Genetic Programming for Skin Cancer Detection in Dermoscopic Images

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*—Development of an effective skin cancer detection system can greatly assist the dermatologist while significantly increasing the survival rate of the patient. To deal with melanoma detection, knowledge of dermatology can be combined with computer vision techniques to evolve better solutions. Image classification can significantly help in diagnosing the disease by accurately identifying the morphological structures of skin lesions responsible for developing cancer. Genetic Programming (GP), an emerging Evolutionary Computation technique, has the potential to evolve better solutions for image classification problems compared to many existing methods. In this paper, GP has been utilized to automatically evolve a classifier for skin cancer detection and also analysed GP as a feature selection method. For combining knowledge of dermatology and computer vision techniques, GP has been given domain specific features provided by the dermatologists as well as Local Binary Pattern features extracted from the dermoscopic images. The results have shown that GP has significantly outperformed or achieved comparable performance compared to the existing methods for skin cancer detection.**

I. Introduction

Computer-Aided Diagnosis (CAD) systems help in early detection of melanoma [1]. Melanoma is the deadliest type of skin cancer as it grows quickly and therefore can be life-threatening. The cause of melanoma is the uncontrolled growth of melanocytes, which are the pigment-producing cells responsible for giving color to skin, eyes and hair. Early detection of melanoma is vital for reducing death rate and treatment costs caused by this disease as well. Dermoscopy has proven ability to significantly increase the diagnostic performance as it magnifies the skin lesion or mole (cite for melanoma) up to 100 times. Hence it allows the dermatologist to have a clear view of several structures inside the lesions that are invisible to the naked eye [2]. There are several existing medical procedures that help dermatologists for classification of melanoma for example, Asymmetry, Border, Color and Diameter (ABCD) rule and 7-point checklist (atypical pigment network and vascular pattern, irregular streaks, dots and pigmentation, regression structures and blue-whitish veil). The rich images provided by dermoscopy allows researchers to apply principles of computer vision and machine learning to the challenge of interpreting dermoscopic images, particularly evolving methods for skin cancer classification [3].

Image classification has gained immense importance in recent years as image processing and machine vision are widely used in daily life applications [4] such as face recognition, medical imaging systems and remote sensing. Genetic Programming (GP) is a method that provides solution to a user-defined problem by evolving a computer program. It is inspired by the Darwinian Principle of natural selection [5] that includes operators like selection, crossover and mutation to evolve diverse solutions for the problem. With the flexible representation, GP has provided promising solutions that hardly can be thought of by humans. GP has been extensively employed for pattern recognition, object detection and classification [4], [6], [7]. Moreover, GP methods have been employed for various tasks such as feature extraction [8] which is a process of transforming images into feature values, feature selection [9] which selects good features among whole set of features, and feature construction [10] which builds new features from existing feature set. Feature selection and construction aim at dimensionality reduction that reduces the search space for GP for evolving good solutions meanwhile speeding up the search process by using less number of features in the reduced search space. To find goodness of a feature or its contribution in classifying instances from different classes, various feature ranking methods have been applied. A mutual information based measure is used in this work to find out features which are most prominent to distinguish between diseased and non-diseased instances.

For evolving a classifier, image data needs to be in a format with which GP can deal with. For this task, image descriptors have been extensively used to convert raw pixel values into features that can easily be incorporated into GP to build a classifier [6]. There are two categories of image descriptors: sparse and dense [11]. Sparse descriptors use a number of regions for extracting features, whereas dense descriptors work in a pixel-by-pixel approach. Local binary patterns (LBP) [12], a dense image descriptor, has been widely used in the field of pattern recognition. For extracting features from dermoscopic images, LBP has been used in this work.

With the ability of GP to select good features that can improve the classification performance and its diverse range of solutions for a single problem, it can be applied to dermoscopic images to evolve computer programs that have the potential to classify the diseased images effectively. In the field of medicine, the balance between sensitivity and specificity varies depending, for instance, on whether you are doing screening or confirming a diagnosis. Incorrectly diagnosing a disease is worse than not diagnosing the disease at all [13].

978-1-5090-4601-0/17/$31.00

Therefore, we aim to build a classification model using GP to identify diseased images using LBP and domain specific features provided by the dermatologist.

A. Objectives

The overall aim of this study is to evolve a GP-based classifier, particularly for skin cancer detection using features extracted from dermoscopic images and features based on domain specific knowledge. This study aims at providing answers for the following research questions:

- Which features (among domain specific, LBP and the combination of both) perform better for classifying diseased and non-diseased instances?
- Would the GP approach perform better than other well-known classification methods for skin cancer detection?
- For dimensionality reduction, whether the classification accuracy can be improved when the features selected by GP are used?
- Which features get selected by GP during its evolutionary process to achieve better performance and why?

B. Organization

The rest of the paper is organized as follows. Section II describes the relevant previous work, discusses the LBP method and mutual information calculation. Section III presents the GP system including the program representation, search space, fitness function and describes the evaluation procedure. Section IV describes the experiments performed, dermoscopic dataset, GP parameters and methods for comparison used. Section V presents the results of these experiments and discusses how well they address the research questions, and examines four good evolved solutions. Section VI concludes the paper with the achievements of current work, and provides some possible future directions.

II. Background

This section describes the existing methods on skin cancer diagnosis and GP related classification and feature selection work. Moreover, LBP as directly related to this work and mutual information calculation are also discussed.

A. Literature Survey

For the detection of melanoma, a classification method has been proposed based on Artificial Neural Networks (ANNs) in [14]. The method consists of four stages: preprocessing, pigment network extraction, feature extraction and classification. Preprocessing involves removal of noise such as air bubbles, artefacts caused by applying gel on skin before capturing the image, and hair on skin. The thresholding and directional Gabor filter is applied to the blue component of images for the first stage. For pigment network detection, again the Gabor filter is applied with different thresholding values. For feature extraction, mean and standard deviation are computed on the pixel values of the sub-images with a window of size 2×2 or 4×4 pixels. These statistical parameters are taken due to the fact that cancerous pigmented network has irregular distribution and thickened lines as compared to normal benign lesions. Classification is then performed using ANNs fed with the extracted features. The performance is assessed by the commonly used classification accuracy measure and the method achieved 94% accuracy. In medical diagnosis, when the data available is mostly unbalanced (having different number of instances from different classes), using the classification accuracy as the fitness evaluation measure is however not recommended. It is due to the fact that high performance will be achieved by correctly classifying only the non-disease instances (which often outnumber the diseased instances in medical data); however, performing poorly on diseased instances which are the main concern. Hence, a different measure is used in this study to deal with the data imbalance problem.

Piccolo et al. [15] focus on validating the use of digital dermoscopy by comparing melanoma classification diagnosis of experienced dermatologists with computer-aided diagnosis based on ANNs and also with diagnosis provided by minimal trained clinicians. The results are given in terms of sensitivity and specificity of 92% and 99%, respectively, for the trained dermatologist, 69% and 94%, respectively, for the clinician, and 92% and 74%, respectively, for the computer analysis. Sensitivity is the true positive rate and specificity is the true negative rate. According to the results obtained, the authors have suggested computer analysis must be developed in order to assist and not to replace physicians in the diagnosis of skin cancer lesions as the best diagnostic results can be achieved by using both trained computer classifier and experienced dermatologist diagnosis. Hence, inspired by their work, we have used both domain specific features and LBP features to automatically evolve a good classifier by GP.

Variation in color of melanoma is a major discriminative aspects for dermatologists that is studied in [2]. This paper evaluates the importance of color in key-points detection steps of the bag-of-features model for the classification of melanoma images based on k-Nearest Neighbor(k-NN). Furthermore, gray-scale and color sampling methods using Harris Laplace detector and its color extensions are compared. The performance of scale-invariant feature transform (SIFT) and Color-SIFT patch descriptors is also analyzed. The method achieved 85% sensitivity, 87% specificity and 87% balanced accuracy. These results cannot be compared with our results as they have used reduced dataset. To the best of our knowledge, GP has not yet been studied for skin cancer detection in dermoscopic image data. Hence, this will be the first time to utilize GP for detecting skin cancer in dermoscopic images.

In [13], a method for brain tumour classification on magnetic resonance imaging (MRI) is proposed based on statistical methods for preprocessing, fuzzy c-means for brain image segmentation and GP for tumour classification that achieved 97% accuracy. Early detection of defective nodules in lung computed tomography (CT) images increase the survival rate of the patients by 50%, hence, a GP-based nodule detection method is developed in [16] and achieved 92% detection rate.

Earlier in 1996, Poli [17] described a set of requirements

for terminal set, function set and fitness function in GP to evolve efficient optimal filters for the tasks of feature detection and image segmentation, and studied their behaviour in brain MRI and X-ray coronarograms. They have compared their results with ANNs and reported that GP has outperformed the competitor method. ANNs gave 31.7% sensitivity and 92.2% specificity, whereas GP achieved 61.5% sensitivity and 99.2% specificity. With better results obtained by GP, the author has further elaborated that GP has far better ability for image analysis as compared to other existing methods. Therefore, motivated by their findings, GP is employed for skin cancer classification in this study.

B. Local Binary Patterns

The local binary patterns (LBP) image descriptor proposed by Ojala et al. [12] is a dense image descriptor that has been used extensively for feature extraction in a wide range of computer vision applications. LBP works by scanning the image in a pixel-by-pixel fashion using a sliding window of fixed radius, where the value of the central pixel is computed based on the intensities of neighbouring pixels that lie on the radius as depicted in Figure 1. It also generates a histogram (i.e. feature vector) based on the computed values. The LBP operator is formally defined as:

$$LBP_{p,r} = \sum_{i=0}^{p-1} t(v_i - v_c)2^i \tag{1}$$

where r is the radius, p is the number of neighbouring pixels, v_i and v_c are the intensity values of the i^{th} neighbour and central pixel, respectively. Here $t(x)$ returns 1 if $x \geq 1$ and 0 otherwise. The value computed from above expression is assigned to central pixel and corresponding bin of histogram is incremented by 1. The value of b^{th} bin of a histogram H computed on an image of size $m \times n$ is given as:

$$H(b) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (LBP_{p,r}(V_{i,j}) = b) \tag{2}$$

where the value of b ranges between 0 and $B-1$, B is the maximum number of bins in the histogram, $V_{i,j}$ is the value of the pixel at coordinate (i, j).

Furthermore, the LBP codes are divided into two categories: *uniform* and *non-uniform*. A code is uniform if circularly it does not have more than two bitwise transitions from 0 to 1 or 1 to 0. For example, the codes 00000110, 01111110, and 00001000 are uniform, whilst the codes 00110011, 11001110, and 01010101 are non-uniform. The size of feature vector can be reduced from 2^p bins to $p(p-1)+3$ bins by omitting non-uniform codes. Moreover, using only uniform codes, allows to detect various texture primitives such as corners, edges, line ends, dark spots and flat regions. In the dermoscopic images, uniform codes can help in detection of pigmented network, streaks and blobs which can largely increase the classification performance.

In our experiment, a histogram of uniform codes is generated; hence, there are 59 $(= 8 \times (7) + 3)$ LBP features for a single image. The window size of 3×3 pixels and a radius of 1 pixel ($LBP_{8,1}$) is used.

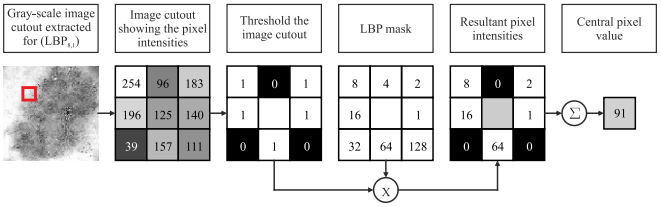


Fig. 1. Step-by-Step procedure to generate $LBP_{8,1}$ code for image cut-out (having 8 neighbouring pixels and radius = 1) and get a decimal value of the central pixel.

C. Mutual Information

Entropy and mutual information are used to measure the information of random variables [18]. Entropy, $H(X)$ measures the uncertainty of random variables (features) and Mutual information, $I(X, Y)$ shows the shared information between two variables. Conditional entropy, $H(X|Y)$ is the average uncertainty about a feature X after observing knowledge of class label Y.

$$H(X) = -\sum_{x \in \mathcal{X}} (P(x))\ log_2\,(P(x) \tag{3}$$

$$H\,(X|Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P\,(x, y)\, log_2 P\,(x|y) \tag{4}$$

where $P(x)$is the probability of feature X being x, $P(x, y)$ is the joint probability of feature X and Y being x and y, respectively. $P(x|y)$ is the posterior probability of X given Y. Mutual Information is calculated as the difference between entropy of a feature X (given in Equation (3)) and conditional entropy of that feature (given in Equation (4)), with respect to the class label Y. Formally $I(X, Y)$ is given by:

$$\begin{aligned} I\,(X, Y) &= H\,(X) - H\,(X|Y) \\ &= -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P\,(x, y)\, log_2 \frac{P\,(x, y)}{P\,(x)\,P\,(y)} \end{aligned} \tag{5}$$

When feature X is closely related to class Label Y, mutual information will be high showing X as a powerful feature with good discriminating ability between classes. For validating why GP selects particular features during the evolution process and improves performance, we used mutual information as a feature ranking method.

III. The GP Method

This section discusses the proposed GP method. The method uses two types of features; one extracted from images using LBP, and the other non-image domain specific features. This section describes the function set, terminal set and fitness function. The dermoscopic images are converted to gray-scale images before applying LBP.

A. Terminal Set

The terminal set consists of LBP features (extracted from LBP) and domain features (provided by the dermatology experts along with the dataset). There are a total of 71 numeric features (= 59 (LBP) +12 (domain)). However, three scenarios are used in using these features: firstly, by using only domain features; secondly, using only LBP features; and thirdly, by using both LBP and domain features to check the performance of the system. The value of the i^{th} feature is indicated as F_i.

B. Function Set

The function set consists of four arithmetic operators, one conditional and two trigonometric functions $\{+, -, \times, /, if, sin, cos\}$, where the first three and last two operators have the same arithmetic and trigonometric meaning; whereas division is protected that returns zero when the denominator is 0. The if operator takes four input arguments and returns the third if the first is greater than the second; otherwise, it returns the fourth argument [7].

C. Fitness Function

As the dataset is imbalanced (having very different number of instances in different classes), therefore the standard overall classification accuracy (given in Equation (6)) defined as the ratio between correctly classified instances $(N_{correct})$ and total number of instances (N_{total}), cannot be used as the performance measure for our data imbalance problem.

$$Accuracy = \frac{N_{correct}}{N_{total}} \tag{6}$$

Instead, we used a fitness function proposed by Patterson et al. [19] which gives equal importance to both classes without any bias. Mathematically it is represented as:

$$Fitness = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{7}$$

where TP refers to true positive which is the number of correctly classified diseased instances, TN refers to true negative which is the number of correctly classified non-diseased instances, FP refers to false positive which is the number of incorrectly classified non-diseased instances as diseased, and FN refers to false negative which is the number of incorrectly classified diseased instances as non-diseased. The fitness value ranges between 0 and 1, where 1 is the *ideal* case representing all the instances being correctly classified. In the paper, the results are represented as percentage, which is achieved by taking product of obtained fitness value and 100.

IV. Experiment Design

A. Dataset

A dataset of dermoscopic images namely PH^2 [1] acquired from Pedro Hispano Hospital Portugal, is used in the experiments. The dataset includes images of skin lesions, their clinical diagnosis, their binary masks and information of domain specific features provided by the dermatologists, based on 7-point checklist Menzies method. Hence, according to clinical diagnosis there are 80 common nevi, 80 atypical nevi, and 40 melanomas among the 200 melanocytic lesions. In dermatology, common nevi refers to non-disease lesion, atypical nevi refers to a lesion that is currently non-disease, but can develop melanoma at a later stage in patient's life, and melanoma is the diseased lesion. Samples of the three categories of skin lesions are presented in Figure 2. For our experiments of binary classification, 80 common nevi and 80 atypical nevi are used as one class denoted as "non-melanoma", and 40 melanoma are used as the other class denoted as "melanoma".

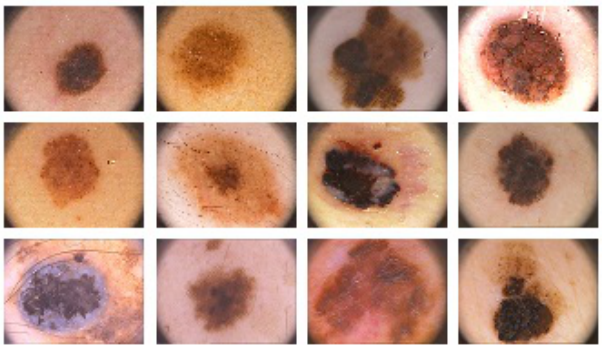


Fig. 2. Some Images of dermoscopic dataset, with common nevi (row 1), atypical nevi (row 2) and melanomas (row 3).

The dermoscopic images were obtained from Tuebinger Mole Analyzer system with a magnification of 20× and resolution of 768×560 pixels. Dermoscopy includes using an optical instrument having powerful lighting system to examine skin lesions in a higher magnification. Before taking the image, a gel is placed on the lesion that enables the dermatoscope (instrument) to capture morphological structures and patterns in inner layers of human skin. Hence, such images are rich enough to investigate them for presence of skin cancer. The images are 8-bit RGB (red, green and blue) color images. An expert dermatologist evaluated each image by considering these important classification parameters; manual segmentation of skin lesion, clinical and Histopathology diagnosis, and dermoscopic criteria based on 7-point checklist (Asymmetry, Pigment network, Dots/Globules, Streaks, Regression areas, Blue-whitish veil and presence/absence of Colours; white, red, light-brown, dark-brown, blue-gray, black). Therefore, a total of 12 dermoscopic features are provided for each image. This dataset has been used in [20], [2] for detecting pigment network using directional filters and melanoma classification using a bag-of-features model, respectively.

The dataset is split into training and test with 70% instances for training and 30% instances for testing. The dataset is randomly split five times (referred to as Split-1, Split-2, ..., Split-5) to get five training sets and five test sets. For making the results unbiased of splitting (which specific instances are present in training and which in test), we have reported results from five different splits instead of one. Each pair of training and test sets has the same ratio of instances of each class

TABLE I
PARAMETER SETTINGS OF THE GP METHOD.

Parameter	Value	Parameter	Value
Generations	50	Crossover Rate	0.80
Population Size	1024	Mutation Rate	0.19
Initial Population	Ramped half-and-half	Reproduction Rate	0.01
Selection type	Tournament	Tree minimum depth	2
Tournament size	7	Tree maximum depth	8

as present in the original dataset. For illustration, the first split "Split-1" has 140 (= 70% of 200) instances for training and 60 (= 30% of 200) instances for testing. Furthermore, among 120 training instances, 112 are non-melanoma and 28 are melanoma, keeping the original ratio (4:1) of class distribution same. Similarly, among the 60 test instances, 48 are non-melanoma and 12 are melanoma.

B. GP Settings

Parameter settings for GP are given in Table I. For each independent run, the number of generations is 50. For finding good solutions, the population size is set to 1024 with "Ramped half-and-half" method for generating the initial population. For the evolution process, crossover, mutation and reproduction rates are set to 0.80, 0.19 and 0.01, respectively. Tournament selection is used with size 7. The depth of the trees is ranging between 2 and 8 levels.

The number of GP runs is 30, therefore, the average of 30 runs is computed and the standard deviation from these values are also calculated for each of the 5 splits.

C. Methods for Comparison

To check the effectiveness of our method, we have compared the performance of GP with other non-GP methods: k-Nearest Neighbor (k-NN) where $k = 1$ (the closest neighbour), Support Vector Machines (SVM) with linear kernel, Naïve Bayes (NB), Decision Trees (J48), Random Forest (RF) and Multilayer Perceptron (MLP). The learning rate, momentum and training time for MLP are set to 0.1, 0.2, and 500, respectively and it has one hidden layer with three neurons. These parameters are specified empirically as they gave the best performance amongst other settings. The implementations of all these methods are taken from the most commonly used Waikato Environment for Knowledge Analysis (WEKA) software [21]. The implementation of GP method is done using the Evolutionary Computing Java-based (ECJ) package [22].

D. Classification Tasks

We performed two tasks for checking the classification performance of GP and the non-GP methods.

1) Classification using all features: For the first experiment, we have used all the features for classification, the training process of GP is repeated for 30 times on each of the 5 splits. Here, the number of features for both, domain and LBP are 71, 12 and 59, respectively. To ensure a different starting point for each of the 30 independent runs, a different *seed* value is used each time. Hence, for GP there are 30 runs $\times$ 5 Splits $\times$ 3 feature sets $\times$ 1 GP method = 450 independent runs. Each non-GP method is evaluated only once on each of the 5 splits, therefore, for the non-GP methods, there are 5 splits $\times$ 3 feature sets (both, domain and LBP) $\times$ 6 non-GP methods = 90 independent runs.

TABLE II
THE ACCURACY (%) ON THE TEST SET USING ALL FEATURES (RESULT FOR GP IS REPRESENTED IN TERMS OF MEAN STANDARD DEVIATION ($\bar{x} \pm s$)).

	GP							
Features	best	$\bar{x} \pm s$	k-NN	NB	SVM	J48	RF	MLP
Both1	97.92	91.28 ± 4.25	81.25 +	83.33 +	90.63 =	94.79 -	73.96 +	87.50 +
Both2	95.83	90.76 ± 3.75	89.58 +	90.63 =	90.63 =	96.88 -	83.33 +	90.63 =
Both3	97.92	89.65 ± 4.30	85.42 +	80.21 +	81.25 +	87.50 +	75.00 +	85.42 +
Both4	93.75	87.92 ± 4.66	82.29 +	87.50 =	89.58 =	86.46 +	78.13 +	92.71 -
Both5	93.75	88.33 ± 4.41	86.46 =	94.79 -	81.25 +	73.96 +	90.63 -	91.67 -
Domain1	97.92	90.52 ± 5.20	73.96 +	98.96 -	88.54 +	87.50 +	91.67 =	90.63 =
Domain2	95.83	90.55 ± 4.76	89.58 =	95.83 -	93.75 -	93.75 -	97.92 -	90.63 =
Domain3	94.79	87.77 ± 4.48	85.42 +	88.54 =	85.42 +	89.58 =	85.42 +	78.13 +
Domain4	94.79	87.40 ± 3.53	75.00 +	96.88 -	89.58 -	87.50 =	90.63 -	86.46 =
Domain5	93.75	84.31 ± 4.63	87.50 -	96.88 -	81.25 +	81.25 +	89.58 -	80.21 +
LBP1	81.25	65.69 ± 7.24	87.50 -	62.50 +	65.63 +	70.83 -	65.63 =	71.88 -
LBP2	77.08	67.08 ± 6.10	78.13 -	73.96 -	62.50 +	61.46 +	70.83 -	78.13 -
LBP3	83.33	67.29 ± 5.20	64.58 +	58.33 +	61.46 +	61.46 +	62.50 +	68.75 =
LBP4	80.21	67.01 ± 5.02	60.42 +	70.83 -	69.79 -	67.71 =	57.29 +	67.71 =
LBP5	84.38	70.49 ± 7.73	75.00 -	77.08 -	58.33 +	60.42 +	66.67 +	70.83 =

2) Classification using GP-Selected features: With the help of crossover and mutation operators, GP is capable of evolving good solutions. In other words, GP is good at searching for good features during the evolution process. These dominant selected features can be used to achieve better classification performance by classifying with GP and non-GP methods. Hence, to check this property of GP, we have used the features selected by GP during the first experiment to perform the second experiment. Here, we are using 30 feature subsets obtained from the 30 solutions/trees evolved in the previous experiment (as selected by the best GP tree from each of the 30 runs). In this task, the number of independent runs for GP is 13500 (= 3 (feature sets) $\times$ 5 (splits) $\times$ 30 (selected feature subsets) $\times$ 30 (runs) $\times$ 1 (classifier)). However, the number of fitness evaluations in GP is huge, which is calculated as product of independent runs (13500), the number of generations (50) and the population size (1024), and comes out to be 7.05×10^8. The total number of independent runs for the non-GP methods is 2700 (= 3 (feature sets) $\times$ 5 (splits) $\times$ 30 (selected feature subsets) $\times$ 6 (classifiers)).

V. RESULTS AND DISCUSSIONS

The results of the two experiments are presented in Table II and Table III, respectively. Vertically, the tables comprise of blocks where each correspond to one feature-set (using only domain features, using only LBP features or using both of them), while horizontally they consist of 9 columns where the first lists feature-set, second shows the highest GP fitness achieved among its 30 runs (for the first experiment in Table II) and 900 runs (for the second experiment in Table III), and 7 for different classification methods. The values of GP are the average accuracy and the standard deviation ($\bar{x} \pm s$) computed from 30 and 900 GP runs for the first and second experiment, respectively. For the other classifiers in our first experiment, only accuracy is given as they are run only

TABLE III
THE ACCURACY (%) ON THE TEST SET USING GP-SELECTED FEATURES (RESULT ARE GIVEN IN TERMS OF MEAN AND STANDARD DEVIATION ($\bar{x} \pm s$)).

Features	GP (best)	GP ($\bar{x} \pm s$)	k-NN	NB	SVM	J48	RF	MLP
Both1	100.0	90.94 ± 4.21	89.41±4.92 =	87.43±3.45 +	91.56±1.53 =	87.85±6.44 +	88.65±4.39 +	88.65±3.41 +
Both2	97.92	90.80 ± 3.68	89.55±4.44 =	93.40±2.67 -	93.75±0.00 -	91.77±4.51 =	91.94±2.95 =	91.91±2.39 -
Both3	98.96	89.50 ± 4.33	85.45±4.39 +	82.50±6.54 +	91.22±4.40 -	88.06±4.61 =	83.37±5.56 +	85.24±3.64 +
Both4	97.92	89.00 ± 3.99	85.52±4.62 +	88.06±3.49 =	89.58±0.00 -	85.14±4.96 +	85.17±4.55 +	86.98±5.12 =
Both5	96.88	86.33 ± 4.94	84.79±4.47 =	91.91±2.84 -	87.57±2.48 -	84.44±3.80 +	88.58±3.53 -	86.77±4.57 =
Domain1	97.92	89.59 ± 5.39	84.48±6.97 +	96.39±2.49 -	88.68±2.60 =	87.22±3.03 +	92.12±4.02 -	86.77±4.80 +
Domain2	97.92	91.65 ± 4.26	91.46±3.16 =	95.76±0.76 -	93.75±0.00 -	93.75±0.00 -	97.64±2.12 -	93.96±3.30 -
Domain3	97.92	89.25 ± 4.41	85.35±2.23 +	89.24±1.46 =	86.36±3.01 +	89.83±1.80 =	85.97±2.18 +	81.08±3.70 +
Domain4	96.87	88.45 ± 4.58	82.78±6.74 +	95.28±2.03 -	89.44±1.19 -	88.13±0.95 =	91.67±2.54 -	87.01±2.15 +
Domain5	94.79	83.59 ± 5.36	87.50±4.39 -	93.40±3.38 -	82.71±2.92 +	81.74±1.82 +	86.53±3.96 -	80.66±3.64 +
LBP1	81.25	64.85 ± 6.38	72.12±6.48 -	61.28±5.35 +	50.69±1.55 +	64.93±4.41 =	65.42±0.78 =	69.90±4.95 -
LBP2	89.58	68.18 ± 6.90	73.33±5.82 -	63.99±9.45 =	50.56±1.42 +	62.50±7.28 +	65.38±5.66 +	73.40±4.30 -
LBP3	82.29	65.54 ± 5.61	63.72±4.61 +	62.64±3.44 +	52.92±3.90 +	61.84±3.99 +	61.60±3.57 +	68.06±4.93 -
LBP4	82.29	65.41 ± 6.71	62.29±5.09 +	70.31±4.06 -	50.97±1.76 +	60.00±5.59 +	58.06±2.93 +	69.03±3.85 -
LBP5	91.67	67.74 ± 7.13	72.85±4.72 -	70.42±6.23 =	50.14±0.75 +	64.55±5.16 +	65.31±3.60 +	69.90±5.64 -

once. However, for the second experiment, the values of non-GP methods are represented as mean accuracy and standard deviation computed over 30 independent runs. For making a clear comparison between GP and non-GP methods, the results are also tested using the Wilcoxon signed-rank test with a significance level of 5%. The statistical test has been applied once on the test results to check whether GP can compete with other more powerful classifiers. The symbols "+", "-" and "=" are used to represent significantly better, significantly worse and not significantly different performance, respectively.

Using GP as a classifier, it is evident from the results in Table II, that using both features in all splits are giving better classification performance as compared to the domain and LBP features. Also the standard deviation is smaller as compared to the domain and LBP features in most of the splits. The LBP features are experiencing a high standard deviation which can be the result of bad generalization on the test set as it is still achieving good highest accuracy (84.38%). The domain features are providing good performance, however, not as good as the combination of both features and the standard deviation is also worse than both features in most cases. Our results validate the statement concluded by [15] in their experiments that using knowledge from both domains (dermatology and computer vision), we can achieve better classification performance in identifying diseased and non-diseased instances. Among both features in our experiments, the knowledge of dermatology comes from domain features and knowledge of computer vision comes from LBP features.

From the results given in Tables II and III, we see that GP has evolved good classifiers specifically when using both features. GP with selected features is most prominent by successfully classifying all diseased and non-diseased instances giving 100% in some GP runs. In the Tables, a fitness of 95.83% (calculated using Equation (7)) shows that only one melanoma instance is misclassified (11 out of 12) and all non-melanoma instances are correctly classified (48 out of 48). Instead, if the commonly used classification accuracy (given in Equation (6)) was used as the fitness measure, the accuracy achieved would be higher i.e. 98.33% (59 out of 60) however it is not appropriate for data imbalance problem.

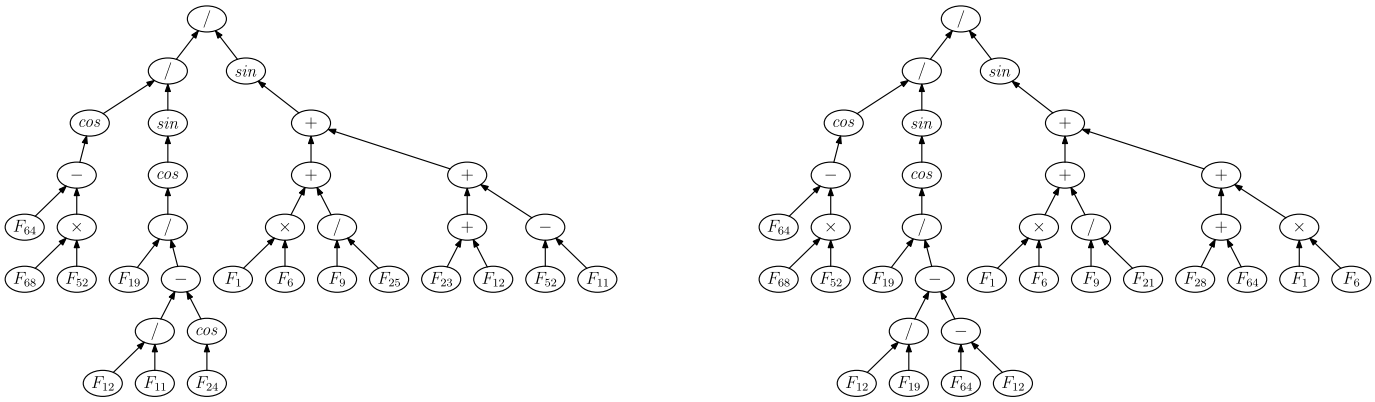
A. Comparison with other Classifiers

In most cases, using both features GP has better and equal performance than other non-GP methods. The reason for RF's bad performance using both features is due to the fact that RF randomly selects a feature subset among all features, then by applying an impurity measure it selects the best feature among the feature subset and assigns it to its tree node. Here, as the number of LBP features is nearly five times more than domain features, hence, there are more chances of selecting LBP features in the subset than domain features. According to the feature ranking (discussed in Section I), we see that domain features have more mutual information as compared to LBP features. Hence, randomly selecting more LBP features results in reduced performance of RF. As can be seen from Table II, GP using all features has shown better or equal performance in most cases (24 out of 30) among both features.

The results obtained in the second experiment are listed in Table III. The accuracy is calculated as the mean of 30 accuracies, where each value is obtained from the classifier applied to one feature subset. The accuracy and standard deviation of RF for Split-2 in case of both features is 91.94 ± 2.95 as presented in Table III. Furthermore, according to Wilcoxon test, here the "=" symbol indicates that there is no statistically significant difference between RF and GP. Using SVM with GP selected features (Table III) has outperformed GP using both features, however, using all features (Table II) GP performance remains better or comparable to SVM. Moreover, SVM using GP selected features has achieved 0.0 standard deviation, which is the best result (Table III). Hence GP has shown good performance as a feature selection method as well.

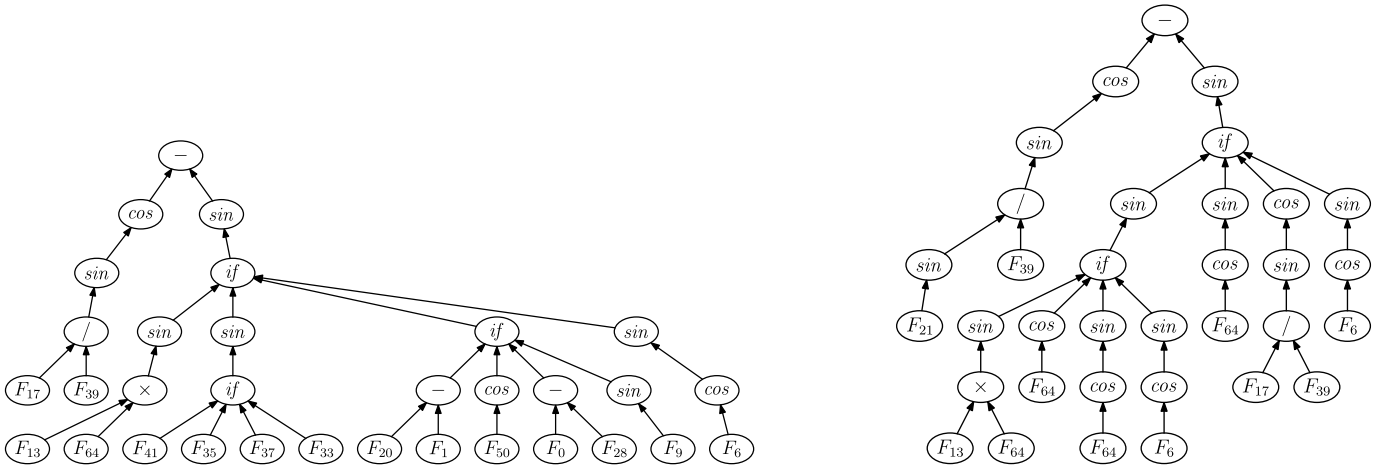
B. Program Analysis

Analyzing the evolved GP trees can provide useful understanding of how GP learns to solve a specific problem. We examine four of the best evolved classifiers using both feature-set from different splits. The four programs are depicted in Figures 3 and 4, where each figure has two evolved programs from the same GP run. Part (a) of these figures show an evolved program at an earlier generation whereas part (b) shows the evolved program at last generation. The input



(a) (b)

Fig. 3. Good Evolved GP trees from Split1 using both features, 20^{th} (same) run with 83.03% and 90.18% accuracy, respectively.



(a) (b)

Fig. 4. Good Evolved GP trees from Split2 using both features, 25^{th} (same) run with 83.03% and 91.51% accuracy, respectively.

TABLE IV
FOUR GOOD EVOLVED PROGRAMS FOR BOTH FEATURE SET.

Figure	Infix Expression
Fig. 3(a)	$((cos(F_{64} - (F_{68} \times F_{52}))/sin(cos(F_{19}/((F_{12}/F_{11}) - cos(F_{24})))))/sin(((F_1 \times F_6) + (\underline{F_9/F_{25}})) + ((\underline{F_{23}} + \underline{F_{12}}) + \underline{(F_{52} - F_{11})})))$
Fig. 3(b)	$((cos(F_{64} - (F_{68} \times F_{52}))/sin(cos(F_{19}/((F_{12}/F_{19}) - (F_{64} - F_{12})))))/sin(((F_1 \times F_6) + (\underline{F_{19}/F_{21}})) + ((\underline{F_{28}} + \underline{F_{64}}) + \underline{(F_1 \times F_6)})))$
Fig. 4(a)	$(cos(sin(\underline{F_{17}/F_{39}})) - sin(if(\underline{sin(F_{13} \times F_{64})}, sin(if(F_{41}, F_{35}, F_{37}, F_{33})), \underline{if(F_{20} - F_1, cos(F_{50}), F_0 - F_{28}, sin(F_9))}, \underline{sin(cos(F_6))}))))$
Fig. 4(b)	$(cos(sin(sin(F_{21})/F_{39})) - sin(if(sin(\underline{if(sin(F_{13} \times F_{64}), cos(F_{64}), sin(\underline{cos(F_{64})}), sin(cos(F_6)))}), sin(\underline{cos(F_{64})}), \underline{cos(sin(F_{17}/F_{39}))}, sin(cos(F_6)))))$

features correspond to F_0–F_{70} in these programs where LBP features range between F_0–F_{58} and domain features between F_{59}–F_{70}. The Infix expressions of these four trees, which we have analyzed, are given in Table IV. Vertically, this table comprises four rows each of which is showing the Infix expressions of two GP trees evolved in the same run. Furthermore, the first column of the table lists the figure number of the corresponding tree representation.

The first two programs we analyze are shown in Figure 3(a) and (b). The solution in Figure 3(a) has achieved 83.03% accuracy. The second program we analyze is from the same GP run in a later generation, (as shown in Figure 3(b)) scored 90.18% and is identical to the GP tree in Figure 3(a) except for some differences in selecting features for its nodes which are underlined in column two, Table IV for Figure 3(a) and (b). These differences are responsible for the variation in performance between the two solutions. These figures show that the overall structure of the evolved GP programs can be decomposed to examine how the learned GP classifiers solve a particular problem. These evolved solutions use a series of operators $\{+, -, \times, /, if, sin, cos\}$ within the tree. Similarly, two more GP trees have been examined from a different run and a different split. The structure of these programs are also same except one major and the rest, minor differences. The major difference is that the nested if operator has moved from its third argument to first argument with altered branches, which are underlined in Table IV for Figure 4(a) and (b). The tree in Figure 4(b) is bigger than the tree in Figure 4(a) as it has increased its tree depth from 6 to 8. These changes have improved the performance from 83.03% to 91.51%.

For analysing why GP selects a particular feature that results in its improved performance, we have observed from Figures 3 and 4 that features having more mutual information values contribute towards GP's better performance. Examining features according to mutual information, we have seen that domain features have high mutual information values as compared to LBP features, specifically F_{64} and F_{68}. These two domain features correspond to presence or absence of "blue-whitish veil" and "dark-brown color" in the dermoscopic images, respectively. These features appear more often in the trees shown in Figures 3(b) and 4(b), which have higher accuracy than those in Figures 3(a) and 4(a).

VI. Conclusion

Inspired by the promising results of genetic programming in computer vision, we have investigated how well GP can perform for skin cancer detection in dermoscopic images. GP has achieved good results and has the potential to provide efficient and effective solutions for real-world problems such as cancer detection. It has been seen that using knowledge from both domains (dermatology and computer vision), GP has achieved better or comparable performance in most cases as compared to other methods (k-NN, SVM, Naïve Bayes, Decision Trees, Random Forest, and Multilayer Perceptron). With all features, highest accuracy achieved is 97.92% and with GP-selected features, even 100% accuracy is achieved in some of GP runs on the unseen data. Early detection of skin cancer is curable and largely increases the survival rate of the patient. Therefore, such CAD systems can help dermatologists to assist their decision. We have used GP to evolve solutions for this binary classification problem of detecting skin cancer. We have used LBP to extract features from images and also used features provided by domain experts. Three scenarios are considered in using different feature sets; using domain features, using LBP features and using a combination of both. We have also examined GP as a feature selection method and did experiments where GP-Selected features are used to evolve solutions using GP and other state-of-the-art classification methods.

In the future, we would like to investigate the impact of employing preprocessing techniques (to remove reflection artefacts in the images due to the presence of gel and noise due to hair) before feature extraction. We are also interested in checking the system performance by using a different dataset from another origin and also focus on the computation time to make it effective for real-world applications like cancer diagnosis. As "blue-whitish veil" and "dark-brown" color domain features are most prominent in diagnosing diseased images, we can utilize color descriptors to generate better features than those generated from gray-scale images.

References

[1] 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*, 2013, pp. 5437–5440.

[2] C. Barata, J. S. Marques, and J. Rozeira, "Evaluation of color based keypoints and features for the classification of melanomas using the bag-of-features model," in *Proceedings of the 9th International Symposium on Advances in Visual Computing*. Springer, 2013, pp. 40–49.

[3] L. K. Ferris, J. A. Harkes, B. Gilbert, D. G. Winger, K. Golubets, O. Akilov, and M. Satyanarayanan, "Computer-aided classification of melanocytic lesions using dermoscopic images," *Journal of the American Academy of Dermatology*, vol. 73, no. 5, pp. 769–776, 2015.

[4] H. Al-Sahaf, A. Song, K. Neshatian, and M. Zhang, "Two-tier genetic programming: towards raw pixel-based image classification," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12 291–12 301, 2012.

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

[6] H. Al-Sahaf, M. Zhang, and M. Johnston, "Binary image classification using genetic programming based on local binary patterns," in *Proceedings of the 28th International Conference on Image and Vision Computing New Zealand*. IEEE, 2013, pp. 220–225.

[7] W. A. Tackett, "Genetic programming for feature discovery and image discrimination," in *Proceedings of the 5th International Conference on Genetic Algorithms*. Morgan Kaufmann, 1993, pp. 303–311.

[8] H. Al-Sahaf, K. Neshatian, and M. Zhang, "Automatic feature extraction and image classification using genetic programming," in *Proceedings of the 5th International Conference on Automation, Robotics and Applications*. IEEE, 2011, pp. 157–162.

[9] S. Ahmed, M. Zhang, and L. Peng, "Enhanced feature selection for biomarker discovery in LC-MS data using GP," in *Proceedings of the 2013 IEEE Congress on Evolutionary Computation*. IEEE, 2013, pp. 584–591.

[10] K. Neshatian, M. Zhang, and P. Andreae, "A filter approach to multiple feature construction for symbolic learning classifiers using genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 16, no. 5, pp. 645–661, 2012.

[11] J. Chen, G. Zhao, V. Kellokumpu, and M. Pietikäinen, "Combining sparse and dense descriptors with temporal semantic structures for robust human action recognition," in *IEEE International Conference on Computer Vision Workshops*. IEEE, 2011, pp. 1524–1531.

[12] 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.

[13] S. Shen, W. Sandham, M. Granat, M. Dempsey, and J. Patterson, "A new approach to brain tumour diagnosis using fuzzy logic based genetic programming," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, 2003, pp. 870–873.

[14] N. Alfed, F. Khelifi, A. Bouridane, and H. Seker, "Pigment network-based skin cancer detection," in *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2015, pp. 7214–7217.

[15] D. Piccolo, A. Ferrari, K. Peris, R. Daidone, B. Ruggeri, and S. Chimenti, "Dermoscopic diagnosis by a trained clinician vs. a clinician with minimal dermoscopy training vs. computer-aided diagnosis of 341 pigmented skin lesions: a comparative study," *British Journal of Dermatology*, vol. 147, no. 3, pp. 481–486, 2002.

[16] W.-J. Choi and T.-S. Choi, "Computer-aided detection of pulmonary nodules using genetic programming," in *Proceedings of the 2010 IEEE International Conference on Image Processing*, 2010, pp. 4353–4356.

[17] R. Poli, "Genetic programming for feature detection and image segmentation," in *AISB Workshop on Evolutionary Computing*. Springer, 1996, pp. 110–125.

[18] C. E. Shannon and W. Weaver, "The mathematical theory of information," 1949.

[19] G. Patterson and M. Zhang, "Fitness functions in genetic programming for classification with unbalanced data," in *Proceedings of the 2007 Australasian Joint Conference on Artificial Intelligence*. Springer, 2007, pp. 769–775.

[20] C. Barata, J. S. Marques, and J. Rozeira, "A system for the detection of pigment network in dermoscopy images using directional filters," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2744–2754, 2012.

[21] 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.

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