Automatically Diagnosing Suspicious Regions in Skin Images Using Two-Stage Genetic Programming

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

Abstract—Developing a computer-aided diagnostic system for detecting various skin malignancies from images has attracted many researchers. Unlike many machine learning approaches such as Artificial Neural Networks, Genetic Programming (GP) automatically evolves models with flexible representation. GP successfully provides effective solutions using its intrinsic ability to select prominent features (i.e. feature selection) and build new features (i.e. feature construction). Existing approaches have utilized GP to construct new features from the complete set of original features and the set of operators. However, the complete set of features may contain redundant or irrelevant features that do not provide useful information for classification. This study aims to develop a two-stage GP method where the first stage selects prominent features, and the second stage constructs new features from these selected features and operators such as multiplication in a wrapper approach to improve the classification performance. To include local, global, texture, color, and multi-scale image properties of skin images, GP selects and constructs features extracted from local binary patterns and pyramid-structured wavelet decomposition. The accuracy of this GP method is assessed using two real-world skin image datasets captured from the standard camera and specialized instruments, and compared with commonly used classification algorithms, three state-of-the-art, and an existing embedded GP method. The results reveal that this new approach of feature selection and feature construction effectively helps improve the performance of the machine learning classification algorithms. Unlike other black-box models, the evolved models by GP are interpretable, therefore the proposed method can assist dermatologists to identify prominent features, which has been shown by further analysis on the evolved models.

Index Terms—Image Classification, Genetic Programming, Skin Cancer Images, Feature Selection, Feature Construction.

I. INTRODUCTION

In the United States, more than 5 million new skin cancer cases are diagnosed every year which makes it a challenging public health problem [1]. Among skin cancers, Melanoma is the most dangerous form, which can be fatal if not detected early [2]. The incidence of skin cancer has been increasing globally with over 104,350 estimated cases including almost 11,650 deaths according to the Cancer Statistics Report in 2019 [1]. While the mortality is quite high, the survival rate of melanoma exceeds 95% when diagnosed in earlier stages [2]. The drastic spike in the prevalence of skin cancer, invasive biopsy tests and immense treatment expenses have rendered its early diagnosis a key public health concern.

For examining a skin lesion, dermatologists commonly follow the ABCD (Asymmetry, Border irregularity, Color

variation and Dermoscopic structure) rule of dermoscopy [3]. This rule calculates a score by measuring these four lesion properties to effectively divide various types of skin cancers [4]. Another regularly utilized clinical methodology is the 7-point check-list strategy (Pigment Network, Streaks, Asymmetry, Regression areas, Dots, Blue-whitish veil, and presence/absence of six colors: black, white, dark-brown, light-brown, blue-gray, red) [5]. Identifying the properties of ABCD rule and 7-point check-list strategy in a skin lesion image requires domain expert knowledge, which can be expensive to employ. These significant visual properties and access to an enormous number of skin images have interested numerous researchers to present computer aided diagnostic (CAD) frameworks that can help dermatologist in early identification.

Skin lesion images come with various artifacts, e.g., gel, reflection, and hair, which make feature extraction process very difficult. Moreover, automatic skin cancer image classification is a very difficult task due to a number of factors such as the different location of lesion in an image, the immense intra-class variations of melanomas, and the large inter-class similarity between different kinds of skin cancers [6]. Along these lines, it is necessary to define strategies that can capture and/or construct informative features, which are, by one way or another, fit to imitate these clinical properties and, henceforth, utilize different local, global, texture, and color features.

Feature extraction can be applied to sub-images to extract local features while its application to the whole image extracts global features [7]. Such diagnostic systems or classification methods are potentially useful, which can accurately classify a particular skin cancer in actual circumstances [8]. In addition to classifying correctly, identifying significant features concurrently which can show critical visual patterns to a dermatologist is another key contribution. Moreover, features directly extracted from images may not have discriminating information about images of different classes for accurate classification. In other words, the presence of various artifacts often leads to redundant or irrelevant extracted features which may lead to poor classification performance. Two feature manipulation techniques, *feature selection* and *feature construction*, can be employed in such cases which help to pick important features and construct new informative features provided the original set of features, respectively, to improve classification accuracy [9]–[11]. Feature extraction, on the other hand, transforms the original set of features into a reduced representation set [12]. Most recently, convolutional neural networks (CNNs) have

gained immense popularity in dermoscopy image analysis. Codella et al. [13] performed feature extraction from skin cancer images by utilizing the Caffe architecture. Esteva et al. [14] tried to produce a performance as good as a human expert by training an Inception network on a huge private dataset with both dermoscopy and clinical images. However, with limited size of the available medical datasets, it is usually infeasible to train a CNN effectively from scratch [15].

Genetic programming (GP) is an evolutionary computation method which solves a particular problem at hand by automatically evolving computer programs/solutions (often represented as trees) [16]. GP utilizes genetic operations such as crossover, mutation, and reproduction on a current population of solutions to produce a new population of solutions [17]. The success of GP relies on its algorithmic characteristics: 1) no explicit assumption about the problem, 2) flexible to combine with existing approaches to get benefit from the best features of different methods, 3) robust by using a population-based search mechanism and randomized options, 4) makes GP less likely to get trapped in sub-optimal solutions, and 5) capable of providing unpredictable solutions which humans cannot presume useful for design domains [18]. Moreover, in GP, there is less demand for sample data compared to some other deep neural network based approach. As not all features are necessary for classification, GP uses its built-in filtering ability to select the important features at its leaf nodes (terminals), which makes GP an effective method for feature selection. These selected features usually have better ability to distinguish images of different classes, which greatly help achieve performance gains. A program evolved by GP can be considered as a newly constructed feature (CF) that can help improve the classification accuracy; hence, GP is an effective feature construction method. In addition to classification, GP has been explored extensively for feature selection and construction [9]–[11]. In image analysis, a wide range of applications have utilized GP, including object detection [19], feature extraction [12], feature construction [11], and classification [20], [21]. Recently, a GP method [11] has been developed using texture features for skin cancer detection from dermoscopy images. This method can only solve binary classification problem and its goodness has been evaluated on a single dataset. Though the method was fast being an embedded approach, it cannot achieve good classification performance.

Given the evaluation criteria, feature selection algorithms can be classified into three categories: wrapper, filter, and embedded approaches. While a wrapper approach incorporates a learning (classification) method in evaluating the feature subset, a filter approach does not utilize any classification method [22]. An embedded approach integrates classifier learning and feature selection into a solitary procedure [22].

Unlike the existing classification methods that produced viable outcomes for a single image modality, the proposed method aims to work well for skin images captured from both the standard camera and specialized instruments. Existing approaches automatically construct new informative features from the complete set of original features. However, new features built from selected features have not been investigated. Performing feature selection first to identify prominent

features and constructing new features from these selected features has the potential to improve the classification performance.

A. Contributions

This study develops a new two-stage GP system for skin image classification in a wrapper approach (2SGP-W). The aim of the first stage of GP is to filter out the redundant or irrelevant features and pick only prominent features with high discriminating ability between classes. The aim of the second stage of GP is to perform classification in a wrapper approach using only the features selected in stage-1.

The main contributions of this study are as follows:

- We design a new two-stage GP method for feature selection and construction in skin cancer detection from images which provides better classification performance compared to the existing methods.
- Unlike existing approaches which either include color or texture features, the proposed method includes various texture, color, and frequency-based features to increase the classification performance. The proposed method achieves much better performance than four state-of-the-art skin cancer classification methods and six widely used machine learning classification algorithms on two real-world skin image datasets captured from different optical devices.
- The proposed method utilizes very little training time and can predict a class label to a test image in fractions of a second, which is efficient for real-world problems such as skin cancer detection.

II. LITERATURE REVIEW

A. Background

1) *Local Binary Pattern (LBP)*: Ojala et al. [23] developed an image descriptor for feature extraction in computer vision applications that has been extensively researched. LBP uses a sliding window having a fixed radius to scan an image from top to bottom and left to right in a pixel-by-pixel fashion. It assigns the value to the central pixel according to the intensity values of the adjacent pixels situated on the radius as shown in Fig. 1. With the computed values, an LBP histogram (feature vector) is generated.

LBP are classified into two patterns; uniform and non-uniform. A non-uniform pattern consists of two or more bitwise transitions circularly from 0 to 1 or 1 to 0. For example, the code (01011100) shown in Fig. 1 is non-uniform. On the other hand, a uniform pattern consists of at most one such bitwise transition. For example, the codes 00011110, 01111000, and 10000000 are uniform. The size of the feature vector is 2^b where b is the number of adjacent pixels. This size can be further reduced to $b(b-1) + 3$ bins if only the uniform patterns are considered. All the non-uniform patterns are combined into one bin.

In skin lesions, uniform patterns help detect corners (lesion boundary), dots (flat regions), and line ends (streaks), which can help in distinguishing different types of skin cancers. In this work, two sets of LBP features are extracted:

Fig. 1. The process of extracting LBP features from a skin image by creating a LBP histogram.

- 59 uniform LBP features extracted from gray-scale skin images, termed as L_G features.
- Uniform LBP features extracted from red (R), green (G) and blue (B) color channels of the skin image, termed as L_R features. From each color channel, 59 uniform LBP features are extracted. Hence, these three features vectors are concatenated together to form a single feature vector with 177 features.

2) *Wavelet Features*: Texture analysis can reflect the visual characteristics of a skin lesion, which forms the basis of clinical diagnosis (e.g. ABCD rule of dermoscopy) [24]. The pyramid-structured wavelet analysis [25] provides internal structure and detailed texture characteristics (local features), as well as overall properties (global features) of the skin lesion. Three-level pyramid-structured wavelet decomposition is used to extract the frequency-based features from four color channels; luminance, red, green, and blue. The luminance color channel is calculated as:

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

Eight statistical measures and ratios are extracted from the wavelet coefficients. These measures are mathematically represented in Table I, where i is an index of wavelet tree nodes (n), X_i is a $J_i \times K_i$ matrix of the i^{th} node, X_i' is its transpose, x_{jk} is the jk^{th} element, and $eig(X_i)$ are the eigenvalues. J and K are dimensions (resolution) of the matrices (images) over which wavelet decomposition is applied. These statistical measures are extracted first from the original image and are further divided by a factor of two at each decomposition level.

Fig. 2(a) shows a skin lesion image and Fig. 2(b) shows its pyramid-structured wavelet decomposition. Fig. 2(c) displays the three levels in a wavelet tree structure where ovals represent nodes. The wavelet tree consists of a total of 13 nodes where the top parent node is the original image, and 4 nodes in each of the three subsequent decomposition levels ($3 \times 4 = 12$) are the wavelet coefficients. On each tree node, the eight statistical measures as given in Table I, are applied to give $8 \times 13 = 104$ features. Since we computed these features on four color channels, the total number of wavelet features become 416 ($= 8 \text{ measures} \times 13 \text{ nodes} \times 4 \text{ color channels}$) in this study.

TABLE I
STATISTICAL MEASURES APPLIED TO THE WAVELET COEFFICIENTS [24].

Measure	Formula
Energy	$E_{n_i} = \frac{\sum_{j=1}^J \sum_{k=1}^K x_{jk}^2}{J \times K}$
Kurtosis	$K_{n_i} = \frac{\sum_{j=1}^J \sum_{k=1}^K \left(\frac{x_{jk} - M(n_i)}{Std(n_i)} \right)^4}{J \times K}$
Mean	$M_{n_i} = \frac{\sum_{j=1}^J \sum_{k=1}^K x_{jk}}{J \times K}$
Norm	$N_{n_i} = \max \left(\sqrt{eig(X_i \times X_i')} \right)$
Standard Deviation	$Std_{n_i} = \sqrt{\frac{\sum_{j=1}^J \sum_{k=1}^K (x_{jk} - M_{n_i})^2}{J \times K}}$
Entropy	$H_{n_i} = \frac{\sum_{j=1}^J \sum_{k=1}^K \left(x_{jk} \times \log(x_{jk}) \right)}{J \times K}$
Skewness	$S_{n_i} = \frac{\sum_{j=1}^J \sum_{k=1}^K \left(\frac{x_{jk} - M(n_i)}{Std(n_i)} \right)^3}{J \times K}$
Average Energy	$AvgE_{n_i} = \frac{\sum_{j=1}^J \sum_{k=1}^K x_{jk} }{J \times K}$

B. Related Work

Recently, the ensembles of CNNs have been utilized for classifying skin images, providing promising results. Harangi et al. [26] developed an ensemble of VGGNet, GoogLeNet, and AlexNet to classify skin cancer images, and prove that the ensemble-based approach has provided better results than all the three of its member CNNs. Valle et al. [27] combined the concepts of transfer learning and ensembles of CNNs. Their results show that their method with an ensemble of CNNs model outperformed existing methods with unstable sequential designs. Xie et al. [28] developed a melanoma detection methods using artificial neural network (ANN) based ensemble model. The method achieved high classification accuracy mainly due to the new border features employed and the proposed ANN architecture. However, training a model effectively with these deep learning approaches generally requires a huge number of images. Moreover, deep learning approaches have a “black-box” model, which hinders the insights of prominent features.

To solve the multi-class classification problem of skin images, a hierarchical classification approach has been adopted by many researchers. Ballerini et al. [29] designed a hierarchical k -Nearest Neighbors (k -NN)-based model for non-melanoma classification from standard camera images (non-dermoscopy). This system relied on expert knowledge as it required hand-crafted texture and color features which is usually difficult to extract when dealing with large image datasets. Shimizu et al. [30] also used a hierarchical system and extracted several color, texture, and sub-region features to classify four skin cancer classes. The hierarchical structures in [29], [30] produced a better performance compared to the standard non-hierarchical classification algorithms. Recently, Barata et al. [31] performed a hierarchical diagnosis for skin cancer multi-class classification using a pre-trained DenseNet-161 architecture. They also investigated the significance of color normalization and lesion segmentation. In general, the use of a pre-trained CNN requires pre-processing of a dataset to the same input settings for which CNN was originally developed, such as fixed-size images and RGB or gray-scale images, which increases the computation time and reduces the

Fig. 2. Three-level pyramid-structured wavelet decomposition (shown in (b)) on a skin image (shown in (a)) with a schematic three-level wavelet tree with in oval (shown in (c)).

versatility of using any image size. In addition, decreasing the size of a skin image will eventually distort the aspect ratio, resulting in the loss of informative features.

Garnavi et al. [24] developed a melanoma detection method by employing various border, geometrical, and texture features. This method utilized a filter approach (gain-ratio-based feature selection) to generate an optimal feature set. Successfully applying feature selection, this diagnostic system produced an overall accuracy of 91.26%. However, the method lacks an appropriate way of using various types of features concurrently. Recently, Alfed et al. [32] proposed a bag-of-features approach with new texture and color features for melanoma detection. The authors successfully demonstrated the effectiveness of histogram of gradients (HoG) and histogram of lines features in skin cancer detection using dermoscopy and standard skin image datasets.

Kawahara et al. [33] performed 10-class classification using the Dermofit dataset by employing filters from a pre-trained CNN. Using a standard overall classification accuracy for a highly imbalanced dataset may produce biased results towards the majority class, which is a limitation of their work. From the confusion matrix shown in [33], the overall accuracy is 81.80%, whereas the balanced accuracy is 60.12% for the 10-class classification problem. Most researchers have used overall accuracy until the International Skin Imaging Collaboration (ISIC) 2018¹ challenge started collecting balanced accuracy along with other evaluation measures.

Li et al. [34] developed a deep learning based lesion indexing network (LIN) to detect and classify skin cancer from images. Their method extracted informative features which resulted in good overall performance. The authors concluded

that with better segmentation, their method could achieve better results. Hasan et al. [35] developed a CNN based method for skin cancer detection utilizing various feature extracting techniques to extract features from dermoscopic images. They achieved 89.5% accuracy on test data, which needed improvement. Moreover, the classification model experienced overfitting which is a limitation of their method.

Tschandl et al. [36] developed a CNN based skin cancer detection method and checked its performance on pigmented melanocytic lesions from dermoscopic images. However, their method could not achieve sufficient accuracy. They found that distinguishing between non-melanocytic and non-pigmented skin cancers is a difficult task. Saba et al. [8] developed a three-step deep CNN based skin cancer detection method. The first step performs color transformation to enhance contrast. The second step uses CNN to extract lesion boundaries. The last step uses transfer learning to extract deep features. Their method provided good results on only a small dataset, but could not achieve good performance on other datasets.

Jafari et al. [37] developed a CNN based model to detect melanoma from skin cancer images. Their method incorporated pre-processing and post-processing images before and after segmentation, respectively. Their method used both local and global contextual information concurrently to segment the lesion regions. Their method achieved very good performance but did not mention computation time which is important in real-world cancer detection problems. Le et al. [38] developed a transfer learning model, termed ResNet50, for skin cancer image classification. Their method did not use any pre-processing steps or handcrafted feature selection. The authors concluded that results can be improved by using pre-processing steps and appropriate feature extraction and selection methods before classification. Behzad et al. [39] utilized deep convolution networks to develop a skin lesion segmentation method using an unsupervised learning approach. Their method produced a global map for skin lesions with effective segmentation results. Ali et al. [40] compared two deep learning approaches; a supervised and an unsupervised, for the task of skin lesion segmentation. Their results showed that the supervised approach performed better than the unsupervised approach in terms of dice coefficient and jaccard index. However, the segmentation results shown by supervised approach missed the lesion area where the skin images include many artifacts.

In the literature, GP has been extensively utilized for image analysis [7], [19], [41]–[43]. In addition to image analysis, GP has successfully achieved promising results in scheduling, classification, and symbolic regression [44]. Earlier in 2003, Zhang et al. proposed an object detection method using GP [19]. Their method is capable of locating multiple objects in large images and predicting class label of each detected object. The results evaluated on three datasets of varying difficulty demonstrated the ability of the evolved GP program to object detection and multi-class classification. Ryan et al. [42] proposed a Stage-1 breast cancer detection procedure using GP. The procedure identifies suspicious malignant regions by employing multiple stages including pre-processing, segmentation of breast region and feature extraction. Results

¹<https://challenge2018.isic-archive.com/>

showed that the solutions provided by GP for this difficult cancer detection task are human-readable.

Al-Sahaf et al. [7] developed a new GP-based image descriptor for texture image classification. The method automatically generates a feature vector without any human intervention. Experiments on nine image descriptors and seven image datasets depicted the effectiveness of their multi-class classification method. Later, the algorithm in [7] was utilized to perform transfer learning by Iqbal et al. [43], to deal with even more difficult texture image classification tasks. This transfer learning approach has the ability to solve complex tasks that most other algorithms remain unable to tackle, as shown by the results. Most recently, Bi et al. [12] proposed a GP method to learn novel features automatically, and simultaneously evolve an ensemble for image classification. This method uses commonly used classification algorithms and image-related operators such as Gabor filter, Laplacian filter, LBP, and HoG, to evolve ensembles of classifiers for classification. This method has provided promising results on several image datasets. However, the generated models formed by various classifiers are complex and challenging to interpret.

For the problem of skin image classification, a GP-based binary classification method was designed which combined biomedical (domain-specific) and LBP (domain-independent) features to achieve good results [45]. Later, they utilized feature selection and feature construction abilities of GP using local and global features for melanoma detection in a binary classification problem [11]. They developed a binary classification method in [46] by employing a multi-tree GP in an embedded approach for melanoma detection. They further developed a multi-class classification method in [15] using a wrapper approach to effectively discriminate even 10 classes of skin cancer images. With human-readable GP evolved models, they identified prominent skin image features that are potentially helpful to a dermatologist to identify particular visual patterns and effectively diagnose skin cancers in real-world circumstances.

Most of the existing approaches have tested the goodness of their method(s) by using single-source images which are captured from one optical device. In real-world settings, however, images are captured from several instruments and, thus, these techniques can not be extended to other images captured from different instruments or may work poorly. Accordingly, there is a need for classification methods for skin cancer images with: 1) the ability to provide good classification results without using expert knowledge, 2) sufficient information regarding local, global, color and texture properties required to achieve good classification performance, 3) the ability to be applied to multi-source images, 4) the ability to be automatically generated without the need of setting a huge number of parameters, 5) the ability to take images of different sizes as input, and 6) the ability to easily interpret and to identify prominent features necessary to guide the dermatologist well enough to classify different skin cancers.

III. THE PROPOSED METHOD

The proposed two-stage wrapper-based GP method (2SGP-W) is described in this section. Figs. 3 and 4 present the

overall structure of the training and testing/evaluation processes, respectively. The method starts by converting the image datasets to feature vectors by applying a feature extraction method as described in Section II-A. During stage-1, GP takes these features as input. In this work, GP utilizes its traditional tree-like representation where an individual consists of one tree. GP has the intrinsic ability to select features during the evolutionary process. GP usually picks the most prominent features at its leaf nodes, as not all features can provide good between images of different cancer types. GP creates a tree with prominent features using genetic operators such as crossover, mutation, and elitism. We expect that the selected features have high discriminating ability between classes. A classification algorithm, such as a decision tree (J48), takes these prominent features as input for classification. GP is run for multiple (10) times to get the best evolved tree. To this end, we get a GP tree whose selected features when provided to the classification algorithm, have achieved the highest classification performance among all the GP runs. Stage-1 ends here, and the features showing up in the best individual (evolved tree) with the best classification results 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 and classification. Here, again GP is run for multiple (30) times in a wrapper approach using the selected features (i.e. the output from stage-1) only. It is expected that evolving a tree from the selected features has more potential as compared to evolving a tree from the original set of features. This is because the original set of features consists of both relevant features with good discriminating ability, and irrelevant or redundant features with least distinguishing ability, so stage-1 tries to get rid of those irrelevant features and pick the prominent relevant features. Though stage-1 can still select some irrelevant features, but most of them are not selected. Hence, in stage-2, we have more relevant features in the feature set which results in good constructed feature. The highest performing tree evolved on the training data is selected which is considered as a single constructed feature. Using the selected features (computed in stage-1) and the constructed feature (computed in stage-2) are concatenated together to form a final feature vector. This feature vector is given to a classification algorithm such as J48 (the same algorithm as in stage-1) for classification. The main aim of 2SGP-W is to improve the feature subset selection (stage-1) and the constructed feature (stage-2) during the evolutionary process while providing good classification performance by a machine learning classification algorithm such as a J48.

For illustration, let us take the example of the 59 LBP (L_G) features. We provide a complete set of 59 features to the proposed method. In stage-1, GP uses these 59 features to evolve a tree. In its evolved tree, GP selects important features, let us assume GP selects 15 features in one run. These 15 selected features are provided to a machine learning classification algorithm such as J48 to perform classification. This is the process in stage-1 and is repeated 10 times to get 10 different classification accuracies. Since in each GP run, GP starts evolving its population with a different seed,

Fig. 3. Training process of the proposed 2SGP-W method.

Fig. 4. Test process of the proposed 2SGP-W method.

thus, we get different evolved populations in each run. From each run, we select the highest performing tree as the best GP tree. Since the proposed method uses 10 GP runs, we get 10 best trees at the end of stage 1. Among these 10 trees, the tree with the highest performance is picked. For example, that this best tree consists of 13 LBP features at its leaf nodes. These selected features are used to make the feature subset for stage-2. In stage-2, GP evolves a tree based on these 13 selected features only, and not the whole set of 59 features. The number of GP runs in stage-2 is 30. In one GP run, for example, GP selects only 8 features at its leaf nodes from the given set of 13 features. Also, the GP tree with itself is a constructed feature. The value of this constructed feature is computed using the values of the selected features at the leaf nodes and the mathematical operators selected in the GP tree. At this point, a feature vector is formed using these 8 selected features and the one constructed feature. This newly formed feature vector is provided to the same classification algorithm such as J48 to get classification accuracy. GP runs for 30 times (each time using 13 features) and gets 30 accuracies, which are averaged to get the final classification accuracy.

To classify a test image, the methodology is shown in Fig. 4. A test image is transformed to a feature vector using a feature extraction method described in Section II-A. We utilize the best trees from stage-1 and stage-2 to create a feature vector of GP-selected and GP-constructed features. This feature vector is provided to a classification algorithm to predict the class label.

A. Fitness Function

The overall standard classification accuracy is defined as the number of correctly classified images divided by the total number of images in a dataset. Using this accuracy as a fitness is not suitable when there is a class imbalance problem. Class imbalance refers to different number of images in different classes in a dataset. This accuracy may lead to results biased towards the majority class. In such a class imbalance scenario, using balanced accuracy as a fitness function is appropriate which is defined as

$$fitness = \frac{1}{z} \sum_{i=1}^z \frac{TP_i}{TP_i + FN_i} \quad (2)$$

where z represents the total number of classes, TP represents the true positives, and FN represents the false negatives. The ratio $\frac{TP_i}{TP_i + FN_i}$ shows the true positive rate of class i . From the Equation 2, balanced accuracy takes into account the accuracies of all classes in a dataset.

B. Terminal Set

The terminal set consists of three types of features derived from the feature extraction methods mentioned in Section II-A. The details of these features are as follows;

- 1) L_G : Gray-level skin images are used to extract 59 LBP features as described in Section II-A1, following the procedure shown in Fig. 1.
- 2) L_R : From each of the red, green, and blue color channels, 59 LBP features are extracted which are concatenated in a single feature vector with $177 (= 3 \text{ channels } 59 \times \text{LBP features})$ L_R features.

- 3) Wavelet: The local and global properties are included by using three-level pyramid-structured wavelet decomposition as described in Section II-A2. Eight statistical measures defined in Table I are extracted from each node of the wavelet decomposition to get a total of 416 wavelet features.

The value of the i^{th} feature for the above three types of features is indicated by F_i , as shown by the GP individuals in Figs. 12 and 13. For the L_G and L_R features, a window size of 3×3 pixels and a radius of 1 pixel ($LBP_{8,1}$) is adopted, which are the fundamental and widely used settings for extracting LBP features.

C. Function Set

The function set consists of the following three kinds of operators;

- 1) Arithmetic: $\{+, -, \times, /\}$, where addition, subtraction, and multiplication have the original arithmetic meaning, whereas division is protected which means when a number is divided by zero it returns zero.
- 2) Trigonometric: $\{\sin, \cos\}$
- 3) Conditional: $\{if\}$ operator takes four input values. If the first value is greater than the second value, it returns the third value; else it returns the fourth value.

IV. EXPERIMENT DESIGN

This section discusses the design of the experiments. It covers the details of the datasets, the benchmark techniques for comparison, the experiments, and the parameter settings.

A. Datasets

The proposed 2SGP-W method is evaluated on two skin image datasets of varying difficulty. Details of these datasets are given in Table II. The PH² dataset is publicly available², whereas the Dermofit dataset is not³. The datasets vary in terms of the number of classes, the number of images, the size of images and the optical device with which the images in a dataset are captured.

1) PH²: A dataset of specialized dermoscopic images namely PH² [47] is acquired from Pedro Hispano Hospital Portugal. Dermoscopy involves using an efficient illumination system and an optical tool to view skin lesions at a higher magnification. A liquid gel/solution is placed on the lesion before capturing the image, which allows the dermatoscope (device) to obtain morphological patterns in the inner layers of human skin. These images are thus informative enough to examine them for skin cancer detection. The images are 8-bit RGB (red, green, and blue) images.

The dataset includes 8-bit RGB (red, green, and blue) images of skin lesions, their binary masks and their clinical diagnosis. The dataset consists of three types of skin lesion images: common nevi, atypical nevi, and melanoma. For multi-class classification experiments, PH² has these three

TABLE II
REAL-WORLD SKIN CANCER DATASETS.

Name	Classes	#Instances	Image size
PH ²	Melanoma	40	$764 \times 576 - 768 \times 576$
	Common Nevi	80	$763 \times 553 - 769 \times 577$
	Atypical Nevi	80	$764 \times 575 - 768 \times 576$
Dermofit	Melanoma	76	$367 \times 439 - 3055 \times 1630$
	Melanocytic Nevus / Mole	331	$177 \times 189 - 857 \times 828$
	Squamous Cell Carcinoma (SCC)	88	$269 \times 273 - 1341 \times 1097$
	Basal Cell Carcinoma (BCC)	239	$189 \times 206 - 1341 \times 1130$
	Intraepithelial Carcinoma (IC)	78	$565 \times 265 - 2176 \times 2549$
	Actinic Keratosis (AK)	45	$193 \times 221 - 777 \times 702$
	Seborrheic Keratosis (SK)	257	$189 \times 229 - 1825 \times 1329$
	Pyogenic Granuloma (PG)	24	$292 \times 235 - 1870 \times 1834$
	Dermatofibroma (Df)	65	$436 \times 338 - 1498 \times 1492$
	Haemangioma (Hg)	96	$328 \times 193 - 914 \times 890$

classes. Samples of the three categories of skin lesions are presented in Fig. 5(a). In dermatology, there are some non-malignant (benign) moles which may have chances to develop malignancy over time. Such moles are grouped in atypical nevi class. For binary classification task, the atypical nevi and benign classes are combined together to make a single benign class against the melanoma class.

2) Dermofit Image Library: The University of Edinburgh provides a standard camera image library of 1300 high-quality skin lesions [29]. The images are captured under standardized conditions. The dataset is divided into 10 classes based on opinions from expert dermatologists and dermatopathologists. An image sample from each of the ten skin cancer types in this dataset is shown in Fig. 5(b). The details of the dataset including class names, number of images, and range of image sizes in each class have been detailed in Table II.

For evaluating the binary classification experiments, two classes are used; Melanocytic Nevus (mole) as *benign*, and Malignant Melanoma as *malignant*. For multi-class classification experiments, dermofit has ten classes.

B. Benchmark Techniques

The proposed 2SGP-W method is compared to six commonly used classification methods to check its effectiveness: Naïve Bayes (NB), SVM with a Radial Basis Function (RBF) kernel, k -NN where $k = 5$, J48, Random Forest (RF), and Multilayer Perceptron (MLP). The widely applied Waikato Environment for Knowledge Analysis (WEKA) package version 3.8 [48] is utilized to implement these methods. In k -NN, k is set to 5 to avoid noisy instances while still being efficient. For RF, the maximum depth of a tree and the number of trees are set to 5 and 10, respectively. In MLP, the momentum, training epochs, learning rate and the number of units in one hidden layer are set to 0.2, 60, 0.1, and 20, respectively. These parameters are taken from an earlier work [15] where they are empirically defined, since they produce the best results amongst other settings.

Moreover, we also compare 2SGP-W with the recently developed state-of-the-art methods for the PH² and Dermofit datasets:

- Patino et al. [49] developed a multi-class classification method using the PH² dataset where different morphological operations are designed to encompass asymmetry,

²<https://www.fc.up.pt/addi/ph2%20database.html>

³<https://licensing.edinburgh-innovations.ed.ac.uk/i/software/dermofit-image-library.html>

(a) PH²: each row represents images from one class.

(b) Dermofit: each image belongs to one class.

Fig. 5. Image samples from the two datasets.

color and border features. Three classification methods are used; SVM, logistic regression and a fully connected neural network. An average accuracy of 86.5% is achieved by the neural network while outperforming the other two methods.

- Ain et al. [11] developed a 2Stage-GP method in an embedded approach using LBP gray and color features. They have evaluated the performance of their method on the PH² dataset only. Their method can only solve binary classification problem and provided 78.17% balanced classification accuracy.
- Kawahara et al. [33] extracted features from a pre-trained CNN which are provided to a logistic regression classifier using Dermofit dataset to classify its 10 classes. Their method achieved 81.80% overall accuracy, which equals 60.12% balanced accuracy calculated from the confusion matrix provided.
- Fisher et al. [50] developed a hierarchical decision tree, where a different k -NN is trained for each decision node. 2500+ features are extracted using generalized co-occurrence texture matrices and lesion specific characteristics. On the Dermofit dataset, a standard overall accuracy of 78.10% and a balanced accuracy of 70.50% has been achieved.

C. Experiments

For performing the experiments, 10-fold cross validation is used using random stratified sampling. This is because PH² dataset is very small (200 images) and some classes in Dermofit have very small number of images (Pyogenic Granuloma with 24 images). The dataset is divided into ten folds where training uses nine folds and the remaining one fold is used for the testing process. For all the different

TABLE III
PARAMETER SETTINGS OF THE PROPOSED 2SGP-W 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

combinations of folds, this cycle is repeated ten times and the results are recorded as the mean of the fitness values.

During stage-1 and stage-2, GP is executed for 10 and 30 times, respectively. After stage-1, among the 10 runs, the best tree with highest performance on the training data is used to create a feature vector of GP-selected features. Using these GP-selected features, GP is executed 30 times during stage-2. It is worth mentioning here that at both stages the test folds remain unseen to prevent feature selection and feature construction biases. In each set of experiments, the random seeds for the 30 runs are all different. The GP method is implemented using the Evolutionary Computing Java-based (ECJ) package [51].

D. Parameter Settings

The parameter settings of our proposed 2SGP-W method are listed in Table III. In both stages, the evolutionary process stops when the classification algorithm such as a J48 achieves 100% accuracy or a maximum of 50 generations is reached.

V. RESULTS AND DISCUSSIONS

A. Overall Results

The results of the proposed method are presented in Tables IV and V for binary and multi-class classification problems, respectively. Vertically, these tables comprise of three blocks which corresponds to the results of using L_G , L_R and wavelet features, respectively. Horizontally, the table consists of 5 columns where first lists the classification algorithm, second and third show respectively the test performances on the PH² and the Dermofit datasets using all features, represented by “All”. Similarly, the rest of the columns show test performances using the proposed 2SGP-W method on the two datasets. The values of the results using all features is the mean of applying 10-fold cross validation to the dataset. Since 2SGP-W is repeated 30 times, hence, we get 30 accuracies for each classifier which are represented as mean and standard deviation ($\bar{x} \pm s$) in Tables IV and V.

To get a clear comparison between using different methods, the results are also tested using *One-sample t-test*. It is applied to compare 2SGP-W to the other deterministic methods. This statistical test has been applied to the test results to check which method has better ability to discriminate between different classes of skin cancers. The symbols “ $\uparrow$ ”, “ $\downarrow$ ” and “=” are used to represent significantly better, significantly worse and not significantly different performance, respectively, of the 2SGP-W compared to all features. For example, on the PH² dataset, in Table IV, the test performance of SVM with 2SGP-W using L_R features is represented as “89.84 $\pm$ 1.41 $\uparrow$ ”

TABLE IV
RESULTS OF *Binary Classification*: THE ACCURACY (%) ON THE TEST SET USING ALL FEATURES, AND 2SGP-W (RESULTS ARE REPRESENTED IN TERMS OF MEAN ACCURACY AND STANDARD DEVIATION ($\bar{x} \pm s$)).

		PH ²		Dermofit	
		All	2SGP-W	All	2SGP-W
Average number of features					
		59	24.43	59	34.73
L _G	NB	63.44	79.22 $\pm$ 1.88 $\uparrow$	60.39	72.32 $\pm$ 0.86 $\uparrow$
	SVM	70.94	87.03 $\pm$ 0.78 $\uparrow$	55.95	76.22 $\pm$ 0.78 $\uparrow$
	k-NN	70.62	79.53 $\pm$ 1.31 $\uparrow$	60.99	69.67 $\pm$ 1.96 $\uparrow$
	J48	61.56	81.87 $\pm$ 2.34 $\uparrow$	62.93	71.61 $\pm$ 1.20 $\uparrow$
	RF	62.81	90.62 $\pm$ 2.04 $\uparrow$	57.03	75.83 $\pm$ 2.53 $\uparrow$
	MLP	67.81	79.53 $\pm$ 1.16 $\uparrow$	58.85	74.56 $\pm$ 1.84 $\uparrow$
Average number of features					
		177	35.95	177	40.64
L _R	NB	76.25	84.54 $\pm$ 0.16 $\uparrow$	66.37	81.92 $\pm$ 0.58 $\uparrow$
	SVM	75.00	89.84 $\pm$ 1.41 $\uparrow$	54.77	83.17 $\pm$ 1.37 $\uparrow$
	k-NN	74.02	81.88 $\pm$ 0.63 $\uparrow$	61.39	77.81 $\pm$ 2.64 $\uparrow$
	J48	73.13	87.19 $\pm$ 2.32 $\uparrow$	58.43	85.60 $\pm$ 1.78 $\uparrow$
	RF	75.94	93.65 $\pm$ 0.83 $\uparrow$	55.01	86.23 $\pm$ 1.59 $\uparrow$
	MLP	76.88	83.13 $\pm$ 0.63 $\uparrow$	61.82	78.38 $\pm$ 2.78 $\uparrow$
Average number of features					
		416	36.57	416	42.08
Wavelet	NB	75.00	86.88 $\pm$ 1.37 $\uparrow$	96.77	99.11 $\pm$ 0.34 $\uparrow$
	SVM	70.63	97.50 $\pm$ 1.41 $\uparrow$	84.61	99.08 $\pm$ 1.26 $\uparrow$
	k-NN	75.63	67.18 $\pm$ 4.23 $\uparrow$	87.97	87.14 $\pm$ 1.50 $\uparrow$
	J48	68.78	91.88 $\pm$ 2.32 $\uparrow$	93.46	98.29 $\pm$ 2.36 $\uparrow$
	RF	73.44	92.18 $\pm$ 2.80 $\uparrow$	91.63	99.34 $\pm$ 1.22 $\uparrow$
	MLP	78.75	71.25 $\pm$ 3.42 $\uparrow$	96.27	92.14 $\pm$ 3.79 $\uparrow$

where the “ $\uparrow$ ” sign represents that 2SGP-W significantly outperformed using all features.

B. Dimensionality Reduction

Analyzing the effect of dimensionality reduction achieved by the 2SGP-W method, it has been seen that in case of the PH² dataset, GP selects (on average 24) even less than the half of the total 59 L_G features in its tree with a tree depth of 6 as shown in Table IV. Here the number of features is 24.43 computed as the average number of features appeared in the 30 GP runs during stage-1. The number of features have significantly reduced in case of L_R features (from 177 to around 35.57) and wavelet features (from 416 to around 38.16). A similar trend in dimensionality reduction has been observed in the 2SGP-W evolved programs for multi-class classification. Using L_G, L_R, and wavelet features, the average number of selected features are 30.36, 39.66, and 40.70, reduced from a total of 59, 177, and 416 features, respectively. In the proposed 2SGP-W results, all the classification algorithms have achieved better performance compared to using “All” features in non-GP classification algorithms. This has clearly demonstrated that GP has pushed most of the classification algorithms to achieve good performance with its feature selection and construction ability even with a reduced number of features.

C. Binary Classification

Table IV shows the results of binary classification, or in other words identifying melanoma from benign images. The

results show that 2SGP-W has provided much better results than the non-GP classification algorithms which use the whole set of features. The proposed method provides the highest results on the PH² dataset using *wavelet* features with SVM reaching 97.50% average accuracy. On the Dermofit dataset, wavelet features remain prominent in providing the highest average accuracy, i.e., 99.34% with RF. This shows that wavelet features which capture both local and global properties of skin images have the most potential in distinguishing melanoma from benign images. Since, these wavelet features are extracted from multiple color channels, they provide color information as well. These characteristics of wavelet features make them more informative compared to the LBP features which encompass only local texture information. The result of the statistical tests show that 2SGP-W (with an “ $\uparrow$ ” sign in Table IV) has significantly outperformed all the commonly used classification algorithms on both datasets.

D. Multi-class Classification

Table V shows the results of multi-class classification. The non-GP methods using all features have achieved 71.50% and 64.46% highest accuracy with MLP on PH² and Dermofit datasets, respectively. It is evident that 2SGP-W has been effective in providing much better results compared to the non-GP classification algorithms using all features. Among the two datasets, 2SGP-W provides good results on the PH² dataset with three classes (relatively easy task) achieving 97.24% average accuracy with SVM. Similarly, for the difficult task of distinguishing between ten types of skin cancers in Dermofit dataset, the performance is also very good. Here, RF achieved the highest test performance using wavelet features producing 85.67% average accuracy. The result of the statistical tests show that 2SGP-W (with an “ $\uparrow$ ” sign in Table V) has significantly outperformed all the commonly used classification algorithms on both datasets.

E. Comparison with the State-of-the-Arts

For PH², the most recent state-of-the-art has been developed by Patino et al. [49] which has achieved 86.5% balanced accuracy in the multi-class classification problem using 10-fold cross validation. Ain et al. [11] developed a two stage GP method for melanoma detection in a binary classification task. This method [11] is tested on only the PH² dataset using 10-fold cross validation and produced a balanced accuracy of 78.17%. Since the experimental setups in both methods [11], [49] are the same as our proposed 2SGP-W method, we can make a direct comparison. In binary classification, 2SGP-W with 97.50% performance outperformed the first method [49] by providing an increase of nearly 10% accuracy. In multi-class classification, 2SGP-W with 97.24% average accuracy outperformed the second method [11] with an improvement of nearly 19% accuracy.

To the best of our knowledge, Kawahara et al. [33] provide the state-of-the-art result on Dermofit using CNNs. Their method achieved an overall accuracy of 81.80% with an experiments set of 5-fold cross validation. However, according to the confusion matrix given in the study, this overall accuracy

TABLE V
RESULTS OF *Multi-class Classification*: THE ACCURACY (%) ON THE TEST SET USING ALL FEATURES, AND 2SGP-W (RESULTS ARE REPRESENTED IN TERMS OF MEAN ACCURACY AND STANDARD DEVIATION ($\bar{x} \pm s$)).

		PH ²		Dermofit	
		All	2SGP-W	All	2SGP-W
Average number of features					
		59	28.33	59	31.52
L _G	NB	51.00	82.87 $\pm$ 4.58 $\uparrow$	24.77	55.12 $\pm$ 3.16 $\uparrow$
	SVM	57.50	77.34 $\pm$ 3.42 $\uparrow$	38.69	63.33 $\pm$ 3.81 $\uparrow$
	k-NN	31.62	79.60 $\pm$ 2.65 $\uparrow$	53.50	48.25 $\pm$ 5.36 $\uparrow$
	J48	47.00	74.02 $\pm$ 1.04 $\uparrow$	22.38	58.50 $\pm$ 2.64 $\uparrow$
	RF	54.50	83.78 $\pm$ 2.22 $\uparrow$	33.62	59.44 $\pm$ 5.12 $\uparrow$
	MLP	48.50	75.46 $\pm$ 3.92 $\uparrow$	39.31	60.20 $\pm$ 4.22 $\uparrow$
Average number of features					
		177	34.68	177	44.26
L _R	NB	55.00	82.33 $\pm$ 1.46 $\uparrow$	25.15	53.27 $\pm$ 2.87 $\uparrow$
	SVM	60.00	80.44 $\pm$ 2.80 $\uparrow$	42.38	49.33 $\pm$ 5.56 $\uparrow$
	k-NN	60.00	81.36 $\pm$ 1.25 $\uparrow$	33.08	51.22 $\pm$ 4.75 $\uparrow$
	J48	47.50	84.02 $\pm$ 2.45 $\uparrow$	27.23	56.71 $\pm$ 3.52 $\uparrow$
	RF	59.50	86.97 $\pm$ 2.02 $\uparrow$	34.85	69.62 $\pm$ 4.62 $\uparrow$
	MLP	61.00	79.68 $\pm$ 3.70 $\uparrow$	44.38	62.30 $\pm$ 3.31 $\uparrow$
Average number of features					
		416	39.18	416	43.55
Wavelet	NB	69.50	93.66 $\pm$ 2.35 $\uparrow$	46.85	84.97 $\pm$ 3.64 $\uparrow$
	SVM	67.00	97.24 $\pm$ 2.93 $\uparrow$	59.92	85.43 $\pm$ 3.75 $\uparrow$
	k-NN	68.00	86.87 $\pm$ 1.87 $\uparrow$	53.15	75.29 $\pm$ 3.51 $\uparrow$
	J48	61.00	96.97 $\pm$ 2.02 $\uparrow$	45.85	83.88 $\pm$ 2.58 $\uparrow$
	RF	69.00	96.38 $\pm$ 0.58 $\uparrow$	55.92	85.67 $\pm$ 4.62 $\uparrow$
	MLP	71.50	95.00 $\pm$ 2.57 $\uparrow$	64.46	72.30 $\pm$ 3.31 $\uparrow$

equals 60.12% balanced accuracy. Recently, Fisher et al. [50] used the Dermofit dataset to test the performance of their hierarchical tree approach to classify skin cancer images. The authors reported a balanced accuracy of 70.50% using leave-one out cross validation. Since comparison cannot be done directly with these two state-of-the-arts [33], [50] (5-folds vs 10-folds, and leave-one out vs 10-folds), here we try to give a general estimate of accuracy achieved by the current state-of-the-arts on Dermofit dataset.

VI. FURTHER ANALYSIS

A. Overall Analysis

To explore the effectiveness of employing two stages instead of following the traditional approach of employing one stage, we have further analyzed the evolutionary process of stage-1 and stage-2 as depicted in Figs. 6 to 9. These plots are generated for both the binary and multi-class classification experiments using L_R features. Though there are 50 generations in both stages but for comparison purposes, here we have shown stage-1 executed till 100 generations (51st generation to 100th generation is, therefore, shown with dotted line). This is to examine the effect of having the second stage, i.e. whether running stage 1 for 100 generations can achieve better performance of the proposed 2-stage method with 50 generations in each stage (and stage-2 uses the features selected in stage-1 as the input). In other words, is stage-2 really necessary or needed? By doing so, we would like to see the difference in training performance among the 51st to 100th generations in stage-1 and the 1st to 50th generations

in stage-2. To make this obvious from the graphs, we have plotted stage-2 from 51 generation onwards on x-axis.

From the plots in Figs. 6 to 9, a general trend can be observed; in the start of the evolutionary process, GP tries to explore the search space and makes larger jumps, regardless of whether it has been provided with all the original features or only the selected features. To get a clear understanding of how stage-2 is effective, we observe the stage-2 starts from a higher average accuracy most of the time as compared to the average accuracy of 51st generation in stage-1. For example, in Fig. 6(a) on the PH² dataset with NB as a wrapper classification algorithm, stage-2 starts at 86.84% average accuracy (shown in red color), whereas stage-1 at its 51st generation reaches 84.52% average accuracy. This trend is not always true. In a few cases, stage-2 with the selected features starts with a lower average accuracy compared to the stage-1. Such an example is given in Fig. 6(b) with SVM as a wrapper classification algorithm. However, whether stage-2 starts with a lower or a higher average accuracy compared to stage-1, it always provides better average accuracy at the end of the evolutionary cycle. We can clearly see this in Fig. 6(b), where stage-2 starts with 82.16% average accuracy, cuts the stage-1 line at 85.78%, and keeps improving after that by making larger jumps to end at a better average performance of 93.46% compared to stage-1 ending at 87.95%. Hence, we can say that the selected features have the potential to push GP make bigger jumps and help the classification algorithm learn better to achieve good training performance.

B. Computation time

The average training time required for 2SGP-W to execute the two stages and to test their performances on the test data in the binary and multi-class classification tasks using wavelet features is depicted in Figs. 10, and 11, respectively. Various factors affect the amount of time it takes to train a classification algorithm such as 1) how big is a dataset? 2) which feature selection approach (filter or wrapper) is adopted? and 3) how many features are used to evolve an individual? While the proposed method seems expensive to implement with two GP stages and a wrapper method, creating a solution will not take longer than 18 minutes on average.

In Fig. 10, NB takes the minimum time to train a model among the six wrapper binary classification algorithms for binary classification on both datasets. Overall, RF remains prominent being the fastest and highest-performing wrapper classification algorithm on the PH² and Dermofit datasets. RF spends 2.78 and 4.17 hours, respectively, to evolve good individuals during stage-1 and stage-2. With these trained evolved individuals, testing an unseen skin image after converting it to an original set of feature vector takes only on average 0.60 and 8.00 milliseconds. Therefore, we may conclude here that the proposed 2SGP-W method for binary classification is very efficient for identifying melanoma in real-time clinical circumstances. It can allow dermatologists to determine whether or not a biopsy is needed in skin image diagnosis.

For multi-class classification, Figs. 11(a) and (b) show that the computation time increases many folds while training a

Fig. 6. Convergence plots for the **PH²** dataset in the *binary classification* task where blue line represents stage-1 and red line represents stage-2.

Fig. 7. Convergence plots for the **Dermofit** dataset in the *binary classification* task where green line represents stage-1 and pink line represents stage-2.

Fig. 8. Convergence plots for the **PH²** dataset in the *multi-class classification* task.

Fig. 9. Convergence plots for the **Dermofit** dataset in the *multi-class classification* task.

large (such as Dermofit) dataset with 1300 images as compared to training a small (such as PH²) dataset with 200 images. A comparison of Figs. 10 and 11 illustrates that binary classification tasks take less training time as compared to multi-class classification tasks. Similar to the test time in binary classification, using these trained evolved individuals during stage-1 and stage-2, an unseen skin image can be tested in fractions of a second as shown in Fig. 11(b).

C. Evolved GP individuals

In addition to feature selection and feature construction, the intuition behind using GP is its ability to evolve interpretable solutions. This section describes the two good GP individuals evolved during stage-1 and stage-2 as shown in Figs. 12 and 13. They have been taken from Dermofit experiments using RF in a multi-class classification task. From the evolved GP individual shown in Fig. 12, GP has selected only 11 features among a total of 416 wavelet features during stage-1, effectively reducing dimensionality many folds. Using these

Fig. 10. The average computation time for *binary classification* using 2SGP-W method on the datasets.

eleven features as input to stage-2, GP further selects most prominent features to build an informative GP constructed feature as shown in Fig. 13. These selected (eleven features from stage-1) and constructed (one feature in stage-2) features are provided to RF, where RF keeps improving the classification performance during the evolutionary cycle.

Fig. 11. The average computation time for *multi-class classification* using 2SGP-W method on the datasets.

TABLE VI

THE DETAILS OF THE WAVELET FEATURES SELECTED IN THE GP INDIVIDUAL SHOWN IN FIG. 12 IN TERMS OF STATISTICAL MEASURE, COLOR CHANNEL, LEVEL OF DECOMPOSITION AND NODE.

Feature	Measure	Channel	Level	node
W ₅	Norm	Red	0	—
W ₄₆	Entropy	Red	1	3
W ₆₄	Energy	Red	2	2
W ₉₅	Average Energy	Red	2	1
W ₁₂₅	Norm	Red	3	1
W ₁₃₇	Mean	Green	1	3
W ₁₄₂	Entropy	Green	1	3
W ₂₃₄	Standard Deviation	Blue	1	3
W ₂₅₈	Standard Deviation	Blue	2	2
W ₂₈₈	Energy	Blue	3	2
W ₃₀₉	Norm	Blue	3	4

The wavelet features selected by the GP individual shown in Fig. 12 are listed in Table VI. We draw the following conclusions from this table:

- Three out of the eleven features belong to the third-level nodes, reflecting our use of up till three levels of wavelet decomposition. This illustrates that further decomposition does not have informative features for classification purposes;
- Texture features obtained from red, green, and blue color channels are more informative and are chosen to construct this individual;
- The node column in Table VI shows that the selected features are extracted from the low and the middle frequency channels; and
- Norm, entropy, energy and standard deviation are prominent selected features among the eight statistical measures.

Moreover, the sub-trees “(F₄₆–F₁₄₂)” and “if (F₉₅, F₆₄, F₃₀₉, F₂₃₄)” appear four and nine times, respectively, which show the potential of these sub-trees selected multiple times to generate this GP tree.

D. Constructed Feature

To see why the selected and constructed features can achieve good performance, we take an example of a constructed feature shown in Fig. 13 evolved in a GP run during stage-2 on the Dermofit dataset in the multi-class classification task. This tree is constructed from eight features. The values of this constructed feature are plotted in Fig. 14 where the ten boxes represent the ten classes in the Dermofit dataset. It can be seen

that the range of values for one class differs from the range of values for the other classes. For example, the values of the SCC class range between 0.11 and 0.52, whereas the values of the PG class range between 0.67 and 1.0. Similarly, more than 50% values of the SCC class differ from the Melanoma, Mole, BCC, IC, SK, and DF classes. On the other hand, the values of the SCC class overlap completely on the AK and Hg classes. This shows that this constructed feature alone cannot distinguish SCC instances from all the other classes; both the selected and constructed features are needed. The results show that GP not only selects important features but also constructs informative features with high discriminating ability.

E. Feature Appearance in Evolved GP Individuals

GP selects relevant features from the original set of features to automatically construct new informative features. In this work, this built-in ability of GP to feature selection has also been investigated on the Dermofit experiments for the binary and multi-class classification tasks as shown in Figs. 15 and 16. The bars show the frequency of occurrence of each feature in the evolved tree after stage-1 among the 30 GP runs. From these plots, it is obvious that certain features are more frequently selected than the other features. For example, in Figs. 15(a) and 16(a), the last feature F₅₈ in L_G features appear almost thrice compared to the other 58 features in both bar plots. Similarly, F₁₀₂ and F₅₈ appeared the most among the L_R, and wavelet features in the binary classification task as given in Figs. 15(b) and (c), respectively. In multi-class classification task, F₁₂₂ and F₂₄₃ appeared the most among the L_R, and wavelet features as given in Figs. 16(b) and (c), respectively.

We have dug deeper into the texture, color, local and global properties of these important and most frequently occurring wavelet features; F₅₆, F₂₄₃, and F₂₄₈ in binary classification, and F₄₈, F₂₄₃ and F₃₄₆ in multi-class classification. We have observed that both F₅₆, and F₂₄₈ are the energy measures of the red and the blue channels, respectively. F₂₄₃ represents standard deviation calculated from level 1 wavelet coefficients in the blue color channel which is a significant feature selected most of the times in both the binary and multi-class classification tasks. F₄₈, and F₃₄₆ represent energy and standard deviation calculated from level 1 wavelet features in the red and luminance color channels. The selection of these features confirms that statistical measures such as energy and standard deviation capture most of the lesion properties at level 1 of the wavelet coefficients which are useful to differentiate between different skin cancers.

VII. CONCLUSIONS

In this paper, a novel two-stage GP method in a wrapper approach has been developed for skin cancer image classification. The proposed method aims at including both color and texture information from skin images in order to provide effective discrimination between images of different classes. The features are extracted using LBP and pyramid structured wavelet decomposition from various color channels of skin

[illegible]

(a) L_G (b) L_R (c) Wavelet

Box plot showing the distribution of the number of genes with expression fold change greater than 2 for various cancer types. The y-axis represents the number of genes (0 to 1.0). The x-axis lists cancer types: Melanoma, Mole, SCC, BCC, IC, AK, SK, PG, Df, and Hg. Each box plot shows the median (red line), interquartile range (blue box), and range (dashed whiskers). Outliers are marked with red dots.

Cancer Type	Min	Q1	Median	Q3	Max	Outliers
Melanoma	0.31	0.35	0.42	0.47	0.50	0.21, 0.65
Mole	0.10	0.30	0.52	0.75	1.00	
SCC	0.10	0.18	0.25	0.31	0.40	0.52
BCC	0.40	0.52	0.68	0.78	0.92	0.18
IC	0.32	0.35	0.50	0.60	0.83	0.05, 0.88
AK	0.05	0.08	0.22	0.38	0.60	0.25, 0.70
SK	0.62	0.62	0.75	0.87	1.00	0.25, 0.35
PG	0.82	0.85	0.90	0.95	0.98	0.68
Df	0.45	0.58	0.68	0.80	0.88	0.35
Hg	0.05	0.15	0.32	0.45	0.58	0.75

Figure 1 consists of three bar charts labeled (a), (b), and (c), each showing the frequency distribution of features for a different model. The y-axis for all charts is 'Frequency' and the x-axis is 'Feature'.

- (a) L_G : The y-axis ranges from 0 to 200, and the x-axis ranges from 0 to 40. The bars are red. The frequency is relatively stable, mostly between 50 and 100, with a slight increase towards the end of the feature range.
- (b) L_R : The y-axis ranges from 0 to 60, and the x-axis ranges from 0 to 150. The bars are blue. The frequency is highly variable, with many peaks reaching between 40 and 60.
- (c) Wavelet: The y-axis ranges from 0 to 40, and the x-axis ranges from 0 to 400. The bars are black. The frequency is also highly variable, with several prominent peaks reaching between 30 and 40.

lesion images. These domain independent features have sufficient local pixel-based RGB and global color variation information and help achieve good results without expert intervention. The proposed two-stage GP method is utilized to perform feature selection during stage-1 and feature construction from the selected features only during stage-2. The selected and constructed features are provided to a classification algorithm such as J48 which has produced effective results for skin cancer binary and multi-class image classification.

increased the classification performance. The selected features by GP have high discriminating ability compared to the original set of features, which is evident from the convergence plots for stage-1 and stage-2. In other words, the classification performance of selected and constructed features is always better compared to the classification performance of the original set of features. We have also found that wavelet features with their detailed internal structure and color information remain prominent in providing the highest classification accuracy for both binary and multi-class classification on both datasets.

Although the proposed method has provided effective and efficient solutions to the complex problem of skin cancer image classification, there are some limitations which can be addressed in the future. The proposed method has not used any domain knowledge, but GP has the ability to include domain knowledge as well. This will be explored in the future to possibly improve performance with using both the domain independent and domain-specific knowledge. Real-world images often come with a lot of noise, which hinders accurate image classification. In skin images, various kinds of noises are present in the form of skin hairs, dark corners due to camera calibrations, bubbles due to the layer of gel, ink and ruler markers, and color map to estimate the lesion diameter. We will improve the proposed method to handle these issues.

REFERENCES

- [1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: A Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.
- [2] N. H. Matthews, W. Q. Li, A. A. Qureshi, M. A. Weinstock, and E. Cho, "Epidemiology of melanoma," in *Cutaneous Melanoma: Etiology and Therapy [Internet]*. Codon Publications, 2017.
- [3] W. Stolz, A. Riemann, A. B. Cagnetta, L. Pillet, W. Abmayr, D. Holzel, P. Bilek, F. Nachbar, and M. Landthaler, "ABCD rule of dermatoscopy: a new practical method for early recognition of malignant-melanoma," *European Journal of Dermatology*, vol. 4, no. 7, pp. 521–527, 1994.
- [4] R. Kasmi and K. Mokrani, "Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule," *IET Image Processing*, vol. 10, no. 6, pp. 448–455, 2016.
- [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," *Archives of 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 Transactions on Medical Imaging*, vol. 36, no. 4, pp. 994–1004, 2017.
- [7] H. Al-Sahaf, A. Al-Sahaf, B. Xue, M. Johnston, and M. Zhang, "Automatically evolving rotation-invariant texture image descriptors by genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 1, pp. 83–101, 2017.
- [8] T. Saba, M. A. Khan, A. Rehman, and S. L. Marie-Sainte, "Region extraction and classification of skin cancer: A heterogeneous framework of deep cnn features fusion and reduction," *Journal of medical systems*, vol. 43, no. 9, pp. 1–19, 2019.
- [9] S. Ahmed, M. Zhang, L. Peng, and B. Xue, "Multiple feature construction for effective biomarker identification and classification using genetic programming," in *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*. ACM, 2014, pp. 249–256.
- [10] B. Tran, B. Xue, and M. Zhang, "Genetic programming for feature construction and selection in classification on high-dimensional data," *Memetic Computing*, vol. 8, no. 1, pp. 3–15, 2015.
- [11] Q. U. Ain, B. Xue, H. Al-Sahaf, and M. Zhang, "Genetic programming for feature selection and feature construction in skin cancer image classification," in *Proceedings of the 15th Pacific Rim International Conference on Artificial Intelligence*, ser. Lecture Notes in Computer Science. Springer, 2018, pp. 732–745.
- [12] Y. Bi, B. Xue, and M. Zhang, "Genetic programming with a new representation to automatically learn features and evolve ensembles for image classification," *IEEE Transactions on Cybernetics*, 2020.
- [13] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," in *Proceedings of the International Workshop on Machine Learning in Medical Imaging*. Springer, 2015, pp. 118–126.
- [14] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [15] Q. U. Ain, H. Al-Sahaf, B. Xue, and M. Zhang, "Generating knowledge-guided discriminative features using genetic programming for melanoma detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2020.
- [16] J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection*. MIT press, 1992, vol. 1.
- [17] J. R. Koza and R. Poli, "A genetic programming tutorial," 2003.
- [18] A. E. Eiben and J. Smith, "From evolutionary computation to the evolution of things," *Nature*, vol. 521, no. 7553, p. 476, 2015.
- [19] M. Zhang, V. B. Ciesielski, and P. Andreae, "A domain-independent window approach to multiclass object detection using genetic programming," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 8, p. 206791, 2003.
- [20] W. Choi and T. Choi, "Computer-aided detection of pulmonary nodules using genetic programming," in *Proceedings of the 2010 IEEE International Conference on Image Processing*. IEEE, 2010, pp. 4353–4356.
- [21] W. A. Tackett, "Genetic programming for feature discovery and image discrimination," in *Proceedings of the 5th International Conference on Genetic Algorithms*, 1993, pp. 303–311.
- [22] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2016.
- [23] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.
- [24] R. Garnavi, M. Aldeen, and J. Bailey, "Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1239–1252, 2012.
- [25] T. Chang and C.-C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Transactions on Image Processing*, vol. 2, no. 4, pp. 429–441, 1993.
- [26] B. Harangi, A. Baran, and A. Hajdu, "Classification of skin lesions using an ensemble of deep neural networks," in *Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2018, pp. 2575–2578.
- [27] E. Valle, M. Fornaciari, A. Menegola, J. Tavares, F. V. Bittencourt, L. T. Li, and S. Avila, "Data, depth, and design: Learning reliable models for melanoma screening," *arXiv preprint arXiv:1711.00441*, 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 Transactions on Medical Imaging*, vol. 36, no. 3, pp. 849–858, 2017.
- [29] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical k-NN approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*. Springer, 2013, pp. 63–86.
- [30] K. Shimizu, H. Iyatomi, M. E. Celebi, K.-A. Norton, and M. Tanaka, "Four-class classification of skin lesions with task decomposition strategy," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 274–283, 2015.
- [31] C. Barata and J. S. Marques, "Deep learning for skin cancer diagnosis with hierarchical architectures," in *Proceedings of the International Symposium on Biomedical Imaging*, vol. 2. IEEE, 2019.
- [32] N. Alfed and F. Khelifi, "Bagged textural and color features for melanoma skin cancer detection in dermoscopic and standard images," *Expert Systems with Applications*, vol. 90, pp. 101–110, 2017.
- [33] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *Proceedings of the 13th International Symposium on Biomedical Imaging*. IEEE, 2016, pp. 1397–1400.
- [34] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *Sensors*, vol. 18, no. 2, p. 556, 2018.
- [35] M. Hasan, S. D. Barman, S. Islam, and A. W. Reza, "Skin cancer detection using convolutional neural network," in *Proceedings of the 5th International Conference on Computing and Artificial Intelligence*, 2019, pp. 254–258.
- [36] P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, A. Blum, R. P. Braun, H. Cabo, J.-Y. Gourhant, J. Kreusch, A. Lallas *et al.*, "Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks," *JAMA dermatology*, vol. 155, no. 1, pp. 58–65, 2019.
- [37] M. H. Jafari, N. Karimi, E. Nasr-Esfahani, S. Samavi, S. M. R. Soroushmehr, K. Ward, and K. Najarian, "Skin lesion segmentation in clinical images using deep learning," in *Proceedings of the 23rd International conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 337–342.
- [38] D. N. Le, H. X. Le, L. T. Ngo, and H. T. Ngo, "Transfer learning with class-weighted and focal loss function for automatic skin cancer classification," *arXiv preprint arXiv:2009.05977*, 2020.
- [39] B. Bozorgtabar, S. Sedai, P. K. Roy, and R. Garnavi, "Skin lesion segmentation using deep convolution networks guided by local unsupervised learning," *IBM Journal of Research and Development*, vol. 61, no. 4/5, pp. 6–1, 2017.
- [40] A.-R. Ali, J. Li, and T. Trappenberg, "Supervised versus unsupervised deep learning based methods for skin lesion segmentation in dermoscopy images," in *Canadian Conference on Artificial Intelligence*. Springer, 2019, pp. 373–379.
- [41] R. Poli, "Genetic programming for image analysis," in *Proceedings of the 1st Annual Conference on Genetic Programming*. MIT Press, 1996, pp. 363–368.
- [42] C. Ryan, K. Krawiec, U. O'Reilly, J. Fitzgerald, and D. Medernach, "Building a stage 1 computer aided detector for breast cancer using genetic programming," in *Proceedings of the European Conference on Genetic Programming*. Springer, 2014, pp. 162–173.
- [43] M. Iqbal, B. Xue, H. Al-Sahaf, and M. Zhang, "Cross-domain reuse of extracted knowledge in genetic programming for image classification," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 4, pp. 569–587, 2017.

- [44] H. Al-Sahaf, Y. Bi, Q. Chen, A. Lensen, Y. Mei, Y. Sun, B. Tran, B. Xue, and M. Zhang, "A survey on evolutionary machine learning," *Journal of the Royal Society of New Zealand*, vol. 49, no. 2, pp. 205–228, 2019.
- [45] Q. U. Ain, B. Xue, H. Al-sahaf, and M. Zhang, "Genetic programming for skin cancer detection in dermoscopic images," in *Proceedings of the 2017 Congress on Evolutionary Computation*, 2017, pp. 2420–2427.
- [46] —, "A multi-tree genetic programming representation for melanoma detection using local and global features," in *Proceedings of the 31st Australasian Joint Conference on Artificial Intelligence. Lecture Notes in Computer Science*. Springer, 2018, pp. 111–123.
- [47] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "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*. IEEE, 2013, pp. 5437–5440.
- [48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *Special Interest Group on Knowledge Discovery and Data Mining Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
- [49] D. Patiño, A. M. Ceballos-Arroyo, J. A. Rodríguez-Rodríguez, G. Sanchez-Torres, and J. W. Branch-Bedoya, "Melanoma detection on dermoscopic images using superpixels segmentation and shape-based features," in *Proceedings of the 15th International Symposium on Medical Information Processing and Analysis*, vol. 11330. International Society for Optics and Photonics, 2020, p. 1133018.
- [50] R. B. Fisher, J. Rees, and A. Bertrand, "Classification of ten skin lesion classes: Hierarchical knn versus deep net," in *Annual Conference on Medical Image Understanding and Analysis*. Springer, 2019, pp. 86–98.
- [51] S. Luke, *Essentials of metaheuristics*, 2nd ed. Lulu, 2013, [Online] Available: <http://cs.gmu.edu/~sean/book/metaheuristics/>.