Generating Knowledge-Guided Discriminative Features Using Genetic Programming for Melanoma Detection

Qurrat Ul Ain, *Student Member, IEEE*, Harith Al-Sahaf, *Member, IEEE*, Bing Xue, *Member, IEEE*, and Mengjie Zhang, *Fellow, IEEE*

***Abstract*—Melanoma is the deadliest form of skin cancer that causes around 75% of deaths worldwide. However, most of the skin cancers can be cured, especially if detected and treated early. Existing approaches have employed various feature extraction methods, where different types of features are used individually for skin image classification which may not provide sufficient information to the classification algorithm necessary to discriminate between classes, leading to sub-optimal performance. This study develops a novel skin image classification method using multi-tree genetic programming (GP). To capture local information from gray and color skin images, Local Binary Pattern is used in this work. In addition, for capturing global information, variation in color within the lesion and the skin regions, and domain-specific lesion border shape features are extracted. GP with a multi-tree representation is employed to use multiple types of features. Genetic operators such as crossover and mutation are designed accordingly in order to select a single type of features at terminals in one tree of the GP individual. The performance of the proposed method is assessed using two skin image datasets having images captured from multiple modalities, and compared with six most commonly used classification algorithms as well as the standard (single-tree) wrapper and embedded GP methods. The results show that the proposed method has significantly outperformed all these classification methods. Being interpretable and fast in terms of the computation time, this method can help dermatologist identify prominent skin image features, specific to a type of skin cancer in real-time situations.**

***Index Terms*—Genetic programming, feature selection, feature construction, image classification, melanoma detection.**

Manuscript received May 24, 2019; revised October 10, 2019 and February 10, 2020; accepted March 1, 2020. This work was supported in part by the Marsden Fund of New Zealand Government under Contracts VUW1509 and VUW1615, in part by the Science for Technological Innovation Challenge (SfTI) fund under Grant E3603/2903, in part by the University Research Fund at Victoria University of Wellington under Grants 216378/3764 and 223805/3986, and in part by MBIE Data Science SSIF Fund under the Contract RTVU1914. *(Corresponding author: Qurrat Ul Ain.)*

The authors are with the School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6140, New Zealand (e-mail: qurrat.ul.ain@ecs.vuw.ac.nz; harith.al-sahaf@ecs.vuw.ac.nz; bing.xue@ecs.vuw.ac.nz; mengjie.zhang@ecs.vuw.ac.nz).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TETCI.2020.2983426

I. Introduction

Skin cancer is the most common form of cancer, accounting for nearly 40% of occurrences globally [1]. Melanoma incidence has risen enormously over the past 30 years [2]. In 2019, there will be an estimated number of 96,480 new skin cancer cases which may lead to 7,230 estimated deaths in the US [2]. When diagnosed early, skin cancer is highly curable with a survival rate of nearly 92% [2]. However if not detected early, it spreads to other parts of the body which can be fatal [2]. Early diagnosis of skin cancer has become a top priority of public health due to the rapid increase in incidence rate of skin cancer particularly melanoma. New developments in the areas of computer vision provide improved computer aided diagnostic (CAD) systems which facilitate earlier diagnosis of various skin cancers that require no biopsy.

Dermatologists generally follow a scoring method called the ABCD rule of dermoscopy [3] which quantify four lesion characteristics; Asymmetry, Border, Color, and Dermoscopic structure, which help effectively separate different kinds of skin images [4]. Another commonly used approach is the 7-point check-list method (Asymmetry, Regression areas, Dots, Streaks, Pigment network, Blue-whitish veil and presence/absence of six colors; black, white, light-brown, dark-brown, red, and blue-gray) [5]. These fundamental medical properties and availability of skin images have attracted many researchers to make efficient and effective CAD systems that can greatly help in early diagnosis. Researchers are interested to formulate methods that can capture informative features similar to these medical properties and, hence, incorporate both local and global features. Local features extract information from a part of an image whereas global features capture information from the whole image. However, automated skin image classification incorporates various challenges mainly due to 1) the high inter-class variation of melanomas, 2) the high intra-class similarity among various types of skin cancers, 3) varying location of lesion in skin images, and 4) presence of various artifacts in skin images, e.g., hair, gel, and reflection [6]. According, both local and global features are often needed.

Genetic programming (GP) is an evolutionary computation method that evolves computer programs (models or trees) to solve a particular problem [7]. GP applies genetic operations such as crossover and mutation to transform a population of computer programs iteratively into a new generation of programs [7]. GP typically represents a computer program in a tree-like arrangement where terminal nodes and internal nodes are made up of features and functions, respectively. Due to the fact that all the features are not important for classification, GP utilizes its implicit feature selection ability to automatically select the important features as its terminals. The evolved trees are the new constructed features from the original set of features with high discriminating ability between classes, which greatly helps in achieving good performance. In image analysis, GP has been widely explored for a broad range of applications such as object detection [8], feature extraction [9], feature construction [9], [10], evolving texture image descriptors [11], [12], and classification [13], [14].

With its flexible representation, GP can evolve multiple trees in a single individual, referred as multi-tree GP (MTGP) [15]. In this work, as we are interested to encompass the different local, global, texture, and color image properties of the lesion images in our classification model, we have employed MTGP to effectively evolve multiple trees (constructed features) each based on a specific property, e.g., one tree for gray-scale features, one for pixel-based color features and another for border shape features. On the other hand, in a MTGP approach evolving multiple trees based on all different type of features may not result in meaningful constructed features. Similarly, evolving these constructed features individually in a single-tree GP approach will use only one specific property of skin images (e.g. based on either local features or global features) and, hence, may not provide sufficient information necessary for classification. Moreover, using all these different features together to evolve a single-tree GP-constructed feature has resulted in poor performance. Therefore, MTGP is applicable where different image properties (local, global, texture, and color information) encompassed in different sets of features are necessary to evolve good solution in order to get sufficient informative features in terminal set. In the literature, MTGP has been studies for automatically evolving image descriptors for texture image classification [11], constructing features to create benchmark datasets [16], self-assembling swarm robots [17], and multi-class classification [18]. Based on the evaluation criteria, feature selection algorithms are categorized into three groups: filter, wrapper and embedded approaches. A wrapper approach includes a classification/learning algorithm in the feature subset evaluation whereas a filter approach is independent of any classification algorithm [19]. An embedded approach combines feature selection and classifier learning into a single process [19]. Generally, filter-based methods ignore the performance of the selected features on a classification algorithm while wrapper-based methods evaluate the feature subsets based on the classification performance, resulting in improved performance.

Unlike existing approaches, the proposed MTGP method constructs informative features in a wrapper approach which are provided to a machine learning classification algorithm (such as $k-$nearest neighbor or decision trees) for classification. This feature construction ability of our MTGP method generates knowledge-guided features which help the classification algorithm to produce good results.

In real-world situations, various optical devices are used in the hospitals and medical centers to record skin images which include specialized dermatoscope and standard camera. Images taken from different optical devices might possess different visual characteristics. In addition, with different camera settings, characteristics like illumination, scale, and reflection might differ. Therefore, which feature extraction method can extract more informative features for a specific kind of images (taken from different devices) still needs thorough investigation. Different from existing approaches which mainly focus on designing a classification method for a single image modality (images captured from one instrument), this study aims at developing a robust skin cancer classification method which can produce good results across multiple image modalities.

Therefore, accounting all the important factors discussed above, we become interested in developing a method for real-world skin image classification by designing a MTGP approach, with multiple constructed features each of which evolves using a particular set of features.

A. *Goals*

This work focuses on developing a novel method using the multi-tree approach in GP while enhancing classification performance using a wrapper approach for skin image classification. This work aims at automatically generating a classification model that uses features constructed from a variety of local and global features having sufficient information to discriminate images of different classes. Different from most existing approaches which can only provide effective results for a single image modality, this method is developed for images captured from multiple image modalities (devices). The following objectives will be explored in this study:

- Developing a new multi-tree based GP method with a wrapper approach having sufficient informative features for binary and multi-class skin image classification problems.
- Assessing the performance of the proposed classification method quantitatively and comparing it to six commonly used classification algorithms and eight single-tree GP methods on two real-world skin image datasets.
- Investigating the effectiveness of the proposed method on datasets having images captured from multiple modalities (specialized instruments and standard cameras).
- Investigating the efficiency of the proposed method in terms of analyzing computation time to train the proposed method and test its performance on the test images.
- Analyzing the interpretability of evolved GP-constructed features.
- Investigating the different types of prominent features necessary to provide sufficient information for the diagnosis of skin images.

B. *Organization*

The rest of the paper is organized as follows. The background and the related work are described in Section II. The proposed

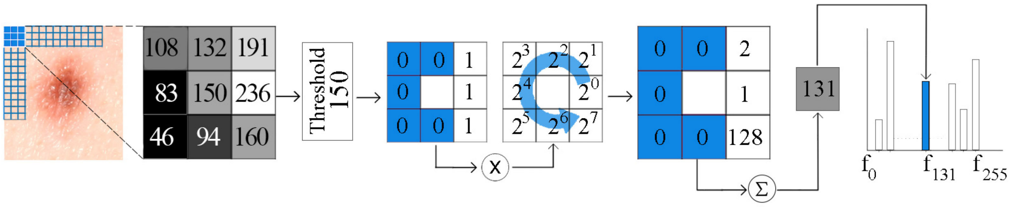


Fig. 1. Process of generating LBP histogram with $LBP_{8,1}$ code for sub-image having radius = 1, and neighboring pixels = 8.

MTGP method is described in detail in Section III. The experiment design is presented in Section IV. Section V presents and discusses the results of binary and multi-class classification tasks. Section VI further analyses the results in terms of computation time, thoroughly examining a GP individual evolved by the proposed method. Section VII provide conclusions obtained from this work and highlights some future directions.

II. Literature Review

A. Background

1) Local Binary Pattern: Local binary patterns (LBP) is a dense image descriptor that has been widely studied for feature extraction in computer vision applications [20]. LBP scans an image in a pixel-by-pixel manner using a window of fixed radius. The intensity value of the central pixel is evaluated based on the intensity values of neighboring pixels lying on the radius. From these computed central pixel values, it then generates a histogram (i.e. feature vector). Radius defines the distance (in terms of number of pixels) between the central pixel and the neighboring pixels. The process of generating LBP histogram is shown in Fig. 1, where the radius and neighboring pixels are set to 1 and 8, respectively. The central pixel value 150 is compared to the eight neighboring pixel values 108, 132, 191, 83, 236, 46, 94, and 160. When the central pixel value is greater than the neighboring value, it assigns "0," else assigns "1". It then follows the pixels along a circle, i.e. clockwise or anti-clockwise to get an 8-digit binary code i.e., 11000001 which is converted to decimal i.e., 131 and the size of the corresponding bin in the histogram is increased by one. It then moves the sliding window and repeats the process until the entire image is scanned to generate the complete LBP histogram for an image.

Furthermore, LBP codes are divided into two categories: *uniform* and *non-uniform*. A uniform LBP code does not have more than two bit-wise transitions circularly from 1 to 0 or 0 to 1. For example, 11110000, and 10000111 are uniform codes, whereas 00100110, 01011110, and 01011100 are non-uniform codes. The size of the feature vector can be reduced from 2^b bins to $b(b-1)+3$ bins by using only uniform codes and omitting non-uniform codes. Uniform codes identify the presence of various texture patterns such as flat areas, line ends, corners, edges, and dark spots in images. In skin images, uniform codes can help detect blobs (flat areas) and streaks (line ends) which may help distinguish between different classes of skin lesions.

2) Color Variation Features: Color characteristics inside the lesion area are often used by dermatologists to identify the type of skin cancer. According to dermatologists, melanoma skin lesions are categorized by variegated coloring. The presence of different colors, especially in the form of irregular patches or veils (such as blue-whitish veil), induces high variance in the red, green, blue (RGB) color space. In this work, the red, green and blue color channel data of the pixels in the lesion area and skin area are stored as $\text{Lesion}_{\text{Color}}$ features. From each color channel, mean μ and variance σ are calculated and represented as μR, μG, μB and σR, σG, σB. To include complex color distributions inside the lesion area, mean ratios of the mean values are calculated which are represented as $\frac{\mu_R}{\mu_G}$, $\frac{\mu_R}{\mu_B}$, $\frac{\mu_G}{\mu_B}$. Variations in color of the skin lesion as compared to the surrounding skin is also evaluated. These features are calculated as $\frac{\mu_R}{\gamma_R}$, $\frac{\mu_G}{\gamma_G}$, $\frac{\mu_B}{\gamma_B}$, where γ is the mean value of the skin-only area around the lesion. These 12 color variation features have been adopted from [21].

3) Geometry-Based Features: The geometrical shape of the border of the lesion is an important characteristic that provide diagnostic information about a type of skin cancer. According to the ABCD rule of dermoscopy [3], asymmetry is assigned with the highest weight among the four features of asymmetry, border irregularity, color variation and dermoscopic structure. This work includes some standard geometry-based features adopted from [22] and [23]. The details of these 11 geometry-based features are listed in Table I. Moreover, all the images within each dataset in this study have similar spatial resolution, which allows us to extract these geometrical features i.e., area and perimeter, without any scale issue. When images in a single dataset are not captured under standardized conditions such as magnification and resolution, such datasets usually require a normalization procedure before extracting features.

B. Related Work

Earlier in 1994, an artificial neural network (ANN) approach was designed to classify skin images as malignant or benign [24]. Using lesion shape and lesion color features, this approach obtained good performance. However, it requires a dermatologist to identify the lesion boundaries manually, which makes this system expensive to implement. In [22], a CAD system is developed for melanoma classification which selects an optimal set of features from different types of features such as texture, border-based, and geometrical shape. To classify melanoma and benign images, four classification algorithms (Naïve Bayes, support vector machine (SVM), random forest, and hidden logistic model tree) are employed. Though this diagnostic system produced very good results (91.26% with 23 features), it utilizes

TABLE I
GEOMETRICAL BORDER SHAPE FEATURES

Name	Description
Area	Number of pixels of the lesion.
Perimeter	Number of pixels along the detected boundary.
Greatest Diameter	The length of the line which connects the two farthest boundary points and passes across the lesion centroid.
Shortest Diameter	The length of the line which connects the two nearest boundary points and passes across the lesion centroid.
Circularity Index	The shape uniformity expressed as $CRC = 4\pi A/P^2$
Irregularity Index A	$IrA = P/A$
Irregularity Index B	$IrB = P/GD$
Irregularity Index C	$IrC = P \times \big((1/SD) - (1/GD)\big)$
Irregularity Index D	$IrD = GD - SD$
Major and Minor Asymmetry Indices	These indices are defined as the area difference between the two halves of the lesion, taken the principal axes [obtained by (19)] as the major symmetry axis, and its 90° rotation as the minor axes of the symmetry.
Asymmetry Index	This index is measured by $AI = (A_D/A) \times 100$ where A_D represents the difference between the number of pixels of the two halves in a lesion.

different types of features individually and lacks an appropriate way to combine them.

The robustness of a CAD system is one of the most important characteristics for dermoscopy images [25]. It is difficult to develop a robust system for multi-source images acquired under different conditions, such as varying illumination and different acquisition devices. Hence, it has been suggested to use the color constancy algorithms and the results of SVM have shown increased performance using RGB histograms as features. For effective feature learning from color images, a quaternion-based grassmann average network (QGANet) is developed [26]. The experiment results proved the goodness of the method on three histopathological color image datasets. Since the QGANet algorithm embeds the grassmann average network (GANet) into a principal component analysis network (PCANet), the computational complexity of QGANet is four times more than the baseline GANet.

Identifying the score of the ABCD rule of dermoscopy has been recently studied [4]. In pre-processing, Gabor filters and active contours are utilized to detect lesion boundaries. The extracted features, according to the ABCD rule, are used to compute the total dermoscopy score, which is then used for binary classification. The method has produced good sensitivity and specificity results and revealed the potential of extracted features in building a good classification model.

Adjed *et al.* [27] developed a binary classification method for melanoma detection through fusion of texture and structure features. The method extracts texture features from different variants of LBP, and structure features from curvelet and wavelet transforms. SVM classifier produced good results in terms of sensitivity (78.93%) and specificity (93.25%).

Recently, Xie *et al.* [28] proposed an ANN-based ensemble model for melanoma detection from skin images. The algorithm works by first extracting the lesion area with a self-generating neural network. Various types of features such as border, texture, and color are extracted, which are then given to a neural network ensemble method for binary classification. The results revealed the goodness of the new border features, which played a vital role in achieving improved accuracy.

For melanoma detection from skin images, Yu *et al.* [6] developed a 2-stage convolutional neural network (CNN) architecture. The first stage performs lesion segmentation using a fully convolutional residual network and the second stage performs classification with a very deep residual network. Their results revealed the potential of very deep CNNs, even with limited training data to solve such a complex task of melanoma detection.

GP has been widely explored for image analysis [8], [9], [12], [29], [30]. In 1996, Poli developed a GP-based method for image segmentation and feature detection [29]. A set of requirements for fitness function, terminal set, and function set in GP has been outlined necessary to generate effective optimal filters in X-ray coronarograms and brain MRI.

Ryan *et al.* [30] described a GP-based fully automated system to detect Stage-1 breast cancer. The method detects suspicious regions called regions-of-interest (ROI) and outputs the likelihood of malignancy. It is a seven stage method, where the first five stages implement pre-processing, breast segmentation and feature extraction, while the last two stages employ a multi-objective GP approach for building and testing the classifier. Results have revealed the ability of GP to produce human-readable solutions, and capable of examining the GP individuals.

Zhang *et al.* proposed a domain-independent approach to the problem of multi-class object detection using GP [8]. The aim of their method is to locate a number of objects of different classes that are contained in a large image, and predict the class label of each of the detected objects. The method is tested using three datasets of increasing difficulty. The evolved program is capable of performing object detection and multi-class classification tasks.

Al-Sahaf *et al.* [12] developed a GP-based method to automatically generate an image descriptor i.e., a feature vector, for texture image classification. The feature vector generated in their approach is quite similar to LBP [20]. However, a domain-expert designs the formulas in LBP, whereas these formulas are automatically generated by GP in their work. Experiments revealed the goodness of the proposed method in comparison to other GP and non-GP methods. Iqbal *et al.* [31] improved the structure of the algorithm in [12] to perform transfer learning, to cope with difficult texture image classification tasks. The results proved the effectiveness of their method, showing ability to solve even more difficult tasks which most other algorithms cannot solve. Lensen *et al.* [9] developed a GP-based method capable of performing multiple tasks in a single evolved GP individual; region detection, feature extraction and binary image classification. Their results have shown improved classification performance compared to the existing GP approaches.

Recenlty Ain *et al.* [32] tackled the problem of skin image classification using GP with a combination of biomedical (domain-specific) and LBP (domain-independent) features. They have also designed a feature selection and feature construction method by using local and global features in GP for the task of melanoma detection [10]. Using a multi-tree GP in an embedded approach, Ain *et al.* [33] proposed a binary classification method to effectively identify melanoma in images. Their method works by evolving multiple trees in a GP individual on the training data, where each tree operates as a binary classifier. The tree producing the best performance on the training data is used to test the performance on the test data. They have identified the important image features by analyzing the good evolved GP programs, which can help dermatologists to make diagnosis in real-world situations.

Several existing methods [6], [24], [28] have developed CNNs for skin image classification which have shown good results, but have some limitations. Most of these CNN architectures are developed as a black-box; therefore, they are not interpretable. Such classification models lack the ability to identify prominent features in classifying skin images. Moreover, the performance of a CNN is severely limited by amount of data needed to train a classification model. Usually, CNNs require a large number of training images to achieve sufficiently good results. Generally, the medical data available in real world applications is limited. Consequently, training a model with a large dataset requires long computation time and hence, large computing resources. Some existing approaches [21], [22], [27], [28], [34] developed classification methods for melanoma detection where various features are extracted from skin images. These methods assessed the goodness of these features individually using different machine learning classification algorithms. However, they lack using a combination of different types of features concurrently to achieve performance gains. Performance can be improved by utilizing all these features concurrently by designing an effective way of combining these different types of features. Most existing methods have used only one image modality (images captured from a single instrument) to test the performance of their method(s). However, in real-world situations, there are images captured from different instruments and hence, these methods, developed for a single image modality, cannot be applied to or may perform poorly on other image modalities. Hence, there is a need for a classification method for skin images which, having sufficient informative features, has the ability to be applied to multiple image modalities, easily interpretable in order to guide the dermatologist, and able to discriminate between various classes of skin cancers.

III. The Proposed Method

This section provides a detailed description of the proposed MTGP wrapper method, which starts by presenting an overview of the algorithm to evolve a GP individual in order to highlight the key components of our proposed method, and how the constructed features from the evolved individual are used for classification. Then the program structure, i.e., the terminal and the function sets, the crossover and mutation operators, and the fitness function, are discussed.

The proposed method operates on a set of predefined/extracted features which include local and global information about the skin images. The local features are extracted with the help of LBP descriptor which works with the pixel values and can significantly capture informative features about various skin properties such as lines/streaks, blobs, homogeneous regions, and irregular border patterns. The global features are extracted by focusing on shape and color variation characteristics of skin lesions. These features are defined in [21] and [22]. These features are of utmost importance because without using these human crafted features, it is difficult to achieve good performance for such a difficult task as skin image classification. These global features capture the properties of asymmetry, border, color and diameter (ABCD) rule of dermoscopy, which plays a vital role for the dermatologist in distinguishing malignant from benign images. Hence, incorporating these informative features help the classifier learn better and produce an effective model.

A. The Overall Algorithm

The overall structure of the MTGP in a wrapper approach for skin image classification is shown in Fig. 2. First, the four types of features are extracted from each image of a dataset. Hence, one image is represented by four feature vectors. Then the dataset is divided into training and test sets. The MTGP algorithm runs on the training set of the dataset to select a subset of relevant features for each type of features among the four feature types. It then constructs four features from these selected features. In other words, GP evolves four trees in a single individual based on the four types of features, which is the evolutionary training process. Then the training set and the test set are transformed to a new training set and a new test set by constructing new features from the four trees evolved during the training process. A classification algorithm (such as a decision tree) is then trained on the transformed training set. The learned classifier is then applied to the transformed test set to obtain the final test classification performance.

B. GP Program Representation

In a GP individual, each tree is constructed from elements of the terminal and the function sets. In this study, tree-based GP is used to represent an individual. Furthermore, an individual consists of four trees. The intuition behind evolving four trees is that four different types of features are used which have been extracted using different feature extraction methods described in Section II-A. Each tree in a GP individual is generated using a single type of features. For illustration (as shown in Fig. 3), the

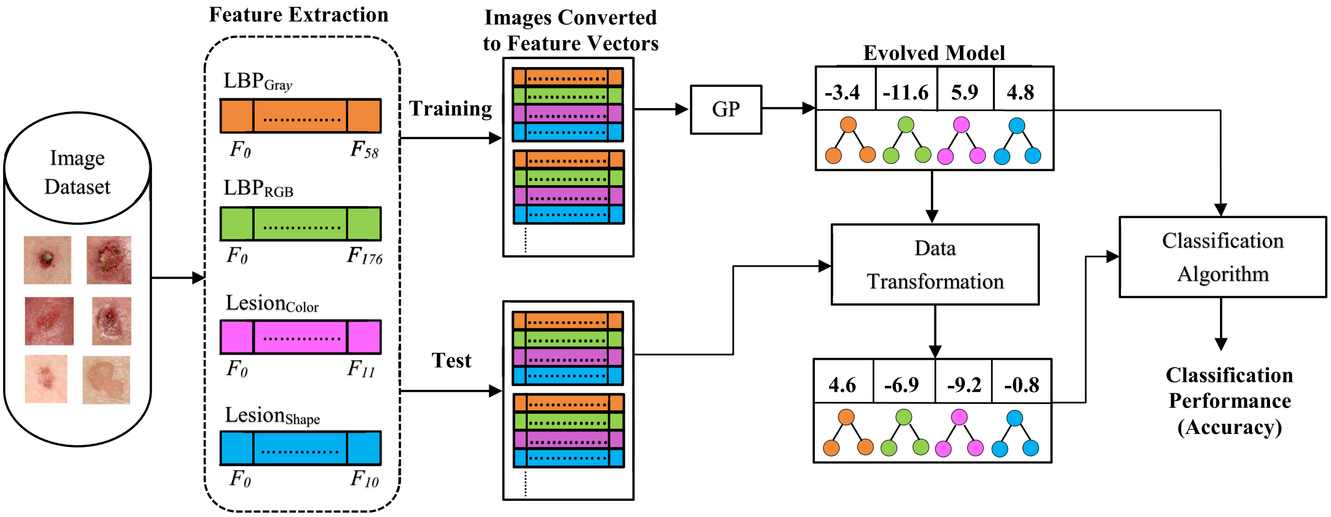


Fig. 2. Overview of the proposed method: Each image in the dataset is input to four feature extraction methods to obtain four feature vectors, namely $\mathrm{LBP_{Gray}}$, $\mathrm{LBP_{RGB}}$ $\mathrm{Lesion_{Color}}$, and $\mathrm{Lesion_{Shape}}$, for each image. The training set is given to GP to evolve four trees each based on a single feature vector. Using these four trees (constructed features), the training and test sets are transformed into new training and test sets. The transformed training set is provided to a classification algorithm (such as a decision tree) to evolve a classification model. The learned classification model is applied to the transformed test set to obtain the test classification performance.

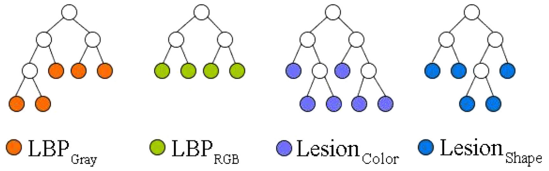


Fig. 3. A GP individual with four trees, each evolved with a single type of features.

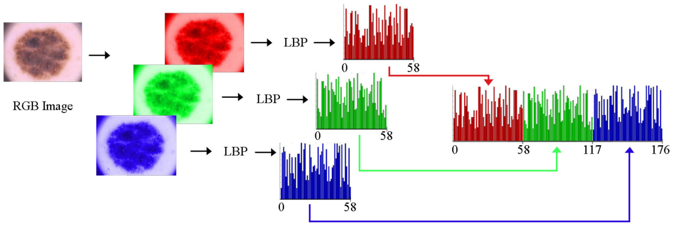


Fig. 4. Step-by-Step procedure to generate the $\mathrm{LBP_{RGB}}$ feature vector from a color image.

terminal set of the first tree consists of $\mathrm{LBP_{Gray}}$ features only. Similarly, the terminal sets of the second, third and fourth trees consist of $\mathrm{LBP_{RGB}}$, $\mathrm{Lesion_{Color}}$, and $\mathrm{Lesion_{Shape}}$ features, respectively. However, all the trees share the same function set that consists of seven operators as described in Section III-D.

C. Terminal Set

The terminal set comprises of four types of features, which are extracted from the feature extraction methods described in Section II-A. The four types of features include;

1) $\mathrm{LBP_{Gray}}$: 59 LBP features extracted from gray-level skin images as described in Section II-A1(1), and using the process shown in Fig. 1.
2) $\mathrm{LBP_{RGB}}$: 59 LBP features are extracted from from the three color channels i.e., red, green, blue. They are concatenated to get a single feature vector having 177 (= 59 LBP features × 3 channels) $\mathrm{LBP_{RGB}}$ features as illustrated in Fig. 4.
3) $\mathrm{Lesion_{Color}}$: The variation in color across the skin image is calculated by 12 $\mathrm{Lesion_{Color}}$ features as described in Section II-A(2).
4) $\mathrm{Lesion_{Shape}}$: The geometrical border shape properties are calculated by 11 $\mathrm{Lesion_{Shape}}$ features as described in Section II-A(3).

The value of the ith feature for the above four types of features is indicated by Gi, Ri, Ci, and Si, respectively, as shown by the GP individual in Fig. 10. For $\mathrm{LBP_{RGB}}$ and $\mathrm{LBP_{Gray}}$ features, a window size of 3×3 pixels and a radius of 1 pixel ($\mathrm{LBP_{8,1}}$) is used, which are the simplest and the most commonly used settings for extracting LBP features.

D. Function Set

The function set comprises the seven most commonly used operators. There are four arithmetic operators $\{+, -, \times, /\}$, two trigonometric $\{\sin, \cos\}$ operators, and one conditional operator $\{if\}$. Among the four arithmetic operators, the first three operators works with their usual arithmetic meaning, however, the last operator i.e., division is protected, which means it returns zero when the denominator is zero. The conditional if operator works with four input values and outputs the third input if the first input is greater than the second input; else, it outputs the fourth input.

E. Crossover and Mutation

To meet the objective of having only one type of features in a single GP tree, genetic operators, such as crossover and

Algorithm 1: Same-Index Crossover.

```
function CROSSOVER $P^1$, $P^2$          ▷ Two GP
Individuals (parents), each having four trees
    for $i = 1$ to 4 do
        XOVER($P^1_i$, $P^2_i$)          ▷ Crossover between
        trees having same
    type of features as terminals
    end for
    return $C^1$, $C^2$          ▷ The two children obtained
    after XOVER
end function
```

Algorithm 2: Same-Index Mutation.

```
function MUTATION($P^1$)          ▷ One GP Individual
(parent) having four trees
    for i = 1 to 4 do
        $P^1 \leftarrow init\ (T_i)$          ▷ Generate a new tree with
                  a single type of features
        MUTATE($P_i$, $P^1$)          ▷ Mutate the tree
        from parent
        individual with the new generated tree,
        both having the same type of features
    end for
    returnC$^1$          ▷ One child obtained after
    MUTATE
end function
```

mutation, are designed accordingly, which is called *same-index-crossover/mutation* [16]. The step-by-step process is given in Algorithms 1 and 2. This crossover/mutation guarantees that the GP individual evolved at the end of the evolutionary process, consists of four trees where each tree evolves from a single type of features. For example, in case of crossover having two parents, the tree generated using $\mathrm{LBP}_{\mathrm{RGB}}$ features in the first parent can only crossover with the tree generated using the same $\mathrm{LBP}_{\mathrm{RGB}}$ features in the second parent, and it is ensured that it cannot crossover with $\mathrm{Lesion}_{\mathrm{Shape}}$, $\mathrm{LBP}_{\mathrm{Gray}}$ or $\mathrm{Lesion}_{\mathrm{Color}}$ features as described in Algorithm 1. Similarly, for example, in case of mutation having one parent, a newly created tree generated using $\mathrm{Lesion}_{\mathrm{Color}}$ features can only mutate with a previously generated tree in parent from $\mathrm{Lesion}_{\mathrm{Color}}$ features as described in Algorithm 2.

The traditional GP evolves one tree in its individual, hence, for the crossover operation, one node from the tree is randomly picked. The computational complexity of crossover in the traditional GP approach is $\theta(n)$, where n denotes the number of trees. In this work, since a GP individual has four trees, the computational complexity of the same-index crossover (Algorithm 1) will be four times as the traditional GP, i.e., $\theta(4)$. Similarly, the computational complexity of the same-index mutation (Algorithm 2) will be four times more than the traditional GP with one tree.

TABLE II
REAL-WORLD SKIN CANCER DATASETS

Name	Classes	No. of Images
PH^2	Melanomas	40
	Common Nevi	80
	Atypical Nevi	80
Dermofit	Melanoma	76
	Melanocytic Nevus / Mole	331
	Actinic Keratosis	45
	Seborrhoeic Keratosis	257
	Basal Cell Carcinoma	239
	Pyogenic Granuloma	24
	Squamous Cell Carcinoma	88
	Dermatofibroma	65
	Intraepithelial carcinoma	78
	Haemangioma	96

F. Fitness Function

The balanced classification accuracy is used as the fitness function, which is defined as

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

where m is the number of classes, TP refers to the true positive, FN refers to the false negative, and the ratio $\frac{TP_i}{TP_i+FN_i}$ represents the true positive rate of a class. When there are different number of instances in different classes (a class imbalance problem), using balanced accuracy is more suitable than the standard overall accuracy, which is the ratio between the number of correctly classified instances and the total number of instances. Using this fitness (Equation (1)), all four trees (constructed features) are allowed to improve themselves during the evolutionary process.

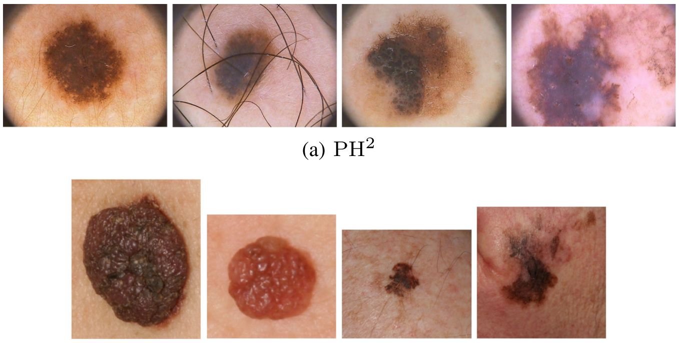
IV. EXPERIMENT DESIGN

The aim and design of the experiments are discussed in this section. The discussion also includes the datasets, the other classification methods used for comparison, the experiments and the parameter settings.

A. Datasets

The proposed MTGP method is evaluated using two skin image datasets. The two datasets are different from each other in terms of size of images, instruments with which the images are captured, presence of hair, reflection and gel artifacts, etc. The details of these datasets are discussed below.

1) PH^2: A dataset of dermoscopic images, namely PH^2 [35], is used in this work. This dataset includes skin lesion images, their binary masks, and their clinical diagnosis. The details of image classes and number of images in each class is given in Table II. In dermatology, atypical nevi refers to currently non-malignant lesions which may develop cancerous cells later, common nevi refers to non-malignant lesions, and melanomas refer to malignant lesions. Samples of this dataset are presented in Fig. 5(a). For the binary classification experiments, atypical nevi class and common nevi class are together considered as one



(a) PH^2

(b) Dermofit

Fig. 5. Image samples from the two datasets.

class and denoted as "non-melanoma," and melanoma class are denoted as "melanoma".

The dermoscopic images were captured from a Tuebinger Mole Analyzer system with a resolution of 768×560 pixels and a magnification of $20\times$. Dermoscopy implies using an optical device with a strong lighting system to thoroughly examine skin lesions at a higher magnification. To capture the morphological structures inside the lesion area, a gel is applied which helps capture these patterns from the inner layers of the skin. Therefore, these images are rich enough to allow one to detect presence/absence of skin cancer. However, some samples in the dataset are clogged with hair and some have reflection artifacts as shown in the second and fourth sample in Fig. 5(a), respectively, which make the task of classification more difficult. The images are 8-bit RGB color images. A dermatologist examined each image and provided these classification parameters; manual segmentation of the skin lesions, Histopathology and clinical diagnosis, and dermoscopic criteria based on the 7-point check-list method. This dataset has been used in [36], [37].

2) Dermofit: The Dermofit Image Library has a total of 1,300 skin lesion images. The images are normal RGB captured with a quality SLR camera under controlled (ring flash) indoor lighting. There are ten categories of lesions in the dataset as listed in Table II. Each image is studied by field experts: dermatologists and dermatopathologists to provide a gold standard diagnosis. Images consist of the lesion surrounded by normal skin. The dataset also provides a binary mask with each lesion that denotes the lesion area. For our experiments of detecting melanoma in a binary classification setup, Melanocytic Nevus / Mole (ML) and Melanoma (MEL) classes are used to explore a dataset of 407 total images. For multi-class classification, we have used the 10 classes.

B. Methods for Comparison

To evaluate the performance of the proposed MTGP method, six classification methods are used: Naïve Bayes (NB), k-Nearest Neighbor (k-NN) where $k = 5$, Support Vector Machines (SVMs) with a Radial Basis Function (RBF) kernel, Decision Trees (J48) where the minimum number of instances per leaf equals 2, Random Forest (RF), and Multilayer Perceptron (MLP). These methods are implemented through the commonly used Waikato Environment for Knowledge Analysis package [38]. In a study [39] on kernel functions in SVM, it has been shown that a non-linear kernel has the ability to achieve similar or better performances than a linear kernel on numerical data. For RF, the number of trees and the maximum depth of a tree are set to 10 and 5, respectively. For MLP, the number of units in a single hidden-layer, the momentum, learning rate, and training epochs and are set to 20, 0.2, 0.1, and 60, respectively. These parameters are adopted from a previous study [10] where they are specified empirically as they show the best performance amongst other settings.

C. Experiments

In this study, two sets of experiments are conducted each of which aim at investigating a specific task. The first set of experiments are designed for the task of melanoma detection, which aims at distinguishing melanoma images from all the given set of images; basically it is a binary classification task. We also investigated the effectiveness of the proposed method for multi-class classification in the second set of experiments. The first task is relatively easy (two classes) as compared to the second task (three classes in case of PH^2 and ten classes in Dermofit). For both sets of experiments, the results are compared with other classification methods as described in Section IV-B. The results of binary classification method are also compared with the existing embedded approach for melanoma detection [33]. For both sets of experiments, GP is wrapped with six classification algorithms namely NB, SVM, k-NN, J48, RF, and MLP (each executed individually) to check which classification algorithm works best for these skin image classification tasks.

The datasets are divided into training and test by *10-fold cross validation* using stratified random sampling. The number of GP runs is 30 and the results are represented as the mean and the standard deviation of the accuracy values. T0 evolve an individual with four trees (four constructed features) on the training data (9 folds), the fitness given in Equation (1) is used for a classification algorithm such as NB, SVM, k-NN, J48, RF, and MLP, which computes the balanced accuracy among all the classes. These four constructed features are used to transform the test data (one$-$fold). Using the different combinations of folds, this procedure is repeated 10 times to get the accuracy for $10-$*fold cross validation*. Therefore, for the 30 GP runs, we get 30 accuracy values each for training and test sets.

For the two classification tasks, the number of independent runs for GP is 7200 (=12 (GP wrapper methods) $\times$ 30 (runs) $\times$ 10 (folds) $\times$ 2 (datasets)). However, the number of fitness evaluations in GP is huge, since it is calculated as the product of the population size (1024), the number of generations (50) and the number of independent runs (7200), and comes out to be 3.69×10^8 evaluations. For each of the 30 GP runs, different random seeds are used. The MTGP method is implemented using "Evolutionary Computing in Java" (ECJ) package [40].

D. Parameter Settings

The parameter settings of our proposed MTGP method are listed in Table III. The evolutionary process stops when the

TABLE III
PARAMETER SETTINGS OF THE PROPOSED MTGP METHOD

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

classification algorithm such as a decision tree achieves 100% accuracy or a maximum of 50 generations is reached.

V. RESULTS AND DISCUSSIONS

The results of the experiments are presented and discussed in this section. The results are represented as the mean and the standard deviation ($\bar{x} \pm s$) among the 30 GP runs as shown in Tables IV and V, where the value of one GP run is computed as the mean of applying *10-fold cross validation*. The deterministic methods are run once, therefore, their results are represented as the mean of applying *10-fold cross validation*.

To assess the significance of the proposed method, *one sample t-test* and *Wilcoxon signed-rank test* both with a significance level of 5% are used. On each dataset, the overall best performance is made **bold**.

- *One sample t-test* is applied to check the significance of the proposed stochastic method against the non-GP deterministic methods. The symbol "↑" appears next to the deterministic method that has been significantly outperformed by the proposed method, and a "↓" is used to indicate that the corresponding method has significantly better performance than that of the proposed method.
- *Wilcoxon signed-rank test* is applied on the test results of the proposed stochastic method against the other GP stochastic methods to identify which method has better ability to correctly classify the image instances. Two symbols "+," and "−" represent that the proposed method significantly outperforms and does not significantly outperform the other method, respectively.

A. Binary Classification

The binary classification results are presented in Table IV. Vertically, the table has five blocks where the first block shows the results of the proposed MTGP wrapper method and the existing MTGP embedded method, the second block shows results of other non-GP methods, the third block shows results of single-tree GP wrapper methods each using one type of features, and the fourth shows results of single-tree GP embedded methods. Horizontally, the table consists of five columns where the first lists the classification algorithm, the second and third columns show the training and test performances on the PH^2 dataset, respectively, and the fourth and fifth columns show the performances on the Dermofit dataset.

Among the six wrapper classification algorithms in our proposed method (row 1 in Table IV), it has been observed that RF and J48 achieved very good classification performance on the training data. However, on the test data, RF achieved the highest performances with $89.75 \pm 1.55\%$ and $95.35 \pm 0.83\%$ on PH^2 and Dermofit datasets, respectively.

From the results of the statistical significance test presented in Table IV, it is evident that the proposed MTGP method in a wrapper approach not only outperformed all the non-GP methods but also outperformed all the single-tree GP wrapper and embedded methods which proves the authenticity and effectiveness of the MTGP method for melanoma detection.

B. Multi-Class Classification

The multi-class classification results are presented in Among the six wrapper classification algorithms in our proposed method (row 1 in Table IV), it has been seen that RF achieved the highest classification performance on the training as well as test data. J48 also produced nearly the same training performance, however, remain far behind in producing similar results on test data. RF has achieved $96.42 \pm 1.45\%$ and $80.74 \pm 1.24\%$ test performances on the PH^2 and Dermofit datasets, respectively. It is important to note that PH^2 dataset consists of three classes and Dermofit consists of ten classes. Most of these classifiers produce good results for a 3-class problem (in case of the PH^2 dataset) such as J48 producing $80.64 \pm 2.24\%$ accuracy, however, only RF performed well enough for the 10-class problem (in case of the Dermofit dataset).

From the statistical significance test given in Table V, it is clear that the proposed method significantly outperformed all the non-GP methods for this difficult task as well, which shows the potential of our proposed MTGP method for skin image classification problems.

C. Comparison With Other Classification Methods

The results in Tables IV and V show that our proposed method has significantly outperformed all the non-GP classification methods. In particular, the "↑" sign appears next to the classification performances of the six classification methods in case of both datasets for both tasks i.e., binary and multi-class classification. For melanoma detection in a binary classification task, the highest performance is given by MLP (78.44%) in case of PH^2 dataset among these classification methods. However, our MTGP approach has outperformed all the six classification methods by providing an accuracy of $89.75 \pm 1.55\%$ on average on the unseen data. Similarly, the highest performance is given by J48 (73.98%) in case of Dermofit dataset. However, the proposed method significantly outperformed all the six classification algorithms reaching an accuracy of $95.35 \pm 0.83\%$. For the complex task of multi-class classification, the highest performance among the non-GP classification methods is given by MLP on both datasets. However, our proposed method has significantly outperformed MLP by achieving $96.42 \pm 1.45\%$ and $80.74 \pm 1.24\%$ accuracy on average on PH^2 and Dermofit dataset, respectively. These non-GP classification methods have used all the four sets of features with a total of 259 (=177 (LBP_{RGB}) + 59 (LBP_{Gray}) + 12 ($Lesion_{Color}$) + 11 ($Lesion_{Shape}$)) features. Even then, these methods remain unable to provide good performance. This is the main reason

TABLE IV
RESULTS OF THE PROPOSED MULTI-TREE GP METHOD FOR BINARY CLASSIFICATION. COMPARISON BETWEEN THE PROPOSED MTGP METHOD, THE EXISTING MTGP EMBEDDED METHOD, THE SINGLE-TREE GP METHODS, AND THE NON-GP CLASSIFICATION METHODS: ACCURACY (%) ON THE TRAINING AND THE TEST SETS OF BOTH DATASETS

		PH^2		Dermofit	
		training	test	training	test
Multi-tree GP Wrapper	NB	92.21 ± 0.60	85.70 ± 2.65	89.40 ± 0.70	80.45 ± 2.18
	SVM	89.51 ± 0.79	81.52 ± 3.58	88.64 ± 0.74	80.33 ± 2.71
	k-NN	93.91 ± 0.55	61.26 ± 4.05	91.68 ± 0.60	69.27 ± 2.89
	J48	99.71 ± 0.15	85.18 ± 3.72	98.33 ± 0.25	84.18 ± 4.11
	RF	99.87 ± 0.08	**89.75 ± 1.55**	98.00 ± 0.26	**95.35 ± 0.83**
	MLP	88.86 ± 0.74	73.87 ± 1.52	90.45 ± 0.63	80.47 ± 2.02
Multi-tree GP Embedded	−	79.69 ± 1.35	78.87 ± 2.92 +	75.63 ± 0.99	74.57 ± 1.86 +
Non-GP Methods	NB	93.85	77.81 ↑	86.42	72.26 ↑
	SVM	89.62	70.00 ↑	95.16	70.02 ↑
	k-NN	77.33	75.00 ↑	74.35	64.17 ↑
	J48	97.05	71.25 ↑	97.09	73.98 ↑
	RF	88.99	75.63 ↑	84.28	68.12 ↑
	MLP	78.92	78.44 ↑	79.83	73.00 ↑
single-tree GP Wrapper	LBP_{Gray}	94.18 ± 0.11	85.25 ± 3.10 +	79.07 ± 1.48	85.58 ± 2.06 +
	LBP_{RGB}	95.60 ± 0.29	88.25 ± 3.02 +	80.76 ± 0.86	86.75 ± 2.03 +
	$\text{Lesion}_{\text{Color}}$	93.92 ± 0.64	82.00 ± 2.03 +	89.22 ± 0.33	92.31 ± 0.92 +
	$\text{Lesion}_{\text{Shape}}$	91.76 ± 0.65	84.00 ± 1.22 +	87.57 ± 0.66	90.29 ± 1.32 +
single-tree GP Embedded	LBP_{Gray}	82.84 ± 1.35	65.96 ± 3.96 +	73.41 ± 1.87	59.91 ± 3.57 +
	LBP_{RGB}	84.42 ± 1.43	73.87 ± 2.34 +	75.52 ± 1.62	63.26 ± 3.19 +
	$\text{Lesion}_{\text{Color}}$	81.59 ± 2.31	65.70 ± 3.61 +	81.06 ± 1.31	74.13 ± 2.67 +
	$\text{Lesion}_{\text{Shape}}$	78.06 ± 1.97	49.89 ± 5.34 +	74.74 ± 2.67	61.74 ± 7.06 +

TABLE V
RESULTS OF THE PROPOSED MULTI-TREE GP METHOD FOR MULTI-CLASS CLASSIFICATION: ACCURACY (%) ON THE TRAINING AND THE TEST SETS OF BOTH DATASETS

		PH^2		Dermofit	
		training	test	training	test
Multi-tree GP Wrapper	NB	76.25 ± 0.51	75.01 ± 1.76	44.17 ± 0.41	49.23 ± 1.51
	SVM	75.71 ± 0.57	77.17 ± 2.00	37.60 ± 0.78	38.69 ± 1.34
	k-NN	83.37 ± 0.43	57.43 ± 2.40	55.88 ± 0.46	41.13 ± 0.91
	J48	95.29 ± 0.25	80.64 ± 2.24	80.09 ± 0.24	69.25 ± 1.41
	RF	95.56 ± 0.36	**96.42 ± 1.45**	84.93 ± 0.59	**80.74 ± 1.24**
	MLP	95.51 ± 0.23	47.47 ± 2.42	49.16 ± 0.57	31.88 ± 1.18
Non-GP Methods	NB	91.30	56.25 ↑	59.14	30.99 ↑
	SVM	98.47	58.33 ↑	80.56	33.73 ↑
	k-NN	72.89	64.00 ↑	60.64	41.08 ↑
	J48	96.44	49.58 ↑	87.78	33.91 ↑
	RF	88.39	55.50 ↑	61.12	47.54 ↑
	MLP	83.80	64.17 ↑	63.01	50.30 ↑

why these different types of features cannot achieve good performance without using a suitable way of combining them. Hence, we designed the MTGP approach which automatically evolves good constructed features to help the classification algorithm learn effectively to discriminate between multiple classes.

D. Comparison With Single-Tree GP Methods

In comparison to the single-tree GP methods for melanoma detection, the MTGP method has more ability to discriminate between images of different classes. Our MTGP method constructs four features, each from LBP_{RGB}, LBP_{Gray}, $Lesion_{Color}$ and $Lesion_{Shape}$ feature sets. These four constructed features are then given to wrapper classification algorithm for classification, e.g., RF. However, in case of single-tree GP method, GP evolves a single tree (one feature) using one type of features as terminals, which is given to the classification algorithm. From single-tree GP results (Table IV), we can see that a classification algorithm remained unable to perform well with only a single feature. We selected RF as a wrapper classification algorithm for single-tree GP methods as it produced highest results on all of the MTGP approaches.

Furthermore, among the two datasets, it has been shown that a particular type of features are important for classifying images that belong to a particular dataset. For the PH^2 dataset, the LBP_{RGB} features have produced the highest classification accuracy (88.25 ± 3.02%) among the four single-tree GP wrapper methods. Hence, we can say that the LBP_{RGB} has the most potential to distinguish "*benign*" and "*malignant*" skin lesion images taken from a dermatoscope (as shown in PH^2

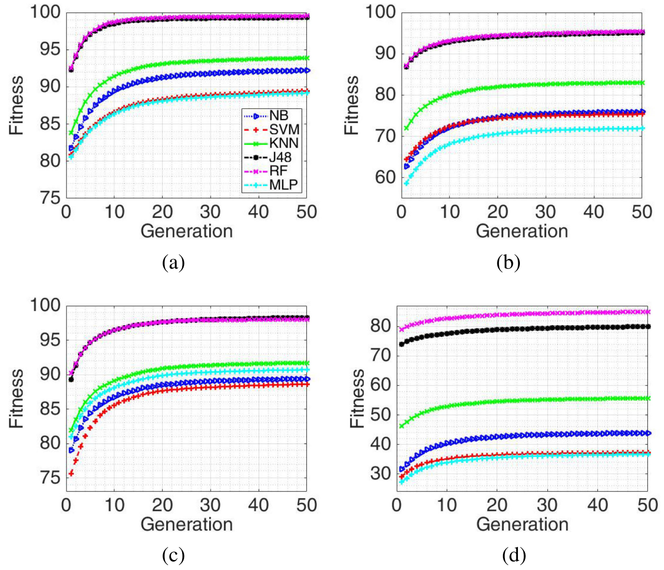


Fig. 6. The average fitness value per generation on PH^2 dataset for (a) binary classification, and (b) multi-class classification, and on Dermofit dataset for (c) binary classification, and (d) multi-class classification.

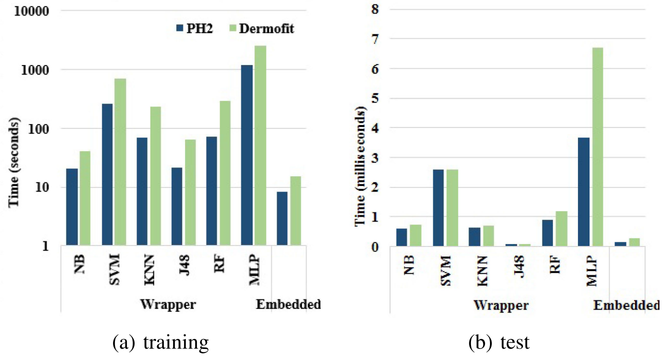


Fig. 7. The average computation time for *binary classification* using MTGP wrapper and embedded approaches on the two skin image datasets.

dataset) whereas, the $Lesion_{Color}$ features have produced the best results (92.31 ± 0.92%) among the four type of features on standard camera images (as shown in Dermofit dataset). It is evident from the results of single-tree GP embedded and wrapper methods that selection of a suitable feature extraction method largely impacts on achieving good performance. Hence, we can say that images taken from a particular device requires a particular feature extraction method necessary to obtain informative features. We observe a similar pattern while generating a classification model using our multi-tree approach.

The existing skin image classification methods using GP [10], [32] have employed the standard single-tree GP methods using an embedded approach and test their performance on only a single dataset. Moreover, GP has been used as an embedded method for performing feature selection as well as classification in the existing MTGP approach for melanoma detection [33]. All these existing works aim at only melanoma detection, i.e., a binary image classification task. To the best of our knowledge, this is the first time that MTGP is used in a wrapper approach, which is effective both for the binary and multi-class skin image classification tasks. Furthermore, our MTGP method in a wrapper approach for binary classification has outperformed all of these three existing methods in terms of classification performance as discussed earlier in this section.

VI. Further Analysis

A. Overall Analysis

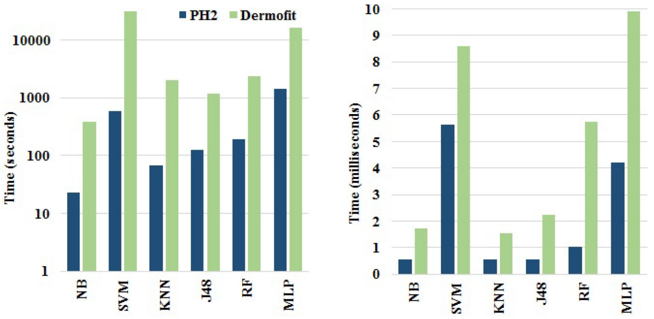
The average fitness value per generation of the 30 independent runs (each having 10 independent runs for the 10 folds in $10-$*fold cross validation*) using different seed values on the training data of the two datasets is depicted in Fig. 6. Fig. 6(a) and (b) show these plots for binary and multi-class classification on the PH^2 dataset, respectively, and Fig. 6(c) and d) show these plots on the Dermofit dataset.

For binary classification task (Fig. 6(a) and (c)), these graphs show that on average the programs make larger jumps in the first few generations than in the later generations. This trend has been observed in all the six wrapped classifiers (NB, SVM, k-NN, J48, RF, and MLP). In case of PH^2 dataset using RF classifier, as shown in Fig. 6(a), the fitness value has increased from 92.41% to 99.08% in the first 20 generations compared to the increase in fitness from 99.08% to 99.34% over the remainder 30 generations.

The plots for multi-class classification task (Fig. 6(b) and (d)) show different behavior compared to the binary classification task (Fig. 6(a) and b)). There is an abrupt increase in the first few generations (around 10) which becomes slightly insignificant in later generations (last 40 generations). This trend is more visible in case of the PH^2 dataset as compared to Dermofit dataset where a significant increase in fitness is only seen among the first 5 generations. In comparing RF and J48, we have seen that both these classifiers have shown similar training curves except in the case of multi-class classification on dermofit dataset where RF outperforms J48 by a relevant margin as can be clearly seen in Fig. 6(d).

B. Computation Time

The average training time needed for the proposed MTGP method and to test its performance on the unseen data for solving binary and multi-class classification tasks is presented in Figs. 7 and 8. We have also analyzed the training and test time required for the existing single-tree GP approaches using a single type of features, as shown in Fig. 9. Clearly, the time required to train a classification algorithm is affected by the number of images and classes in a dataset, the number of trees in the evolved GP individual, the number of features used to evolve an individual, and whether a wrapper or an embedded approach is adopted. This happens because evaluating a population of individuals having four trees necessarily requires more time as compared to evaluating a population of individuals having one tree. Similarly, during the evolutionary process, the *same-index crossover/ mutation* is applied on four trees which has more computational complexity, thereby takes more time as compared



(a) training (b) test

Fig. 8. The average computation time for *multi-class classification* using MTGP wrapper approaches on the two skin image datasets.

to the simple crossover in case of the single-tree approaches. Although the proposed method is more expensive, it does not take more than 15 seconds on average to evolve a solution. Furthermore, the embedded approaches are less expensive as compared to wrapper approaches. In our wrapper approach, after constructing new features, the original training and test datasets are transformed to a new dataset (with the help of the constructed features) which are then used to train (and test) a classification algorithm (NB, SVM, k-NN, J48, RF, and MLP). Hence, training a classification algorithm with the new constructed features results in increased computation time.

In Fig. 7, among the six wrapper binary classification algorithms, NB is the fastest to train a model. Overall, the fastest and highest-performing average training time using the proposed MTGP wrapper approach is given by RF on PH^2 and Dermofit datasets and takes only 71.7 and 297.5 seconds, respectively. Similarly, having these trained methods at hand, they take only 0.9 and 1.2 milliseconds on average to test an unseen skin image. Therefore, we can say that our proposed binary classification method is very effective and efficient for melanoma detection in real-time clinic situations and help dermatologists to decide whether a biopsy is required or not in diagnosis of skin images.

For multi-class classification, Fig. 8(a) and (b) depict that training a dataset with ten classes (Dermofit dataset) increases the computation time by many folds as compared to training a dataset with three classes (PH^2 dataset). Since multi-class classification methods require more training time as compared to binary classification methods, this behavior can easily be observed while comparing Figs. 7 and 8. However, an unseen image can be tested in fractions of a second using these trained models as shown by the test time depicted in Fig. 8(b).

We have also analyzed the computation time taken by the single-tree GP methods for binary classification as presented by the bar plots in Fig. 9. Clearly, the wrapper approaches take more time to train a classification method as compared to embedded approaches. Similarly, the bigger dataset (Dermofit) takes more time as compared to the smaller dataset (PH^2), regardless of which approach (wrapper or embedded) is used. Having these trained methods at hand, it takes only fractions of milli-seconds to test their performance on the test image as shown in Fig. 9(b). Overall, the embedded approaches are taking more test time as compared to the wrapper approaches. This is due to the fact that the evolved models in the embedded approaches have bigger trees with larger number of features, and hence, have more function nodes, which slightly increases computation time.

C. An Evolved GP Individual

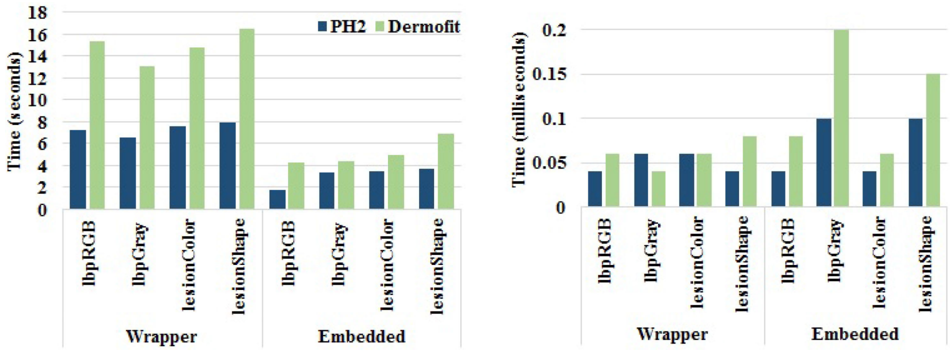
GP evolves models that can be interpretable. To see why our proposed MTGP in a wrapper approach can achieve good classification results, we have analyzed a good GP individual with four trees in Fig. 10 from the PH^2 binary classification experiments. The four constructed features have given 87.5% accuracy on the test data. GP found this perfect solution giving 100% accuracy on the training data, just after 24 generations. In Fig. 10, colored nodes show terminals, whereas white nodes show functions. As discussed earlier in Section V, $\mathrm{LBP}_{\mathrm{RGB}}$ features have the most potential compared to other feature types to classify images in PH^2 dataset. Since LBP captures local pixel-based properties of an image, these features with gray and color information can incorporate good discriminative information regarding the presence or absence of melanoma in a skin image. Furthermore, ($\mathrm{Lesion}_{\mathrm{Shape}}$ and $\mathrm{Lesion}_{\mathrm{Color}}$) features, which capture the global properties such as geometrical border shape and color variation between the lesion region and the skin region, respectively cannot provide as good performance as LBP feature.

In the $\mathrm{LBP}_{\mathrm{Gray}}$ tree from Fig. 10(a), the features G_{10} and G_{28} are selected twice and thrice, respectively. In addition, the expression $if(G_{28}, G_{52}, G_{24}, G_{51})$ is selected twice. This illustrates that these features possess good distinguishing ability between classes. Among the 177 $\mathrm{LBP}_{\mathrm{RGB}}$ features, only six prominent features (R_1, R_{12}, R_{26}, R_{40}, R_{116}, and R_{136}) are used to construct a tree (Fig. 10(b)). Similarly, only six features (G_{10}, G_{24}, G_{28}, G_{30}, G_{51}, and G_{52}) among the 59 $\mathrm{LBP}_{\mathrm{Gray}}$ features have been selected to build the tree in Fig. 10(a). The $\mathrm{Lesion}_{\mathrm{Color}}$ tree in Fig. 10(c) has been built from only two features among the 12 $\mathrm{Lesion}_{\mathrm{Color}}$ features, and the $\mathrm{Lesion}_{\mathrm{Shape}}$ tree in Fig. 10(d) has been constructed from six features among the 11 $\mathrm{Lesion}_{\mathrm{Shape}}$ features. Hence, the feature selection and construction ability of GP has provided discriminative constructed features as input to the decision tree classification algorithm, which helps this classification method achieve promising results.

In $\mathrm{Lesion}_{\mathrm{Color}}$ tree (Fig. 10(c)), C_0 and C_5 representing the mean of red color channel (μR) and the variance of blue color channel (σB), are combined to produce a significant constructed feature. In $\mathrm{Lesion}_{\mathrm{Shape}}$ tree (Fig. 10(d)), S_0, S_1, S_5, S_6, S_8, and S_9 are selected which correspond to the geometrical shape features: area, perimeter, irregularity indices A, C and D, and the major asymmetry index. These shape features can assist the dermatologist in real-time situations by providing significant knowledge about the lesion geometrical properties and hence, making a diagnosis much easier.

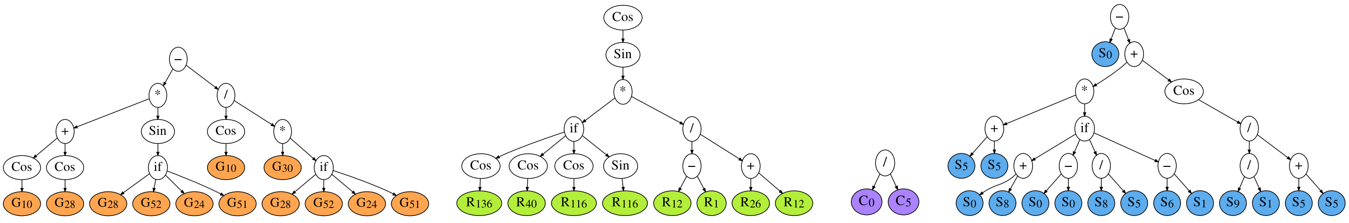
D. Feature Appearance in Constructed Features

GP automatically constructs new features by selecting more relevant and discriminative features among the whole set of original features. We have also explored and analysed this intrinsic



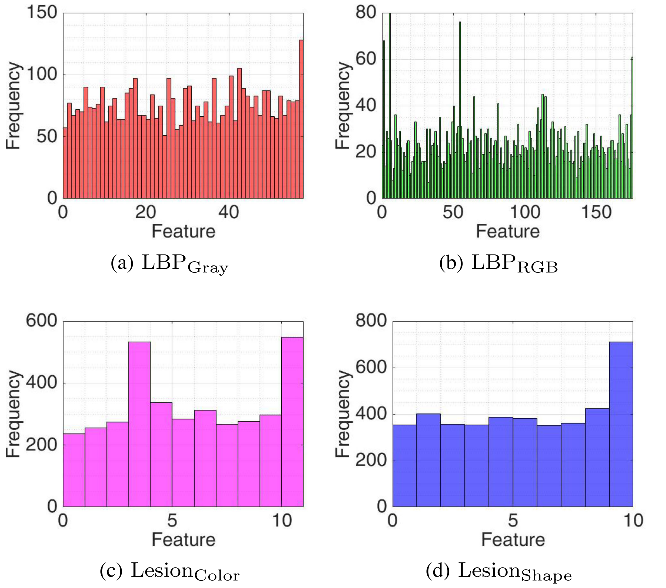
(a) training (b) test

Fig. 9. The average computation time for binary classification on (a) the training data, and (b) the test data, using different single-tree GP approaches based on a single type of features on the two skin image datasets.



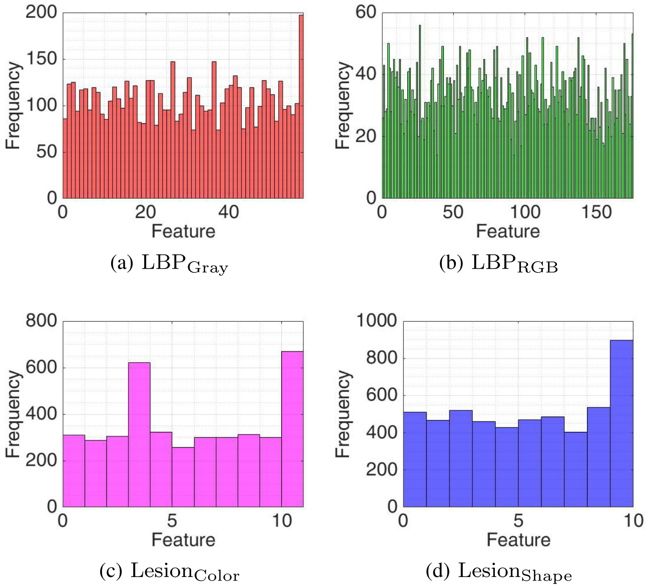
(a) (b) (c) (d)

Fig. 10. A good evolved GP individual for PH^2 dataset using a) $\mathrm{LBP}_{\mathrm{Gray}}$, b) $\mathrm{LBP}_{\mathrm{RGB}}$, c) $\mathrm{Lesion}_{\mathrm{Color}}$, and d) $\mathrm{Lesion}_{\mathrm{Shape}}$ features producing 87.5% accuracy on the test data in the binary classification task.



(a) $\mathrm{LBP}_{\mathrm{Gray}}$ (b) $\mathrm{LBP}_{\mathrm{RGB}}$ (c) $\mathrm{Lesion}_{\mathrm{Color}}$ (d) $\mathrm{Lesion}_{\mathrm{Shape}}$

Fig. 11. The average frequency of features in trees, each evolved with a single type of features on the PH^2 dataset in the *binary classification* task using RF as a classifier.



(a) $\mathrm{LBP}_{\mathrm{Gray}}$ (b) $\mathrm{LBP}_{\mathrm{RGB}}$ (c) $\mathrm{Lesion}_{\mathrm{Color}}$ (d) $\mathrm{Lesion}_{\mathrm{Shape}}$

Fig. 12. The average frequency of features in trees, each evolved with a single type of features on the PH^2 dataset in the *multi-class classification* task using RF as a classifier.

ability of GP to feature selection. Figs. 11 and 12 show the bars for the average number of times each feature appears in the constructed features among the 30 GP runs in the PH^2 experiments for binary and multi-class classification using RF as a classifier, respectively. It is evident from these plots that there are some features which are selected more frequently as compared to other features, e.g., G_{58}, the last feature among $\mathrm{LBP}_{\mathrm{Gray}}$ features in Fig. 11(a) appears almost twice as frequently as the other 58 $\mathrm{LBP}_{\mathrm{Gray}}$ features. Similarly, R_5, C_{11} and S_{10} have the highest frequency of occurrence among the $\mathrm{LBP}_{\mathrm{RGB}}$, $\mathrm{Lesion}_{\mathrm{Color}}$, and $\mathrm{Lesion}_{\mathrm{Shape}}$ features as shown in Fig. 11(b), (c) and (d), respectively. We have seen a similar pattern while having a closer look at Fig. 12 for multi-class classification task in the PH^2 experiments, where these features have the highest frequency again except R_5. This shows that these features have significant discriminative ability between classes, not only for binary classification task but also for multi-class classification. R_{26} which also represents the same structural properties as R_5, has the highest frequency among $\mathrm{LBP}_{\mathrm{RGB}}$ in the multi-class classification task.

For a deep analysis of these significant features (G_{58}, C_{11} and S_{10} for both tasks, R_5 for binary classification task alone, and R_{26} for multi-class classification task), digging further into the local and global properties of these features, we see that G_{58} are the non-uniform LBP features combined in one bin for gray-scale images. Though non-uniform features are not considered to have discriminative properties for texture analysis (that is why they are binned together in one bin), however, in our dataset, the number of times these non-uniform features appear in one class of images is quite different from their appearance in other classes, which makes them highly significant. For $\mathrm{LBP}_{\mathrm{RGB}}$ features, R_5 and R_{26} corresponds to the two 3×3 LBP windows which both represent edges present in an image. Inside the skin lesion, the different structures such as dots, streaks and regression areas with varying colors are highlighted by these pixel-level edge properties. Among the different classes, these structures vary and hence, these edge detecting $\mathrm{LBP}_{\mathrm{RGB}}$ features become prominent in distinguishing between classes. Among $\mathrm{Lesion}_{\mathrm{Color}}$ features, C_{11} corresponds to $\frac{\mu_B}{\bar{\mu}_B}$, which shows the ratio between mean of blue color channel of the lesion region and its surrounding skin region. Among $\mathrm{Lesion}_{\mathrm{Shape}}$ features, S_{10} is the most significant as its frequency is almost double as compared to the other 11 $\mathrm{Lesion}_{\mathrm{Shape}}$ features (Figs. 11(d) and 12(d)). It corresponds to asymmetry index, which provides the necessary information about the shape, particularly being computed from the asymmetry axes and area of the lesion. As described earlier in Section II-A(3), our analysis also confirms that asymmetry plays the most important and essential role in making a diagnosis for the binary and multi-class classification of skin cancer images.

VII. Conclusion

In this work, a novel method for skin cancer image classification using MTGP in a wrapper approach has been developed. The proposed method utilizes various local and global features extracted from skin cancer images. These features have sufficient information related to pixel-based RGB and gray-level properties, domain-specific geometrical shape characteristics, and variation in color within the lesion and the skin areas. These four type of pre-extracted features are given to multi-tree GP to generate four trees in a single GP individual by designing suitable genetic operators such as *same-index-crossover/mutation*. This type of crossover/mutation guarantees that each tree evolves from a single type of features. These trees are considered as constructed features, which are provided to a wrapper classification method to generate a classification model.

Our MTGP method has significantly produced better results than all the six commonly used classification algorithms (NB, k-NN, SVM, J48, RF, and MLP), the eight single-tree GP methods, and an existing MTGP embedded method. This shows the evidence of effective feature construction, which results in achieving good binary and multi-class classification results. We have also analyzed an interesting behavior to select a suitable feature extraction method in order to classify well a particular type of images taken from a specific optical instrument. We found that the local pixel-based features provide good discriminating knowledge to classify specialized (dermoscopy) images. On the other hand, global color variation and border shape features have more potential to discriminate images captured from a standard camera.

Although the proposed method has dealt well with the problems of using various types of features effectively and has provided good performance, it still has some limitations that will be addressed in the future. One of the limitation of this method is that it needs a binary mask usually provided by an expert dermatologist along with the images. The proposed method has used gray-scale and color features which have local and global information, however, it can further improve its performance by incorporating more information from frequency-based features such as wavelet-based features. In the future, we will explore employing pre-processing techniques to remove the various artifacts from skin images such as dark corners, ink markers, bubbles due to presence of gel, color chart used for measuring the diameter, ruler marks and skin hair. It is a challenging task to reduce noise from skin images without losing informative features necessary to achieve performance gains. Moreover, we would like to extend this work where GP will also be used for feature extraction before employing feature selection and feature construction.

References

[1] B. W. Stewart *et al.*, "World cancer report 2014," *Health*, 2014.

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

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

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

[5] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *Arch. Dermatology*, vol. 134, no. 12, pp. 1563–1570, 1998.

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

[7] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, vol. 1. Cambridge, MA, USA: MIT Press, 1992.

[8] M. Zhang, V. B. Ciesielski, and P. Andreae, "A domain-independent window approach to multiclass object detection using genetic programming," *EURASIP J. Advances Signal Process.*, vol. 2003, no. 8, pp. 841–859, 2003.

[9] A. Lensen, H. Al-Sahaf, M. Zhang, and B. Xue, “Genetic programming for region detection, feature extraction, feature construction and classification in image data,” in *Proc. Euro. Conf. Genetic Program.*, 2016, pp. 51–67.

[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 *Proc. 15th Pacific Rim Int. Conf. Artif. Intell.*, 2018, pp. 732–745.

[11] H. Al-Sahaf, B. Xue, and M. Zhang, “A multitree genetic programming representation for automatically evolving texture image descriptors,” in *Proc. Asia-Pacific Conf. Simulated Evolution Learn.*, 2017, pp. 499–511.

[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 Trans. Evol. Comput.*, vol. 21, no. 1, pp. 83–101, Feb. 2017.

[13] W.-J. Choi and T.-S. Choi, “Computer-aided detection of pulmonary nodules using genetic programming,” in *Proc. Int. Conf. Image Process.*, 2010, pp. 4353–4356.

[14] W. A. Tackett, “Genetic programming for feature discovery and image discrimination,” in *Proc. 5th Int. Conf. Genetic Algorithms*, 1993, pp. 303–311.

[15] M. Oltean and D. Dumitrescu, “Multi expression programming,” Dept. Comput. Sci., Babes-Bolyai University, Cluj-Napoca, Romania, Tech. Rep. 1, 2006.

[16] A. Lensen, B. Xue, and M. Zhang, “Generating redundant features with unsupervised multi-tree genetic programming,” in *Proc. Eur. Conf. Genetic Program.*, 2018, pp. 84–100.

[17] J.-H. Lee, C. W. Ahn, and J. An, “An approach to self-assembling swarm robots using multitree genetic programming,” *Scientific World J.*, vol. 2013, 2013, Art. no. 593848.

[18] D. P. Muni, N. R. Pal, and J. Das, “A novel approach to design classifiers using genetic programming,” *IEEE Trans. Evol. Comput.*, vol. 8, no. 2, pp. 183–196, Apr. 2004.

[19] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A survey on evolutionary computation approaches to feature selection,” *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.

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

[21] 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 J. Translational Eng. Health Medicine*, vol. 5, pp. 1–17, Jan. 2017.

[22] R. Garnavi, M. Aldeen, and J. Bailey, “Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis,” *IEEE Trans. Inf. Technol. Biomedicine*, vol. 16, no. 6, pp. 1239–1252, Nov. 2012.

[23] I. Maglogiannis and C. N. Doukas, “Overview of advanced computer vision systems for skin lesions characterization,” *IEEE Trans. Inf. Technol. Biomedicine*, vol. 13, no. 5, pp. 721–733, Sep. 2009.

[24] F. Ercal, A. Chawla, W. V. Stoecker, H.-C. Lee, and R. H. Moss, “Neural network diagnosis of malignant melanoma from color images,” *IEEE Trans. Biomed. Eng.*, vol. 41, no. 9, pp. 837–845, Sep. 1994.

[25] C. Barata, M. E. Celebi, and J. S. Marques, “Improving dermoscopy image classification using color constancy,” *IEEE J. Biomed. Health Informat.*, vol. 19, no. 3, pp. 1146–1152, May 2015.

[26] J. Shi, X. Zheng, J. Wu, B. Gong, Q. Zhang, and S. Ying, “Quaternion Grassmann average network for learning representation of histopathological image,” *Pattern Recognit.*, vol. 89, pp. 67–76, 2019.

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

[28] 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 Trans. Med. Imag.*, vol. 36, no. 3, pp. 849–858, Mar. 2017.

[29] R. Poli, “Genetic programming for image analysis,” in *Proc. 1st Annu. Conf. Genetic Program.*, 1996, pp. 363–368.

[30] C. Ryan, K. Krawiec, U.-M. OReilly, J. Fitzgerald, and D. Medernach, “Building a stage 1 computer aided detector for breast cancer using genetic programming,” in *Proc. Eur. Conf. Genetic Program.*, 2014, pp. 162–173.

[31] M. Iqbal, B. Xue, H. Al-Sahaf, and M. Zhang, “Cross-domain reuse of extracted knowledge in genetic programming for image classification,” *IEEE Trans. Evol. Comput.*, vol. 21, no. 4, pp. 569–587, Aug. 2017.

[32] Q. U. Ain, B. Xue, H. Al-sahaf, and M. Zhang, “Genetic programming for skin cancer detection in dermoscopic images,” in *Proc. Congr. Evol. Comput.*, 2017, pp. 2420–2427.

[33] Q. Ul Ain, H. Al-Sahaf, B. Xue, and M. Zhang, “A multi-tree genetic programming representation for melanoma detection using local and global features,” in *Proc. 31st Australas. Joint Conf. Artif. Intell.*, 2018, pp. 111–123.

[34] K. Shimizu, H. Iyatomi, M. E. Celebi, K.-A. Norton, and M. Tanaka, “Four-class classification of skin lesions with task decomposition strategy,” *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 274–283, Jan. 2015.

[35] 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 *Proc. 35th Annu. Int. Conf. IEEE Eng. Medicine Biol. Soc*, 2013, pp. 5437–5440.

[36] C. Barata, J. S. Marques, and J. Rozeira, “A system for the detection of pigment network in dermoscopy images using directional filters,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 10, pp. 2744–2754, Oct. 2012.

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

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

[39] S. S. Keerthi and C.-J. Lin, “Asymptotic behaviors of support vector machines with gaussian kernel,” *Neural Comput.*, vol. 15, no. 7, pp. 1667–1689, 2003.

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

Qurrat Ul Ain (Student Member, IEEE) received the B.Sc. degree in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2009 and the M.S. degree in computer science from International Islamic University, Islamabad, Pakistan, in 2013. She joined the Victoria University of Wellington, Wellington, New Zealand, in July 2016 where she is currently working toward the Ph.D. degree in computer science. Her current research interests include evolutionary computation, particularly genetic programming, computer vision, pattern recognition, machine learning, feature manipulation including feature selection, extraction and construction, and transfer learning. She is a member of the IEEE Computational Intelligence Society and has been serving as a reviewer of international journals and conferences. She is also a member of the Evolutionary Computation Research Group and Feature Analysis, Selection, and Learning in Image and Pattern Recognition with the Victoria University of Wellington.

Harith Al-Sahaf (Member, IEEE) received the B.Sc. degree from Baghdad University, Baghdad, Iraq, in 2005, the master’s and Ph.D. degrees form the Victoria University of Wellington (VUW), Wellington, New Zealand, in 2010 and 2017, respectively, all in computer science. He joined VUW, New Zealand in July 2007. In October 2016, he has been joined the School of Engineering and Computer Science, VUW as a Postdoctoral Research Fellow and as a Full-Time Lecturer since September 2018. His current research interests include evolutionary computation, particularly genetic programming, computer vision, pattern recognition, evolutionary cybersecurity, machine learning, feature manipulation including feature detection, selection, extraction and construction, transfer learning, domain adaptation, one-shot learning, and image understanding.

Dr. Al-Sahaf is a member of the IEEE CIS ETTC Task Force on Evolutionary Computer Vision and Image Processing, the IEEE CIS ETTC Task Force on Evolutionary Computation for Feature Selection and Construction, the IEEE CIS ISATC Task Force on Evolutionary Deep Learning and Applications, and the IEEE CIS ISATC Intelligent Systems for Cybersecurity.

Bing Xue (Member, IEEE) received the B.Sc. degree from the Henan University of Economics and Law, Zhengzhou, China, in 2007, the M.Sc. degree in management from Shenzhen University, Shenzhen, China, in 2010, and the Ph.D. degree in computer science from the Victoria University of Wellington (VUW), Wellington, New Zealand, in 2014. She is currently an Associate Professor with the School of Engineering and Computer Science, VUW. She has more than 200 papers authored in fully refereed international journals and conferences and her research focuses mainly on evolutionary computation, feature selection, feature construction, image analysis, and transfer learning.

Dr. Xue is currently the Chair of IEEE Computational Intelligence Society (CIS) Data Mining and Big Data Analytics Technical Committee, the Vice Chair of the IEEE Task Force on Evolutionary Feature Selection and Construction, the Vice Chair of IEEE CIS Task Force on Transfer Learning & Transfer Optimization, and the Vice Chair of IEEE CIS Task Force on Evolutionary Deep Learning and Applications. She is also an Associate Editor for six international journals, including IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE, *ACM Transactions on Evolutionary Learning and Optimization*, and *Journal of Royal Society of New Zealand*.

Mengjie Zhang (Fellow, IEEE) received the B.E. and M.E. degrees from Artificial Intelligence Research Center, Agricultural University of Hebei, Hebei, China, and the Ph.D. degree in computer science from RMIT University, Melbourne, VIC, Australia, in 1989, 1992, and 2000, respectively. He is currently a Professor of Computer Science, the Head of the Evolutionary Computation Research Group, and an Associate Dean (Research and Innovation) in the Faculty of Engineering. He has authored more than 500 research papers in refereed international journals and conferences. His current research interests include evolutionary computation, particularly genetic programming, particle swarm optimization, and learning classifier systems with application areas of image analysis, multi-objective optimization, feature selection and reduction, job shop scheduling, and transfer learning.

Dr. Zhang is a Fellow of Royal Society of New Zealand and have been a Panel member of the Marsden Fund (New Zealand Government Funding). He was the Chair of the IEEE CIS Intelligent Systems and Applications Technical Committee, and the Chair for the IEEE CIS Emergent Technologies Technical Committee and the Evolutionary Computation Technical Committee, and a member of the IEEE CIS Award Committee. He is the Vice Chair of the IEEE CIS Task Force on Evolutionary Feature Selection and Construction, the Vice Chair of the Task Force on Evolutionary Computer Vision and Image Processing, and the founding Chair of the IEEE Computational Intelligence Chapter in New Zealand.