Automatically Diagnosing Skin Cancers From Multimodality Images Using Two-Stage Genetic Programming

Qurrat Ul Ain, *Member, IEEE*, Harith Al-Sahaf, *Member, IEEE*, Bing Xue, *Senior 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 multiscale 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*—Feature construction, feature selection, genetic programming (GP), image classification, skin cancer images.**

Manuscript received November 2, 2021; revised March 6, 2022 and May 1, 2022; accepted June 4, 2022. This work was supported in part by the Marsden Fund of New Zealand Government under Contract VUW1913 and Contract VUW1914; in part by the Science for Technological Innovation Challenge (SfTI) Fund under Grant E3603/2903; in part by the University Research Fund at Victoria University of Wellington under Grant 223805/3986; in part by the MBIE Data Science SSIF Fund under Contract RTVU1914; in part by the National Natural Science Foundation of China (NSFC) under Grant 61876169; in part by the Huayin Medical under Grant E3791/4165; and in part by the MBIE Endeavor Research Programme under Contract C11X2001. This article was recommended by Associate Editor L. Jiao. *(Corresponding author: Qurrat Ul Ain.)*

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

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCYB.2022.3182474.

Digital Object Identifier 10.1109/TCYB.2022.3182474

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 has rendered its early diagnosis a key public health concern.

For examining a skin lesion, dermatologists commonly follow the asymmetry, border irregularity, color variation, and dermoscopic structure (ABCD) 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 checklist strategy (pigment network, streaks, asymmetry, regression areas, dots, blue-whitish veil, and presence/absence of six colors: 1) black; 2) white; 3) dark-brown; 4) light-brown; 5) blue-gray; and 6) red) [5]. Identifying the properties of the ABCD rule and 7-point checklist 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 the dermatologist in early identification.

Skin lesion images come with various artifacts, for example, gel, reflection, and hair, which make the 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 intraclass variations of melanomas, and the large interclass 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, 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 subimages to extract local features while its application to the entire 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: 1) *feature selection* and 2) *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], [10]. Feature extraction, on the other hand, transforms the original set of features into a reduced representation set [11]. Most recently, convolutional neural networks (CNNs) have gained immense popularity in dermoscopy image analysis. Codella *et al.* [12] performed feature extraction from skin cancer images by utilizing the Caffe architecture. Esteva *et al.* [13] 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 the limited size of the available medical datasets, it is usually infeasible to train a CNN effectively from scratch [14].

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) [15]. GP utilizes genetic operations, such as crossover, mutation, and reproduction, on a current population of solutions to produce a new population of solutions. The success of GP relies on its algorithmic characteristics: 1) no explicit assumption about the problem; 2) flexible to combine with the existing approaches to obtain 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 suboptimal solutions; and 5) capable of providing unpredictable solutions which humans cannot presume useful for design domains [16]. 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], [10]. In image analysis, a wide range of applications have utilized GP, including object detection [17]; feature extraction [11]; feature construction [10]; and classification [18], [19]. Recently, a GP method [10] 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: 1) wrapper; 2) filter; and 3) 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 [20]. An embedded approach integrates classifier learning and feature selection into a solitary procedure [20].

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 accuracy.

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.

1) 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.
2) Unlike existing approaches which either include color or texture features, the proposed method includes 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 taken from different optical devices.
3) 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. Feature Extraction

1) Local Binary Pattern: Ojala *et al.* [21] developed an image descriptor for feature extraction in computer vision applications that has been extensively researched. Local binary pattern (LBP) uses a sliding window having a fixed radius

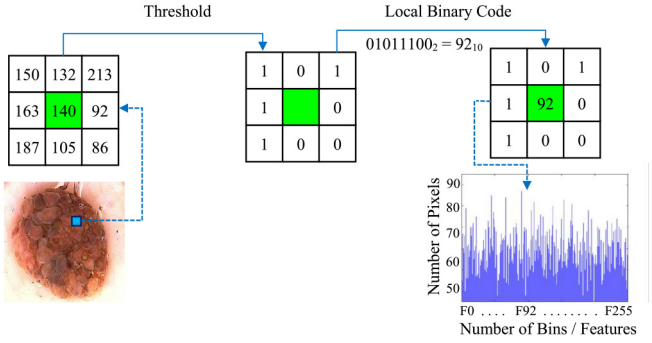

Fig. 1. Process of extracting LBP features from a skin image by creating an LBP histogram.

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: 1) uniform and 2) nonuniform. A nonuniform 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 nonuniform. 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 nonuniform 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.

1) Fifty nine uniform LBP features extracted from grayscale skin images, called L_G features.
2) Uniform LBP features extracted from red (R), green (G), and blue (B) color channels of an image, called L_R features. From each channel, 59 uniform LBP features are extracted. Hence, concatenating the three feature vectors forms 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) [22]. The pyramid-structured wavelet analysis [23] 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

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

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 ith node, X_i' is its transpose, x_{jk} is the jkth 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.

TABLE I
STATISTICAL MEASURES APPLIED TO THE WAVELET COEFFICIENTS [22]

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	$H_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\left(x_{jk}^2 \times log(x_{jk}^2)\right)}{J \times K}$
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	$AvgE_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\lvert x_{jk}\rvert}{J \times K}$

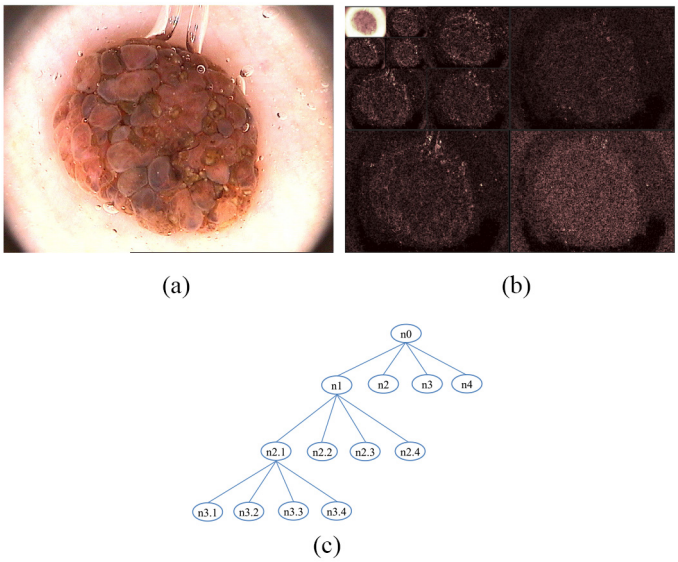

Fig. 2. Three-level pyramid-structured wavelet decomposition [shown in (b)] on a skin image [shown in (a)] with a schematic three-level wavelet tree with in oval [shown in (c)].

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 four 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 becomes 416 (= 8 measures × 13 nodes × 4 color channels).

B. Related Work

Recently, the ensembles of CNNs have been utilized for classifying skin images, providing promising results. Harangi *et al.* [24] developed an ensemble of VGGNet, GoogLeNet, and AlexNet to classify skin cancer images, and proved that the ensemble-based approach has provided better results than all the three of its member CNNs. Valle *et al.* [25] combined the concepts of transfer learning and ensembles of CNNs. Their results show that their method with an ensemble of CNNs model outperformed the existing methods with unstable sequential designs. Xie *et al.* [26] developed a melanoma detection methods using the 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 multiclass classification problem of skin images, a hierarchical classification approach has been adopted by many researchers. Ballerini *et al.* [27] designed a hierarchical k-nearest neighbors (k-NN)-based model for nonmelanoma classification from standard camera images (nondermoscopy). This system relied on expert knowledge as it required handcrafted texture and color features which is usually difficult to extract when dealing with large image datasets. Shimizu *et al.* [28] also used a hierarchical system and extracted several color, texture, and subregion features to classify four skin cancer classes. The hierarchical structures in [27] and [28] produced a better performance compared to the standard nonhierarchical classification algorithms. Barata and Marques [29] performed a hierarchical diagnosis for skin cancer multiclass classification using a pretrained DenseNet-161 architecture. They also investigated the significance of color normalization and lesion segmentation. Recently, Mahajan *et al.* [30] developed a skin disease classification method using few-shot learning strategies based on metalearning. Their results demonstrated the effectiveness of using group equivariant convolutions to improve disease classification. However, the experimental setup requires the images to be resized to a fixed resolution irrespective of their original sizes, which may lead to loss of texture information, and biased results in datasets with varied image sizes. In general, the use of a pretrained CNN requires preprocessing of a dataset to the same input settings for which CNN was originally developed, such as fixed-size images and RGB or grayscale images, which increases the computational time and reduces the 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.* [22] 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 and Khelifi [31] 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.* [32] performed 10-class classification using the Dermofit dataset by employing filters from a pretrained CNN. Using a standard overall classification accuracy for a highly imbalanced dataset may produce biased results toward the majority class, which is a limitation of their work. From the confusion matrix shown in [32], the overall accuracy is 81.80%, whereas the balanced accuracy is 60.12% for the 10-class classification problem. Most researchers have used the overall accuracy until the International Skin Imaging Collaboration (ISIC) 2018[1] challenge started collecting balanced accuracy along with other evaluation measures.

Li and Shen [33] 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.* [34] 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.* [35] 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 nonmelanocytic and nonpigmented 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.* [36] developed a CNN-based model to detect melanoma from skin cancer images. Their method incorporated preprocessing and postprocessing 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 computational time, which is important in real-world cancer detection problems. Le *et al.* [37] developed a transfer learning model, called ResNet50, for skin cancer image classification. Their method did not use any preprocessing steps or handcrafted feature selection. The authors concluded that results can be improved by using preprocessing steps and appropriate feature extraction and selection methods before classification. Bozorgtabar *et al.* [38] utilized

[1] https://challenge2018.isic-archive.com/

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.* [39] compared two deep learning approaches: 1) a supervised and 2) 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 the 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], [17], [40]–[42]. In addition to image analysis, GP has successfully achieved promising results in scheduling, classification [43], and symbolic regression [44]. Earlier, Zhang *et al.* [17] proposed an object detection method using GP. 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 multiclass classification. Ryan *et al.* [41] proposed a Stage-1 breast cancer detection procedure using GP. The procedure identifies suspicious malignant regions by employing multiple stages, including preprocessing, segmentation of breast region, and feature extraction. Results 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 multiclass classification method. Later, the algorithm in [7] was utilized to perform transfer learning by Iqbal *et al.* [42], 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.* [11] 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 the 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 construction abilities of GP using local and global features for melanoma detection in a binary classification problem [10]. They developed a binary classification method in [46] by employing a multitree GP in an embedded approach for melanoma detection. They further developed a multiclass classification method in [14] using a wrapper approach to discriminate even ten 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 multisource 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. Proposed Method

The proposed 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 obtain 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.

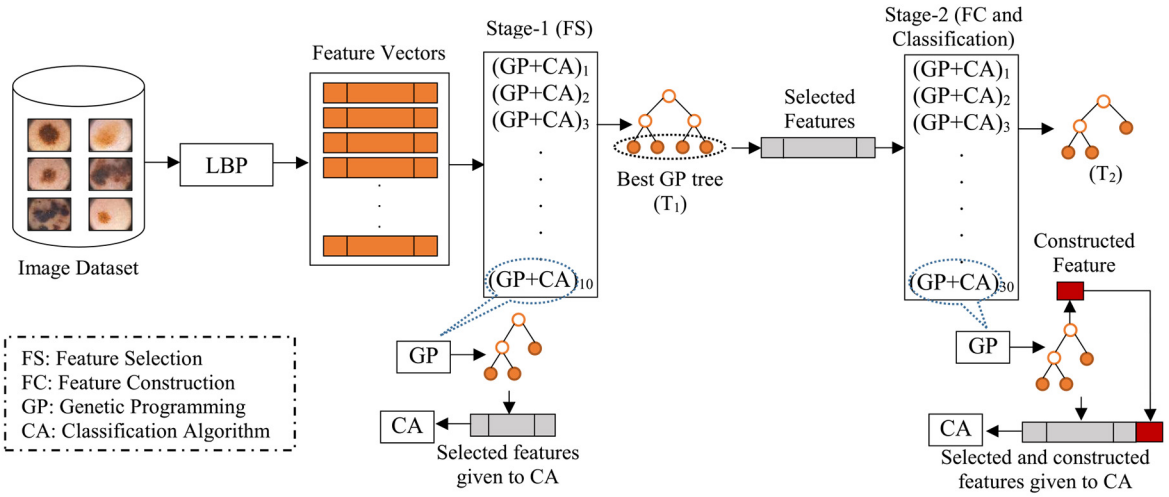

Fig. 3. Training process of the proposed 2SGP-W method.

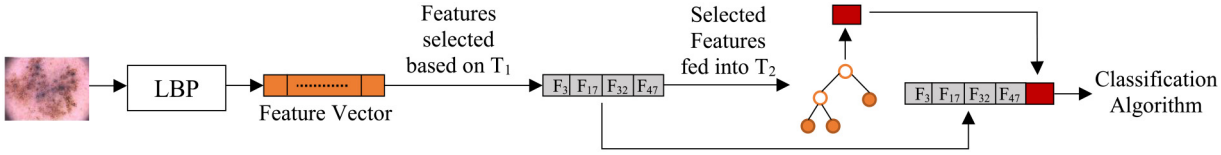

Fig. 4. Test process of the proposed 2SGP-W method.

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 CF. The highest performing tree evolved on the training data is selected which is considered as a single CF. The selected features (computed in stage-1) and the CF (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 CF (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 ten times to obtain ten different classification accuracies. Since in each GP run, GP starts evolving its population with a different seed; thus, we obtain 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 ten GP runs, we obtain ten best trees at the end of stage 1. Among these ten trees, the tree with the highest performance is picked. For example, 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 entire set of 59 features. The number of GP runs in stage-2 is 30. In one GP run, for example, GP selects only eight features at its leaf nodes from the given set of 13 features. Also, the GP tree with itself is a CF. The value of this CF 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 eight selected features and the one CF. This newly formed feature vector is provided to the same classification algorithm such as J48 to obtain classification accuracy. GP runs for 30 times (each time using 13 features) and obtains 30 accuracies, which are averaged to obtain 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-CFs. This feature vector is given 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 toward the majority class. In such a class imbalance scenario, using balanced accuracy as a fitness function is appropriate

which is defined as

$$\text{fitness} = \frac{1}{z}\sum_{i=1}^{z}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{2}$$

where z represents the total number of classes, TP represents the true positives, and FN represents the false negatives. The ratio $(\text{TP}_i/\text{TP}_i + \text{FN}_i)$ shows the true positive rate (TPR) of class i. From (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 (RGB) color channels, 59 LBP features are extracted which are concatenated in a single feature vector with 177 (=3 channels 59 $\times$ 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 obtain a total of 416 wavelet features.

The value of the ith feature for the above three types of features is indicated by F_i, as shown by the GP individuals in Figs. 10 and 11. For L_{G} and L_{R} features, a window size of 3×3 pixels and a radius of 1 pixel ($\text{LBP}_{8,1}$) is adopted, which are the fundamental and widely used settings for LBP.

C. Function Set

The function set consists of three kinds of operators as follows.

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 PH^2 dataset is publicly available,[2] whereas the Dermofit dataset is not.[3] The datasets vary in terms of the number of classes, the number of images, the size of images, and the image capturing optical device.

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	$193 \times 221 - 777 \times 702$
	Seborrhoeic 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	$328 \times 193 - 914 \times 890$

1) PH^2: A dataset of specialized dermoscopic images, namely, PH^2 [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 dataset includes 8-bit RGB images of skin lesions, their binary masks, and clinical diagnosis. The dataset consists of three types of skin lesion images: 1) common nevi; 2) atypical nevi; and 3) melanoma. For multiclass classification experiments, PH^2 has these three classes. Samples of the three categories of skin lesions are presented in Fig. 5(a). In dermatology, there are some nonmalignant (benign) moles which may have chances to develop malignancy over time. Such moles are grouped in atypical nevi class. For a 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 [27]. The images are captured under standardized conditions. The dataset is divided into ten 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: 1) Melanocytic Nevus (mole) as *benign* and 2) Malignant Melanoma as *malignant*. For multiclass classification experiments, Dermofit has ten classes.

B. Experiments

For performing the experiments, 10-*fold cross validation* is used using random stratified sampling. This is because the

[2]https://www.fc.up.pt/addi/ph2%20database.html

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

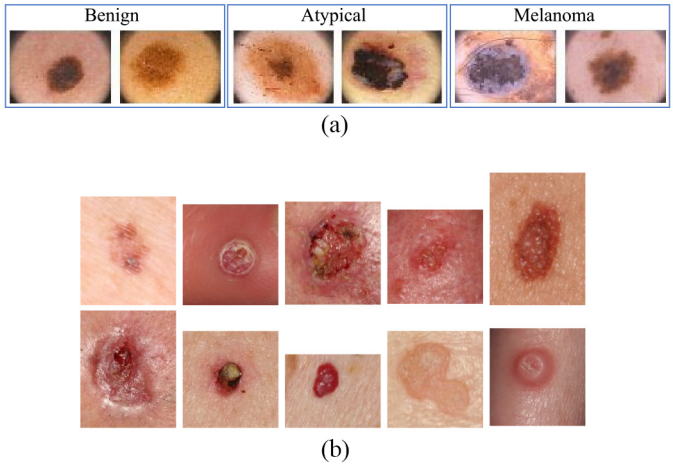

Fig. 5. Image samples from the two datasets. (a) PH2: two image samples from the three classes. (b) Dermofit: each image belongs to one class.

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

PH^2 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 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 ten 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 Evolutionary Computing Java-based package is used to implement GP [48].

C. 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.

D. 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 package version 3.8 [49] 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 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 [14] where they are empirically defined, since they produce the best results amongst other settings.

Moreover, we also compare 2SGP-W with recently developed state-of-the-arts for the PH^2 and Dermofit datasets, which are compared in Section V-E, and discussed as follows.

1) Patiño *et al.* [50] developed a multiclass classification method using the PH^2 dataset where different morphological operations are designed to encompass asymmetry, color, and border features. Three classification methods are used: a) SVM; b) logistic regression; and c) a fully connected neural network. The neural network outperformed the other two methods with an average accuracy of 86.5%.
2) Ain *et al.* [10] developed a 2Stage-GP method in an embedded approach using LBP features. They have evaluated performance of their method on the PH^2 dataset only. Their method can only solve binary classification problem and provided 78.17% balanced classification accuracy.
3) Alkarakatly *et al.* [51] designed a 5-layer CNN to classify the 3-class classification problem on the PH^2 dataset. Their method produced an overall accuracy of 90%. Using overall accuracy for imbalanced classification problems leads to a bias toward the majority class. Moreover, they reported a TPR of 83% which is clearly lower than the TPR of expert dermatologists that is, 90% [51].
4) Kawahara *et al.* [32] extracted features from a pre-trained CNN which are provided to a logistic regression classifier using the Dermofit dataset to classify its ten classes. Their method achieved 81.80% overall accuracy, which equals 60.12% balanced accuracy on the Dermofit dataset.
5) Fisher *et al.* [52] 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. This method achieved 78.10% overall accuracy and 70.50% balanced accuracy on the Dermofit dataset.

V. RESULTS AND DISCUSSION

A. Overall Results

The results of the proposed method are presented in Tables IV and V for binary and multiclass classification problems, respectively. Vertically, these tables comprise of three blocks, which correspond to the results of using L_G, L_R, and wavelet features, respectively. Horizontally, the table consists of five columns where first lists the classification algorithm, second and third show, respectively, the test performances on the PH^2 and the Dermofit datasets using all features, represented by "*All*." Similarly, the rest of the columns show test

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

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

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 obtain 30 accuracies for each classifier which are represented as mean and standard deviation ($\bar{x} \pm s$) in Tables IV and V.

To obtain a clear comparison between using different methods, the results are also tested using the *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 "↑," "↓," 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 PH^2 dataset, in Table IV, the test performance of SVM with 2SGP-W using L_R features is represented as "89.84 ± 1.41 ↑" where the ↑ 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 PH^2 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 multiclass 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 entire set of features. The proposed method provides the highest results on PH^2 using *wavelet* features with SVM reaching 97.50% average accuracy. On Dermofit, wavelet features remain prominent by providing 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 ↑ sign in Table IV) has significantly outperformed all the commonly used classification algorithms on both datasets.

D. Multiclass Classification

Table V shows the results of multiclass classification. The non-GP methods using all features have achieved 71.50% and 64.46% highest accuracy with MLP on PH^2 and Dermofit, respectively. It is evident that 2SGP-W has effectively provided much better results compared to the non-GP classification algorithms using all features. Among the two datasets,

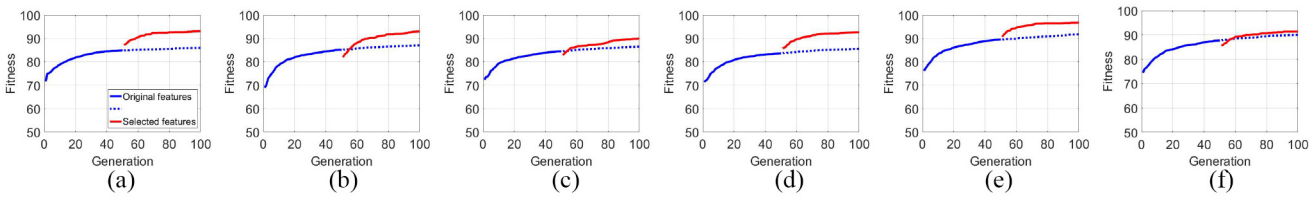

Fig. 6. Convergence plots for the PH^2 dataset in the *binary classification* task where the blue line represents stage-1 and the red line represents stage-2. (a) NB. (b) SVM. (c) k-NN. (d) J48. (e) RF. (f) MLP.

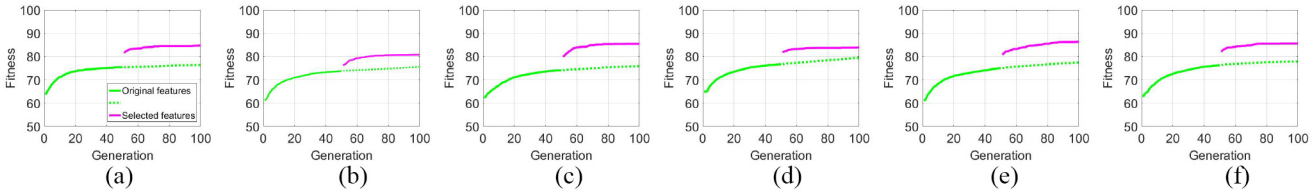

Fig. 7. Convergence plots for the Dermofit dataset in *binary classification* task where the green line represents stage-1 and the pink line represents stage-2. (a) NB. (b) SVM. (c) k-NN. (d) J48. (e) RF. (f) MLP.

TABLE VI
COMPARISON WITH THE STATE OF THE ARTS ON THE TWO DATASETS

Method	Dataset(s)	Strategy	Results
Patino et al. [50]	PH^2	neural network	86.5%
Ain et al. [10]	PH^2	2-stage GP embedded	78.17%
Alkarakatly et al. [51]	PH^2	5-layer CNN	90.00% (overall)
Kawahara et al. [33]	Dermofit	pre-trained AlexNet	60.12%
Fisher et al. [52]	Dermofit	decision tree	70.50%
2SGP-W	Dermofit, PH^2	2-stage GP wrapper	**85.67%, 97.24%**

2SGP-W provides good results on the PH^2 dataset with three classes (relatively easy task) achieving 97.24% average accuracy with SVM. Similarly, for the difficult task of classifying ten types of skin cancers in Dermofit, 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 ↑ sign in Table V) has significantly outperformed all the commonly used classification algorithms on both datasets.

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

Table VI compares 2SGP-W and existing methods regarding datasets used, strategies applied, and results achieved. For PH^2, the most recent state of the art has been developed by Patiño *et al.* [50] which has achieved 86.5% balanced accuracy in the multiclass classification problem using 10-fold cross validation. Ain *et al.* [10] developed a two-stage GP method for melanoma detection in a binary classification task. This method [10] is tested on only PH^2 using 10-fold cross validation and produced a balanced accuracy of 78.17%. Since the experimental setups in both methods [10], [50] 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 [50] by providing an increase of nearly 10% accuracy. In multiclass classification, 2SGP-W with 97.24% accuracy outperformed the second method [10] with an improvement of nearly 19% accuracy. In comparison to [51], 2SGP-W outperformed this CNN method achieving nearly 7% increased performance.

Kawahara *et al.* [32] provided an overall accuracy of 81.80% on Dermofit using pretrained CNN with an experiment set of 5-fold cross validation. However, according to the confusion matrix given in the study, this overall accuracy equals 60.12% balanced accuracy. Recently, Fisher *et al.* [52] provided state-of-the-art results on the Dermofit dataset. Kawahara *et al.* [32] and Fisher *et al.* [52] reported 70.50% balanced accuracy using leave-one out cross validation. Since comparison cannot be done directly with these two methods (5-folds versus 10-folds, and leave-one out versus 10-folds), here we try to give a general estimate of accuracy achieved by the current state-of-the-arts on the 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 and 7. These plots are generated for both the binary and multiclass classification experiments using L_R features. The plots for multiclass classification are given in the supplementary material. 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, that is, 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 onward on the x-axis.

From the plots in Figs. 6 and 7, a general trend can be observed; in the start of the evolutionary process, GP tries

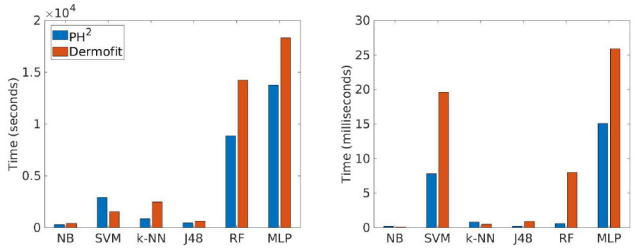

Fig. 8. Average computational time for *binary classification* using 2SGP-W method on the datasets. (a) Training time. (b) Test time.

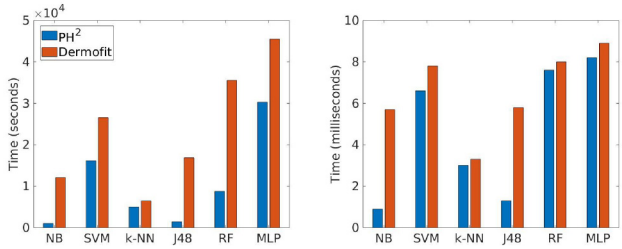

Fig. 9. Average computational time for *multiclass classification* using 2SGP-W method on the datasets. (a) Training time. (b) Test time.

to explore the search space and makes larger jumps, regardless of whether it works with all the original features or only the selected features. To obtain a clear understanding of how stage-2 is effective, we observe the stage-2 starts from a higher average accuracy most of the time compared to the average accuracy of 51st generation in stage-1. For example, in Fig. 6(a) on the PH^2 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. This is shown in Fig. 6(b), where stage-2 starts with 82.16% average accuracy, cuts the stage-1 line at 85.78%, and keeps improving afterward by making larger jumps to end at a better average accuracy of 93.46% compared to stage-1 ending at 87.95%. Hence, we conclude that selected features have potential to push GP make bigger jumps and help classification algorithm learn better to achieve good training performance.

B. Computational 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 multiclass classification tasks using wavelet features is depicted in Figs. 8, and 9, 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 min on average.

In Fig. 8, 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 PH^2 and Dermofit datasets. RF spends 2.78 and 4.17 h, 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 ms. 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 during diagnosis.

For multiclass classification, Fig. 9(a) and (b) shows that the computational time increases many folds while training a large (such as Dermofit) dataset with 1300 images as compared to training a small (such as PH^2) dataset with 200 images. A comparison of Figs. 8 and 9 illustrates that binary classification tasks take less training time as compared to multiclass 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. 9(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. 10 and 11. They have been taken from Dermofit experiments using RF in a multiclass classification task. From the evolved GP individual shown in Fig. 10, GP has selected only 11 features among a total of 416 wavelet features during stage-1, effectively reducing dimensionality many folds. Using these eleven features as input to stage-2, GP further selects most prominent features to build an informative GP CF as shown in Fig. 11. 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.

The wavelet features selected by the GP individual shown in Fig. 10 are listed in Table VII. We conclude the following from this table.

1) Three out of the 11 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.
2) Texture features obtained from red, green, and blue color channels are more informative and are chosen to construct this individual.

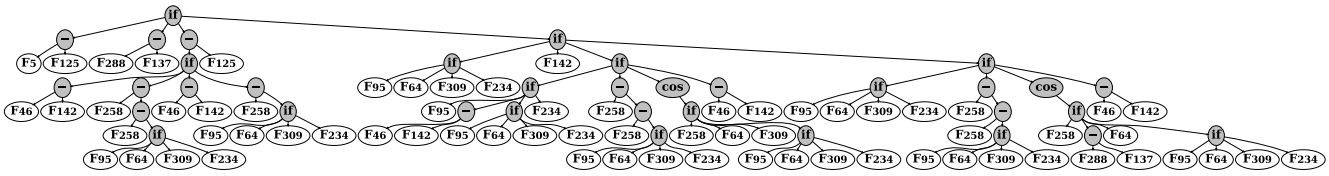

Fig. 10. Good GP individual evolved in stage-1 on the Dermofit dataset in the *multiclass classification* problem. This GP tree selects only eleven features from a total of 416 wavelet features constructively applying dimensionality reduction. These eleven features are selected in stage-1 and given as input to stage-2 which evolves another GP tree as shown in Fig. 11.

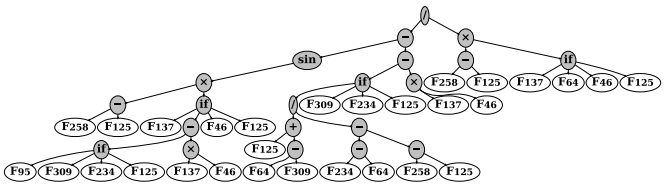

Fig. 11. Good GP individual evolved in stage-2 on the Dermofit dataset in the *multiclass classification* problem. This GP tree selects eight wavelet features from eleven input wavelet features (selected during stage-1). This GP is a CF and provided along with the eleven selected features to RF for classification.

TABLE VII
DETAILS OF THE WAVELET FEATURES SELECTED IN THE GP INDIVIDUAL SHOWN IN FIG. 10 IN TERMS OF STATISTICAL MEASURE, COLOR CHANNEL, AND 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

3) The node column in Table VII shows that the selected features are extracted from the low and middle-frequency channels.
4) Norm, entropy, energy, and standard deviation are prominent selected features among the statistical measures.

Moreover, the subtrees "(F_{46}–F_{142})" and "if (F_{95}, F_{64}, F_{309}, F_{234})" appear four and nine times, respectively, which show the potential of these subtrees to generate this GP tree.

D. Constructed Feature

To see why the selected and CFs can achieve good performance, we take an example of a CF shown in Fig. 11 evolved in a GP run during stage-2 on the Dermofit dataset in the multiclass classification task. This tree is constructed from eight features. The values of this CF are plotted in Fig. 12 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.

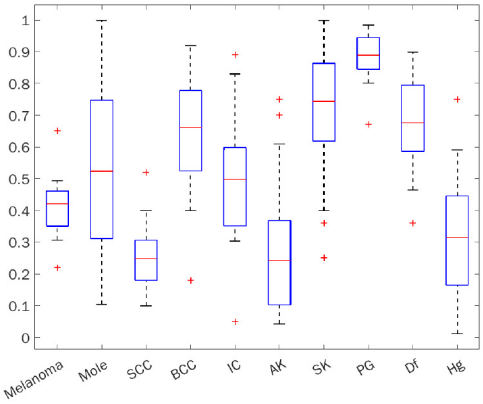

Fig. 12. Values of the CF shown in Fig. 11 on the Dermofit dataset with ten classes.

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 CF alone cannot distinguish SCC instances from all the other classes; both the selected and CFs are needed. The results show that GP not only selects important features but also constructs informative features with high discriminating ability.

VII. CONCLUSION

In this article, a novel two-stage GP method in a wrapper approach has been developed for skin cancer image classification. The proposed method includes both color and texture information to effectively discriminate between images of different classes. The features were extracted using LBP and pyramid structured wavelet decomposition from various color channels of skin 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 performs feature selection during stage-1 and feature construction from the selected features only during stage-2. The selected and CFs are provided to a classification algorithm such as J48, which has produced effective results for skin cancer binary and multiclass image classification.

The proposed method has produced significantly better results as compared to the commonly used classification algorithms (NB, *k*-NN, SVM, J48, RF, and MLP), the four state-of-the-art methods, and an existing embedded GP method on two real-world skin image datasets. The method successfully provided effective results without using any domain knowledge. This showed the evidence of effective feature

selection and feature construction by GP. With very less number of selected and CFs as compared to the number of original features, the proposed method has significantly 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, selected and CFs always perform better compared to the original set of features. We have also found that wavelet features with detailed internal structure and color information remain prominent in providing the highest classification accuracy for binary and multiclass 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. The current method is specifically designed for skin cancer image classification. Future work can be done to significantly extend the method for domain generalization in medical image classification tasks. Real-world images often come with a lot of noise, which hinders accurate image classification. 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 Cancer J. Clin.*, 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]*. Brisbane, QLD, Australia: Codon Publ., 2017.

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

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

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

[6] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 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 Trans. Evol. Comput.*, vol. 21, no. 1, pp. 83–101, Feb. 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," *J. Med. Syst.*, 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 *Proc. Annu. Conf. Genet. Evol. Comput.*, 2014, pp. 249–256.

[10] Q. U. Ain, B. Xue, H. Al-Sahaf, and M. Zhang, "Genetic programming for feature selection and feature construction in skin cancer image classification," in *Proc. 15th Pac. Rim Int. Conf. Artif. Intell.*, 2018, pp. 732–745.

[11] Y. Bi, B. Xue, and M. Zhang, "Genetic programming with a new representation to automatically learn features and evolve ensembles for image classification," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1769–1783, Apr. 2021.

[12] 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 *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2015, pp. 118–126.

[13] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[14] Q. U. Ain, H. Al-Sahaf, B. Xue, and M. Zhang, "Generating knowledge-guided discriminative features using genetic programming for melanoma detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 4, pp. 554–569, Aug. 2021.

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

[16] A. E. Eiben and J. Smith, "From evolutionary computation to the evolution of things," *Nature*, vol. 521, no. 7553, pp. 476–482, 2015.

[17] M. Zhang, V. B. Ciesielski, and P. Andreae, "A domain-independent window approach to multiclass object detection using genetic programming," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 8, pp. 1–19, 2003.

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

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

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

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

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

[23] T. Chang and C.-C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Trans. Image Process.*, vol. 2, pp. 429–441, 1993.

[24] B. Harangi, A. Baran, and A. Hajdu, "Classification of skin lesions using an ensemble of deep neural networks," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 2575–2578.

[25] E. Valle *et al.*, "Data, depth, and design: Learning reliable models for melanoma screening," 2017, *arXiv:1711.00441*.

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

[27] 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*. Dordrecht, The Netherlands: Springer, 2013, pp. 63–86.

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

[29] C. Barata and J. S. Marques, "Deep learning for skin cancer diagnosis with hierarchical architectures," in *Proc. Int. Symp. Biomed. Imag.*, vol. 2, 2019, pp. 841–845.

[30] K. Mahajan, M. Sharma, and L. Vig, "Meta-DermDiagnosis: Few-shot skin disease identification using meta-learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 730–731.

[31] N. Alfed and F. Khelifi, "Bagged textural and color features for melanoma skin cancer detection in dermoscopic and standard images," *Expert Syst. Appl.*, vol. 90, pp. 101–110, Dec. 2017.

[32] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *Proc. 13th Int. Symp. Biomed. Imag.*, 2016, pp. 1397–1400.

[33] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *Sensors*, vol. 18, no. 2, p. 556, 2018.

[34] M. Hasan, S. D. Barman, S. Islam, and A. W. Reza, "Skin cancer detection using convolutional neural network," in *Proc. 5th Int. Conf. Comput. Artif. Intell.*, 2019, pp. 254–258.

[35] P. Tschandl *et al.*, "Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks," *JAMA Dermatol.*, vol. 155, no. 1, pp. 58–65, 2019.

[36] M. H. Jafari *et al.*, "Skin lesion segmentation in clinical images using deep learning," in *Proc. 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 337–342.

[37] 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," 2020, *arXiv:2009.05977*.
[38] B. Bozorgtabar, S. Sedai, P. K. Roy, and R. Garnavi, "Skin lesion segmentation using deep convolution networks guided by local unsupervised learning," *IBM J. Res. Develop.*, vol. 61, nos. 4–5, pp. 1–8, Jul/Sep. 2017.
[39] A.-R. Ali, J. Li, and T. Trappenberg, "Supervised versus unsupervised deep learning based methods for skin lesion segmentation in dermoscopy images," in *Proc. Can. Conf. Artif. Intell.*, 2019, pp. 373–379.
[40] R. Poli, "Genetic programming for image analysis," in *Proc. 1st Annu. Conf. Genet. Program.*, 1996, pp. 363–368.
[41] C. Ryan, K. Krawiec, U.-M. O'Reilly, J. Fitzgerald, and D. Medernach, "Building a stage 1 computer aided detector for breast cancer using genetic programming," in *Proc. Eur. Conf. Genet. Program.*, 2014, pp. 162–173.
[42] M. Iqbal, B. Xue, H. Al-sahaf, and M. Zhang, "Cross-domain reuse of extracted knowledge in genetic programming for image classification," *IEEE Trans. Evol. Comput.*, vol. 21, no. 4, pp. 569–587, Aug. 2017.
[43] Q. U. Ain, H. Al-Sahaf, B. Xue, and M. Zhang, "Genetic programming for automatic skin cancer image classification," *Expert Syst. Appl.*, vol. 197, Jul. 2022, Art. no. 116680. [Online]. Available: https://doi.org/10.1016/j.eswa.2022.116680
[44] H. Al-Sahaf *et al.*, "A survey on evolutionary machine learning," *J. Roy. Soc. 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 *Proc. Congr. Evol. Comput.*, 2017, pp. 2420–2427.
[46] Q. U. Ain, B. Xue, H. Al-sahaf, and M. Zhang, "A multi-tree genetic programming representation for melanoma detection using local and global features," in *Proc. 31st Australas. Joint Conf. Artif. Intell. Lecture Notes Comput. Sci.*, 2018, pp. 111–123.
[47] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH^2—A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2013, pp. 5437–5440.
[48] S. Luke, *Essentials of Metaheuristics*, 2nd ed. Morrisville, NY, USA: Lulu, 2013. [Online]. Available: http://cs.gmu.edu/ ∼sean/book/metaheuristics/
[49] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *Special Interest Group Knowl. Discovery Data Min. Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.
[50] D. Patiño, A. M. Ceballos-Arroyo, J. A. Rodriguez-Rodriguez, G. Sanchez-Torres, and J. W. Branch-Bedoya, "Melanoma detection on dermoscopic images using superpixels segmentation and shape-based features," in *Proc. 15th Int. Symp. Med. Inf. Process. Anal.*, vol. 11330, 2020, Art. no. 1133018, doi: 10.1117/12.2545300.
[51] T. Alkarakatly, S. Eidhah, M. Al-Sarawani, A. Al-Sobhi, and M. Bilal, "Skin lesions identification using deep convolutional neural network," in *Proc. Int. Conf. Adv. Emerg. Comput. Technol. (AECT)*, 2020, pp. 1–5.
[52] R. B. Fisher, J. Rees, and A. Bertrand, "Classification of ten skin lesion classes: Hierarchical KNN versus deep net," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, 2019, pp. 86–98.

Qurrat Ul Ain (Member, IEEE) received the B.Sc. degree in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2009, the M.S. degree in computer science from International Islamic University, Islamabad, Pakistan, in 2013, and the Ph.D. degree in computer science from the Victoria University of Wellington, Wellington, New Zealand, in 2020.

She joined the Victoria University of Wellington in July 2016, where she is currently working as a Research Assistant. Her current research interests include evolutionary computation, particularly genetic programming, computer vision, medical image analysis, pattern recognition, machine learning, feature selection, extraction and construction, and transfer learning.

Dr. Ul Ain is a member of the IEEE Computational Intelligence Society and the Evolutionary Computation Research Group and Feature Analysis, Selection, and Learning in Image and Pattern Recognition, Victoria University of Wellington. She has been serving as a reviewer for more than ten international journals and conferences.

Harith Al-Sahaf (Member, IEEE) received the B.Sc. degree in computer science from Baghdad University, Baghdad, Iraq, in 2005, and the M.Comp.Sc. and Ph.D. degrees in computer science from the Victoria University of Wellington (VUW), Wellington, New Zealand, in 2010 and 2017, respectively.

In October 2016, he has joined the School of Engineering and Computer Science, VUW, as a Postdoctoral Research Fellow and has been a Full-Time Lecturer since September 2018. His current research interests include evolutionary computation, particularly genetic programming, computer vision, pattern recognition, evolutionary cybersecurity, machine learning, feature manipulation, including feature detection, selection, extraction, and construction, transfer learning, domain adaptation, one-shot learning, and image understanding.

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

Bing Xue (Senior Member, IEEE) received the B.Sc. degree from the Henan University of Economics and Law, Zhengzhou, China, in 2007, the M.Sc. degree in management from Shenzhen University, Shenzhen, China, in 2010, and the Ph.D. degree in computer science from the Victoria University of Wellington (VUW), Wellington, New Zealand, in 2014.

She is currently a Professor of Artificial Intelligence and the Deputy Head of the School of Engineering and Computer Science, VUW. She has over 300 papers published in fully refereed international journals and conferences. Her research focuses mainly on evolutionary computation, machine learning, classification, symbolic regression, feature selection, evolving deep NNs, image analysis, transfer learning, and multiobjective machine learning.

Dr. Xue is currently the Chair of the IEEE CIS Evolutionary Computation Technical Committee and IEEE CIS Task Force on Evolutionary Deep Learning and Applications, and the Editor of *IEEE CIS Newsletter*. She has also served as an Associate Editor for several international journals, such as *IEEE Computational Intelligence Magazine*, IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, and *ACM Transactions on Evolutionary Learning and Optimization*.

Mengjie Zhang (Fellow, IEEE) received the B.E. and M.E. degrees from the Artificial Intelligence Research Center, Agricultural University of Hebei, Baoding, China, in 1989 and 1992, respectively, and the Ph.D. degree in computer science from RMIT University, Melbourne, VIC, Australia, in 2000.

He is currently a Professor of Computer Science, the Head of the Evolutionary Computation Research Group, and the Associate Dean (Research and Innovation) of the Faculty of Engineering, Victoria University of Wellington, Wellington, New Zealand. He has published over 700 research papers in refereed international journals and conferences. His current research interests include machine learning, evolutionary computation, genetic programming, image analysis, multiobjective decision making, feature selection and reduction, scheduling and combinatorial optimization, and evolutionary deep learning and transfer learning.

Prof. Zhang was the Chair of the IEEE CIS Intelligent Systems and Applications Technical Committee, the IEEE CIS Emergent Technologies Technical Committee, and the IEEE CIS Evolutionary Computation Technical Committee. He is currently the Chair of the IEEE CIS PubsCom Strategic Planning Committee and the IEEE CIS Outstanding Ph.D. Dissertation Award Committee, and the Founding Chair of the IEEE Computational Intelligence Chapter in New Zealand. He is a Fellow of the Royal Society of New Zealand, a Fellow of Engineering New Zealand, and an IEEE Distinguished Lecturer.