Multi-tree Genetic Programming with A New Fitness Function for Melanoma Detection

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*—The occurrence of malignant melanoma had enormously increased since past decades. For accurate detection and classification, not only discriminative features are required but a properly designed model to combine these features effectively is also needed. In this study, the multi-tree representation of genetic programming (GP) has been utilised to effectively combine different types of features and evolve a classification model for the task of melanoma detection. Local binary patterns have been used to extract pixel-level informative features. For incorporating the properties of ABCD (asymmetrical property, border shape, color variation and geometrical characteristics) rule of dermoscopy, various features have been used to include local and global information of the skin lesions. To meet the requirements of the proposed multi-tree GP representation, genetic operators such as crossover and mutation are designed accordingly. Moreover, a new weighted fitness function is designed to evolve better GP individuals having multiple trees influencing each other's performance during the evolution, in order to get overall performance gains. The performance of the new method is checked on two benchmark skin image datasets, and compared with six widely used classification algorithms and the single tree GP method. The experimental results have shown that the proposed method has significantly outperformed all these classification methods.**

I. Introduction

Malignant melanoma is one of the deadliest form of skin cancers. Fortunately, malignant melanoma can be treated successfully if diagnosed at an early stage [1]. The continuous increase in incidence of melanoma in recent years, its painful biopsy procedures, high mortality rate, and their huge medical cost have made its early diagnosis an important priority of public health. Since this type of cancer is visible on the skin, it can be monitored easily. In recent years, there has been rising interest in developing computer aided diagnostic (CAD) systems for automated detection and diagnosis of skin cancer, specifically malignant melanoma [2]–[9].

Dermoscopy is a non-invasive diagnostic technique which includes using an optical instrument having powerful lighting system to examine skin lesions in a higher magnification [10]. The dermoscopic images are rich enough for extracting useful information. This has attracted many researchers in the field of computer vision and pattern recognition to develop CAD systems that can assist dermatologist in early detection. Dermatologists use the ABCD (asymmetry, border irregularity, color variation and dermoscopic structure) rule of dermoscopy which is a scoring method to quantify these four lesion properties and effectively separate melanoma from benign lesions [11]. Automated melanoma recognition from images is a very challenging task due to 1) the low contrast of skin lesions, 2) the presence of many artifacts in the image, 3) the huge intra-class variation of melanomas, and 4) the high degree of inter-class visual similarity between melanoma and non-melanoma skin lesions [3].

Image data often requires to be transformed into a form that can be given directly to a classification algorithm for the task of image classification. An image descriptor is a method that transforms an image into useful information, which we call features. There are many different methods for extracting various types of features, each of which is often suitable for different domains. However, it is often challenging to know which type of features are good for a particular image classification task, due mainly to the complexity of the images themselves and the lack of domain knowledge. Meanwhile, features are often not equally important for classification, and irrelevant or redundant features may even reduce the classification performance due to the large search space and interactions between features. Feature selection which aims to select a subset of informative and complementary features is often necessary. Developing new features from the existing set of features is called feature construction.

Genetic Programming (GP), which is an evolutionary algorithm, has been extensively used for feature selection and feature construction [10] and recently used to automatically evolve image descriptors [12]. GP searches for solutions to a user-defined problem by evolving a computer program, often in a tree-like structure where terminal nodes consist of features and internal nodes consist of functions [13]. GP applies genetic operators like crossover, mutation and reproduction during its evolutionary process to evolve diverse solutions. The main aim of feature selection and construction is to improve the classification performance by reducing the search space for evolving better solutions (GP individuals) meanwhile speeding up the search process by using a smaller number of features.

Different from single-tree GP which evolves one tree in an individual, GP can have more than one trees to solve a particular problem, which is termed as multi-tree GP (MTGP) [14]. MTGP has been used for self-assembling swarm robots [15], multi-class classification [16], and automatically evolving image descriptors [17]. For melanoma detection, it is important to have enough informative features in the terminal set to get good GP individuals having better discriminative ability between classes. Hence, different kinds of features which include

color, texture, border shape and geometrical characteristic properties are required. To achieve good performance, images captured from different instruments might have different visual properties such as scale, illumination, and reflection. Hence, it is hard to decide beforehand which type of features are suitable for which type of images (captured from different instruments). However, different kinds of features can be used to have enough informative features that can be used for feature selection and construction to design a powerful and robust GP model, which can perform well for images taken from different acquisition devices (such as specialized instruments and standard camera). Images captured from different instruments might have different visual properties such as scale, illumination, and reflection, hence, we donot know beforehand which type of features are suitable for which type of images (captured from different instruments). Therefore, a multi-tree GP approach having multiple trees, each evolved using a different type of informative features, seems to be a promising approach.

A. Objectives

The overall aim of this study is to develop a multi-tree GP representation based method for the task of melanoma classification from skin cancer images, by using different types of features as the terminal sets. Different from most existing approaches, this work focuses on evolving a GP individual where different color, texture, border and geometrical shape features are used in a suitable way to achieve performance gains. A new fitness function is proposed which allows multiple trees (each evolved with one type of features) among a GP individual to influence each other's behavior rather than evolving without any interaction (as the case in [7]), hence evolving better classification models. This study aims at finding answers to the following research questions:

- Whether the new weighted fitness function provides better performance as compared to the existing fitness function and why?
- How well is the multi-tree GP approach as compared to single-tree GP approach across different datasets?
- Whether the proposed GP method can outperform the other non-GP classification algorithms?
- Having images captured from different acquisition devices, which type(s) of features are most prominent in providing better discriminating ability between benign and malignant images?

II. Background

A. Related Work

Earlier in 1994, Ercal et al. [1] designed a neural network approach to the automated detection of melanoma from three benign categories of tumors. Their approach used features, based on lesion shape and relative lesion color. These features were supplied to an artificial neural network (ANN) for classification of skin cancer images as malignant or benign. This approach obtained 80% accuracy of the malignant and benign tumors on real skin tumor images. However, the boundaries of a lesion are required to be identified manually by a dermatologist, which makes this system expensive to implement.

Garnavi et al. [8] presented a novel CAD system for melanoma classification. The work aims at the selecting an optimal set of features and integrating these features, which are derived from textural, border-based, and geometrical properties of the lesion. The texture features are extracted using wavelet-decomposition, the border features are extracted by constructing a boundary-series model of the lesion border while analyzing it in spatial and frequency domains, and the geometry features are extracted from lesion shape indexes. The gain-ratio method is used for feature selection. Four machine learning classification algorithms (support vector machine (SVM), random forest, naïve bayes, and hidden logistic model tree) are used to classify melanoma and benign images. This diagnostic system achieved an accuracy of 91.26%, with only 23 features. This approach described the advantage gained in manually combining texture with border and geometry features, compared to using only texture features. Furthermore, in the optimized feature set, texture feature have the highest contribution among the three types of feature sets. Though this diagnostic system achieved good classification performance, but the method lack a suitable way of combining different types of features.

In [11], automatic scoring of the ABCD rule for dermoscopy lesions is implemented. The images are first pre-processed to remove hair artefacts using Gabor filters and boundaries are detected using active contours. Then features are extracted for the characteristics of ABCD rule by using existing and newly designed methods. To classify a lesion as melanoma or benign, the total dermoscopy score (TDS) is calculated. The experimental results have shown good performance in terms of sensitivity and specificity. Moreover, the results demonstrate that the extracted features can be used to build a good classifier for melanoma detection.

Recently, Adjed et al. [2] introduced fusion of texture and structural features for classifying malignant melanoma. The textural features are extracted from different variants of local binary pattern operators, whereas the structural features are extracted from wavelet and curvelet transforms. The method used SVM as the classifier and showed encouraging performance with sensitivity of 78.93%, specificity of 93.25% and accuracy of 86.07%.

Shimizu et al. [5] extracted 828 features grouped into three categories: color, texture, and sub-region. Two classification models are designed: a layered based on a task decomposition strategy, and flat models. The method is developed to classify 964 dermoscopy images belonging to four skin cancer classes. The layered model outperformed the flat models, achieving detection rate of 90.48%, for melanoma.

Xie et al. [4] developed an ANN based ensemble model for classifying melanocytic tumors as benign or malignant. The algorithm has three stages; 1) lesions are extracted with the help of a self-generating neural network (SGNN); 2) color, texture and border features are extracted from the lesion area;

and 3) lesions are classified using a classifier based on a neural network ensemble model. The results have shown that the new border features and the proposed classifier model has significantly improved the classification accuracy.

Yu et al. [3] proposed a two-stage deep convolutional neural network (CNN) architecture for melanoma recognition. The authors constructed a fully convolutional residual network for lesion segmentation and integrated it with a very deep residual network for classification. This study produced good results and demonstrated that very deep CNNs can be employed to solve complicated medical image analysis tasks, even with limited training data.

The existing methods [1], [3], [4] have used CNNs for skin cancer image classification. Although these methods have shown good classification performance, they are implemented as a black-box, hence, are not interpretable. Such classification models cannot clearly suggest which features are more prominent in classifying skin cancer images. Furthermore, the performance of a CNN is usually constrained by data and requires sufficient training examples to provide good classification performance. Training a model using a large dataset needs long time and requires large computing resources. Some existing approaches [2], [4], [5], [8], [18] extracted various kinds of features from skin cancer images and compared the performance of these features for image classification using commonly used machine learning classification algorithms. However, they remain unable to design an effective way of combining different types of features, and necessary to improve performance gain.

B. Feature Extraction

1) Local Binary Patterns Features: Ojala et al. [19] developed an image descriptor termed as local binary patterns (LBP). It is an image descriptor that has been used commonly in a wide range of computer vision applications for the task of feature extraction. LBP works by scanning the image pixel-by-pixel using a sliding window of fixed radius.The value of the central pixel is computed based on the values of the neighboring pixels lying on the radius as depicted in Fig. 1. It then generates a histogram (i.e. feature vector) from these computed values. The LBP operator is defined as:

$$LBP_{p,r} = \sum_{i=0}^{p-1} z(a_i - a_c)2^i \tag{1}$$

where p is the number of neighboring pixels, r is the radius, a_i and a_c are the intensity values of the i^{th} neighbor and central pixel, respectively. Here, $z(x)$ returns 1 if $x \geq 0$, else it returns 0. The value computed from Equation (1) is assigned to the central pixel and the corresponding bin in the histogram is incremented by 1. The value of k^{th} bin of a histogram H computed on an image of size $w \times h$ is given as:

$$H(k) = \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} (LBP_{p,r}(V_{i,j}) = k) \tag{2}$$

where the value of k ranges between 0 and $K-1$, K being the maximum number of bins in the histogram, and $V_{i,j}$ is the value of the pixel at the coordinate (i, j). The LBP codes are divided into two categories: *uniform* and *non-uniform*. A code is said to be uniform if it does not have more than two bitwise transitions circularly from 0 to 1 or 1 to 0. For example, the codes 0011110, 01111000, and 10000001 are uniform, whereas the codes 00101011, 11010110, and 01010001 are non-uniform. The size of feature vector can be reduced from 2^p bins to $p(p-1)+3$ bins by combining all non-uniform codes into a single bin. Uniform codes detect various texture primitives such as corners, edges, line ends, flat regions and dark spots [19]. In a skin cancer image, uniform codes can help in detection of streaks, blobs and pigmented network, hence extracting highly informative features, resulting in improved classification performance.

In our experiments, we generate a histogram of uniform codes; hence, there are 59 $(= 8 \times (7) + 3)$ LBP features for a single image.

2) Color contrast features: Color is a significant component of the ABCD rule [20], often used by dermatologists to classify skin lesions. Most CAD systems have incorporated color features to enhance their classification performance [1], [2], [4], [5], [8], [21]. Melanoma lesions are categorized by variation in color across the lesion area. This color variation leads to high variance in the red, green, blue (RGB) color space. Hence, highly discriminative features can be extracted from RGB color channels. In this work, the pixels in the segmented skin lesion of red, green and blue color channels are used to extract color contrast features. These features are adopted from [18]. The mean (μ) and variance (σ) of each channel is computed and denoted as μR, μG, μB and σR, σG, σB. Features based on complex non-uniform color distributions within the skin lesion region are extracted by computing mean ratios of the mean values, such as $\frac{\mu_R}{\mu_G}$, $\frac{\mu_R}{\mu_B}$, $\frac{\mu_G}{\mu_B}$. Variations among the color of the skin lesion and the surrounding skin is also computed; $\frac{\mu_R}{\overline{\mu}_R}$, $\frac{\mu_G}{\overline{\mu}_G}$, $\frac{\mu_B}{\overline{\mu}_B}$, where $\overline{\mu}$ represents the mean value of surrounding skin area. These 12 features are denoted as $\mathrm{Lesion}_{\mathrm{color}}$ features.

3) Geometrical shape features: Border irregularity and geometrical characteristics of the shape of a lesion provide significant diagnostic information for detecting melanoma. Asymmetry is given the highest score among the four characteristics; asymmetry, border irregularity, color, and diameter of the ABCD rule of dermoscopy [20]. In this work, we have used standard geometry features (area, perimeter, greatest diameter, circularity index, irregularity index A, irregularity index B, and asymmetry index) adopted from [22] and some other shape features (shortest diameter, irregularity index C, irregularity index D, and major and minor asymmetry indices) adopted from [8]. In this study, images within each dataset have fairly similar spatial resolution; hence, there has been no scale issues for area and perimeter features. Here, we have a set of 11 geometrical shape features from each skin lesion image denoted by $\mathrm{Lesion}_{\mathrm{shape}}$ features.

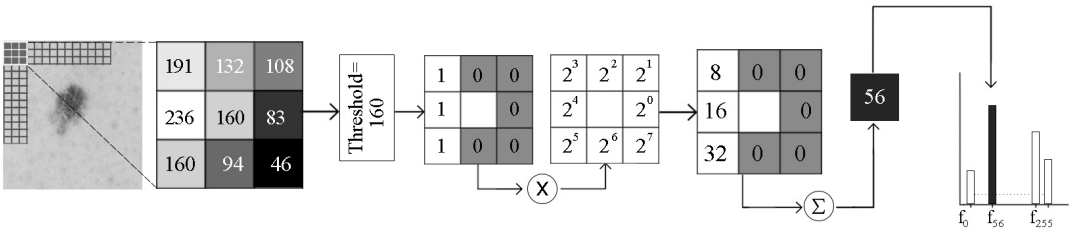

Fig. 1. Step-by-Step procedure to generate $LBP_{8,1}$ code for image cut-out (having 8 neighboring pixels and radius = 1) and get a decimal value of the central pixel.

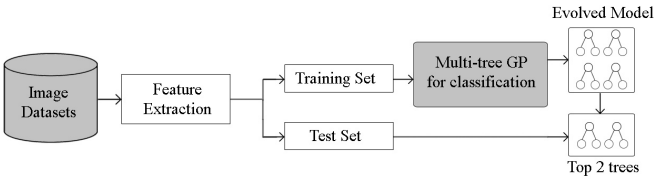

Fig. 2. The overview of the proposed method.

TABLE I
NUMBER OF EACH TYPE OF FEATURES.

Type of feature	No. of features	Type of feature	No. of features
$\text{Lesion}_{\text{color}}$	12	LBP_{RGB}	177
$\text{Lesion}_{\text{shape}}$	11	LBP_{gray}	59

III. THE MULTI-TREE GENETIC PROGRAMMING METHOD

This section describes the proposed MTGP method in detail. An individual in this MTGP approach consists of four trees. Each tree is constructed and evolved using one type of features. The four sets of features are $\text{Lesion}_{\text{color}}$, $\text{Lesion}_{\text{shape}}$, LBP_{gray} and LBP_{RGB} as illustrated in Section II-B. Fig. 2 depicts the structure of the proposed method. An example of evolved model having four trees is later presented in Fig. 5. This section also describes the terminal set, the function set, crossover and mutation operators, and the new fitness function in the new method.

A. Terminal Set

The terminal set consists of four sets of features, extracted from four different feature extraction methods as discussed in Section II-B. These features and the number of each type of features are summarised in Table I.

To extract LBP_{RGB} and LBP_{gray} features, LBP is used with a window size of 3×3 pixels and a radius of 1 pixel ($\text{LBP}_{8,1}$). LBP_{gray} features are extracted from gray-scale skin cancer images, whereas LBP_{RGB} features are extracted from the three color channels (red, green, blue) which are then concatenated to get a single feature vector. The value of the i^{th} feature for the $\text{Lesion}_{\text{color}}$, $\text{Lesion}_{\text{shape}}$, LBP_{gray} and LBP_{RGB} features is indicated as C_i, S_i, R_i, and G_i, respectively (an example is shown later in Fig. 5).

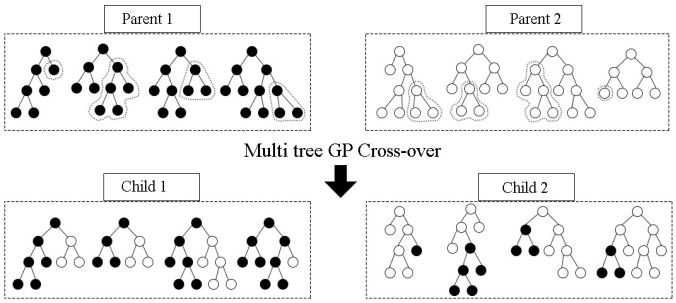

Fig. 3. The proposed same-index-crossover operator.

B. Function Set

The function set consists of seven most commonly used operators; four arithmetic operators $\{+, -, \times, /\}$, one conditional operator $\{if\}$, and two trigonometric $\{sin, cos\}$ operators. Among the arithmetic operators, the first three operators have the normal arithmetic meaning, whereas division is protected which returns 0 when divided by 0. The if operator takes four inputs and returns the third input if the second input is smaller than the first input; otherwise, it returns the fourth input.

C. Crossover and Mutation

In order to retain only one type of features in one tree of a GP individual, we have designed the genetic operators, such as crossover and mutation, accordingly, which is termed as *same-index-crossover/mutation*. This is illustrated by Fig. 3. The tree evolved from $\text{Lesion}_{\text{color}}$ features in Parent 1 can only crossover/mutate with the tree evolved from the same $\text{Lesion}_{\text{color}}$ features in Parent 2, and it cannot crossover/mutate with any of the other three trees evolved from $\text{Lesion}_{\text{shape}}$, LBP_{gray} or LBP_{RGB} features. Hence, this type of crossover/mutation ensures that at the end of the evolutionary process, the evolved GP individual consists of four trees, each evolved using a single type of features to avoid different types of features with similar ability causing confusion to the GP system.

D. Fitness Function

For evaluating each individual in the proposed multi-tree GP approach, we proposed a weighted fitness function, where

the weights are assigned based on the classification accuracy of each tree in one GP individual. The fitness is defined as

$$fitness = \sum_{i=1}^{n} (W_i \times accuracy(t_i)) \quad (3)$$

$$W_i = \frac{accuracy(t_i)}{\sum_{i=1}^{n} accuracy(t_i)} \quad (4)$$

$$accuracy(t_i) = \frac{1}{2}\left(\frac{TP_i}{TP_i + FN_i} + \frac{TN_i}{TN_i + FP_i}\right) \quad (5)$$

here n is the number of trees and t_i is the i^{th} tree in a GP individual, W_i is the weight assigned to the i^{th} tree, and $accuracy\,(\cdot)$ is the balanced accuracy among the two classes given by Equation (5). TP, TN, FP, and FN refers to true positive, true negative, false positive, and false negative, respectively. Each tree in an individual also works as a simple classifier that can classify binary problem: if an instance x has a negative value on the constructed high-level feature, GP will classify x to "*benign*" class; otherwise to "*malignant*" class.

Equation (5) is more appropriate to use balanced accuracy than standard overall accuracy since it can cope well with the class imbalance problem. Using Equation (3) as the fitness function, we allow all the trees to be able to evolve during the evolutionary process and the tree having higher accuracy would contribute more towards the fitness of that individual, via being allocated a higher weight. In [7], average accuracy of the trees is used as a fitness function in the multi-tree representation, which allows all the trees to grow while giving equal importance to all the four trees. However the performance of one tree has no influence on the performance of other trees. In other words, the interaction between trees during the evolutionary process was quite limited. Therefore, we designed a new fitness function in this work to evolve better GP individuals, where trees influence each other's performance and interacts during the evolutionary process. It is important to note here that the interaction between trees is not in terms of genetic operators (crossover and mutation), but via the weighted fitness function, which encourages the GP method to search for an individual with all the four trees having high classification accuracy, not only one tree like in [7]. Furthermore, after getting an evolved model on the training data, each tree in a GP individual often produces a different accuracy on the training data. Among these trees, we take the top two highest performing trees on the training data and use them classify unseen test data. This is to use the power of two models (two trees) to increase the confidence of the prediction.

IV. Experiment Design

A. Datasets

1) PH^2 *dataset:* This dataset [23] contains dermoscopy images captured from a specialised instrument for skin cancer images called dermatoscope. Such high quality images are rich enough for skin cancer classification. The dataset consists of 200 images of three classes: common nevi (80 instances), atypical nevi (80 instances), and melanomas (40 instances). In dermatology, common nevi refers to non-disease lesion (mole), atypical nevi refers to a currently non-disease lesion, but may develop malignancy later, whereas melanoma is the diseased lesion. For the experiments on binary classification, 80 common nevi and 80 atypical nevi are used as "*benign*" class, and 40 melanoma are used as "*malignant*" class. Samples of the two classes are shown in Fig. 4.

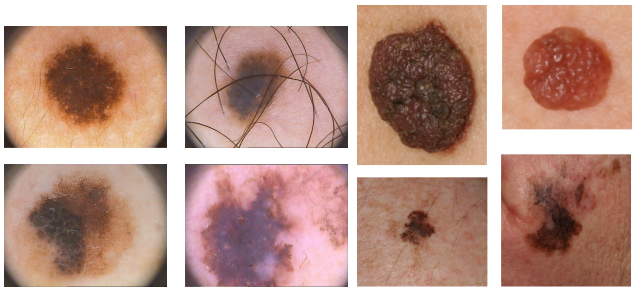

Fig. 4. Samples: first 2 columns are from PH^2, and second 2 columns are from Dermofit.

TABLE II
Parameter Settings of the GP method.

Parameter	Value	Parameter	Value
Generations	100	Crossover Rate	0.80
Population Size	1024	Mutation Rate	0.19
Initial Population	Ramped half-and-half	Elitism	0.01
Selection type	Tournament	Tree minimum depth	2
Tournament size	7	Tree maximum depth	6

2) Dermofit dataset: The Dermofit Image Library [24] is a set of 1300 high quality skin lesion images collected under standardized conditions with internal color standards, captured from a standard camera. The lesions span across 10 different classes, where each image has a gold standard diagnosis. Images consist of a snapshot of the lesion surrounded by normal skin. For evaluating the binary classification methods, we have used two classes; 1) Melanocytic Nevus (mole) with 331 images as "*benign*", and 2) Malignant Melanoma with 76 images as "*malignant*". Samples of the two classes are shown in Fig. 4.

B. GP Settings

The parameter settings of the proposed multi-tree GP method are listed in Table II. The evolutionary process keeps evolving until a stopping criterion is met, which is either reaching a maximum of 100 generations or when a perfect individual with accuracy 100% is found.

10-fold cross validation is used in the experiments on these two datasets, where the ratio of instances of each class in each fold remains the same as in the original dataset. For each GP training process, 9 folds are used. The evolved final GP individual consists of four trees, where the top two trees (in terms of training accuracy) are selected and are used to classify the test data (1-fold). This procedure is repeated 10 times using all the different combinations of folds to get the average result of *10-fold cross validation*. This gives the result for a single GP run. Each GP method has been conducted 30 runs on each dataset, so the above procedure is repeated

30 times. At the end, we get 30 accuracy values each for training and test instances. The single tree GP and MTGP methods are implemented using the Evolutionary Computing Java-based package [25].

C. Classification Methods for Comparison

In order to check the performance of our proposed multi-tree GP method, we have used six classification algorithms: Naïve Bayes (NB), k-Nearest Neighbor (k-NN) where $k = 1$, SVMs, Decision Trees (J48), Random Forest (RF), and Multilayer Perceptron (MLP). These methods are implemented through the Waikato Environment for Knowledge Analysis (WEKA) package [26]. Similar to the existing approaches [10], [12], [27], we have used a Radial basis Function (RBF) kernel instead of the default linear kernel in WEKA. The RBF kernel helps derive complex relations between the skin lesion classes and complex nonlinear skin lesion data represented as a feature vector space [18]. For MLP, the learning rate, momentum, training epochs and the number of hidden layers are set to 0.1, 0.2, 60, and 20, respectively. These parameters are taken from the previous studies [7], [10] where they are specified empirically as they gave best performance among other settings.

V. Results and Discussions

A. Overall Results

The results of the experiments are presented in Table III. Vertically, the table is divided into three blocks where the first shows the results of the proposed multi-tree GP method (MTGP) and the baseline method [7] ($\mathrm{MTGP_{old}}$), the second shows results of the six non-GP classification methods, and the third shows results of single tree GP methods each using one type of features. Horizontally, the table is divided into five columns where first lists the classification algorithm, second and third show, respectively, the training and test accuracies for the $\mathrm{PH^2}$ dataset, and fourth and fifth show these performances for the Dermofit dataset. The values of these results are the mean and standard deviation of the 30 runs of results.

In order to compare the performance of different methods, *Wilcoxon signed-rank test* with the significance level of 5% is used here. This statistical test is applied on the test results to check which method has better discriminating ability between benign and malignant classes. The symbols "+", "=" and "−" are used to represent significantly better, not significantly different, and significantly worse performance, respectively, of the proposed MTGP method in comparison with other methods. For example, in case of the $\mathrm{PH^2}$ dataset, the test performance of MLP is represented as "$78.44 \pm 10.96+$", where the "+" sign represents that MTGP significantly outperformed the MLP classification method.

The results of the statistical test has clearly shown the effectiveness of the proposed MTGP method with a weighted fitness function. It has been observed that the proposed method has not only significantly outperformed all non-GP methods, but has also outperformed all single tree GP methods. Furthermore, to highlight the impact of incorporating the new weighted fitness function into the multi-tree representation on finding better solutions, the proposed method has significantly outperformed the baseline ($\mathrm{MTGP_{old}}$) method as presented in Table III.

TABLE III
Comparison between the proposed Multi-tree GP method, the non-GP and single-tree GP Classification methods: Accuracy (%) on the training and test set of the two datasets (represented in terms of mean and standard deviation ($\bar{x} \pm s$)).

		$\mathrm{PH^2}$		Dermofit	
		training	test	training	test
	MTGP	81.62 ± 1.30	**81.08 ± 1.22**	77.33 ± 0.95	**77.30 ± 1.50**
	$\mathrm{MTGP_{old}}$	81.36 ± 1.01	78.61 ± 2.00 +	76.79 ± 1.02	75.03 ± 1.90 +
Non-GP Methods	NB	93.85 ± 1.11	77.81 ± 08.44 +	86.42 ± 0.70	72.26 ± 11.62 +
	SVM	89.62 ± 1.37	70.00 ± 10.29 +	95.16 ± 0.84	70.02 ± 10.34 +
	KNN	100.0 ± 0.00	75.63 ± 14.71 +	100.0 ± 0.00	72.08 ± 09.52 +
	J48	97.05 ± 2.71	71.25 ± 11.08 +	97.09 ± 1.31	73.98 ± 10.65 +
	RF	100.0 ± 0.00	76.56 ± 09.81 +	99.93 ± 0.22	71.30 ± 09.80 +
	MLP	78.92 ± 1.23	78.44 ± 10.96 +	79.83 ± 1.95	73.00 ± 08.51 +
Single-tree GP	$\mathrm{Lesion_{color}}$	83.24 ± 2.57	64.96 ± 3.82 +	81.91 ± 1.41	74.02 ± 2.97 +
	$\mathrm{Lesion_{shape}}$	79.70 ± 2.22	50.20 ± 5.21 +	75.92 ± 2.76	62.51 ± 6.82 +
	$\mathrm{LBP_{RGB}}$	85.64 ± 1.65	73.27 ± 2.30 +	77.02 ± 1.93	63.61 ± 3.14 +
	$\mathrm{LBP_{gray}}$	84.68 ± 1.66	65.96 ± 3.90 +	75.03 ± 2.13	60.02 ± 3.66 +

While comparing the results of the proposed method with the traditional single tree GP methods, it shows that the new MTGP method has more potential to evolve better classification models as compared to single tree GP methods. Using a new weighted fitness function has enabled the different trees in an individual to interact with each other and generate better models than the previous approach [7]. Such an interaction greatly influence the way how a tree in a GP individual searches for a better solution.

Among the two datasets, different types of features are prominent in playing the role of classification. In case of the $\mathrm{PH^2}$ dataset, the $\mathrm{LBP_{RGB}}$ features have shown the highest performance (73.27 ± 2.30) among the four single tree GP methods. In case of Dermofit dataset, the $\mathrm{Lesion_{color}}$ features have produced best results (74.02 ± 2.97) among the four single tree GP methods. From these results, it can be seen that images captured from a dermatoscope (a specialised instrument for skin cancer images, such as in $\mathrm{PH^2}$), $\mathrm{LBP_{RGB}}$ has the highest ability to discriminate between "*malignant*" and "*benign*" classes, whereas for images taken from a standard camera (such as in Dermofit dataset), $\mathrm{Lesion_{color}}$ dominate other types of features to discriminate between classes. Therefore, we can conclude that images captured from different instruments need different feature extraction methods to obtain necessary information important for distinguishing between classes. We have also seen the same behavior in the multi-tree approach while evolving an individual. Among the four trees, on the $\mathrm{PH^2}$ dataset $\mathrm{LBP_{RGB}}$ features gave the highest accuracy most of the cases. In case of the Dermofit dataset, the tree representing $\mathrm{Lesion_{color}}$ features usually has the highest accuracy. However, a tree producing a very good performance on the training data, might not achieve good results on the test data. Therefore, in order to maintain better results on the test data, we cannot

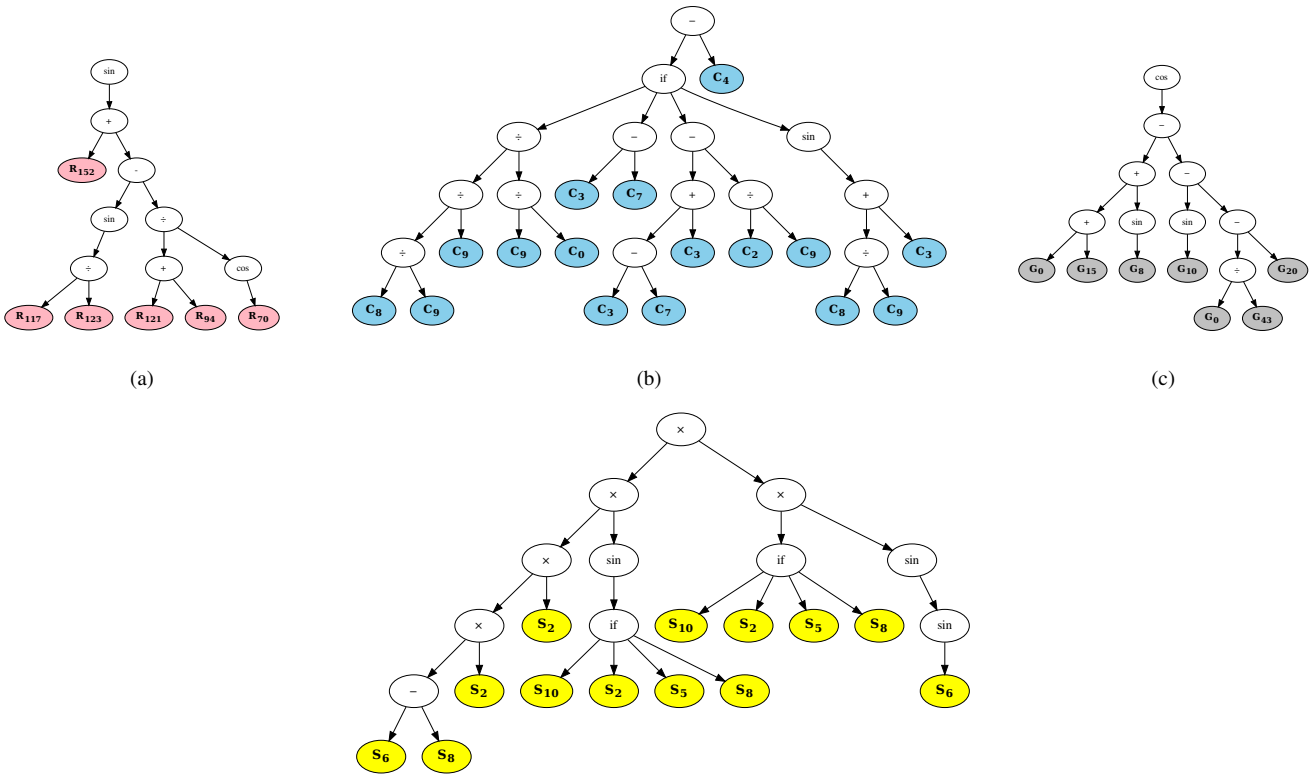

Fig. 5. A good evolved GP individual for PH^2 dataset using a)$\mathrm{LBP}_{\mathrm{RGB}}$, b) $\mathrm{Lesion}_{\mathrm{color}}$, c) $\mathrm{LBP}_{\mathrm{gray}}$, and d) $\mathrm{Lesion}_{\mathrm{shape}}$ features.

rely on using only one tree. Therefore, we used the two highest performing trees to check the performance on the unseen data. In this way, on the PH^2 dataset, if the $\mathrm{LBP}_{\mathrm{RGB}}$ features alone cannot achieve good performance on test data, we can still rely on and achieve better results by using $\mathrm{LBP}_{\mathrm{gray}}$ features having the second highest performance. This is one of the advantage of the MTGP approach, where a single evolved individual consists of more than one tree and it is possible to use more trees to get increased performance. Due to this characteristic of the proposed MTGP method, it has outperformed not only the single tree GP and non-GP classification methods (Table III) but also the existing approach [7].

B. Program Analysis

GP evolved models that can be interpreted as they clearly show which features are potentially prominent in discriminating between classes. To analyse why the proposed MTGP method achieved good performance, a good evolved GP individual is shown in Fig. 5 from the *Dermofit* experiments. The individual has four trees, each evolved using one of the four types of features a)$\mathrm{LBP}_{\mathrm{RGB}}$, b) $\mathrm{Lesion}_{\mathrm{color}}$, c) $\mathrm{LBP}_{\mathrm{gray}}$, and d) $\mathrm{Lesion}_{\mathrm{shape}}$ achieving 77.94% accuracy on the test data. In Fig. 5, white nodes represent functions and colored nodes represent terminals. While evolving this model on the training data, the individual accuracy values for $\mathrm{LBP}_{\mathrm{gray}}$ tree, $\mathrm{LBP}_{\mathrm{RGB}}$ tree, $\mathrm{Lesion}_{\mathrm{color}}$ tree, and $\mathrm{Lesion}_{\mathrm{shape}}$ tree are 66.84%, 79.14%, 85.49% and 80.08%, respectively. As discussed earlier, for *Dermofit* dataset $\mathrm{Lesion}_{\mathrm{color}}$ features has the highest potential to distinguish between melanoma and benign lesions as compared to other types of feature. However, the feature types of $\mathrm{Lesion}_{\mathrm{shape}}$ and $\mathrm{LBP}_{\mathrm{RGB}}$ are also giving good accuracy. Hence, performance can be further improved with the benefit from these trees. Due to the fact that a high performing tree on the training data might not produce good result on the test data (overfitting), relying on just one evolved tree might not produce fruitful results (as has been observed from comparing the results of $\mathrm{MTGP}_{\mathrm{old}}$ to that of the proposed method).

From Fig. 5(b) in the $\mathrm{Lesion}_{\mathrm{color}}$ tree, the features C_9 and C_3 were selected 5 and 4 times, respectively. Also the expressions $(C_3 - C_7)$ and (C_8/C_9) appear 2 times, which shows that these features have high discriminating ability. This is the highest performing tree among the four trees in this individual. Among the total of 177 $\mathrm{LBP}_{\mathrm{RGB}}$ features, a tree in Fig. 5(a) constructed from only six dominant features (R_{152}, R_{117}, R_{123}, R_{121}, R_{194}, R_{70}) has shown 79.14% accuracy on the training data. In $\mathrm{Lesion}_{\mathrm{color}}$ tree in Fig. 5(b), C_3, C_7, C_8 and C_9 (corresponding to σR, $\frac{\mu_R}{\mu_B}$, $\frac{\mu_G}{\mu_B}$ and $\frac{\mu_R}{\mu_R}$) showing the variance of red channel, ratios of mean of 1) the red channel lesion area and the blue channel lesion area, 2) the green

channel lesion area and the blue channel lesion area, and 3) the red channel lesion area and the red channel skin area, are most important. In $\text{Lesion}_{\text{shape}}$ tree in Fig. 5(d), S_2, S_5, S_8, S_9, and S_{10} are selected, which correspond to greatest diameter, irregularity indices A, and B, major asymmetry index, and Asymmetry Index, respectively. These border shape features can provide essential information to the dermatologist in diagnosing melanoma.

VI. Conclusion

This work has developed a new fitness function in a multi-tree GP method for the task of skin cancer image classification. Various local and global features are used, which have information regarding pixel-based gray-level and RGB characteristics, color variation across the image (inside and between lesion and skin regions) and geometrical border shape properties. An individual consists of four trees, each evolved with one type of features. To meet this requirement, genetic operators such as crossover and mutation are designed accordingly, called *same-index-crossover/mutation*. The use of a new weighted fitness function has provided better solutions as compared to the existing method where average accuracy has been used to evolve a GP model. This fitness function allows the four trees to influence each other's performance during the evolutionary process. Moreover, the top two highest performing trees on the training data are selected and used to measure the performance of the unseen test instances. Our method has outperformed all the single-tree GP methods and all the most commonly used classification algorithms, showing evidence of good discriminating ability between "$malignant$" and "$benign$" skin lesions. We have also found an interesting behavior that different types of features are most prominent for different types of images captured from different acquisition devices.

In the future, we would like to investigate GP for feature extraction directly from skin cancer images. Moreover, for real-world images, how to reduce noise without losing discriminative features still requires a lot of research.

References

[1] F. Ercal, A. Chawla, W. V. Stoecker, H.-C. Lee, and R. H. Moss, "Neural network diagnosis of malignant melanoma from color images," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 9, pp. 837–845, 1994.

[2] F. Adjed, S. J. S. Gardezi, F. Ababsa, I. Faye, and S. C. Dass, "Fusion of structural and textural features for melanoma recognition," *IET Computer Vision*, vol. 12, no. 2, pp. 185–195, 2017.

[3] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Transactions on Medical Imaging*, vol. 36, no. 4, pp. 994–1004, 2017.

[4] F. Xie, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik, "Melanoma classification on dermoscopy images using a neural network ensemble model," *IEEE Transactions on Medical Imaging*, vol. 36, no. 3, pp. 849–858, 2017.

[5] K. Shimizu, H. Iyatomi, M. E. Celebi, K.-A. Norton, and M. Tanaka, "Four-class classification of skin lesions with task decomposition strategy," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 274–283, 2015.

[6] C. Barata, M. E. Celebi, and J. S. Marques, "Development of a clinically oriented system for melanoma diagnosis," *Pattern Recognition*, vol. 69, pp. 270–285, 2017.

[7] 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*. Springer, 2018, pp. 111–123.

[8] 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.

[9] W. Abbes and D. Sellami, "High-level features for automatic skin lesions neural network based classification," in *Image Processing, Applications and Systems (IPAS), 2016 International*. IEEE, 2016, pp. 1–7.

[10] Q. U. Ain, B. Xue, H. Al-Sahaf, and M. Zhang, "Genetic programming for feature selection and feature construction in skin cancer image classification," in *Proceedings of the 15th Pacific Rim International Conference on Artificial Intelligence*, ser. Lecture Notes in Computer Science. Springer, 2018, vol. 11012, pp. 732–745.

[11] R. Kasmi and K. Mokrani, "Classification of malignant melanoma and benign skin lesions: implementation of automatic abcd rule," *IET Image Processing*, vol. 10, no. 6, pp. 448–455, 2016.

[12] H. Al-Sahaf, A. Al-Sahaf, B. Xue, M. Johnston, and M. Zhang, "Automatically evolving rotation-invariant texture image descriptors by genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 1, pp. 83–101, 2017.

[13] J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection*. MIT press, 1992, vol. 1.

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

[15] J.-H. Lee, C. W. Ahn, and J. An, "An approach to self-assembling swarm robots using multitree genetic programming," *The Scientific World Journal*, vol. 2013, 2013.

[16] 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, April 2004.

[17] H. Al-Sahaf, B. Xue, and M. Zhang, "A multitree genetic programming representation for automatically evolving texture image descriptors," in *Proceedings of the 11th International Conference on Simulated Evolution And Learning*, ser. Lecture Notes in Computer Science, vol. 10593. Springer, 2017, pp. 499–511.

[18] T. Satheesha, D. Satyanarayana, M. G. Prasad, and K. D. Dhruve, "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.

[19] 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.

[20] W. Stolz, A. Riemann, A. B. Cognetta, L. Pillet, W. Abmayr, D. Holzel, P. Bilek, F. Nachbar, and M. Landthaler, "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.

[21] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 1146–1152, 2015.

[22] 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.

[23] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "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, 2013, pp. 5437–5440.

[24] 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 *Proceedings of the Color Medical Image Analysis*. Springer, 2013, pp. 63–86.

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

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[27] Q. U. Ain, B. Xue, H. Al-sahaf, and M. Zhang, "Genetic programming for skin cancer detection in dermoscopic images," in *Proceedings of the 2017 Congress on Evolutionary Computation*, 2017, pp. 2420–2427.