# Genetic Programming for Feature Selection and Feature Construction in Skin Cancer Image Classification

Qurrat Ul Ain, Bing Xue, Harith Al-Sahaf, and Mengjie Zhang

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

**Abstract.** The incidence of skin cancer, particularly, malignant melanoma, continues to increase worldwide. If such a cancer is not treated at an early stage, it can be fatal. A computer system based on image processing and computer vision techniques, having good diagnostic ability, can provide a quantitative evaluation of these skin cancer cites called skin lesions. The size of a medical image is usually large and therefore requires reduction in dimensionality before being processed by a classification algorithm. Feature selection and construction are effective techniques in reducing the dimensionality while improving classification performance. This work develops a novel genetic programming (GP) based two-stage approach to feature selection and feature construction for skin cancer image classification. Local binary pattern is used to extract gray and colour features from the dermoscopy images. The results of our proposed method have shown that the GP selected and constructed features have promising ability to improve the performance of commonly used classification algorithms. In comparison with using the full set of available features, the GP selected and constructed features have shown significantly better or comparable performance in most cases. Furthermore, the analysis of the evolved feature sets demonstrates the insights of skin cancer properties and validates the feature selection ability of GP to distinguish between benign and malignant cancer images.

**Keywords:** Genetic Programming · Image classification · Dimensionality reduction · Feature selection · Feature construction.

## 1 Introduction

Melanoma is the most serious type of skin cancer, which spreads rapidly to other parts of the body if left untreated. Hence, early detection is essential, as the estimated 5-year survival rate for melanoma decreases from over 99% if detected in earliest stages to about 14% if detected in later stages [7]. New Zealand has the highest melanoma incidence rate in the world having more than 4000 new cases each year. Since this cancer is visible on the skin, it is potentially detectable at a very early stage that can lead to earlier more effective treatment. New computer vision technologies not only allow earlier detection of melanoma, but also reduces the large number of needless, costly and painful biopsy procedures [6]. This work

develops a computational method which may allow medical practitioners and patients to adequately track skin lesions and detect cancer earlier.

A powerful way to achieve skin cancer detection via computer vision is to use dermoscopy images, and form the task as a binary image classification problem, i.e., *benign* and *malignant* classes of images [22]. For skin cancer classification, important characteristics for distinguishing between different cancer types, are based on dermoscopy criteria, specifically, Asymmetry, Border, Colour, and Diameter (ABCD) rule [20], and 7-point check-list method [4] (Asymmetry, Pigment network, Dots/Globules, Streaks, Regression areas, Blue-whitish veil and presence of Colours; white, red, light-brown, dark-brown, blue-gray, black). These are the key medical properties that help dermatologists for classification of various types of cancer. The dermosopic images are huge in size, whereas the relevant information about the disease is confined in a limited number of pixels or features in these images. Hence, there is a need for dimensionality reduction which aims at reducing the number of features and selecting only prominent features having good discriminating ability between classes. This helps reduce computation time as well as increase performance and interpretability of the commonly used classification algorithms. Moreover, in order to find which salient texture patterns or image features in these images are the cause behind a particular cancer type, interpretable methods are required to provide insights of these critical features.

Genetic Programming (GP) is an evolutionary computation (EC) algorithm based on Darwinian principles of biological evolution and natural selection [11]. GP automatically explores the solution space to evolve a computer program (model/solution), often represented by a tree-like structure, for a given problem [11]. GP has the ability to perform implicit feature selection by selecting prominent features at its terminal (leaf) nodes and the goodness of the evolved program is evaluated by a fitness measure [2]. Feature selection (FS) selects a subset of original features while feature construction (FC) creates a new feature(s) from the original set of features [21]. FC involves transforming a given set of input features to generate a new set of more powerful features [17]. FS and FC both can help improve performance by selecting relevant features and constructing new high-level features. Hence, FS and FC are good tools not only to improve performance, but also to reduce the dimensionality and hence provide features which take less computation time while being processed by the classification algorithm. Moreover the medical practitioners are interested in finding the cause of a disease, and a system is highly recommended to have such causal information. With the property of GP evolved programs being interpretable, giving information about which features are prominent in constructing new high-level features, the medical practitioners can gain deep understanding of which specific texture patterns and colour variations are the cause of the disease. However, there is limited work done to FS and FC in dermoscopy image classification.

***Goals:*** This work develops a new FS and FC method using GP for skin cancer image classification problems. Different from most existing methods, the proposed method aims at constructing features only using the GP-selected fea-

tures, which can have the ability to construct more informative features as compared to construct features from all of the original features. GP-selected and GP-selected-constructed features will be applied with common machine learning algorithms for classification. This work aims to address the following questions:

- Which features get selected by GP during the evolutionary process to achieve better classification performance and why?
- Whether GP can construct informative feature from GP-selected features that improve the performance of common classification algorithms for the task of skin cancer image classification?
- Whether GP-selected-constructed features provide better discriminating ability while using colour features or gray-scale features?
- How well this new method works as compared to other existing method?
- While analyzing the pixel-based texture patterns, how the GP-selected features can help identify those patterns which are prominent in effectively contributing to the classification performance?

## 2 Background

### 2.1 Related Work

Over the last two decades, several computer-aided diagnostic (CAD) systems [7] have been developed to help medical practitioners distinguish between *benign* and *malignant* skin lesions [1, 8, 22]. Zortea et al. [22] developed a CAD tool based on a camera with attached dermatoscope, and compared its performance to three experienced dermatologists. The system extracts features related to the asymmetry, colour, border, geometry, and texture of skin lesions, computed from automatically segmented images. With a dataset of 206 skin lesions, the classifier (quadratic discriminant analysis) provided competitive sensitivity (86%) and specificity (52%) compared to the most accurate dermatologist. *Sensitivity* is the accuracy of correctly classified diseased instances and *specificity* is the accuracy of correctly classified non-disease instances.

Abuzaghleh et al. [1] proposed a non-invasive real-time automated skin lesion system for the early detection and prevention of melanoma. This system has two components: 1) a real-time alert to help the users prevent skin burn measured by an equation, and 2) an automated image analysis module capable of capturing and classifying the lesion images. The second module includes image acquisition, hair removal, lesion segmentation, feature extraction, and classification. The method used the standard overall classification accuracy (i.e. the number of correctly classified instances divided by the total number of instances) as a fitness measure which is not suitable for an imbalance dataset [19]. Esteva et al. [7] demonstrated the classification of skin lesions using convolutional neural network (CNN) trained from images, using only pixels and class labels. The CNN was trained using thousands of images, from 2,032 different classes and its performance is tested against 21 dermatologists on biopsy-proven clinical images. The CNN method outperformed all the experts, demonstrating artificial intelligence being capable of classifying skin cancer with a level of competence comparable

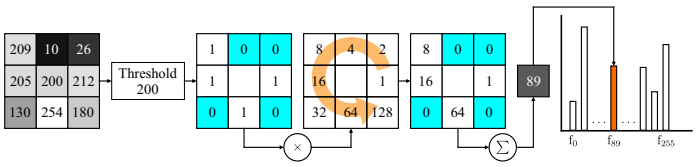



**Fig. 1.** The LBP process.

to dermatologists. Generally, the performance of CNN is primarily constrained by data and can only classify well provided sufficient training examples which leads to long computation time and requires huge computing resources.

Menegola et al. [15] demonstrated transfer learning for automated melanoma screening using deep neural network (DNN). One of the key limitation of using pre-trained DNN is that the images must be resized to match the architecture. In [7] and [15], the images are resized to a smaller size distorting the aspect ratio, which badly effects the texture patterns important to discriminate between classes. In [3], GP has been utilized to automatically evolve a classifier for binary skin cancer image classification based on domain specific features provided by the dermatologists and texture features extracted by local binary patterns. The experiment results and analysis of the evolved programs confirmed the ability of GP for feature selection. Motivated by this feature selection ability, we will use GP as a feature selection and feature construction method in our current work.

Despite extensive research in investigating the diverse presentations and physical characteristics of skin cancer, the clinical diagnostic accuracy remains suboptimal. Skin lesion classification is a very challenging problem for several reasons including relatively poor contrast between the skin and lesion areas, variations in skin tone, presence of artefacts (hairs, ink, gel bubbles, date markers, ruler marks, etc.), non-uniform lighting, physical location of the lesion and most importantly variations in the lesion itself in terms of shape, size, colour, texture and location in the image frame [16]. While designing a robust skin cancer image classification algorithm, these factors must be considered which makes this task harder as compared to other image classification problems.

### 2.2 Local Binary Patterns

Local binary patterns (LBP) is a dense image descriptor developed by Ojala et al. [18] that has been extensively used for feature extraction in a wide range of computer vision tasks. LBP scans the image in a pixel-by-pixel fashion using a sliding window of fixed radius. The central pixel value is computed based on the intensity values of neighbouring pixels lying on the radius as depicted in Fig. 1. It then generates a histogram (i.e. feature vector) based on the computed values. The LBP operator is defined as:

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

where $p$ is the number of neighbouring pixels, $r$ is the radius, $v_i$ and $v_c$ are the intensity values of the $i^{th}$ neighbour and central pixel, respectively. $S(x)$ returns

0 if $x < 0$ and 1 otherwise. The value computed from the above expression is assigned to central pixel and corresponding bin of histogram is incremented by 1. The value of $t^{th}$ bin of a histogram $H$ computed on an image of size $m \times n$ as:

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

where the value of $t$ ranges from 0 to $T-1$, $T$ is the maximum number of bins in the histogram, and $V_{i,j}$ is the value of the pixel at coordinate $(i, j)$. Moreover, there are two kinds of LBP codes: *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 00111000, 00001111, and 10000001 are uniform, whilst the codes 00110110, 01001110, and 01010100 are non-uniform. Uniform codes detect various texture primitives such as corners, edges, line ends, dark spots and flat regions in images. Using only uniform codes, the size of the feature vector can be reduced from $2^p$ bins to $p(p-1)+3$ bins, simply by combining non-uniform codes. In dermoscopy images, uniform codes help in detection of streaks (line ends) and blobs space (flat regions) which may help improve performance.

## 3 The Proposed Method

The proposed GP method is described in this section. It consists of two stages; one for feature selection (stage-1) and second for feature construction (stage-2). The overall structure is depicted in Fig. 2 and Fig. 3. First the images are converted to feature vectors by using LBP as discussed in Section 2.2. Then these features are fed into GP method. GP has the ability of implicit feature selection during its evolutionary process, since not all the features are used as the leaf nodes of a GP tree. The leaf nodes of a GP tree are the selected features. With the help of genetic operators, such as crossover and mutation, GP evolves a classifier/GP tree including informative features. These selected features usually have high discriminating ability between classes. After performing GP for multiple runs, i.e. 10, the features appearing in the best individual (evolved tree) having highest performance on training data are selected.

The selected features which are obtained from stage-1 are used as the input to stage-2 for feature construction. Here again after the 10 individual GP runs, the evolved individual/tree having the highest performance on the training data is selected. The evolved tree is the one constructed feature which will be used along with GP-selected features (computed after stage-1) for classification. To this end, we have the selected features (outcome of stage-1) and a constructed feature (outcome of stage-2). These selected and constructed features are concatenated to form the final feature vector, which will be given to the classification method.

In order to deal with FS bias and FC bias issues, the dataset is divided into 10 folds where 9 folds are used for training and 1 fold for testing, such that only training folds are used for FS and FC and the test fold remain unseen during the learning process. The method used for FS and FC using the training data to evolve selected features (outcome of stage-1) and to evolve constructed feature

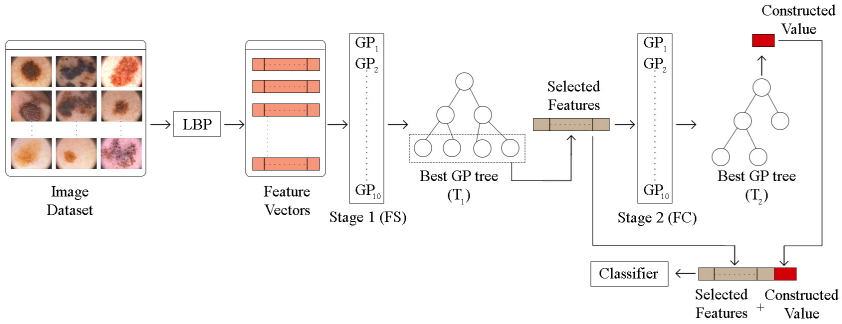



**Fig. 2.** Training process: Training images are first converted to feature vectors using LBP which are given to GP method for feature selection (stage-1). Among the 10 GP runs, the best GP tree having highest performance is selected. Features appearing in this GP tree are called GP-selected features which are then given to GP method for feature construction (stage-2). Again among the 10 GP runs, the best performing GP tree is selected which is called GP-constructed feature. A new feature vector having GP-selected and GP-constructed feature is formed which is given to the classifier.

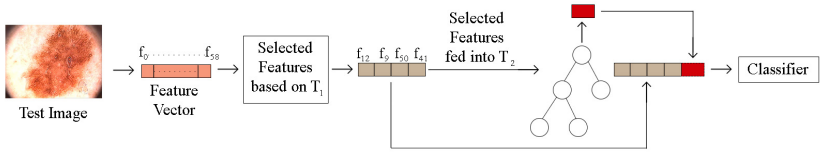



**Fig. 3.** Test process: Each test image is converted to a feature vector using LBP (here features from $f_0$ to $f_{58}$ represents 59 LBP$_{gray}$ features. Based on best GP tree ($T_1$), evolved on training data in stage-1, some of the features are selected (e.g. $f_{12}$, $f_9$, $f_{50}$, $f_{41}$). These feature values are fed into the best GP tree ($T_2$) evolved on training data in stage-2 to get GP-constructed feature value for each test image. GP-selected and GP-constructed features make the final feature vector to be given to the classifier.

(outcome of stage-2) is illustrated in Fig. 2. For getting the transformed feature vectors for the test instances, the method illustrated in Fig. 3 has been adopted in this work.

### 3.1 Fitness Function

Having (very) different number of instances in different classes is commonly referred as a class imbalance problem. In this case, the use of the standard overall classification accuracy, defined as the ratio ($N_{\text{correct}}/N_{\text{total}}$) between correctly classified instances ($N_{\text{correct}}$) and total number of instances ($N_{\text{total}}$), is inappropriate. Alternatively, the balanced classification accuracy has been used as a good measure for imbalance classification problems[19], since it gives equal importance to both classes without any bias. Therefore, we adopted it as the fitness function in this study, which is given in Equation (3):

$$Fitness = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{3}$$

where $TP$, $TN$, $FP$, and $FN$ refers to true positive, true negative, false positive, and false negative, respectively, where *malignant* is the positive class and *benign* is the negative class.

### 3.2 Terminal Set and Function Set

The terminal set consists of uniform LBP features. Gray-level LBP features (referred as $\text{LBP}_{\text{gray}}$) are a total of 59 features and colour LBP features (referred as $\text{LBP}_{\text{rgb}}$) are 177 features. For computing $\text{LBP}_{\text{rgb}}$, a colour image is converted to its red, green and blue channel images and then LBP features are extracted from each of them. Hence there are a total of 177 (= 59 LBP features $\times$ 3 channels) $\text{LBP}_{\text{rgb}}$ features. The value of the $i^{th}$ feature is indicated as $Fi$. The window size of 3×3 pixels and a radius of 1 pixel ($\text{LBP}_{8,1}$) is used.

The function set consists of four arithmetic operators, two trigonometric functions and one conditional operator, which are $\{add, sub, mul, div, sin, cos, if\}$. The first three arithmetic operators and the two trignometric operators have the same arithmetic and trigonometric meaning. However, division is protected that returns 0 when divided by 0. The *if* operator takes four inputs and returns the third if the first is greater than the second; otherwise, it returns the fourth.

## 4 Experiment Design

### 4.1 Dataset

A dataset of dermoscopy images namely $PH^2$ [14] gathered at Pedro Hispano Hospital Portugal, is used in the experiments. Dermoscopy is a non-invasive technique that allows microscopic visualization of inner skin morphological structures not visible to the naked eye [13]. Such images are rich enough to investigate them for presence/absence of skin cancer. The images are 8-bit RGB (red, green and blue) color images. The dataset includes 200 images which belong to three classes: common nevi (80 instances), atypical nevi (80 instances), and melanomas (40 instances). In dermatology, common nevi refers to non-disease lesion, atypical nevi refers to a currently non-disease lesion, but has chances to develop malignancy at a later stage, and melanoma is the diseased lesion. For our experiments focusing on binary classification, 80 common nevi and 80 atypical nevi are used as "*benign*" class, and 40 *melanoma* are used as "*malignant*" class. Samples of the two classes are presented in Fig. 4.

For performing the experiments, *10-fold cross validation* is used. The dataset is divided into ten folds such that nine folds are used for training and one fold for test. In our experiments features are selected and constructed using nine (training) folds and the last (test) fold remains unseen during this FS and FC processes in order to avoid FS and FC biases. This process is repeated ten times for all the different combinations of folds and the results are reported as mean and standard deviation of the fitness value. All the folds are randomly selected but are ensured that the ratio of instances of each class in each fold is the same as in the original dataset.

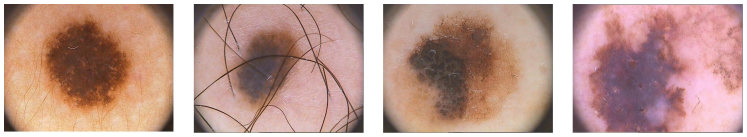


**Fig. 4.** Some dermoscopy images: *benign* lesions (first two), and melanomas (last two).

**Table 1.** Parameter settings of the GP method.

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

### 4.2 GP Parameters

The GP parameters are listed in Table 1. For generating the initial population, "Ramped half-and-half" method is used and the population size is set to 1024. Tournament selection with size 7 is applied to pick good individuals for producing new generations while maintaining population diversity. During the evolutionary process, the percentages for producing new individuals through crossover, mutation and elitism are 0.8, 0.19 and 0.01, respectively. The depth of the trees ranges between 2 and 8. After reaching a maximum of 50 generations, the evolutionary process stops unless a perfect individual with accuracy 100% is found.

For stage-1, the number of individual GP runs is 10. Among these 10 evolved trees, the one having highest performance on the training data is selected and the features appearing in that tree (GP-selected features) are used as input to stage-2 for feature construction. Here in stage-2, GP runs for 10 times and evolves trees. Again the best performing tree among the 10 evolved trees on the training data is selected as the constructed feature. The above procedure is repeated 30 times to get 30 sets of selected and constructed features. Note that the test folds remain unseen during both stages in order to avoid FS and FC biases. In one set of experiments, the random seeds for each of the 10 runs are all different. The implementation of GP method is done using the Evolutionary Computing Java-based (ECJ) package version 23 [12].

### 4.3 Methods for Classification

To check the performance of the feature sets obtained from GP on the test set, six classification methods are applied: Naïve Bayes (NB), $k$-Nearest Neighbor ($k$-NN) where $k = 1$ (the closest neighbor), Support Vector Machines (SVM), Decision Trees (J48), Random Forest (RF), and Multilayer Perceptron (MLP). The implementations of all these methods are taken from the commonly used Waikato Environment for Knowledge Analysis (WEKA) package [9]. In a study [10] on kernel functions in SVM, it has been shown that non-linear kernel can achieve similar or better performance than linear kernel. Hence, a Radial basis Function (RBF) kernel is used instead of the default linear kernel in WEKA.

For MLP, the learning rate, momentum, training epochs and number of hidden layers are set to 0.1, 0.2, 60, and 20, respectively. These parameters are specified empirically as they gave the best performance amongst other settings.

## 5 Results and Discussions

### 5.1 Overall Results

The results of the two experiments using $\text{LBP}_{\text{gray}}$ and $\text{LBP}_{\text{rgb}}$ are presented in Table 2. Vertically, the table comprises of two blocks where first corresponds to the results of using $\text{LBP}_{\text{gray}}$ features and second shows results of $\text{LBP}_{\text{rgb}}$ features. Horizontally, the table consists of 7 columns where first lists the classification algorithm, second and third show respectively the training and test performances using all features represented by "*All*". The rest of the columns show training and test performances using GP-selected, and GP-selected-constructed features. The values of the results using all features is the mean and standard deviation of applying 10-folds cross validation to the dataset. For "GP-selected" and "GP-selected-constructed" columns, the training (Fig. 2) and test (Fig. 3) processes are repeated 30 times, hence we get 30 accuracies for each classifier which are represented as mean and standard deviation ($\bar{x} \pm s$) in Table 2. For making a clear comparison between using different feature-sets, the results are also tested using *Wilcoxon signed-rank test* with a significance level of 5%. The statistical test has been applied on the test results to check which feature-set has better ability to discriminate between *benign* and *malignant* classes. The symbols "+", "−" and "=" are used to represent significantly better, significantly worse and not significantly different performance, respectively, of the two features-sets (GP-selected and GP-selected-constructed) in comparison with all features. For example, in $\text{LBP}_{\text{rgb}}$ block the test performance of SVM using GP-selected features is represented as "76.42 ± 1.35+" where the "+" sign represents that GP-selected features have significantly outperformed all features.

Analysing the effect of dimensionality reduction, it has been seen that while using $\text{LBP}_{\text{gray}}$ features (59 in total), GP selects only half of the features (around 28) in its tree having tree depth of 8. Here the number of features is 28.26 computed as average number of features appeared in 30 evolved GP trees. In case of $\text{LBP}_{\text{rgb}}$, the reduction in number of features is significant (from 177 to around 35). Except $k$-NN, all the classification algorithms have achieved either better or comparable performance for classifying skin cancer images. This shows that GP with its feature selection ability, has pushed most of the classification algorithms achieve good classification performance even with reduced number of features. Moreover, the feature constructed by GP-selected features are more powerful in creating good training models as compared to feature constructed by all set of features. This is evident when comparing GP-selected and GP-selected-constructed results. Our method allows GP to perform implicit feature selection twice during each stage, which helps improve the classification performance.

Variation in colour of *malignant* melanoma is a major discriminative aspect for dermatologists [5] which is validated by the results as well. Comparing the results of gray features and colour features, colour features have shown better

**Table 2.** The accuracy (%) on the training and test set using all features, GP-selected features, and GP-selected-constructed features (results are represented in terms of mean accuracy and standard deviation ($\bar{x} \pm s$)).

| | | All (59 feature) | | GP-selected (28.26 feature) | | GP-selected-constructed (29.26 feature) | |
|---|---|---|---|---|---|---|---|
| | | training | test | training | test | training | test |
| $LBP_{gray}$ | NB | 71.63 ± 2.97 | 63.44 ± 12.2 | 72.12 ± 0.89 | 65.75 ± 1.71+ | 74.32 ± 1.46 | 65.71 ± 2.46+ |
| | SVM | 89.93 ± 1.36 | 70.94 ± 11.9 | 80.79 ± 1.50 | 70.29 ± 1.67− | 84.11 ± 1.79 | 70.94 ± 2.55= |
| | KNN | 100.0 ± 0.00 | 71.25 ± 9.46 | 100.0 ± 0.00 | 67.76 ± 1.94− | 100.0 ± 0.00 | 67.55 ± 1.99− |
| | J48 | 88.89 ± 8.68 | 61.56 ± 13.7 | 85.97 ± 2.79 | 62.94 ± 2.69+ | 92.03 ± 1.73 | 64.84 ± 3.55+ |
| | RF | 100.0 ± 0.00 | 62.81 ± 10.2 | 100.0 ± 0.00 | 64.17 ± 1.48+ | 100.0 ± 0.00 | 65.97 ± 2.02+ |
| | MLP | 74.44 ± 1.53 | 67.81 ± 8.62 | 69.76 ± 1.69 | 66.16 ± 2.15− | 73.81 ± 2.04 | 67.33 ± 2.29= |
| | | (177 feature) | | (34.89 feature) | | (35.95 feature) | |
| | | training | test | training | test | training | test |
| $LBP_{rgb}$ | NB | 79.10 ± 1.62 | 76.25 ± 8.75 | 78.19 ± 0.75 | 76.04 ± 1.81= | 79.82 ± 0.96 | 76.21 ± 1.91= |
| | SVM | 100.0 ± 0.00 | 75.00 ± 13.6 | 85.29 ± 1.14 | 76.42 ± 1.35+ | 87.59 ± 1.44 | 75.77 ± 2.41= |
| | KNN | 100.0 ± 0.00 | 74.69 ± 13.7 | 100.0 ± 0.00 | 73.31 ± 1.88− | 100.0 ± 0.00 | 73.31 ± 1.88− |
| | J48 | 90.87 ± 8.31 | 73.13 ± 9.60 | 83.22 ± 1.59 | 73.39 ± 1.96= | 92.08 ± 1.17 | 72.74 ± 2.84= |
| | RF | 100.0 ± 0.00 | 75.94 ± 9.79 | 100.0 ± 0.00 | 75.34 ± 1.47− | 100.0 ± 0.00 | 75.53 ± 1.72= |
| | MLP | 84.13 ± 1.69 | 76.88 ± 10.1 | 80.32 ± 0.62 | **78.17 ± 0.83**+ | 82.55 ± 1.08 | **77.54 ± 1.67**+ |

performance in almost all cases. According to the overall results, MLP achieved the highest performance, i.e., 78.17% ± 0.83, which is comparatively well enough as compared to the state-of-the-art method [5] having 84.3% balanced accuracy on the same dataset and same fitness measure, considering overhead of preprocessing and manual segmentation, which requires human expertise in [5].

### 5.2 Analysis of the Evolved Features

To see why the GP-selected-constructed features can achieve good performance, we show a good GP tree (Fig. 5) among the 10 GP runs after stage-2 having 90.63% accuracy on the training set. This tree is taken from $LBP_{rgb}$ experiments where the total number of features is 177. In the figure, gray nodes represent functions and white nodes represent terminals. Note that for constructing this tree, features selected by a tree in stage-1 are used only and not the whole feature set. Hence, employing feature selection twice. This tree is constructed from ten $LBP_{rgb}$ features appeared in a tree in stage-1, which are F15, F40, F68, F90, F95, F105, F113, F117, F119, and F154. The values of these 10 selected features (after stage-1) and the constructed feature (after stage-2) are plotted in a bar chart shown in Fig. 6. For analysis of the selected feature, we take the simple example of features F15 and F154. As an example, we take the values of these features for only two instances from each class. The bar plot shows that the values of F15 (shown in black) and F154 (shown in green) for the *benign* instances (B1 and B2) are high as compared to values for *malignant* instances (M1 and M2). Hence, by combining these GP-selected features, the constructed feature divides instances of the two classes into two completely separate intervals as shown by blue colour in Fig. 6. Therefore, using these powerful GP-selected-constructed features from the selected features, the common classification algorithms become able to achieve better discrimination between the *benign* and *malignant* classes, resulting in improved classification performance.

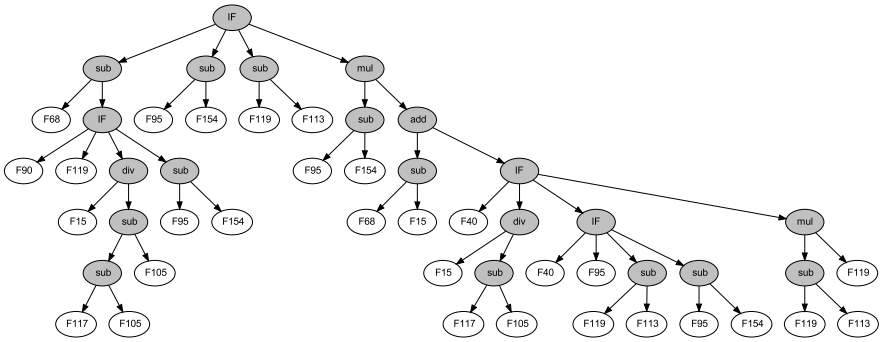



**Fig. 5.** A good evolved GP tree after stage-2 having 90.63% accuracy on training data.

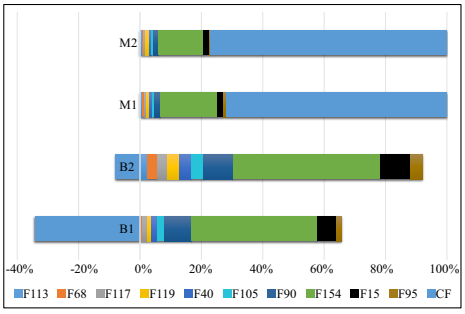
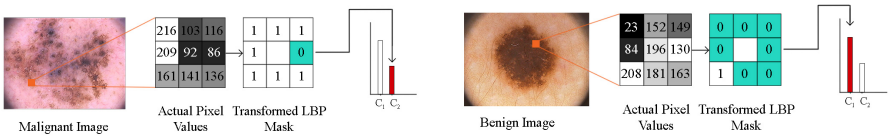



**Fig. 6.** Bar chart showing the values of different selected features after stage-1 and value of constructed feature "CF" after stage-2.

We further analyse the LBP texture pattern of these two features F15 and F154 to match skin cancer image properties like streaks and blobs. Fig. 7(a) shows the extracted 3×3 window for F15, its transformed LBP mask and the histogram showing the given pattern added to the *malignant* class bin represented as $C_2$. This mask shows presence of line ends in the image, which matches the presence of streaks in *malignant* images. According to the bar chart, this value is less for *malignant* images and high for *benign* images, which helps our method to distinguish between the two classes effectively. Similarly, Fig. 7(b) shows the extracted 3×3 window for F154, its transformed LBP mask and the histogram showing the given pattern added to the *benign* class bin represented as $C_1$. This mask shows the presence of corners in an image. Its value for the *malignant* class is lower as compared to the *benign* class. This maps to the structure of the *benign* and *malignant* lesions. The *benign* lesions are often a confined dense structure having less variation in colour, however *malignant* lesions have often sparse structure, spreading over a larger region with no defined boundary and varying colour (refer to Fig. 4 and Fig. 7 for a visual illustration).

## 6 Conclusions

Motivated by the powerful ability of GP in feature selection and feature construction, we developed a GP based two-stage method for feature selection (stage-1)




(a) (b)

**Fig. 7.** Feature analysis (a) *Malignant*, and (b) *Benign*.

and feature construction (stage-2) for the task of skin cancer image classification. The GP selected and constructed features together have shown powerful ability to help common classification algorithms achieve better performance as compared to using the full set of features. Our method constructed new features from GP selected features, hence using the feature selection ability twice, resulting in more powerful constructed features. Using these GP selected and constructed features, the classification algorithms have shown to provide effective solutions for the real-world cancer detection problem. The results have also shown that colour features have more potential to distinguish between *benign* and *malignant* skin lesions as compared to gray features. We further analysed the GP selected features and GP constructed features to get into the insights of skin cancer properties. It has been found that the LBP patterns can be mapped to skin cancer properties, explaining the contribution of the selected features towards their distinguishing behaviour. In the future, we would like to investigate the effect of employing preprocessing techniques to remove noise from the images. We are also interested to further investigate the classification performance of our method by using a different dataset and also focus on the computation time to make it effective for real-world applications like skin cancer diagnosis.

## References

1. Abuzaghleh, O., Barkana, B.D., Faezipour, M.: Noninvasive real-time automated skin lesion analysis system for melanoma early detection and prevention. IEEE Journal of Translational Engineering in Health and Medicine **3**, 1–12 (2015)
2. Ahmed, S., Zhang, M., Peng, L.: Enhanced feature selection for biomarker discovery in LC-MS data using GP. In: Proceedings of the 2013 IEEE Congress on Evolutionary Computation. pp. 584–591. IEEE (2013)
3. Ain, Q.U., Xue, B., Al-Sahaf, H., Zhang, M.: Genetic programming for skin cancer detection in dermoscopic images. In: Proceedings of the 2017 Congress on Evolutionary Computation. pp. 2420–2427. IEEE (2017)
4. Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., Delfino, M.: Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Archives of Dermatology **134**(12), 1563–1570 (1998)
5. Barata, C., Marques, J.S., Celebi, M.E.: Improving dermoscopy image analysis using color constancy. In: Proceedings of the 2014 IEEE International Conference on Image Processing. pp. 3527–3531. IEEE (2014)

6. Carli, P., de Giorgi, V., Chiarugi, A., Nardini, P., Weinstock, M.A., Crocetti, E., Stante, M., Giannotti, B.: Addition of dermoscopy to conventional naked-eye examination in melanoma screening: a randomized study. Journal of the American Academy of Dermatology **50**(5), 683 – 689 (2004)
7. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
8. Ferris, L.K., Harkes, J.A., Gilbert, B., Winger, D.G., Golubets, K., Akilov, O., Satyanarayanan, M.: Computer-aided classification of melanocytic lesions using dermoscopic images. Journal of the American Academy of Dermatology **73**(5), 769–776 (2015)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations Newsletter **11**(1), 10–18 (2009)
10. Keerthi, S.S., Lin, C.J.: Asymptotic behaviors of support vector machines with gaussian kernel. Neural Computation **15**(7), 1667–1689 (2003)
11. Koza, J.R.: Genetic programming: on the programming of computers by means of natural selection, vol. 1. MIT press (1992)
12. Luke, S.: Essentials of metaheuristics. Lulu, 2nd edn. (2013), [Online] Available: http://cs.gmu.edu/ sean/book/metaheuristics/
13. Marghoob, A.A., Malvehy, J., Braun, R.P.: An atlas of dermoscopy. CRC Press (2012)
14. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R.S., Rozeira, J.: Ph2–a dermoscopic image database for research and benchmarking. In: Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5437–5440. IEEE (2013)
15. Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F.V., Avila, S., Valle, E.: Knowledge transfer for melanoma screening with deep learning. arXiv preprint arXiv:1703.07479 (2017)
16. Mishra, N.K., Celebi, M.E.: An overview of melanoma detection in dermoscopy images using image processing and machine learning. arXiv preprint arXiv:1601.07843 (2016)
17. Neshatian, K., Zhang, M., Andreae, P.: A filter approach to multiple feature construction for symbolic learning classifiers using genetic programming. IEEE Transactions on Evolutionary Computation **16**(5), 645–661 (2012)
18. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51–59 (1996)
19. Patterson, G., Zhang, M.: Fitness functions in genetic programming for classification with unbalanced data. In: Proceedings of the 2007 Australasian Joint Conference on Artificial Intelligence. pp. 769–775. Springer (2007)
20. Stolz, W., Riemann, A., Cognetta, A.B., Pillet, L., Abmayr, W., Holzel, D., Bilek, P., Nachbar, F., Landthaler, M.: ABCD rule of dermatoscopy: a new practical method for early recognition of malignant-melanoma. European Journal of Dermatology **4**(7), 521–527 (1994)
21. Xue, B., Zhang, M., Browne, W.N., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation **20**(4), 606–626 (2016)
22. Zortea, M.e.a.: Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists. Artificial Intelligence in Medicine **60**(1), 13–26 (2014)