelet	72.31 ± 2.75 +	88.13 ± 3.58 +

performances with $89.77 \pm 1.84\%$ and $97.26 \pm 1.25\%$ on PH^2 and Dermofit datasets, respectively.

From the results of the statistical significance test presented in Table II, it has been seen that the proposed WGP-5 method not only outperformed EGP-1 (single tree) but also outperformed all the existing WGP-4 and EGP-4 (multi-tree) classification methods, which proves the effectiveness and authenticity of our proposed method for melanoma detection. In comparison with non-GP and WGP-1, WGP-5 has shown either comparable or better performance.

B. Multi-class Classification

The results for the task of multi-class classification for the two datasets are presented in Table III. Among the four classification algorithms in the WGP-4 and WGP-5 methods, J48 achieved the highest classification performance. In case of WGP-4, it has achieved $80.64 \pm 2.24\%$ and $69.25 \pm 1.41\%$ test performance on the PH^2 and Dermofit datasets, respectively, which increases around 5% while using WGP-5 showing the effectiveness of the CF evolved with wavelet features. It is worthwhile to note here that PH^2 dataset has 3 classes and Dermofit has 10 classes (more difficult). For both WGP-4 and WGP-5, most of these classifiers are performing well for a 3-class problem (PH^2 dataset) on the unseen data such as SVM producing $77.17 \pm 2.00\%$ and $84.92 \pm 2.31\%$ accuracy, respectively, however, only J48 performed well enough for the complex 10-class problem (Dermofit dataset) reaching classification performance as high as $74.05 \pm 1.52\%$. From the results of the statistical test presented in Table III, clearly the proposed WGP-4 and WGP-5 methods outperformed all the non-GP methods as well as the WGP-1 methods on the easy (PH^2) and difficult (Dermofit) datasets, which shows its effectiveness for skin cancer image classification problems.

TABLE III
RESULTS OF MULTI-TREE GP METHOD FOR **MULTI-CLASS CLASSIFICATION**: ACCURACY (%) ON THE TEST SET OF BOTH DATASETS.

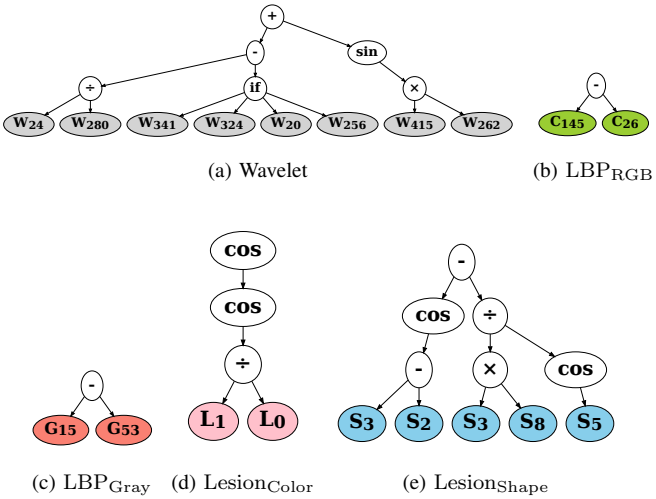
Classification Algorithm		PH^2	Dermofit
Non-GP Methods	NB	71.00 ± 7.68 +	45.92 ± 3.63 +
	SVM	59.50 ± 8.50 +	51.08 ± 4.82 +
	k-NN	65.50 ± 10.36 +	43.54 ± 2.46 +
	J48	58.00 ± 12.49 +	50.08 ± 3.27 +
	RF	71.50 ± 8.67 +	63.85 ± 3.00 +
	MLP	68.50 ± 4.50 +	66.85 ± 4.66 +
WGP-1	LBP_{Gray}	52.00 ± 6.34 +	35.27 ± 1.02 +
	LBP_{RGB}	62.42 ± 4.84 +	41.80 ± 1.94 +
	$Lesion_{Color}$	52.17 ± 3.23 +	43.41 ± 0.00 +
	$Lesion_{Shape}$	51.33 ± 4.37 +	41.28 ± 0.00 +
	Wavelet	67.17 ± 4.78 +	43.48 ± 1.12 +
WGP-4	NB	75.01 ± 1.76 +	49.23 ± 1.51 +
	SVM	77.17 ± 2.00 +	38.69 ± 1.34 +
	k-NN	57.43 ± 2.40 +	41.13 ± 0.91 +
	J48	**80.64 ± 2.24** +	**69.25 ± 1.41** +
WGP-5	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**

C. Overall Results

It is evident from the results of binary and multi-class classification that generating the wavelet-based CFs in the proposed WGP-5 method significantly helps the classification algorithm to build more accurate classifier compared to the existing approaches. The WGP-5 has completely outperformed WGP-4 having around 5% improvement in both the binary and multi-class classification tasks on both datasets. Though, these wavelet features significantly help the non-GP and WGP-1 methods (as well) to acheive performance gains compared to the proposed WGP-5 for binary classification task, these methods remain unable to cope well with multi-class classification task. For illustration, WGP-1 with wavelet features produces $96.94 \pm 1.33\%$ accuracy (Table II) for binary classification, whereas the same method results in $43.48 \pm 1.12\%$ accuracy (Table III) for multi-class classification. This shows that the proposed WGP-5 method works not only well for the easy (binary classification) task but also has the potential to produce good performance for the difficult multi-class classification task leaving other non-GP and existing GP methods far behind.

Different types of features are effective in classifying images captured from different devices. For PH^2 dataset, LBP_{RGB} and wavelet features among WGP-1 and EGP-1 have relatively good performance among the five feature sets as shown in Table II. However, for Dermofit dataset, $Lesion_{Color}$ and wavelet features among WGP-1 and EGP-1 have relatively good performance. For the results of multi-class classification (Table III), similar behavior is shown on PH^2 dataset, whereas such trend is not seen for Dermofit dataset. The difficulty level when moving from binary to multi-class classification for PH^2 dataset is less (2 classes to 3 classes) as compared to Dermofit dataset (2 classes to 10 classes).

It can be observed that images taken from different instruments require different feature extraction methods to get informative features necessary to distinguish between classes. Such a trend has been observed while constructing multiple features using the WGP-4 and WGP-5 methods for binary



(a) Wavelet (b) $\mathrm{LBP}_{\mathrm{RGB}}$

(c) $\mathrm{LBP}_{\mathrm{Gray}}$ (d) $\mathrm{Lesion}_{\mathrm{Color}}$ (e) $\mathrm{Lesion}_{\mathrm{Shape}}$

Fig. 4. A good evolved GP individual for $Dermofit$ dataset using the different types of features producing 98.48% accuracy on the unseen data in the binary classification task.

and multi-class classification. Among all the five trees, on both datasets, wavelet features produced the best results most of the time. However, $\mathrm{LBP}_{\mathrm{RGB}}$ and $\mathrm{Lesion}_{\mathrm{color}}$ features also remain prominent on the PH^2 and Dermofit datasets, respectively. From the results of EGP-1 and WGP-1 methods, it is concluded that selecting a suitable feature extraction method is critical in producing good classification performance.

D. An Evolved GP Individual

To analyze why our proposed WGP-5 method achieved good performance, we show a good GP individual (Fig. 4) with five trees (CFs) evolved using the five types of features with 98.48% accuracy on the unseen data. This individual is taken from the Dermofit experiments for binary classification. In Fig. 4, colored nodes represent terminals (each color represents one type of features) and white nodes represent functions.

VI. Conclusions

This work has developed a novel feature construction method for skin cancer binary and multi-class image classification using multi-tree GP in a wrapper approach. The proposed method incorporates various types of multi-channel and multi-resolution features which possess information related to RGB and gray-level pixel-based image properties, variation in color across the lesion image, as well as geometrical border shape properties. These five types of pre-extracted features are provided to MTGP by utilizing suitable genetic operators such as *same-index-crossover/mutation*. The MTGP method evolves five CFs which are then given to a classification method to generate a model for skin image classification. The proposed method has proved useful for both binary and multi-class skin image classification problems as it has outperformed all the six non-GP classification algorithms, the existing MTGP embedded approach and the single-tree GP methods showing evidence of effective discrimination between classes.

Due to the page limit, we remain unable to dig into the details of prominent features appearing in the CFs of a GP individual which we would like to explore in future. Further, we would investigate the impact of using pre-processing techniques before feature extraction to remove the various artefacts present in skin images such as hair and reflection.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: A Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[2] W. Stolz, A. Riemann *et al.*, "ABCD rule of dermatoscopy: a new practical method for early recognition of malignant-melanoma," *European Journal of Dermatology*, vol. 4, no. 7, pp. 521–527, 1994.

[3] J. R. Koza and R. Poli, "A genetic programming tutorial," 2003.

[4] B. Tran, B. Xue, and M. Zhang, "Genetic programming for multiple-feature construction on high-dimensional classification," *Pattern Recognition*, vol. 93, pp. 404–417, 2019.

[5] M. Oltean and D. Dumitrescu, "Multi expression programming," *Journal of Genetic Programming and Evolvable Machines, Kluwer, second tour of review*, 2002.

[6] D. P. Muni, N. R. Pal, and J. Das, "A novel approach to design classifiers using genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 2, pp. 183–196, 2004.

[7] Q. U. Ain, B. Xue, H. Al-Sahaf, and M. Zhang, "Multi-tree genetic programming with a new fitness function for melanoma detection," in *Proceedings of the Congress on Evolutionary Computation. IEEE*, 2019, pp. 880–887.

[8] H. Al-Sahaf, B. Xue, and M. Zhang, "A multitree genetic programming representation for automatically evolving texture image descriptors," in *Proceedings of the Asia-Pacific Conference on Simulated Evolution and Learning*. Springer, 2017, pp. 499–511.

[9] T. Chang and C.-C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Transactions on Image Processing*, vol. 2, no. 4, pp. 429–441, 1993.

[10] R. Garnavi, M. Aldeen, and J. Bailey, "Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1239–1252, 2012.

[11] T. Satheesha, D. Satyanarayana *et al.*, "Melanoma is skin deep: A 3D reconstruction technique for computerized dermoscopic skin lesion classification," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 5, pp. 1–17, 2017.

[12] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2016.

[13] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[14] I. Maglogiannis and C. N. Doukas, "Overview of advanced computer vision systems for skin lesions characterization," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 721–733, 2009.

[15] A. Lensen, B. Xue, and M. Zhang, "Generating redundant features with unsupervised multi-tree genetic programming," in *Proceedings of the European Conference on Genetic Programming*. Springer, 2018, pp. 84–100.

[16] S. Luke, *Essentials of metaheuristics*, 2nd ed. Lulu, 2013, [Online] Available: http://cs.gmu.edu/ sean/book/metaheuristics/.

[17] T. Mendonça, Ferreira *et al.*, "PH2–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*, 2013, pp. 5437–5440.

[18] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical k-NN approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*. Springer, 2013, pp. 63–86.

[19] M. Hall, E. Frank *et al.*, "The WEKA data mining software: an update," *SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[20] Q. U. Ain, H. Al-Sahaf, B. Xue, and M. Zhang, "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*. Springer, 2018, pp. 111–123.