Genetic Programming based Feature Manipulation for Skin Cancer Image Classification

by

Qurrat Ul Ain

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Computer Science.

Victoria University of Wellington
2020

Abstract

Skin image classification involves the development of computational methods for solving problems such as cancer detection in lesion images, and their use for biomedical research and clinical care. Such methods aim at extracting relevant information or knowledge from skin images that can significantly assist in the early detection of disease. Skin images are enormous, and come with various artifacts that hinder effective feature extraction leading to inaccurate classification. Feature selection and feature construction can significantly reduce the amount of data while improving classification performance by selecting prominent features and constructing high-level features. Existing approaches mostly rely on expert intervention and follow multiple stages for pre-processing, feature extraction, and classification, which decreases the reliability, and increases the computational complexity. Since good generalization accuracy is not always the primary objective, clinicians are also interested in analyzing specific features such as pigment network, streaks, and blobs responsible for developing the disease; interpretable methods are favored. In Evolutionary Computation, Genetic Programming (GP) can automatically evolve an interpretable model and address the curse of dimensionality (through feature selection and construction). GP has been successfully applied to many areas, but its potential for feature selection, feature construction, and classification in skin images has not been thoroughly investigated.

The overall goal of this thesis is to develop a new GP approach to skin image classification by utilizing GP to evolve programs that are capable of automatically selecting prominent image features, constructing new high-level features, interpreting useful image features which can help dermatologist to diagnose a type of cancer, and are robust to processing skin images captured from specialized instruments and standard cameras. This thesis focuses on utilizing a wide range of texture, color, frequency-based, local, and global image properties at the terminal nodes of GP to classify skin cancer images from multiple modalities effectively.

This thesis develops new two-stage GP methods using embedded and wrapper feature selection and construction approaches to automatically generating a feature vector of selected and constructed features for classification. The results show that wrapper approach outperforms the embedded approach, the existing baseline GP and other machine learning methods, but the embedded approach is faster than the wrapper approach.

This thesis develops a multi-tree GP based embedded feature selection approach for melanoma detection using domain specific and domain independent features. It explores suitable crossover and mutation operators to evolve GP classifiers effectively and further extends this approach using a weighted fitness function. The results show that these multi-tree approaches outperformed single tree GP and other classification methods. They identify that a specific feature extraction method extracts most suitable features for particular images taken from a specific optical instrument.

This thesis develops the first GP method utilizing frequency-based wavelet features, where the wrapper based feature selection and construction methods automatically evolve useful constructed features to improve the classification performance. The results show the evidence of successful feature construction by significantly outperforming existing GP approaches, state-of-the-art CNN, and other classification methods.

This thesis develops a GP approach to multiple feature construction for ensemble learning in classification. The results show that the ensemble method outperformed existing GP approaches, state-of-the-art skin image classification, and commonly used ensemble methods. Further analysis of the evolved constructed features identified important image features that can potentially help the dermatologist identify further medical procedures in real-world situations.

Acknowledgments

I would like to take immense pleasure in thanking those whom, without their help, this thesis would not have been possible. Above all, I thank God for His blessings, wellness, health, and abilities He has given me in carrying out this research and learning a little more about His world.

First and foremost, I would like to express my deepest gratitude to my supervisors Prof Bing Xue, Prof Mengjie Zhang, and Dr Harith Al-Sahaf for their valuable advice, unparalleled support, and guidance. Without challenging feedback and insightful suggestions of Assoc Prof Bing Xue, this PhD would not have been achievable. Her profound belief in my abilities and constant encouragement helped me overcome hurdles during my PhD journey. I am indebted to her more than she knows. I am profoundly grateful to Prof Mengjie Zhang for his extensive knowledge and constant support in shaping my leadership and academic skills. I am deeply indebted to Dr Harith Al-Sahaf, who had been a source of inspiration and had spent many dedicated hours and efforts to shape my research abilities and provide valuable and constructive feedback to improve my research work.

Many thanks to all staff members of the School of Engineering and Computer Science. I would like to extend my sincere thanks to all my friends in the Evolutionary Computation Research Group for their valuable feedback and suggestions. Special thanks to my officemates Bach Hoai Nguyen, FangFang Zhang, and Samaneh Azari for their lovely jokes, help, and support.

Words are inadequate in expressing my gratitude and thanks for blessings, tireless efforts, relentless support, and encouragement of my beloved parents (Muhammad Amin and Shahida Parveen). I owe my deepest gratitude to my sisters (Najamus Sehar, Gul-e Rana, and Tehreem Unnisa) for their love, inspiration, and encouragement.

I owe my deepest gratitude to my husband (Atif Ali) for his endless love and continuous support through every thick and thin in life, and to my lovely daughter (Aaminah Atif) for her patience to let her mother focus on her research work.

List of Publications

- **Qurrat Ul Ain**, Bing Xue, Harith Al-Sahaf and Mengjie Zhang. ”Genetic Programming For Skin Cancer Detection in Dermoscopic Images”. Proceedings of 2017 IEEE Congress on Evolutionary Computation (CEC 2017). Donostia - San Sebastian, Spain, 5-8 Jun, 2017. pp.2420-2427. IEEE, 2017.

- **Qurrat Ul Ain**, Bing Xue; Harith Al-Sahaf; and Mengjie 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 (PRICAI 2018), volume 11012, of Lecture Notes in Computer Science, Nanjing, China, 28-31 Aug, 2018, pages 732–745. Springer, 2018.

- **Qurrat Ul Ain**, Harith Al-Sahaf, Bing Xue, Mengjie Zhang. ”A Multi-tree Genetic Programming Representation for Melanoma Detection Using Local and Global Features”. Proceedings of the 31st Australasian Joint Conference on Artificial Intelligence (AI 2018), Lecture Notes in Computer Science. Vol. 11320. Springer. Wellington, New Zealand, 11-14 Dec, 2018. pp. 111-123. Springer, 2018.

- **Qurrat Ul Ain**, Bing Xue, Harith Al-sahaf and Mengjie Zhang. ”Multi-tree Genetic Programming with A New Fitness Function for Melanoma Detection”. Proceedings of 2019 IEEE Congress on Evolutionary Computation (CEC 2019). Wellington, New Zealand, 10-13 Jun, 2019. pp. 880-887. IEEE, 2019.

- **Qurrat Ul Ain**, Bing Xue, Harith Al-sahaf and Mengjie Zhang. ”Genetic Programming for Multiple Feature Construction in Skin Cancer Image Classification”. Proceedings of the 31st International Conference on Image and Vision Computing New Zealand (IVCNZ 2019). Dunedin, New Zealand, 2-4 Dec 2019. pp. 1-6.

- **Qurrat Ul Ain**, Bing Xue, Harith Al-sahaf and Mengjie Zhang. ”Generating Knowledge-Guided Discriminative Features Using Genetic Programming for Melanoma Detection,” in IEEE Transactions on Emerging Topics in Computational Intelligence. DOI: 10.1109/TETCI.2020.2983426.

- **Qurrat Ul Ain**, Bing Xue, Harith Al-sahaf and Mengjie Zhang. ”A Genetic Programming approach to Feature Construction for Ensemble Learning in Skin Cancer Detection”. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2020). Cancun Mexico. 8-12 Jul 2020. pp. 1186-1194.

- **Qurrat Ul Ain**, Harith Al-sahaf, Bing Xue and Mengjie Zhang. “Genetic Programming for Automatic Skin Cancer Image Classification”. Expert Systems with Applications. (under the 2nd round review).

- **Qurrat Ul Ain**, Harith Al-sahaf, Bing Xue and Mengjie Zhang. “Two-stage Genetic Programming for Feature Selection and Feature Construction in Skin Cancer Image Classification”. (status: first draft ready).

Contents

Chapter 1

Introduction

Computer vision is an interdisciplinary field that deals with how "computers" can be made for gaining high-level understanding from digital images or videos. From the perspective of engineering, it seeks to automate tasks that the human visual system can do [28]. An important aspect of image analysis is the extraction of meaningful information (also called features) from images; mainly from digital images by means of digital image processing techniques [187]. Computer vision and image analysis tasks include methods for acquiring, processing, analyzing and understanding digital images, and extraction of high dimensional data from the real world to produce numerical or symbolic information, e.g., in the forms of decisions [134]. Understanding in this context means the transformation of visual images (the input of the retina) into descriptions of the world that can interface with other thought processes and draw out appropriate action.

The importance of image classification is self-evident in the recent years as computer vision and image processing applications are widely spread into our daily lives [16]. Image classification is frequently found in commercial applications as well, such as identifying pedestrians in security surveillance systems [149], categorizing type of cells or detecting anomaly in medical images [31], and differentiating various terrains in satellite imagery applications [161]. One of the most prominent application fields is medical computer vision or medical image processing. This area is characterized by the extraction of information from image data for the purpose of making a medical diagnosis of a patient [52]. Generally, medical image data is in the form of X-rays, microscopy images, angiography images, ultrasonic images, tomography images and dermoscopic images [79]. An example of information which can be extracted from such image data is detection of tumors, or other malign changes. This application area also supports medical research by providing new information, e.g., about the structure of the organ, or about the quality of medical treatments; hence, assisting the medical practitioner in making a decision [70].

1.1 Problem Statement

The incidence of skin cancer, specifically malignant melanoma, continues to increase worldwide. This cancer can strike at any age and is one of the leading causes of loss of life in young individuals. Fair-skinned people, who burn easily and rarely tan, are mostly at risk [69]. Major causes of this disease are [191]: 1) the depletion of ozone layer caused by pollution, and 2) the excessive exposure to sun. Since this cancer is visible on the skin, it is potentially detectable at a very early stage when it is curable. New developments have converged to make fully automatic early skin cancer detection a real possibility [163]. Although some of the new systems reported for these technologies have shown promise in preliminary trials, widespread implementation must await further technical progress in accurate performance and reproducibility [128].

Dermoscopy images are inexpensive to obtain and widely available [37], and provide the most viable option to apply new image processing and machine learning algorithms to skin image classification. Therefore, skin cancer detection, which is in nature an image classification problem, has the most potential to deal with current clinical paradigm that needs to wait until the tumor is at a later stage and then perform an excessive number of costly biopsies [216]. The advent of a fast, accurate and cost-effective on-the-spot technology, i.e., Computer Aided Diagnosis (CAD) Systems are most likely to be afforded by the type of computer analysis of the skin lesion images. Such images come with various artifacts such as presence of gel, hair, and reflection [4] and hence, it is crucial to follow the proper methods to remedy these abnormalities and achieve a correct diagnosis. To deal with removal of such noisy artifacts that hinder accurate classification, the existing methods have used different pre-processing techniques and employed in a multi-stage classification system, which often requires human expertise and human intervention. Moreover, some stages are manually updated in the system, such as manual feature extraction [74], which heavily rely on expert intervention. Hence, there is a need for automatic implementation of these stages in order to save human computation time and produce accurate results.

Clinicians not only focus on correct prediction results, but are also interested in analyzing the insight of the cause of the disease [52], which is only possible if the CAD system is implemented as a "white-box" and is interpretable. More specifically, for skin cancer, dermatologists want to know which particular characteristics such as asymmetric shape or a specific texture pattern in the skin lesion causes a respective cancer stage. Existing machine learning approaches to skin cancer classification [5, 22, 7] like artificial neural networks (ANNs) and support vector machines (SVMs) remain unable to provide such interpretable solutions. Furthermore, the decision trees (J48) classification approach [74] although provides interpretable solutions, but does not perform well when there are complex interactions between features. Moreover, the recently developed classification methods using genetic programming (GP) [43, 44] have mostly used only gray-scale pixel statistics, which may not capture all the information from skin cancer images where variation in color is an important distinguishing characteristic between classes, as evident from the results shown in [167, 95, 154]. GP has been mostly explored for general image classification [16, 44] and object detection [213] problems, and its applicability to domain-specific applications still needs investigation.

Current techniques vary from applying standard machine learning algorithms to medical imaging datasets for developing new approaches adapted for the needs of the field [74]. A key limitation is that the size of a medical image is often large, whereas the relevant information about the disease is confined in a limited number of features in such images. Hence, suitable feature selection and feature construction methods are required to generate informative image features. Furthermore, the recent research [157, 216] reveals that textural, color and geometrical shape information provided from both medical and computer vision domains may help improve the classification performance, which requires a classification method capable of handling different types of medical features and image features together. Moreover, expert knowledge can be utilized to simplify the process of medical image classification, such as experts can provide segmentation of suspicious regions in a medical image. Hence, reducing the complexity of a CAD system by removing this extra step of segmentation for fast processing is an attractive direction to explore.

1.2 Motivations

Skin cancer can spread rapidly and can be life-threatening if left untreated [5, 22, 66, 132]. In 2019, the global incidence of skin cancer was estimated to be over 104,350 cases, with almost 11,650 deaths [182]. Over 4000 people are diagnosed with different types of skin cancers including melanoma every year in New Zealand; that is, around 11 people every day [144]. Melanoma, which is the deadliest type of skin cancer, accounts for nearly 80% of all skin cancer deaths and over 300 New Zealanders die of melanoma every year [144]. The American Cancer Society reported 82,770 new cases of skin cancer in the United States in 2013 which increased to

$83,510$ in 2016, with $12,650$ and $13,650$ melanoma deaths in years 2013 and 2016 respectively, maintaining an increasing trend over last decades [180, 181]. In Australia, melanoma is the most common cancer in people aged between 15 and 44 years [49]. It represents 10% of all cancers and its per-capita incidence is four times higher than in Canada, the UK and the US, with more than $10,000$ cases diagnosed and around 1250 deaths annually [2]. The worldwide continuous increase in incidence of melanoma and other skin cancers in recent years, its high mortality rate and the huge respective medical cost have made its early diagnosis an important priority of public health [7]. Using three decades of cancer registry data (1982 - 2011) from six populations with moderate to high melanoma incidence (US whites and the populations of the United Kingdom, Sweden, Norway, Australia, and New Zealand), a study [204] has described current trends and project future incidence rates and numbers of melanomas out to 2031 based on age-period-cohort models. With the maturity of prevention campaigns and their apparent success in changing behavior, particularly in Australia and New Zealand [49], it is expected to see a decrease in skin cancer incidence and numbers of new cases.

Early detection of skin cancer is critical, as the estimated 5-year survival rate for melanoma decreases from over 99% if detected in earliest stages to about 14% if detected in latest stages [72]. Due to enhancements in skin imaging technology and image processing techniques in the recent years, there has been a significant increase in the development of CAD systems for skin cancer detection [7]. Dermoscopy (also known as dermatoscopy or epiluminescence microscopy) is a method of acquiring a magnified and illuminated image of a region of skin for increased clarity of the spots on the skin [45]. Over the last two decades, a different trend of dermoscopy CAD systems have emerged, where these systems aim to reduce the gap between the medical and engineering knowledge, by trying to mimic the dermatologists behavior when diagnosing a skin lesion [33]. In the last decade, GP has been extensively applied to analysis of molecular data to classify tumor sub-types and characterize the mechanisms of tumor development [207].

GP is an evolutionary computation (EC) algorithm based on Darwinian principles of biological evolution and natural selection that automatically explores the solution space to evolve a computer program (model/solution) for a user-defined problem [103]. Fundamentally, GP consists of a set of operators and a fitness measure that is used to evaluate the performance of an evolved program. The algorithm starts by randomly generating a predefined number of solutions to form an initial population. GP then uses the set of operators to gradually improve these solutions over a number of generations. The ability of GP techniques to handle complex problems represents a key motivation for many researchers to utilize it to perform different tasks such as object detection [9, 38, 183, 211, 212, 213], feature selection [10, 107, 131, 136, 198], feature construction [11, 110, 138, 139, 198], image classification [15, 17, 20, 44, 58, 110, 143, 195], regression [56, 83] and knowledge transfer [51, 90].

Classification tasks often have a large number of features, and irrelevant and redundant features may reduce the performance [209]. In the traditional tree-based GP, a tree-like structure is evolved where features appear at the terminals and operators appear at the internal nodes. GP performs implicit feature selection by selecting features at these terminal nodes and the goodness of the evolved program (having those selected features in it) is calculated using a fitness measure [10]. Feature construction involves transforming a given set of input features to generate a new set of more powerful features [138]. Feature construction can improve classification accuracy by selecting relevant features and constructing high level features. Hence, by reducing the dimensionality through feature construction, better performance can be achieved while using less computation time. Feature selection aims at selecting prominent features, generally have more distinguishing ability between classes, from the complete set of features to improve classification performance. 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 [209]. An embedded approach integrates classifier learning and feature selection into a solitary procedure [209]. GP can be utilized as a feature selection method in a wrapper approach where goodness of feature subsets is evaluated using a fitness measure and multi-class classification is performed using the best feature subset by a machine learning algorithm such as SVM.

Moreover, in the medical domain where clinicians are interested in finding the cause of a disease, a system is highly recommended that provides such causal information. GP evolved programs can be interpreted by analyzing which particular features are selected at the terminal nodes in a tree-based GP representation. With this remarkable property, GP provides the information about which specific features are prominent in constructing new high level features. Hence, clinicians can gain deep understanding of the cause of the disease. Thus, analysis of the evolved GP program provides meaningful insights to the clinicians that can assist in identifying further medical procedures.

While having a close analysis of dermoscopic images, the doctors gather information of various features collectively (similar to feature construction) and then use that gathered knowledge to make the decision of whether further treatment (biopsy) is required or not. Hence in computer vision, mimicking the human ability to capture knowledge from various features and use it collectively are expected to be achieved through feature construction using GP.

1.2.1 Challenges of Skin Cancer Image Classification

For skin cancer classification, clinical guide for distinguishing between various kinds of cancers, e.g., basal cell carcinoma, squamous cell carcinoma, and melanoma, are based on dermoscopic criteria specifically the Asymmetry, Border, Color, and Diameter (ABCD) rule [193] and the 7-point checklist method [24] (Asymmetry, Pigment network, Dots/Globules, Streaks, Regression areas, Blue-whitish veil and presence of colors; white, red, light-brown, dark-brown, blue-gray, and black); these are the key medical properties that help dermatologists for classification of melanoma and other types of cancer. The major limitations and possible chances for better addressing the skin cancer classification tasks are listed below, which form the motivations of this thesis:

1. Usually the size of an image in real-world datasets varies between images of the same dataset. Most of the existing ANN approaches to skin image classification usually resize these varying image sizes to a smaller size which massively distorts the aspect-ratio of the images. This leads to loss of information at the pixel-level and hence, texture patterns extracted from these resized images might not provide sufficient information necessary to distinguish between different types of cancers.

2. The existing methods focus on either one or more visual characteristics (texture, color, and border shape) to extract features, but have not targeted all the important properties [4, 34, 128]. In skin images, the presence of blue-whitish veil makes it easier for the dermatologist to further investigate the stage of cancer present in the lesion. Similarly, the texture analysis such as presence of pigmented network provides evidence of tumor. Abbas et al. [4] focused on only lesion border detection to classify skin images, and Barata et al. [34] converted the RGB color images to gray-scale (completely ignoring color properties) to extract features in pigment network detection of skin images. Moreover, a pre-processing step for hair removal and illumination correction has been essential in the previous systems [100, 108]. Based on these limitations, a system is required that can efficiently extract features based on all important dermoscopic prop-

erties (making sure all or nearly all relevant information from images has been extracted) while eliminating the use of an extra preprocessing step.

3. In medicine, doctors first utilize CAD systems to detect presence/absence of a disease. Once a patient is found positive for a disease, doctors always want to investigate the cause of the disease according to some symptoms. More specifically, a dermatologist wants to specify which particular areas in the lesion and which structures in the respective areas are causing a specific stage of skin cancer. A system that provides such causal information is highly recommended. The existing methods for skin cancer classification have mostly used various classification methods such as SVM [5], Neural Networks, [72, 64, 157, 171], and K-Nearest Neighbor (K-NN) [33, 36]. These systems can only provide the final prediction results to a dermatologist and cannot describe which prominent features or regions in an image is the cause of the disease. With the property of GP evolved programs being interpretable, giving information about which features are prominent in distinguishing different tumor classes, clinician are expected to gain deep understanding of the cause of the disease.

4. According to the needs of dermatologists, it is sometimes more important to diagnose a diseased instance correctly as compared to diagnosing a non-diseased instance correctly [177]. Moreover, most of the medical data is of imbalance; having a different number of instances of each class. The existing approaches to skin cancer classification have often employed overall accuracy as a fitness measure [22, 25]. However, this is inappropriate for unbalanced datasets as it leads to bias towards the majority class. Therefore, a good evaluation/fitness function that avoids bias towards the majority class and classifies well the diseased and non-diseased instances in unbal-

anced datasets should be investigated.

5. Traditional GP approaches require a suitable design in producing good performance for multi-class classification tasks [71]. In the literature, three approaches have been most commonly used: first decomposing the task into binary problems, second a single discriminant function is evolved which separates classes based on multiple thresholds (each class has a predefined threshold interval), and third generating multiple outputs from the individual instead of a single value. In the first approach, the system is run once for each of the classes to be distinguished and in each run, a single-threshold discriminant function is evolved for a particular class. In the second approach, defining a proper threshold interval for each class requires domain knowledge which vary among different experts. A number of researchers have investigated different GP program representations for the third approach, such as Linear GP (LGP) [76], Modi– GP [214] and Probability based GP [186] for the task of multi-class classification. Generally, it is difficult to design a good GP program structure which can effectively and efficiently perform multi-class classification. Existing approaches employ domain independent features which may not give good performance when dealing with complex datasets such as skin cancer images where most people do not have good domain knowledge. To deal with shortcomings of GP for multi-class classification, GP can be employed as a feature selection algorithm in a wrapper approach. In this way, during the evolutionary cycle, GP keeps evolving better individuals with more informative features which helps train a classifier, thereby improves performance many times over the original features.

1.2.2 Why GP?

In the literature, GP has not only been explored for classification, it has been extensively studied in a wide range tasks including feature selection, feature construction and feature extraction. GP is a global search technique and takes the advantage of its most often used tree-based program representation, which other techniques under the EC umbrella, such as particle swarm optimization cannot achieve. Not only this flexible representation makes GP a preferred technique among the other EC techniques, GP but also has other powerful characteristics such as performing multiple tasks simultaneously, automatically evolving models, interpretable solutions, and built-in feature selection add flavors to attract researchers. These GP characteristics motivate to utilize GP for image classification tasks.

- **Automatically Evolving Models.** With its flexibility, GP automatically evolve or generate mathematical models (solutions) to solve a problem with a given set of terminals and functions. During the evolutionary cycle, this tends to develop more suitable models in the subsequent generations where the goodness of the models is measured against a fitness function. Hence, GP keeps evolving models automatically until a stopping criterion is met.

- **Intrinsic Feature Selection.** While evolving models, GP only selects some of the features and does not use the complete set of features. This shows that GP has intrinsic feature selection ability. During the evolutionary process, GP keeps selecting the prominent features where the goodness of the evolved model (or indirectly these features) keeps improving against a pre-defined fitness function. At the end of the evolutionary cycle, GP becomes successful in selecting refined features most suitable to solve the problem at hand.

- **Multiple Tasks.** GP can perform multiple tasks simultaneously. For example, by evolving a computer program, it performs feature selection by selecting prominent features at its terminals, and the evolved tree can be considered as a constructed feature, hence, performs feature construction. For image data, GP can simultaneously perform feature extraction, region detection, feature construction and classification by just evolving one computer program [110].

- **Flexible Representation.** Traditionally GP evolves computer programs, represented in tree structure. However, that is not the only representation GP has. Non-tree GP representations include Linear GP where programs are represented as a sequence of instructions from machine language, and Cartesian GP where programs are encoded using a graph representation [205]. Similarly, a multi-tree GP evolves multiple trees in a GP individual which are initially explored for performing multi-class classification [135]. In a tree-like representation, the programs are dynamically evolved during the evolutionary cycle. In a single population of evolved programs, the size of GP trees varies in terms of tree depth, the number of terminals and internal nodes. This flexible representation suits feature construction and feature selection. A tree can be considered as a constructed feature by computing its value based on the selected features at terminal nodes and operators at the internal nodes. The selected and constructed features are assumed to have a better data quality than the original features.

- **Interpretable Solutions.** Unlike many learning algorithms where the models have a "black box" architecture and cannot be interpreted, GP evolves interpretable models. This allows easy understanding of how GP tackles the problem at hand. This also helps identify the prominent features selected by GP tree and can provide the domain experts with enough knowledge to analyze future data.

1.3 Research Goals

The overall goal of this thesis is to explore feature selection and feature construction abilities of GP by utilizing various GP representations to produce informative feature vectors and effective classification models with sufficient information to discriminate between various types of skin cancers in images. The specific objectives of this research work are described in more detail as follows.

1. *Develop multi-stage GP methods for skin cancer image classification.*

   The domain-specific features provided by the domain experts (based on 7-point checklist method) are highly informative [24]. It is expected that the combination of these domain-specific skin image features and domain-independent texture features when provided to GP can result in good classification models. Existing approaches to feature construction [11, 198] have utilized the whole set of features to construct new features. In contrast, this thesis will explore constructing new features from previously GP-selected features and not the whole set of features. It is expected that feature construction from selected features results in improved performance than feature construction from whole set of features. This thesis will develop multi-stage GP methods using embedded and wrapper approaches for feature selection and feature construction to automatically generate a feature vector of selected and constructed features for classification. It is expected that the generated feature vector will produce more powerful classification models over using all features. Interpretability of the evolved programs will also be investigated. This approach is not only intended to support the skin image domain but can also be extended conveniently to other image and non-image domains.

2. *Develop a multi-tree GP based embedded feature construction approach for melanoma detection.*

   In the literature [35, 96, 72], only features extracted from either gray-scale images or color images have been used for skin image classification. In some other works [117, 216], new skin lesion features are developed which capture asymmetry and geometrical border shape characteristics of a lesion. These existing works did not include different set of features exhibiting different skin image properties such as texture, color, border shape, local and global features. It is obvious that providing a comprehensive set of informative features to a classifier results in generating better classification models than providing incomplete information. This thesis will explore using different sets of features such as texture, color, local, global and domain-specific geometrical border features, where each tree in an individual will be evolved using a single set of features. Using different sets of features will incorporate as much information as possible which plays a vital role in generating good classification models effectively identifying melanoma from benign images. Designing suitable crossover and mutation operators to meet the requirements of the feature construction process will also be explored.

3. *Extending the multi-tree GP based feature construction method using various types of features to a wrapper approach to skin cancer multi-class classification.*

   The wavelet analysis [54] is well-known for capturing both the local (detailed structure and internal texture) and global (overall properties) information of the skin lesion. It is expected that addition of wavelet features extracted from color channel images can greatly impact on achieving good performance. In addition, the multi-tree approach is more convenient to employ for multiple feature construction where each tree is considered as a constructed feature. With the aim of generating each tree using a single set of features, e.g. one tree for gray-scale features, one for color features and another for frequency-based features, it is expected to generate significant features which can effectively discriminate multiple classes of skin images. Multi-tree approaches on non-image classification datasets have been studied previously [135, 136, 197], however, they have not been investigated for complex image classification tasks where different kinds of features (based on local and global information as well as color and texture descriptors) can be used.

4. *Developing an ensemble classification method using the GP framework for skin cancer image classification.*

   Ensembles of classifiers have been proven to be more effective than a single classification algorithm in skin image classification problems [86, 201, 208]. Generally, the ensembles are created using the whole set of original features. However, some original features can be redundant and may not provide useful information in building good ensemble classifiers. To deal with this, existing feature construction methods that usually generate new features for only a single classifier have been developed but they remain unable to provide good classification performance [141]. In the past, either complete set of original features or selected features are provided to ensembles [82, 199], however, multiple constructed features have not been provided to an ensemble of classifiers. This thesis will explore the ability of multiple feature construction to improve performance in ensemble classification.

1.4 Major Contributions

This section presents the major contribution of this thesis. Each contribution presented below is discussed in detail in chapters 3 to 6 of this thesis.

1. This thesis proposes a binary image classification method to explore feature selection ability of GP by using domain dependent skin image features and domain independent texture features. Experimental results show that GP has achieved good results and has the potential to provide efficient and effective solutions for real-world problems such as cancer detection. Using knowledge from both domains (dermatology and computer vision), GP has achieved significantly better

or comparable performance compared to other commonly used machine learning classification methods. This thesis also proposes two multi-stage GP methods (an embedded and a wrapper approach) where prominent features are selected in one stage and new high level features are constructed from the selected features only in the other stage. Experiment results of the embedded and wrapper approaches outperform the existing baseline GP approaches and other machine learning methods. Insights of the evolved programs reveal that GP selects highly significant features which can help dermatologist to make a diagnosis.

Parts of this contribution have been published in:

Qurrat Ul Ain, Bing Xue, Harith Al-Sahaf and Mengjie Zhang. ″Genetic Programming For Skin Cancer Detection in Dermoscopic Images″. Proceedings of 2017 IEEE Congress on Evolutionary Computation. Donostia - San Sebastian, Spain, 5–8 June, 2017. pp. 2420–2427. IEEE, 2017.

Qurrat Ul Ain, Bing Xue, Harith Al-Sahaf, and Mengjie Zhang. ″Genetic Programming for Feature Selection and Feature Construction in Skin Cancer Image Classification″. Proceedings of the 15th Pacific Rim International Conference on Artificial Intelligence, volume 11012, of Lecture Notes in Computer Science, Nanjing, China, 28–31 August, 2018, pages 732–745. Springer, 2018.

Qurrat Ul Ain, Harith Al-sahaf, Bing Xue and Mengjie Zhang. “Two-stage Genetic Programming for Feature Selection and Feature Construction in Skin Cancer Image Classification″. (In preparation for a journal).

2. This thesis utilizes multi-tree GP representation to develop an embedded approach to include various types of local and global features to improve classification performance of skin cancer image classification. Features which have information regarding pixel-

based gray-level and RGB characteristics, variation in color across the image (inside and between the lesion and skin regions) and geometrical border shape properties are provided to multi-tree GP. In order to avoid mixing of different features in one tree, suitable crossover and mutation operators are adopted. This thesis also proposes a new weighted fitness function in the multi-tree approach for skin cancer image classification. This fitness function allows the different trees in a GP individual to influence each other's performance during the evolutionary process. This method outperforms all the most commonly used classification algorithms and the single-tree GP methods showing evidence of good discriminating ability between "$malignant$" and "$benign$" skin lesions. The evolved GP programs identify the important domain specific and domain independent features which help improve the classification performance.

Parts of this contribution have been published in:

Qurrat Ul Ain, Harith Al-Sahaf, Bing Xue, Mengjie Zhang. "A Multi-tree Genetic Programming Representation for Melanoma Detection Using Local and Global Features". Proceedings of the 31st Australasian Joint Conference on Artificial Intelligence, Lecture Notes in Computer Science. Vol. 11320. Springer. Wellington, New Zealand, December 11–14, 2018. pp. 111–123. Springer, 2018.

Qurrat Ul Ain, Bing Xue, Harith Al-sahaf and Mengjie Zhang. "Multi-tree Genetic Programming with A New Fitness Function for Melanoma Detection". Proceedings of 2019 IEEE Congress on Evolutionary Computation. Wellington, New Zealand, 10–13 June, 2019. pp. 880–887. IEEE, 2019.

3. This thesis proposes the first GP approach for skin cancer image classification which utilizes frequency based wavelet features as well as texture, color and geometrical border shape features of skin images. GP automatically constructs multiple features in a single GP indi-

vidual and provides these new informative features to a classification algorithm for classification in a wrapper approach. This method produces significantly better results than the commonly used classification algorithms, eight single-tree GP methods, and an existing multi-tree GP embedded method. This shows the evidence of effective feature construction, which results in achieving good binary and multi-class classification results. The results find an interesting behavior that selecting a suitable feature extraction method is necessary to classify well a particular type of images taken from a specific optical instrument. With the interpretability of evolved GP models, the most-frequently occurring features in the GP trees are identified. These prominent features are associated with skin cancer characteristics which can help dermatologist distinguish between different types of skin cancers.

Parts of this contribution have been published in:

Qurrat Ul Ain, Bing Xue, Harith Al-sahaf and Mengjie Zhang. ”Generating Knowledge-Guided Discriminative Features Using Genetic Programming for Melanoma Detection,” IEEE Transactions on Emerging Topics in Computational Intelligence. DOI: 10.1109/TETCI.2020.2983426.

Qurrat Ul Ain, Bing Xue, Harith Al-sahaf and Mengjie Zhang. ”Genetic Programming for Multiple Feature Construction in Skin Cancer Image Classification”. Proceedings of the 31st International Conference on Image and Vision Computing New Zealand. Dunedin, New Zealand, 2–4 Dec 2019. pp. 1–6.

Qurrat Ul Ain, Harith Al-sahaf, Bing Xue and Mengjie Zhang. “Genetic Programming for Automatic Skin Cancer Image Classification”. Submitted to Expert Systems with Applications. April 2020. pp. 1–23.

4. This thesis develops an ensemble GP classification method where

multiple constructed features evolved by GP are provided to the ensemble for classification. This method combines the powerful techniques of utilizing multi-tree GP for multiple feature construction, and providing constructed features to an ensemble of classifiers to achieve performance gains in skin cancer image classification. Experiments show that the new features constructed for ensemble of classifiers have more distinguishing ability between classes as compared to features constructed for a single classifier. Compared to the existing GP approaches, commonly used classification and ensemble methods, this method significantly outperformed all of them. Moreover, in comparison to the state-of-the-art convolutional neural network (CNN) methods, the evolved constructed features being interpretable identified important features selected from the original set of features. This information can be helpful to the dermatologist in making a diagnosis.

Parts of this contribution have been published in:

Qurrat Ul Ain, Bing Xue, Harith Al-sahaf and Mengjie Zhang. "A Genetic Programming approach to Feature Construction for Ensemble Learning in Skin Cancer Detection". Proceedings of the Genetic and Evolutionary Computation Conference. Cancun Mexico. July 8-12 Jul 2020. pp. 1–9. Accepted on 20 March 2020.

1.5 Organization of the thesis

The remainder of this thesis is organized as follows. Chapter 2 presents the literature survey of the related work. Chapters 3–6 present the main contributions of this thesis as mentioned in Section 1.4, and can be seen in Fig. 2.1. Chapter 7 concludes this thesis.

Chapter 2 describes the essential medical background of skin cancer and the techniques involved in traditional computer-aided diagnos-

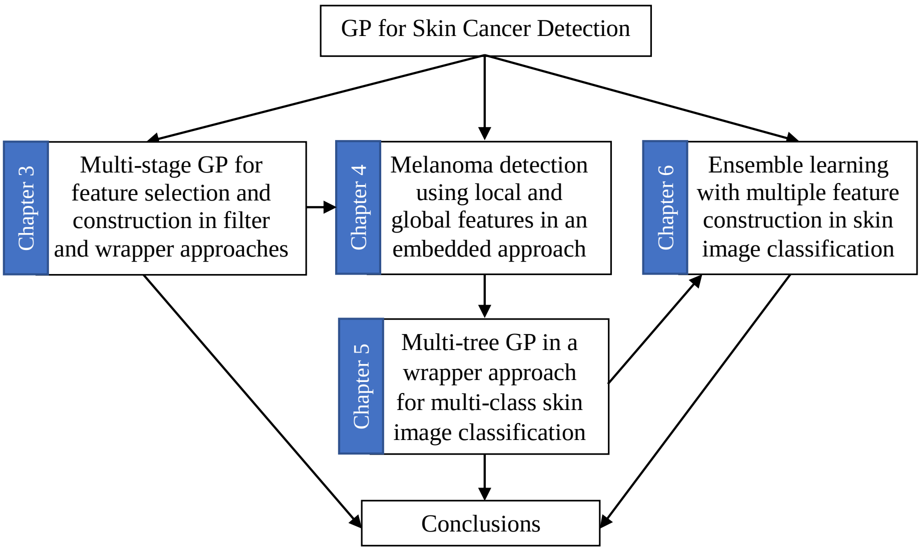

GP for Skin Cancer Detection
Chapter 3
Multi-stage GP for feature selection and construction in filter and wrapper approaches
Chapter 4
Melanoma detection using local and global features in an embedded approach
Chapter 6
Ensemble learning with multiple feature construction in skin image classification
Chapter 5
Multi-tree GP in a wrapper approach for multi-class skin image classification
Conclusions

Figure 1.1: The layout of the main contributions making Chapters 3–6.

tic methods to classify skin cancer images. It describes the basic concepts of computer vision, machine learning, evolutionary computation and Genetic Programming. It then provides an overview of the related work previously developed to deal with the task of skin cancer image classification.

Chapter 3 presents new GP based methods for binary image classification for skin cancer which utilize texture features extracted from skin images and domain-specific features provided by the expert dermatologists. This chapter describes the importance of color as a significant component in skin cancer identification and employs both texture and color information features in one set of experiments. The effectiveness of these methods is experimentally assessed and compared with commonly used classification algorithms. The evolved GP programs are presented and analyzed as well.

Chapter 4 proposes novel embedded based GP approaches to feature selection for melanoma detection in a binary image classification problem. It explains how various texture, color, and geometrical border shape features are extracted from skin images. This chapter then presents how multi-tree representation is used to generate multiple classifiers in a single GP individual. A new fitness function is designed to improve the classification performance of the task at hand. A deeper analysis of the selected texture features reveals that some particular texture patterns occur more frequently in melanoma images than benign images.

Chapter 5 proposes a novel wrapper based binary and multi-class classification approach to feature selection and construction in order to detect various types of skin cancers in dermoscopy and standard camera images. It discusses the various feature extraction methods adopted to extract informative sets of features. This chapter then explains the workflow of the proposed algorithms. A set of experiments are conducted on the real world skin image datasets to examine the performance of the wrapper approach. Comparison to the previous GP embedded approaches is performed in terms of training and test computation time, the frequency of feature appearance in GP trees, and the classification performance.

Chapter 6 presents a novel genetic approach to feature construction in ensemble classification to classify various types of cancers in skin images. This chapter first discusses the limitations of existing ensemble approaches for skin cancer image classification. It then proposes a new ensemble method which relies on constructed features by GP to successfully address these limitations. A set of experiments for binary and multi-class image classification are presented to demonstrate the effectiveness of the proposed method. This chapter then presents the results and compares them with existing GP and the state-of-the-art CNN methods. This chapter further analyzes how each individual classifier in the ensemble improves its performance during the evolutionary cycle.

Chapter 7 concludes the thesis and draws the overall conclusions of the thesis. The main contributions and key research points are emphasized and discussed. This chapter then suggests possible future directions.

Table 1.1: Real-World Skin Cancer Datasets.

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

1.6 Benchmark Datasets

All the skin image classification algorithms proposed in this thesis are evaluated on two skin image datasets of varying difficulty. These datasets are selected because they come up with the binary masks of the lesion area. Since this thesis does not explore image segmentation, and mainly focus on classification, datasets provided along with lesion segmentation (binary masks) are selected. Moreover, one of the datasets also provides domain specific features. Details of these datasets are given in Table 1.1. The PH^2 dataset is publicly available [1], whereas the Dermofit dataset is purchased online [2] with the help of Huawei Industry Fund E2880/3663. 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] https://www.fc.up.pt/addi/ph2%20database.html

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

1.6.1 PH^2

A dataset of dermoscopic images namely PH^2 [123] is acquired from Pedro Hispano Hospital Portugal. The dermoscopic images were obtained from Tuebinger Mole Analyzer system with a magnification of $20\times$ and resolution of around 768×560 pixels. Dermoscopy includes using an optical instrument having a powerful lighting system to examine skin lesions in a higher magnification. Before taking the images, a gel is placed on the lesion that enables the dermatoscope (instrument) to capture morphological structures and patterns in inner layers of human skin. Hence, such images are rich enough to investigate them for presence of skin cancer. The images are 8-bit RGB (red, green and blue) color images.

The dataset includes images of skin lesions, their clinical diagnosis, their binary masks and information of domain specific features provided by the dermatologists, based on the 7-point checklist method. The dataset consists of three types of skin lesion images: common nevi, atypical nevi, and melanoma. Among these three classes, since atypical nevi refers to moles which are currently non-malignant but may develop melanoma later. This atypical nevi class is combined with common nevi which refers to moles, to form one *benign* class. For binary classification experiments, this benign class and the melanoma class are used. For multi-class classification experiments, PH^2 has three classes, which are common nevi, atypical nevi and melanomas. Samples of the three categories of skin lesions are presented in Fig. 1.2.

1.6.2 Dermofit Image Library

The Dermofit Image Library [29] is a set of 1300 high quality skin lesion images collected under standardized conditions with internal color standards, captured from a standard camera. The lesions span across 10 different classes, where each image has a gold standard diagnosis. Images consist of a snapshot of the lesion surrounded by normal skin. There is a huge

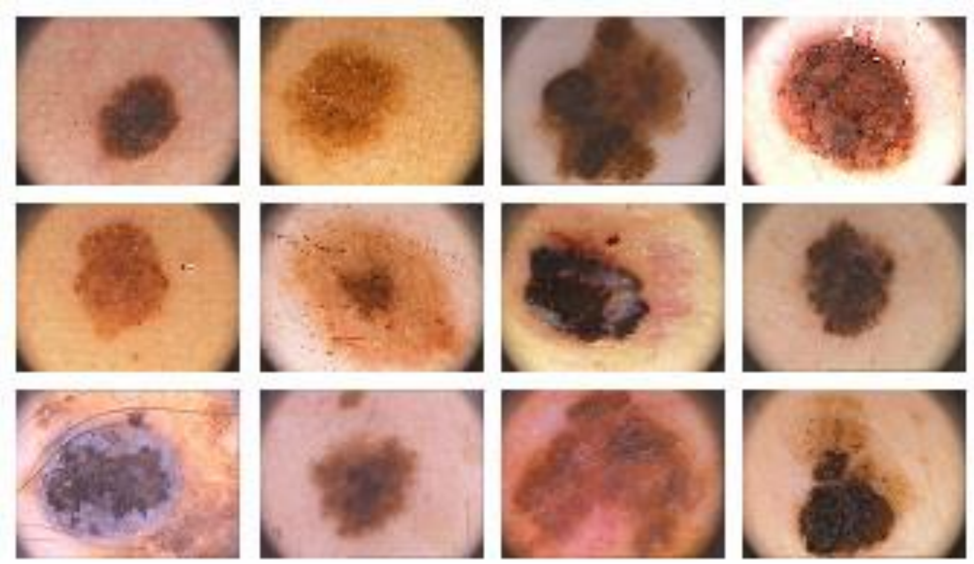


Figure 1.2: Some Images of dermoscopic dataset, with common nevi (row 1), atypical nevi (row 2) and melanomas (row 3).

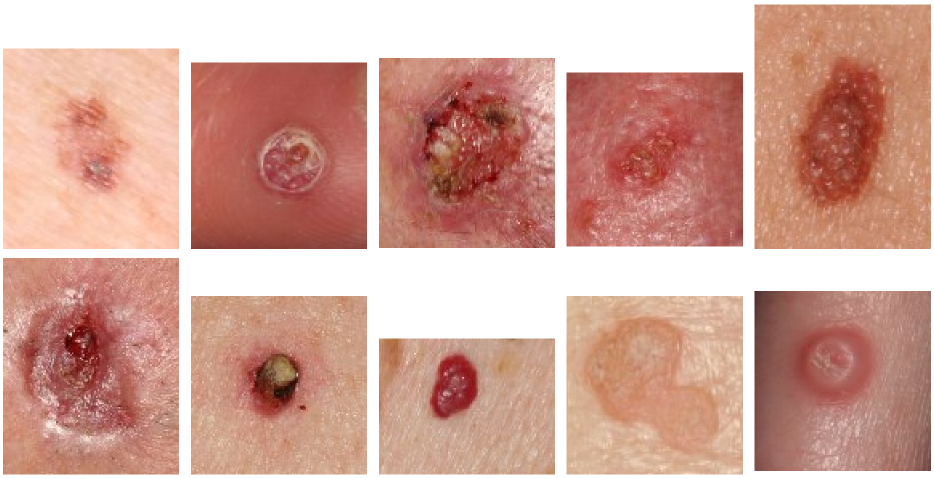


Figure 1.3: Some Images taken from the Dermofit Image Library, each image belongs to one of the ten classes.

variation in image sizes in dermofit dataset ranging between 177×189 and 3055×1630.

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; actinic keratosis, basal cell carcinoma, melanocytic nevus (mole), squamous cell carcinoma, seborrhoeic keratosis, intraepithelial carcinoma, pyogenic granuloma, haemangioma, dermatofibroma, and melanoma as listed in Table 1.1. An image sample from each of the ten skin cancer types in this dataset is shown in Fig. 1.3.

Chapter 2

Literature Review

This chapter introduces the basic concepts of the skin cancer image classification problem and computer vision based techniques. A number of steps involved in skin cancer classification and different kinds of feature extraction methods used in existing work are also described. Then a brief overview on machine learning and different approaches is provided, followed by a review of the Genetic Programming (GP) key components such as the individual representation, evaluation and genetic operators. An overview of the previous work on GP and non-GP based approaches to image classification, specifically skin cancer image classification, closes the discussion of this chapter.

2.1 Computer Vision

Computer vision is concerned with acquiring, understanding, processing and analyzing images [187, 176]. It involves the study and development of algorithms to achieve automatic visual understanding. It also aims at making decisions by extracting numerical or symbolic information from high-dimensional image data from the real-world [176]. The image data can be of various kinds such as video sequences, views from multiple cameras, and multi-dimensional data from a medical scanner. Hence, images play an essential role and represent the key component of this field. Simply, an image can be defined as a two dimensional matrix of values where each value reflects the intensity level of the corresponding pixel. The algorithms of computer vision aim at replicating the human vision system via electronically perceiving and analyzing the content of an image [28].

The introduction of digital cameras and other hand-held devices (e.g., cellphones and other smart devices) has greatly accelerated the task of gathering data in different domains. In order to gain useful information, the obtained data requires processing and analysis. In the field of computer vision, a large number of algorithms and methods have been proposed that aim at performing a variety of tasks such as image enhancement [164], recognition [155, 113], image restoration [27], motion analysis [81], scene reconstruction [101] and object detection [183, 38, 213, 58]. Some basic tasks and their definitions are briefly discussed below as they are closely related to machine learning and computer vision fields.

- **Object Detection**: It is concerned with finding instances of objects in an image against the background [213, 200]. For example, finding a person's face from a set of images of crowd photographs, and detecting tumors from a set of medical images. The task of detection is to locate the coordinates of the central pixel of each object irrespective of its type (i.e., the class label).

- **Object Classification**: Unlike object detection, classification assigns a class label to each object image [200]. In other words, each image is assumed to belong to a single class and the model tries to classify similar (based on shared characteristics) images into one class.

- **Object Recognition**: In the literature, object recognition and object classification have been often used interchangeably. However, here the term object recognition will be used to refer to the task of finding (detecting) and identifying (classifying) objects in an image. In other words, it is the task of performing both object detection and classification operations at the same time [155, 113].

2.1.1 Features in Computer Vision

A feature represents an attribute or a characteristics of an object. Hence, each image is represented by a set of features that are calculated manually or automatically, which can be used as input parameters/variables to perform a task, e.g., classification or regression. In computer vision, the intensity value of a pixel is typically used to represent a single feature which is known as a *low-level feature* [142]. Moreover, a group of pixels can also be used to represent a feature (e.g. the mean of pixel intensities of the eye region in a face image) which is known as a *mid-level feature* [142, 215]. Contrast to local feature, *global features* correspond to extracted features from the entire image. Mid-level features may also be termed as *local features* as they capture local information from the neighboring pixels of the central pixel. An example is local binary patterns (LBPs), which will be discussed in detail in the next section. For illustration, calculating the mean of the pixels only in the eye region is an example of a local feature, whereas calculating the mean from the entire image in an example of a global feature. In both examples, we have calculated the mean, i.e., a mid-level feature. A *high-level feature* in computer vision refers to a more descriptive and abstract features than simple pixel statistics such as the shape of face and eyes [215].

Feature manipulation is an important data pre-processing step, which transforms the input data into informative features to help improve the classification performance. Feature manipulation includes feature selection, feature construction, and feature extraction. One of the widely used feature extraction methods for image classification tasks is the Local binary Patterns (LBP), which is explained in more detail in the next section.

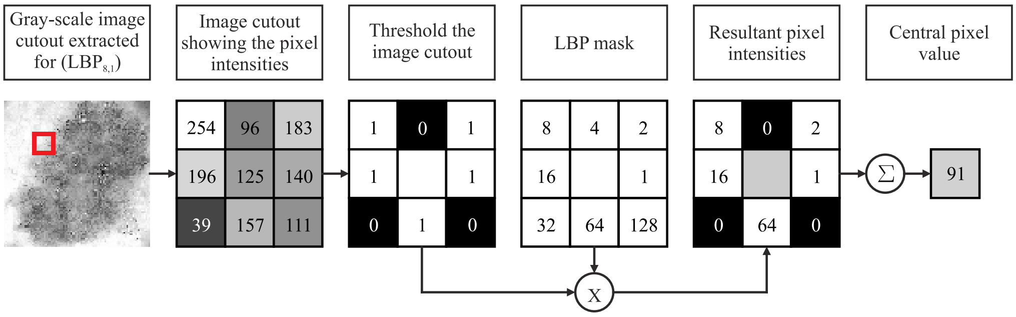

Gray-scale image cutout extracted for (LBP8,1)
Image cutout showing the pixel intensities
Threshold the image cutout
LBP mask
Resultant pixel intensities
Central pixel value
254 96 183
196 125 140
39 157 111
1 0 1
1 1
0 1 0
8 4 2
16 1
32 64 128
8 0 2
16 1
0 64 0
Σ
91
X

Figure 2.1: Step-by-Step procedure to generate $LBP_{8,1}$ code for image cutout (having 8 neighboring pixels and radius = 1) and get a decimal value of the central pixel.

2.1.2 Local Binary Patterns

The local binary patterns (LBP), proposed by Ojala et al. [145], is a dense image descriptor that has been used extensively for feature extraction in a wide range of computer vision applications. LBP works by scanning the image in a pixel-by-pixel fashion using a sliding window of fixed radius, where the value of the central pixel is computed based on the intensities of neighboring pixels that lie on the radius as depicted in Figure 2.1. LBP generates a histogram (i.e. feature vector) based on the computed values. The LBP operator is formally defined as:

$$LBP_{p,r} = \sum_{i=0}^{p-1} t(v_i - v_c)2^i \tag{2.1}$$

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

$$H\left(b\right) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \left(LBP_{p,r}\left(V_{i,j}\right) = b\right) \tag{2.2}$$

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

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

In this thesis, a histogram of uniform codes is generated; hence, there are 59 $(= 8 \times (7) + 3)$ LBP features for a single image. The window size of 3×3 pixels and a radius of 1 pixel ($LBP_{8,1}$) is used, which are the simplest and the most commonly used settings for extracting LBP features.

2.1.3 Skin Cancer Statistics

Skin cancer is the most common form of cancer, accounting for at least 40% of cases globally [1]. The most common type is NMSC (about 80% are BCC and 20% SCC cancers), which occurs in at least $2-3$ million people per year [69]. Basal-cell and squamous-cell skin cancers rarely result in death. In the United States, they were the cause of less than 0.1% of all cancer deaths. In 2012, melanoma occurred in 232,000 people, and resulted in $55,000$ deaths globally [190]. In 2019, there was an estimated number of $96,480$ new skin cancer cases which may lead to $7,230$ estimated deaths in the US [182]. When diagnosed early, skin cancer is highly curable with a survival rate of nearly 92% [182]. The three main types of skin cancer have become more common in the last 20 to 40 years, especially in white people from European origin [162].

Australia and New Zealand have one of the highest rates of skin cancer incidence in the world, almost four times the rates registered in the United States, the UK and Canada [190]. Around $434,000$ people receive treatment for NMSC and $10,300$ are treated for melanoma. Melanoma is the most serious form of skin cancer, which becomes life-threatening if not treated early [121]. Although the mortality is significant, when detected early melanoma survival exceeds 95% [121]. Melanoma is the most common type of cancer in people between $15 - 44$ years in both countries [102]. The incidence of skin cancer has been increasing [1]. In 1995, among Auckland residents of European descent, the incidence of melanoma was 77.7 cases per $100,000$ people per year, and was predicted at that time to increase further in the 21st century due to the effect of local stratospheric ozone depletion and the time lag from sun exposure to melanoma development [93]. According to World Health Organization (WHO) in 2018, the severity level of skin cancer incidence among males and females specially in Australia and New Zealand are shown in Figure 2.2 [3].

2.2 Skin Cancer Diagnosis

2.2.1 Skin cancer

Skin is the largest organ in the human body which consists of two principal layers [173]: epidermis and the dermis (see Figure 2.3). Skin cancers begin from the epidermis and are caused by the development of abnormal cells that have the ability to invade or spread to other parts of the body [93]. There are three main types of skin cancer: basal-cell skin cancer (BCC), squamous-cell skin cancer (SCC) and melanoma. They appear in respective skin cells such as basal cells, squamous cells and melanocytes, as shown in Figure 2.3. BCC and SCC are collectively called non-melanoma skin cancer (NMSC) [162].

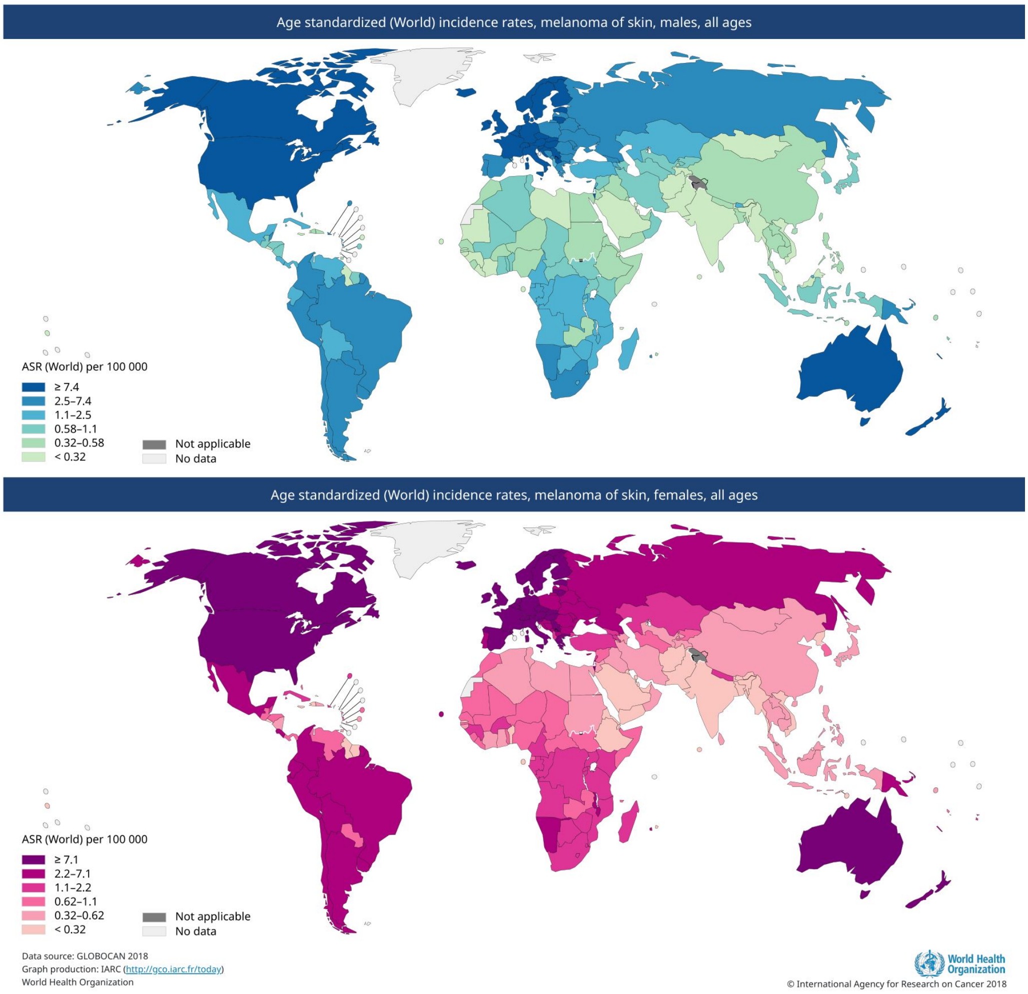

Age standardized (World) incidence rates, melanoma of skin, males, all ages
ASR (World) per 100 000
≥ 7.4
2.5–7.4
1.1–2.5
0.58–1.1
0.32–0.58
< 0.32
Not applicable
No data
Age standardized (World) incidence rates, melanoma of skin, females, all ages
ASR (World) per 100 000
≥ 7.1
2.2–7.1
1.1–2.2
0.62–1.1
0.32–0.62
< 0.32
Not applicable
No data
Data source: GLOBOCAN 2018
Graph production: IARC (http://gco.iarc.fr/today)
World Health Organization
World Health Organization
© International Agency for Research on Cancer 2018

Figure 2.2: Age-standardized death from melanoma in males (upper map) and females (lower map) per $100,000$ inhabitants in 2018 [3].

1. **Basal-cell cancer** grows slowly and often damages the tissue around it, but is unlikely to spread to distant areas or result in death. It appears as a painless raised area of skin, that may be shiny with small blood vessel running over it or may present as a raised area with an ulcer.

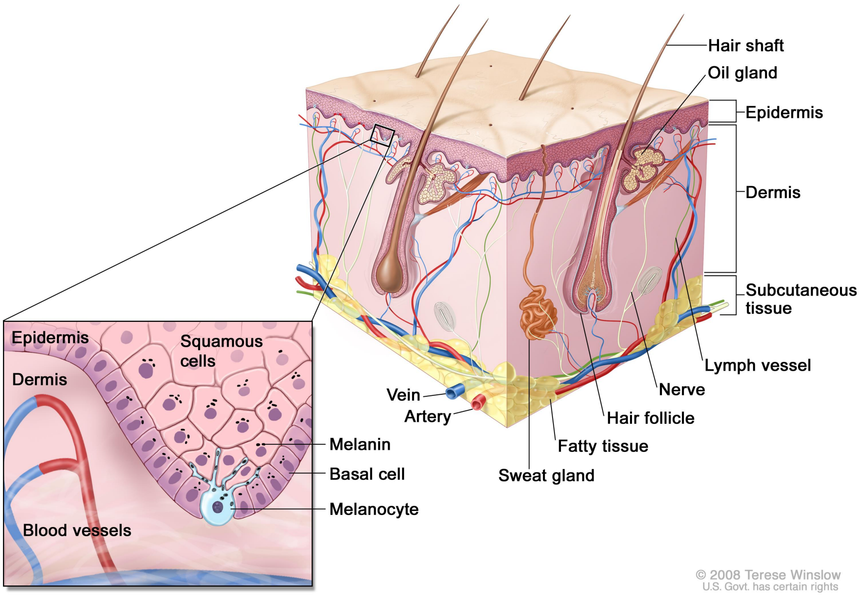


Figure 2.3: Anatomy of the skin, showing the epidermis, the dermis, and subcutaneous (hypodermic) tissue [102]. (Illustration used with permission, copyright 2008 by Terese Winslow.)

2. **Squamous-cell skin cancer** is more likely to spread to other parts and usually appears as a hard lump with a scaly top which can also produce an ulcer.

3. **Melanomas** are the most aggressive form of skin cancer [128, 36, 170, 162, 190, 102]. Signs include a mole that has changed in size, shape, color, has irregular edges, has more than one color, and is itchy and/or bleeds. If it is not diagnosed early, it is likely to invade nearby tissues and spread to other parts of the body.

2.2.2 Risk Factors

According to epidemiological evidence, exposure to ultraviolet (UV) radiation and the sensitivity of an individual's skin to UV radiation are risk factors for skin cancer, though the type of exposure (i.e. high-intensity exposure and short-duration exposure vs. chronic exposure) and pattern of exposure (i.e. continuous pattern vs. intermittent pattern) may differ among the three main skin cancer types [69]. More than 90% of cases are caused by exposure to UV radiation from the sun [190]. Risk factors include the followings [69]:

- Being exposed to natural sunlight or artificial sunlight (such as from tanning beds) over long periods of time.
- Having a fair complexion, which includes 1) fair skin that freckles and burns easily, does not tan, or tans poorly, 2) blue or green or other light-colored eyes, and 3) red or blond hair.
- Having a weakened immune system
- Having certain changes in the genes that are linked to skin cancer.
- Past treatment with radiation or body exposure to arsenic.

2.2.3 Treatment

In cancer detection, a biopsy is performed to determine presence or absence of a disease. A biopsy is defined as a medical test performed by a surgeon, involving extraction of sample cells or tissues for examination [173]. Treatment of skin cancer is generally conducted by surgical removal, but may less commonly involve radiation therapy or topical medications[173]. Treatment of melanoma may involve some combination of surgery, chemotherapy, radiation therapy, and targeted therapy [173]. For all these therapies, the following procedures may be used to treat skin cancer:

1. **Skin exam**: A doctor or a nurse performs visual analysis of the skin for bumps or spots that look abnormal in color, size, shape or texture.

2. **Skin biopsy**: All or part of the abnormal growth is cut from the skin and viewed under a microscope by a pathologist to check for signs of cancer. There are four main types of skin biopsies [190]:
    - **Shave biopsy**: A sterile razor blade is used to "shave-off" the abnormal growth.
    - **Punch biopsy**: A special instrument called a punch is used to remove a circle of tissue from the abnormal growth as shown in Figure 2.4.
    - **Incisional biopsy**: A scalpel is used to remove part of the growth.
    - **Excisional biopsy**: A scalpel is used to remove the entire growth.

2.2.4 Computer Vision Techniques for Skin Cancer Diagnosis

The advances of technologies in the areas of computer vision and machine learning have given us the ability to allow distinction of different skin cancers from the many benign mimics that require no biopsy. These new computer vision technologies not only allow earlier detection of melanoma, but also reduces the large number of needless, costly and painful biopsy procedures [128].

The general approach to developing a computer aided diagnostic (CAD) system for the diagnosis of skin cancer is to find the location of a lesion and also to determine a probability estimate of the incidence of a disease. The inputs to a CAD system are digital images obtained by epiluminescence microscopy (ELM) also referred as digital dermoscopy, with the possibility to add other acquisition systems such as ultrasound or confocal microscopy. Digital dermoscopy is a non-invasive diagnostic technique that permits evaluation of dermoscopic images of skin lesions [163].

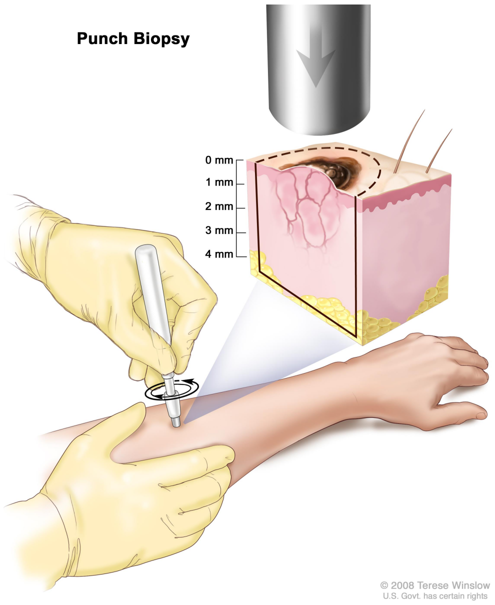


Figure 2.4: Punch biopsy [2]. A hollow, circular scalpel is used to cut into a lesion on the skin. The instrument is turned clockwise and counterclockwise to cut down about 4 millimeters to the layer of fatty tissue below the dermis. A small sample of tissue is taken to be checked under a microscope. Skin thickness is different on different parts of the body.

These images have a high resolution and are captured using a color video camera called dermatoscope, adapted for dermoscopy and connected to a computer [163]. This technique allows computer technology to be utilized for mass storage, indirect evaluation and management of images. In a CAD system, the first phase is the pre-processing of images that allows reducing the ill effects and various artifacts like dark corners due to camera calibrations, ink markers, bubbles due to presence of gel, color chart used for measuring the diameter, ruler marks and skin hair, which may be present in the dermoscopic images. It is followed by the detection of the lesion by an image segmentation technique to segment out the lesion area, which is used as the input to the next stage. Once the lesion is localized

[2] https://www.cancer.gov/types/skin/patient/skin-treatment-pdq

or in other words after *region detection*, different chromatic, gray scale, and *morphological* features will be extracted. These extracted features are then given to a classifier for classification [120].

Distinction of malignant melanoma images demands very fast pre-processing, feature extraction and classification algorithms. It is, therefore, necessary to make the best choice and to set the benchmarks for the diagnostic system development and validation [120]. The following steps are commonly followed in skin cancer diagnosis and classification.

1. **Image Acquisition**: Unaided visual inspection is a substandard approach to diagnosing skin cancer. Numerous imaging modalities are used to determine their suitability in ascertaining a correct diagnosis of skin cancer. These modalities include total cutaneous photography, digital dermoscopy, confocal scanning laser microscopy (CSLM), magnetic resonance imaging (MRI), ultrasound, optical coherence tomography (OCT), and multi-spectral imaging [120].

2. **Pre-processing**: Dermoscopy images often contain artifacts such as uneven illumination, dermoscopic gel, black frames, ink markings, rulers, air bubbles, and intrinsic cutaneous features. Such artifacts greatly hinder in distinguishing different structures, such as blood vessels, hairs, skin lines and texture. Everything that might corrupt the image and consequently affect the outcome of later stages, such as features extraction and classification, must be localized and then removed, masked, or replaced. Many approaches can be used such as image masking, resizing, cropping, hair removal [108, 100], and conversion from RGB color to gray scale image [34]. This pre-processing step is meant to facilitate image segmentation by filtering the image, and feature extraction by enhancing its important features [120].

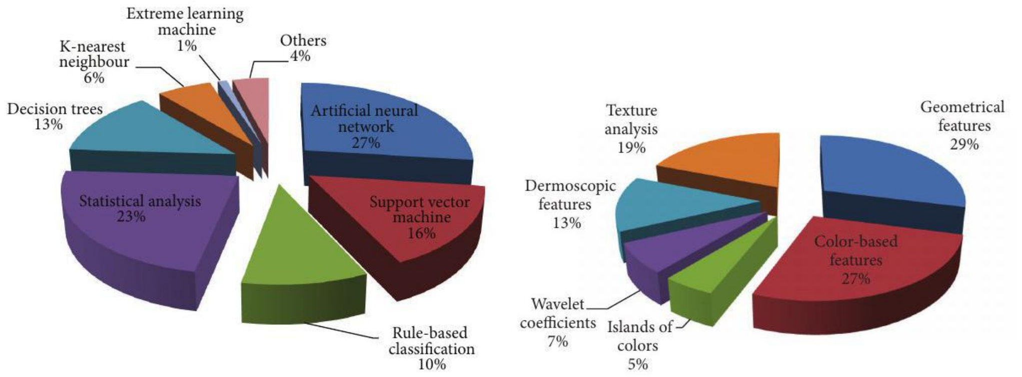
3. **Segmentation**: Segmentation refers to partition an image into disjoint regions that are homogeneous with respect to a particular prop-

erty such as texture, color and luminance [52]. The goal of segmentation is to simplify the representation of an image into something more meaningful and easier to analyze [120]. The main problem of segmentation is that it often fails to detect edges when edges are not enclosing the object completely [5]. Moreover, the border of skin lesions have different characteristics as compared to edges in object detection images. The skin lesion images have morphological structures including tiny blood vessels that make the borders irregular and not having a fine boundary. These properties of lesion images make the segmentation task more difficult.

4. **Feature Extraction**: Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning steps, and in some cases leading to better human interpretations [23]. For image analysis, the goal of feature extraction is dimensionality reduction, and extracting informative features from pixels [16].

5. **Feature Selection**: Feature selection methods aim to select a subset from the entire set of available features. Mainly, feature selection is concerned with selecting relevant features and neglecting redundant, irrelevant or less important features [112]. Feature selection must not be confused with feature extraction. The former is only selecting a subset of the original features and does not perform any transformation on the original values, whereas the latter results in creating new features. It is important to select a reasonably reduced number of useful features while eliminating redundant, noisy, or irrelevant features. However, it is necessary to make sure that such reduction in the number of features does not cost loss of crucial information.

6. **Classification**: The classification phase of the diagnostic system is in charge of making the inferences about the extracted information



(a) (b)

Figure 2.5: Pie Charts showing a) Classification methods used by existing diagnostic systems, and b) Feature distribution used in dermoscopic studies in the literature [120].

in the previous phases in order to produce diagnostic about the input image [64]. According to a study [120] for reviewing the classification algorithms adopted for skin cancer classification, the most commonly used classification algorithms are Support Vector Machines (SVMs), Artificial Neural Network (ANNs) and statistical approaches. This is presented in Figure 2.5(a), which shows the percentage of classification methods used by existing diagnostic systems.

2.2.5 Features in Skin Lesion Images

During clinical examinations, dermatologists use certain criteria to determine the type of skin lesion. The most commonly adopted methods for identifying these lesions during clinical screening procedures by non-dermatologists are the ABCDE criteria [125] and the Glasgow 7-point checklist [116]. It is noted that these methods for diagnosing skin cancer from images are used to determine only whether suspicious lesions could be cancerous. However, the actual diagnosis is carried out by a pathologist, after such suspicious lesions are excised (biopsied). In the literature, various kinds of image features have been used for skin cancer image classification such as wavelet or frequency-based, geometrical, color, and texture features. Figure 2.4(b) illustrates the distribution of these features used in dermoscopic studies, which reveals that color and geometrical features are most widely adopted and studied in the literature [120].

The ABCDE Rule of Dermoscopy

Originally the ABCD criteria, proposed in 1985 by Friedman et al. [77], have been widely adopted in clinical practice, mostly due to its simplicity of use [102]. ABCD defines the diagnosis of a lesion based on its Asymmetry, Border irregularity, color variegation and Diameter.

- *Asymmetry* refers to whether the two halves of the lesion have a similar appearance in terms of color, texture, shape and size. Cancerous lesions are asymmetrical in shape whereas benign lesions are symmetric [25].

- *Border* irregularity describes the firmness of lesion borders whether jagged, blurred or has a fine boundary. Lesions with tumor are characterized by irregular boundary whereas benign lesions are circular shaped [25].

- *Color* variegation suggests presence and absence of some specific colors such as white, red, light-brown, dark-brown, blue-gray and black. Melanoma has this unique feature of color variegation whereas benign lesions have a uniform color [25].

- *Diameter* of a lesion, generally greater than 6 mm (milli meter, 1mm = 10^{-3}m), is considered cancerous.

Later, in 2004, Abbasi et al. [6] expanded the ABCD criteria to ABCDE by incorporating the "E" for an "evolving" lesion over time, which includes changes in features such as size, shape, surface texture and color.

The 7-Point Checklist Method

The 7-point checklist method contains 7 criteria: 3 major (changes in shape, size and color) and 4 minor (diameter $\geq$ 7 mm, crusting or bleeding, inflammation and sensory change). The seven points in this checklist include 1) Atypical pigment network, 2) Gray blue areas, 3) Atypical vascular pattern, 4) Radial streaming (streaks), 5) Irregular diffuse pigmentation (blotches), 6) Irregular dots and globules, and 7) Regression pattern [24].

Some of these characteristics visual appearance are shown in Figure 2.6. As marked with yellow ovals, the *pigment network* is a grid-like network consisting of pigmented lines (brown or black). This structure has a crucial role in the distinction between melanoma and NMSC lesions as discussed in [12]. Dermatologists analyze the images and mark the pigment network as "typical" or "atypical". Marked with red circle in Figure 2.6 is the area highlighting the presence of *dots/globules*. Dots/globules are spherical or oval, different in size, black, brown or gray structures (dots usually smaller than globules). *Streaks* (shown in green circle) are finger-like projections of the pigment network mostly found at the border of the lesion. The presence of *blue-whitish veil* (marked with blue arrow) is a strong malignancy indicator. It is an opaque, irregular blue pigmentation with an overlying, white, ground-glass haze [123]. The dermoscopic structures (streaks, dot/globules, and blue-whitish veil) are labeled as "present" or "absent" in each image of the PH^2 dataset.

2.3 Machine Learning

Machine learning is a broad research field that explores the study and construction of algorithms that can learn from and make predictions on data [126]. It is a sub-field of computer science, evolved from the study of pattern recognition and computational learning theory in artificial in-

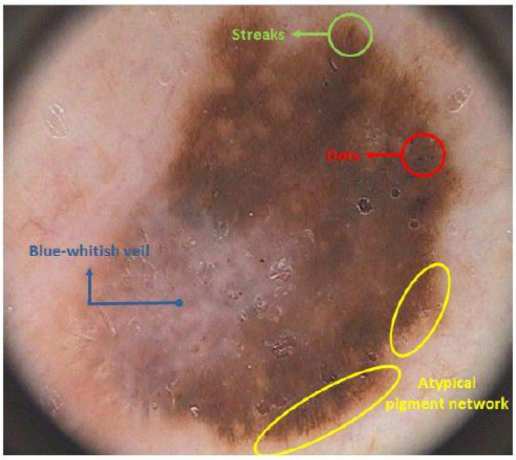


Figure 2.6: Identification of some dermoscopic features based on 7-Point Checklist method [123].

telligence [126]. The key factors of machine learning are representation and generalization [174]. While the former is concerned with the representation of data and various functions evaluated over this data, the latter represents the ability of the system or the model to handle unseen data based on the knowledge gained from the seen examples. Generally, algorithms of machine learning can be divided into the five following groups [23, 137]: (1) supervised learning; (3) semi-supervised learning; (2) unsupervised learning; (4) reinforcement learning; and (5) learning by knowledge transfer.

1. **Supervised Learning**: A supervised learning algorithm aims at defining a generalized function that is capable of predicting an output for unseen data relying on the information of the previously seen data. Classification and regression represent the most well-known example applications of supervised learning [129]. In the case of classification, the system takes the available list of features or description of the inputs (training instances) and predicts the class label for each of them. The desired (known) outputs are used to guide the system during the training phase. Typical methods of this approach include

ANNs, Decision Trees, and Naïve Bayes (NB).

2. **Unsupervised Learning**: Unsupervised learning algorithms are mainly designed to handle situations where the class labels (desired outputs) of the instances are unknown. In contrast to supervised learning, the core objective of unsupervised learning algorithms is to find patterns in the data instead of estimating a generalized function that maps an input to an output [23]. Clustering represents a typical example of this type of machine learning method. The aim of clustering algorithms is to categorize objects into a number of groups (clusters) based upon the similarity and dissimilarity between attributes of those objects. Some of the widely used algorithms to perform clustering include k-means clustering, and hierarchical clustering.

3. **Semi-supervised Learning**: The semi-supervised learning methods combine the schemes of both supervised and unsupervised learning methods, in which unlabeled data along with labeled data are used to train a model. Generally, the number of labeled instances is smaller than the number of unlabeled instances. A typical example of semi-supervised learning methods is transductive support vector machine (TSVM) [92].

4. **Reinforcement Learning**: Reinforcement learning develops an agent (e.g. computer program) that aims at exploring an environment and takes an appropriate action in which some cumulative reward is to be maximized. This learning approach is mainly developed based on the principle of reward and punishment, such that an agent will gain more rewards by taking more correct actions, and penalized whenever an incorrect action is taken. A well-known technique of reinforcement learning is the Temporal Difference (TD) [194] learning method.

5. **Transfer Learning**: The idea of transfer learning came from the hu-

man ability to rely on previously obtained knowledge to learn a new category or to categorize a new object. This is achieved by simply storing incremental new informative knowledge of this object or category. In machine learning, this process is known as knowledge transfer or transfer learning [151]. In contrast to human ability of learning to classify objects, the majority of existing classification algorithms demand abundant examples in order to learn every single category or group of objects [90]. In traditional machine learning, the assumption is that both of the source and target domains or tasks are the same (e.g. drawn from the same distribution). The key point of transfer learning is to define common knowledge among source and target domains that may help improve the performance of the model. In medical domain, where in most cases the data available is limited [170], this approach is favorable as it combines knowledge from both domains to achieve better generalized model for classification of the target domain. More details can be seen from [151].

2.3.1 Classification

Classification is defined as the identification of which of a set of categories a new observation belongs, on the basis of a training set of data. Classification represents an important task in a wide variety of fields such as computer vision and pattern recognition [129]. Image classification aims at categorizing images into different groups based upon their contents such that images of an object or a scene are categorized under one group that are different from instances of other groups. Common to classification problems are the following key terms:

- **Training**: The process by which a learning algorithm uses observations (also called instances) to learn a new classifier is called the *training process* and the observations or instances used in the training process are collectively referred as the *training set*.

- **Testing**: The process of evaluating the performance of the learned classifier is called *testing*. The instances used to test the learned classifier are collectively called the *test set*. It is important to mention that the instances in the test set are from the same problem domain as the training instances, however, they remain unseen during the training process.

- **Validation**: Validation set includes a set of instances used to monitor the training/learning process to control overfitting.

- **Generalization**: Generalization refers to how well the concepts learned by a machine learning algorithm apply to specific instances not seen by the model when it was learning [48]. The goal of a good machine learning model is to generalize well from the training data to test data from the same problem domain. This allows to make predictions in the future on data the model has never seen.

- **Overfitting**: Overfitting refers to a model that learns the training data too well. It happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on test (new) data [48]. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model which do not apply to new data and negatively impact the models ability to generalize.

- **Underfitting**: Underfitting refers to a model that can neither model the training data nor generalize to test (new) data [48]. An underfit machine learning model is not a suitable model and is obvious as it has poor performance on the training data.

- **n-fold Cross Validation**: A dataset is usually re-sampled into a training set and a test set. In this case, a learning algorithm learns different kinds of rules from the training set, and apply these rules to the test set to evaluate the learned model, hence, the algorithm becomes 2-fold. However, many problems (mostly in the medical domain) have a small number of available instances in the dataset and sometimes the dataset is unbalanced. In such datasets, following 2-fold re-sampling method will lead to biased results. Therefore, it is necessary to apply some appropriate re-sampling methods, such as *n-fold cross-validation* [130]. In n-fold cross-validation, the dataset is randomly partitioned into *n* folds (partitions) and the folds are near-equal size. The folds are selected in such a way that the proportion of instances from different classes, remains the same in all folds. Next, a single fold of the *n* folds is retained as the test set, and the remaining *n*-1 folds are used as training set. This process is then repeated *n* times, with each of the *n* folds used exactly once as the test set. After getting results from *n* experiments, the average of these *n* results are taken to make an estimate of classification performance on that dataset. An extreme case of n-fold cross-validation is *leave-one-out cross-validation (LOOCV)*, which uses a single instance from the dataset as the test set, and the remaining instances as the training set. In other words, if the dataset has *k* instances, then the results of *k* experiments are averaged to estimate the classification performance. The advantage of such re-sampling methods is that all instances are used for both training and testing, and each instance is used for testing exactly once.

2.3.2 Learning Paradigms

Generally, there are five learning paradigms, also known as tribes of AI, in machine learning: (1) Instance-based, (2) Induction-based, (3) Connectionist/Neural Learning, (4) Statistical-based, and (5) Evolutionary Learning or Computation (discussed in section 2.4) [126].

Instance-based

Instance-based learning (sometimes referred as memory-based learning) is a family of learning algorithms that, instead of performing explicit generalization, compares new problem (test) instances with instances seen in training, which have been stored in memory [168]. It is called instance-based because it produces hypotheses directly from the training instances. Examples of instance-based learning algorithms are the k-nearest neighbor algorithm, kernel machines, and radial basis function (RBF) networks [129]. These algorithms store (a subset of) their training set when predicting a class for a new test instance; they compute distances or similarities between the instance being evaluated and the training instances to make a decision.

k-Nearest Neighbor (k-NN): The most basic instance-based method is the k-nearest neighbor algorithm [129]. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance [129]. When k-NN is used for classification, it calculates the distances between the test instance and all instances in the training set. Based on this distance calculations, the test instance is classified by assigning the class which is most frequent among the k training samples nearest to that test instance. Some of the most commonly adopted distance measures in k-NN are the Euclidean distance and the Manhattan distance.

Induction-based

Induction-based is a type of learning through observation of different objects or data, building general concepts by observing a set of instances [39].

Decision Trees: Decision tree learning is one of the most widely used methods for inductive inference [129]. It is a method for approximating discrete-valued target functions, where a decision tree represents the learned function. It classifies each instance by sorting it down the tree from the root node to more than one leaf nodes, which provides the classification of that instance. Each node of the tree represents a test of an attribute of the instance, and each branch descending from that node represents one of the possible values for this attribute. An instance is assigned a class by starting at the root node, testing the attribute specified by this node, then selecting that branch which corresponds to the value of the attribute for that particular instance. This procedure is repeated for the sub-tree at the new node. [129]. Classifying a particular instance may involve evaluating only a small number of the attributes depending on the length of the path from the root of the tree to the appropriate leaf node [126]. Decision trees are more likely to face the problem of data over-fitting because it tries to split the data until it makes pure sets. This problem can be resolved by using its extension i.e., *J48* by using Pruning. Pruning reduces the size of decision trees by removing sections of the tree that provide little information to classify instances. Pruning also reduces the complexity of the final classifier, and hence improves predictive accuracy by reducing overfitting.

Decision trees have been successfully applied to a broad range of tasks, such as learning to classify medical patients by their disease, loan applicants by their likelihood of defaulting on payments, and equipment malfunctions by their cause [129, 126]. Decision trees have several advantages; 1) simple to understand and interpret (important insights can be investigated based on experts describing a situation, its alternatives, probabilities, and costs); 2) allows the addition of new possible scenarios, and also helps determine worst, best and expected values for different scenarios, and 3) can be combined with other techniques such as logistic regression [88] and NNs [91]. There are some limitations of decision tree also: 1) unsuitability to predict values of a continuous attribute, 2) possibility of duplicating same sub-tree on different paths, and 3) can become very complex, particularly if many values are uncertain and/or unequally informative attributes (with or without interactions) and irrelevant attributes co-exist [65].

Connectionist/Neural Learning

Neural network learning methods provide a robust approach to approximating real-valued, discrete-valued, and vector-valued target functions [129]. Artificial neural networks (ANNs) are inspired partially by the observation that biological learning systems consists of very complex webs of interconnected neurons in animals brain. An ANN is based on a collection of connected units called *artificial neurons*, which is analogous to axons in a biological brain. Each connection between neurons transmits a signal to another neuron. The receiving neuron can process the signal(s) and then signal downstream neurons connected to it. More often, neurons are organized in layers where different layers may perform different kinds of transformations on their inputs. Signals travel from the first (input) layer, to the last (output) layer, possibly after traversing the layers multiple times. The original aim of the neural network approach was to solve problems in the same way that a human brain would, however, with time, attention deviated on matching specific mental abilities [46]. As a consequence, it leads to deviations from biology such as back-propagation (defined as passing information in the reverse direction) and adjusting the network to reflect that information.

NNs have shown to be successful in many practical problems such as learning to recognize handwritten characters [105], learning to recognize spoken words [104], learning to recognize faces [63], and learning to recognize skin cancer [72] and breast cancer [189] from images.

Multi-Layer Perceptron (MLP): MLP is a category of feedforward ANNs. It consists of at least three layers of nodes, where each node is a neuron that uses a nonlinear activation function, except nodes of the nodes in the first layer which are used directly as inputs. MLP typically uses a supervised learning technique called back-propagation for its training [104]. MLP consists of multiple layers and has nonlinear activation which distinguish it from a linear perceptron. MLP has the ability to classify data that is not linearly separable.

Having more than three layers (an input and an output layer with one or more hidden layers) and nonlinearly activating nodes, MLP is a deep neural network [137]. MLPs are fully connected; each node in one layer connects with a certain weight to every node in the following layer. Learning occurs in the perceptron by changing these connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result [105]. MLPs were a popular machine learning solution in the 1980s, having applications in diverse fields such as image recognition, speech recognition, and machine translation software [202], but thereafter faced strong competition from support vector machines (SVMs). Due to the successes of deep learning, interest in back-propagation networks returned [62].

Statistical Learning

Statistical learning theory is a framework for machine learning which is derived from the fields of statistics and functional analysis [87]. It deals with the problem of finding a predictive function based on given data. It has led to successful applications in fields such as computer vision, bioinformatics and speech recognition [87]. The most common statistical methods in machine learning are Support Vector Machines (SVMs) and Naïve Bayes (NB).

Support Vector machines (SVMs): SVMs are one of the supervised learning methods based on the statistical learning theory. The goal of SVMs is to map the input data to high dimensional space, constructing a hyperplane or a set of hyperplanes, which are used to create decision boundaries for the task of classification [89]. SVMs aim to maximize the distances between the hyperplanes and the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [89]. Instances are classified based on what side of these hyperplanes they fall on. While the basic training algorithm can only construct linear separators, different kernel (i.e., linear, polynomial, radial basis function, and sigmoid) functions can be used to include varying degrees of nonlinearity and flexibility in the model [120].

SVMs have several advantages over the classical classifiers such as decision trees and neural networks. The support vector training mainly involves optimizing a cost function, which eliminates the risk of getting stuck at local minima as in the case of back-propagation neural networks [89]. The main disadvantages of SVMs are the high algorithmic complexity and extensive memory utilization when dealing with large-scale tasks [89].

Naïve Bayes (NB): Bayesian theorems are based on a probabilistic approach to inference [129]. It is based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data. Bayesian reasoning directly manipulates probabilities in order to estimate the behavior of data and provides a framework for analyzing the operation of other algorithms that do not explicitly manipulate probabilities [129]. For example, Bayesian analysis is used to justify a key design choice in neural network learning algorithms: choosing to minimize the sum of squared errors when searching the space of possible neural networks [129]. Similarly, Bayesian perspective is use to analyze the inductive bias of decision tree learning algorithms that favor short decision trees [129].

Naïve Bayes classifiers are the most common bayesian classifiers. NB makes use of the assumption that all input features are conditionally independent, which is why termed as naïve. This assumption can not be applied to many real-world problems, where there are interdependency between the input features, which may cause two-way or multi-way feature interactions. Another practical difficulty is that they typically require initial knowledge of many probabilities. If these probabilities are not known prior to computation, they are estimated based on prior knowledge, previously available data, and assumptions about the form of the underlying distributions [206].

Ensemble Learning

Ensemble learning is a method of learning that builds a collection of base classifiers in the training process for a classification task. Thereafter, base classifier predictions are combined to identify new instances through the application process. It has been shown that an ensemble of classifiers is more effective than any of the base classifiers making up the ensemble.

2.4 Evolutionary Computation

Evolutionary computation (EC), inspired by the theory of natural selection and genetic inheritance, is a sub-field of artificial intelligence and refers to the family of algorithms for global optimization inspired by biological evolution [159]. They are a family of population-based trial and error problem solvers with a stochastic optimization character [126]. The increasingly active field of EC provides valuable tools, to problem solving, machine learning, and optimization [53]. In particular, industrially relevant fields, such as signal and image processing, computer vision, pattern recognition, industrial control, scheduling and timetabling, telecommunication, and aerospace engineering, are using EC techniques to solve complex problems [53]. Under the EC umbrella, there are evolutionary algorithms (EAs), swarm intelligence (SI) algorithms and other optimization techniques such as differential evolution and memetic algorithm.

Evolutionary Algorithms (EAs) is a sub-field of EC, which are population based optimization algorithms. EAs are based on mechanisms inspired by biological evolution including genetic operators like selection, reproduction, mutation and crossover. In EAs, each candidate solution is represented as an individual in the population. The fitness or evaluation measure determines the goodness of each individual. Evolution of the population then takes place after the repeated application of the mentioned genetic operators. Examples of EAs include genetic algorithms (discussed here), genetic programming (discussed in detail in Section 2.5 on page 56), evolution strategy, and evolutionary programming.

Genetic Algorithms

Genetic algorithms (GAs) provide an approach to learning that is based on biological evolution [129]. Candidate solutions are often encoded as bit strings whose interpretation depends on the application. The search for an appropriate solution begins with a randomly generated population, or collection, of initial solutions. Members of the current population give rise to the next generation population by applying operations such as mutation and crossover, which are modeled after processes in biological evolution [129]. At each generation, the solutions in the current population are evaluated based on a measure of fitness, with the most fit solutions selected probabilistically as parents for evolving the next generation.

GAs have been explored widely and applied successfully to a variety of learning and optimization problems [206]. For example, they have been used to learn collections of rules for robot control [119] and to optimize the topology and learning parameters for ANNs [25]. They can search spaces of solutions containing complex interacting parts, where the impact of each part on overall solution's fitness may be difficult to predict [129]. However, GAs tend to be computationally expensive but they can be easily parallelized taking advantage of powerful computer hardware, hence, resulting in decreased costs [129].

Another area of EC is *Swarm Intelligence (SI)* algorithms which are inspired by the collective intelligence of social insects. A swarm is defined as a population of interacting individuals that is capable of optimizing global objectives through collaborative search. Here, the intelligence lies in the networks of interactions among individuals, between individuals and the environment [98]. There is a general stochastic tendency in a swarm for individuals to move towards a center of mass in the population, which results in convergence on an optimal solution [98]. The most common optimization techniques in SI are Particle Swarm Optimization (PSO) [67] and Ant Colony Optimization (ACO) [118].

Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) is a population based stochastic optimization technique inspired by social behavior of birds flocking or fish schooling [99]. In PSO, each candidate solution is encoded as a particle moving in the search space according to simple mathematical formula to update particle's position and velocity. Each particle remembers its local best known position. Hence this collection of particles known as swarm, searches for the optimal solution by updating the position of each particle based on the local best known position of its own and its neighboring particles [99]. PSO is a simple but powerful search technique. It is a metaheuristic as it makes few or no assumptions about the problem being optimized and hence, can search very large spaces of candidate solutions. However, metaheuristics do not guarantee that an optimal solution is always found [98].

2.5 Genetic Programming (GP)

Genetic Programming (GP) is an EC algorithm based on Darwinian principles of biological evolution and natural selection that automatically explores the solution space to evolve a computer program (model/solution) for a given problem [103]. Fundamentally, GP consists of a set of operators and a fitness measure which is used to evaluate the performance of an evolved program.

GP performs an implicit feature selection via randomly selecting features at its terminal nodes during the evolutionary process [122]. The ability of GP to handle complex problems represents a key motivation for many researchers to utilize it to address different tasks, such as object detection [9, 183, 38, 213, 211, 212], feature selection [131, 10, 198, 136, 107], feature construction [198, 138, 110, 11, 139], classification [110, 17, 15, 58, 195, 20, 143, 44, 43], regression [83, 55] and knowledge transfer [90, 51].

The rest of this section explains the key components of GP to clarify the role of each component.

2.5.1 Overview

The main idea of GP process is based on the concept of "Survival of the Fittest" in which a number of individuals (computer programs) evolve gradually to gain improved performance. The process starts by randomly creating a predefined number of initial solutions via using different combinations of the elements in the function and terminal sets. Adopting a carefully designed fitness measure, the performance of each individual is calculated. GP uses reproduction, crossover, and mutation operators to produce new individuals for the population of the next generation from those of the population of the the current generations. Individuals with better fitness values are more likely to be selected for participating in the mating process to generate the population of the subsequent generation. The overall process of GP framework is outlined in the following steps.

1. A predefined number of individuals are randomly created using functions and terminals to create the initial population.

2. The fitness value (performance) of each individual is calculated based on the fitness measure.

3. Following steps are repeated until a termination criterion is met.

    - Select one or more individuals based on their fitness values.
    - Generate one or more new individuals from the selected ones by applying GP operators.
    - Put the newly generated individuals into the next generation.
    - Calculate the fitness value for each of those newly generated individuals.

4. The best evolved individual (the one with the best fitness value) represents the best evolved solution to the problem.

2.5.2 Representation

The most commonly adopted representation for GP individuals is tree-based in which an individual is made up of a root node, a number of internal nodes, and some leaf nodes [160]. The root and non-terminal nodes consist of elements from the function set, whereas the leaf or terminal nodes are taken from the list of terminals. Each function node represents an operation that needs to be performed on the list of input values (i.e. the output values of its child nodes). Function nodes can be as simple as arithmetic operators (e.g. addition, subtraction, and multiplication), or more complex (e.g. loop structures). The terminal nodes represent the leaves of the individual tree, and they do not have inputs. A terminal node is either a randomly generated constant value in a predefined interval, or a value selected from the list of available inputs (feature values). The tree-based representation of the equation$((x_1 \times x_2) - x_1) \div ((x_1 + x_2) \times \sqrt{x_2})$ is presented in Figure 2.7, which has five functions $\{+, -, \times, \div, \sqrt{}\}$ and two terminals $\{x_1, x_2\}$.

The program shown in Figure 2.7 produces a numerical value as it represents the type of output value based on input values at the leaf nodes in the tree. In this example the type of input and output values of all terminal and function node is numerical. However, different applications require the use of different types, such as boolean and string. For example, logical greater than or less than functions compare two numerical values and return a boolean value (either true or false). Hence, a variant of GP representation known as Strongly-typed GP (STGP) [133] has been introduced to address this problem. The use of STGP allows the incorporation of data types and their constraints where different nodes can have different types of input arguments and return different types of outputs. Although tree-based representation is the most common type of representation for GP individuals, it is not the only one. A number of researchers have investigated different types, such as Linear GP (LGP) [76], Cartesian GP (CGP) [127], Grammar-guided GP [165], and Modi-GP [214].

With its flexible representation and different from single-tree GP which evolves one tree in an individual, GP can evolve multiple trees in a single individual to solve a particular problem, referred as multi-tree GP (MTGP) [150]. In the literature, MTGP has been studied for a wide range of applications including automatically evolving image descriptors for texture image classification [19], constructing features to create benchmark datasets [111], self-assembling swarm robots [106], and multi-class classification [135].

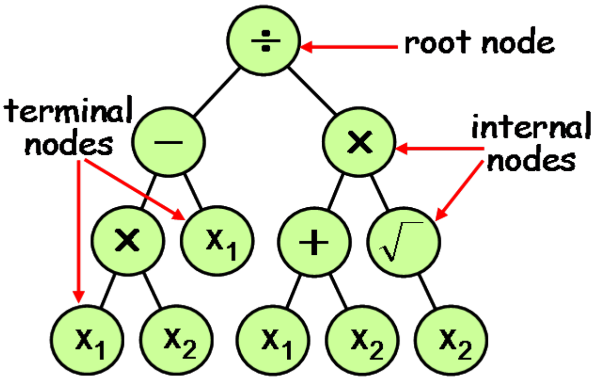


Figure 2.7: GP tree-based representation of the equation $((x_1 \times x_2) - x_1) \div ((x_1 + x_2) \times \sqrt{x_2})$.

2.5.3 Initialization

GP starts the process by randomly generating a number of initial solutions to create the initial population. In tree-based GP, the *minimum-* and *maximum-depth* of the generated individuals imply important restrictions. The *maximum-depth* of a tree is defined as the longest path that starts at the root node and the farthest leaf node. Two of the simplest and commonly known methods that have been extensively adopted in the literature are the *full* and *grow*. The *full* method generates an individual by randomly selecting elements from the function set until the maximum allowed depth of the tree is reached. Then only elements of the terminal set are drawn at random to populate the leaf nodes. The *grow* method is similar to the full method, however, it is allowed to select from both sets (function and terminal) as long as the *maximum-depth* has not been reached. Therefore, the branch stops growing at that point where a terminal node is selected.

To ensure having individuals vary in shape and size, a third method based on the combination of the two has been devised known as *ramped half-and-half* [103, 160]. In this method, half of the individuals in the population are generated using the full method and the other half is generated using the grow method to create the initial population. The ramped half-

and-half method is the most widely used method as compared to other methods [16, 55, 57, 58, 109, 153, 177].

2.5.4 Evaluation

Similar to other EC methods, one of the most important parts of the GP system is the evaluation or fitness measure [160] to evaluate the goodness of an evolved solution. Evaluating the evolved program on a number of fitness cases reflects the goodness of the program to handle different scenarios. The decision of which fitness function to use is critical as GP relies on it to assess the performance of the evolved program. Moreover, the design of this function is heavily dependent on the problem. For example, if the task is to perform classification, then the fitness function can be the accuracy (the percentage ratio of correctly classified instances to the total number of instances). As another example, if the task is to solve a regression problem, then a good fitness measure can be based on how far is the predicted value from the expected value. Moreover, if the dataset is unbalanced, i.e., having a different number of instances of each class, then designing suitable fitness measure is crucial for accurately evaluating the evolved solutions.

2.5.5 Selection Methods

Selection methods decide which individuals will participate in creating individuals of the subsequent generation. The fitness value is used to assess the chances of an individual to be selected. Hence, better individuals (e.g. individuals having good fitness values) are more likely to be selected than inferior ones. One of the earliest selection methods is the *fitness proportionate selection*, also known as *Roulette wheel selection*. This method works by randomly selecting an individual each time based on the distribution of all fitness values of the current generation of individuals. One drawback of the Roulette wheel method is that individuals having bad fitness values may never get a chance to participate in populating the next generation. To deal with this problem, *Tournament selection* has been proposed. In this method, a predefined number (termed as Tournament size) of individuals are randomly taken from the population. Then the individual with the best fitness value among them is selected. This method gives all individuals (regardless of their fitness values) an equal chance to be drawn in the first step. However, this method introduces an extra parameter to be set which is the tournament size. Using a large tournament size will reduce the chances of inferior individuals to be selected; however, the use of a small tournament will increase the probability of inferior individuals to be selected.

2.5.6 Genetic Operators

In GP, the individuals of the current generation are used to populate the subsequent generation through applying a number of operators. Those operators aim at generating new individuals (children) by utilizing the genetic materials of the current population (parents). There are three operators in GP: 1) reproduction, 2) crossover, and 3) mutation. Each of these operators is applied based on a user-defined probabilities, where the sum of the three probabilities is 1.

Reproduction

Reproduction performs a copy operation of a predefined number of the individuals (determined by the rate of this operator) from the current generation to the next one. Reproduction does not ensure the selection of top individuals; it copies the selected individuals by the selection method. However, when it selects the top ranked individuals, it is termed as *elitism*, which aims at maintaining the achieved level of performance and prevent it from degrading in the subsequent generations. This operation ensures that the next generation is at least as good as the current one. When the

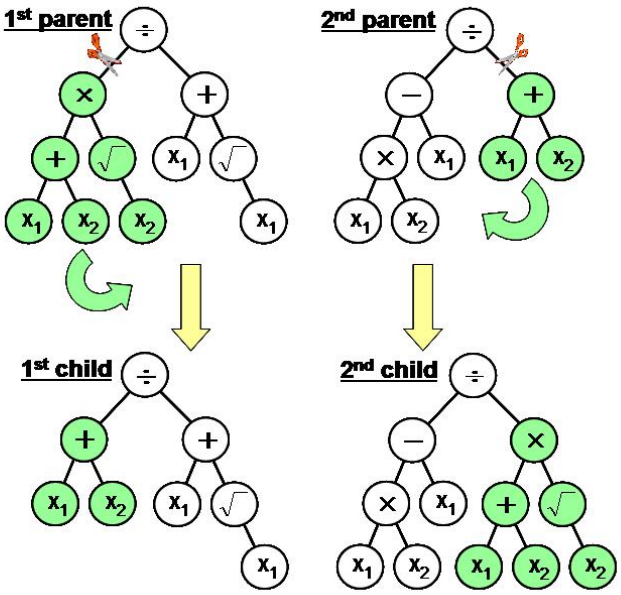


Figure 2.8: Example illustrating the crossover operator.

probability of performing reproduction or elitism operation is set to 1 (i.e. 100%) and both of the crossover and mutation are 0%, the system will copy all the individuals to the subsequent generations, hence, the individuals of the final generation will be identical to those of the first generation. Therefore, the ratio of this operator must be set to a very low value to make the system flexible enough to explore the solution space.

Crossover

The *crossover* operator generates new individuals (children) by exchanging the genetic materials from the two existing individuals (parents). The parent individuals are first selected using one of the selection methods. Next, a crossover-point (indicated with scissors in Figure 2.8) is chosen from each tree, and the sub-trees are swapped at the crossover points. The crossover operator has been widely studied and various methods have been proposed, which includes one-point crossover, two-point crossover, uniform and half-uniform crossover, cut and splice crossover, and three-

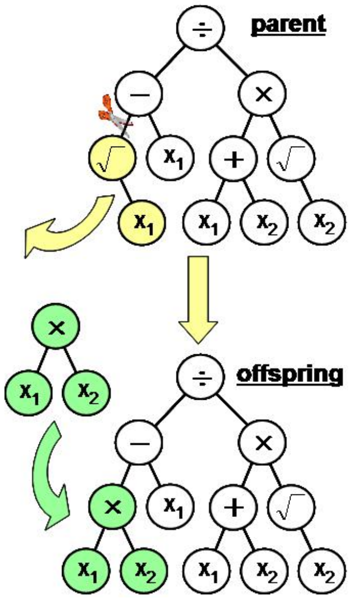


Figure 2.9: Example illustrating the mutation operator.

parent crossover [160].

Mutation

Like crossover, *mutation* uses existing individuals to generate new ones. However, in mutation, only one parent is selected from the current population and the other parent is randomly generated using one of the initialization methods. After selecting the parent from the current population, a mutation-point (indicated with scissors in Figure 2.9) on this parent tree is randomly selected, and the other generated parent tree (shown in green) replaces the sub-tree at that point in order to generate the new child. An example of this operation is shown in Figure 2.9. There is a clear, yet important, difference between crossover and mutation; crossover only makes the system try different combinations of the existing genetic material, whereas mutation allows the system to introduce new genetic materials by using new randomly generated sub-trees.

2.6 Related Work

This section describes the existing work on image classification using machine learning methods, specifically GP. Some related work to cancer image classification such as skin cancer, brain tumor, lung cancer and breast cancer using GP has also been discussed.

2.6.1 Related Work to Image Classification Using GP

Earlier in 1996, Poli [159, 158] described a set of requirements for terminal set, function set and fitness function in GP to evolve efficient optimal filters for the tasks of feature detection and image segmentation, and studied their behavior in brain MRI and X-ray coronarograms. According to the authors, the terminal set must contain a limited number of variables and should capture both fine and broader scale information. Moreover, the functions involved to calculate the terminals during a single run of GP must not be complex and hence computation load for such terminals must be as light as possible. They have compared their results with ANNs and reported that GP has outperformed the competitor method. ANNs gave 31.7% sensitivity and 92.2% specificity, whereas GP achieved 61.5% sensitivity and 99.2% specificity. *Sensitivity* is the true positive rate and *specificity* is the true negative rate. With better results obtained by GP, the authors have elaborated that GP has far better ability for image analysis as compared to other existing methods.

A multi-tier domain-independent GP method for the problem of binary image classification is proposed by Atkins et al. [26]. The main objective was to automatically evolve a classifier that is capable of performing the tasks of image filtering, feature extraction, and classification. Their experiments on two datasets revealed that a comparable performance to the use of domain-specific features has been achieved.

Motivated by the work of Atkins et al. [26], Al-sahaf et al. [16, 20] have proposed and investigated a variety of GP methods for the problem of binary image classification. Evaluating their methods on four datasets of increasing difficulty showed a significantly better performance has been achieved compared to different GP-based and non-GP methods. Al-sahaf et al. [14] have also used GP to evolve an image descriptor, where a special function node is developed using STGP to extract features from pixel values. The results revealed the goodness of the proposed method compared to other GP and non-GP methods. Later, the structure of the algorithm in [14] is further improved to perform transfer learning by Iqbal et al. [90], to cope with difficult texture image classification tasks. This transfer learning method is able to solve difficult tasks that most other algorithms cannot solve, as shown by the results in [90].

Al-sahaf et al. [15] developed a multi-layer approach to feature extraction and image classification using GP. The method is evaluated on four image datasets with varying difficulty and compared with a baseline approach that requires human intervention to perform feature extraction. However, the proposed method does not require human intervention and consists of three layers; bottom layer for filtering, middle layer for feature aggregation and top layer for classification. Results revealed that baseline GP performed better on easy datasets as compared to the proposed method. However the method with no filtering (WNF) performed best on faces dataset and their method with Sobel edge detection (WSED) performed best on cells dataset. On the hardest dataset, WNF outperformed all other methods. However, the method is computationally expensive with a huge function set and large number of layers.

Al-Sahaf et al. [14] developed a GP-based method to automatically generate an image descriptor, i.e., a feature vector, for texture image classification. The feature vector generated in their approach is quite similar to LBP [145]. However, a domain-expert designs the formulas in LBP, whereas these formulas are automatically generated by GP in their work. Experiments revealed the goodness of the proposed method in comparison to other GP and non-GP methods. Iqbal et al. [90] improved the structure of the algorithm in [14] to perform transfer learning, to cope with difficult texture image classification tasks. The results proved the effectiveness of their method, showing ability to solve even more difficult tasks which most other algorithms cannot solve. Lensen et al. [110] developed a GP-based method capable of performing multiple tasks in a single evolved GP individual; region detection, feature extraction and binary image classification. Their results have shown improved classification performance compared to the existing GP approaches.

Earlier in 2003, Smart et al. [185] developed two dynamic-based range selection method for the problem of multi-class image classification in GP. The first method is centered dynamic range selection, and the second is slotted dynamic range selection. The results of evaluating those methods using five datasets of varying difficulty show that both of those methods outperformed the use of the static range selection method.

A GP-based method for the tasks of texture classification and texture segmentation is developed by Song et al. [188]. A bitmap texture dataset that consists of 48 different textures is used to evaluate the proposed method. The results revealed that GP is capable of evolving accurate classifiers. Moreover, their method does not need feature extraction, as a preprocessing step.

A domain-independent approach to the problem of multi-class object detection using GP is proposed by Zhang et al. [213]. The aim of their method is to locate a number of objects of different classes that are contained in a large image, and predict the class label of each of the detected objects. The method is tested using three datasets of increasing difficulty. The method is tested using three datasets of increasing difficulty. The evolved program is capable of performing object detection and multi-class classification tasks.

Fu et al. [78] have elaborated the use of GP for edge detection. As most traditional methods have used local window-based filter approaches, the authors have used the whole image as input and pixels are classified directly into edges and non-edges without the need of doing any preprocessing and post-processing. They have compared their method with the Laplacian and Sobel edge detectors on three sets of images with varying difficulty for edge detection task. Results have shown that detectors evolved by GP outperform the Laplacian detector and compete with the Sobel detector in most cases. The limitation is that only 24 images (4 binary (easy), 4 natural (difficult) and 16 BSD (Berkley Segmentation dataset, very difficult)) are used to evaluate the performance of the proposed method. Also for the BSD dataset, sub-images have been used to reduce the computational cost of training time. However, for medical images, using only sub-images is not advisable as the location of tumor is not known beforehand.

Muni et al. [136] developed a multi-tree GP method for feature selection and multi-class classification. The method is capable of performing two tasks simultaneously; searching for a good feature subset and designing a classifier using the selected features. An evolved individual in a n-class problem had n trees and each tree was initialized using a random feature subset. The method introduced two new crossover operations: homogeneous crossover and heterogeneous crossover were introduced. The method is tested on seven non-image datasets with varying number of features between units to thousands. Comparisons with other methods showed that this method produced better results with selected features as compared to using all features.

Singh et al. [184] proposed a method namely *genetic programming image segmentation (GPIS)* to perform segmentation on biomedical image data and compared their results with an existing GA-based image segmentation tool called *GENIE Pro*. The results are computed on two cells dataset of increasing difficulty. The authors have described image segmentation as a vital step in object detection systems in many application including geosciences, remote sensing, medical image analysis and target detection in security surveillance. In this work, the image analysis operators process the input image in a series and not in a tree like structure. The fitness function is derived from the idea that segmentation can be viewed as a pixel-classification problem, where the value of false positive rate (FPR) and false negative rate (FNR) must be zero for ideal segmentation. An improved fitness function is used that penalizes longer programs. *GPIS* performed better than *GENIE Pro*, however, using a large number of image analysis operators make it expensive at computation time.

Since late 1990's, GP has been widely explored for the task of interest point detector in images. The method proposed by Ebner and Zell [68] is one of the earliest works employing GP to automatically evolve an interest point detector. Interest points are defined as salient image pixels that are unique and distinctive; i.e., they are quantitatively and qualitatively different from other image points, and normally represent a small fraction of the total number of image pixels [146]. Therefore, detection of such points can be useful in recognition tasks. Olague and Trujillo [146, 147, 148] have used GP to evolve interest point detectors taking into consideration the global separability and geometric stability of the detected points. Shao et al. [175] proposed a multi-objective GP method for the task of feature learning in image classification.

Recently, Armand et al. [50] developed a GP based classification method to detect active tuberculosis in raw X-ray images. The process does not perform pre-processing, segmentation, or feature extraction before performing classification, making it fast to train a classification model. Their results outperformed the traditional image classification techniques providing better accuracy while being efficient in computation time. However, the method has utilized a prodigious memory, which makes its adoption in real-world situations difficult.

Benjamin et al. [73] combined the fundamental characteristics of CNNs such as convolution and pooling with GP to generate a model for image classification. They have effectively utilized GP's flexible representation to evolve a model that performs multiple tasks, i.e., learning convolution filters' coefficients, detecting ROIs, extracting features from ROIs, and building a classifier. Though the proposed method has outperformed commonly used classification algorithms, it remains unable to produce better results than CNNs. However, CNNs are more expensive in terms of computation time compared to their method.

In 2020, Stefano et al. [166] developed a GP based auto-encoder for feature learning in 2D images. The method works by generating a partial model in each successive GP generation utilizing the model built in the previous generation. At the end of the GP evolutionary process, the method combines these models to generate a parametric function for reconstructing the training images. The results have shown that the method can precisely and effectively reconstruct the MNIST hand-written digits. Since it is a short paper, the analysis of the evolved models is not presented.

Bi et al. [40] developed a feature construction method using a multi-layer GP approach for image classification. The method uses image-related operators to extract and construct new high-level features. Their method provided good accuracy for binary image classification, but they are not investigated for multi-class image classification. Bi et al. [41] developed a GP method to evolve ensembles for image classification. This method also utilizes image-related operator like their previous work [40], but simultaneously learn new features while developing ensemble of classifiers. However, this method has shown inferior classification results on large image classification datasets. Most recently, Bi et al. [42] 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 datasets. However, the generated models formed by various classifiers are complex and challenging to interpret.

Summary – Most of these GP-based methods combine complex image processing operators such as convolution, derivatives, and filtering, to detect a specific type of structure such as a corner. A system capable of detecting other types of structures such as blobs, streaks, lines, and pigmented network (in dermoscopic images), and automatically constructing new high-level features could be more effective at generating useful information for skin cancer image classification. The above GP methods have used reduced image sizes e.g., 256 × 256. These methods cannot be applied to skin cancer images because reducing the size of a skin image may distort aspect-ratio which results in losing informative features. Some of the above methods have not utilized the remarkable property of GP to analyze the evolved programs, which can help improve the credibility of their work.

2.6.2 Related Work to Cancer Image Classification

Machine Learning Methods for Skin cancer image classification

Earlier in 2002, with the advent of digital dermoscopy, Piccolo et al. [157] focused on validating the use of digital dermoscopy by comparing melanoma classification diagnosis of experienced dermatologists with computer-aided diagnosis based on ANNs and also with diagnosis provided by minimal trained clinicians. The results are given in terms of sensitivity and specificity of 92% and 99%, respectively, for the trained dermatologist, 69% and 94%, respectively, for the clinician, and 92% and 74%, respectively, for the computer analysis. According to the results obtained, the authors have suggested computer analysis must be developed in order to assist and not to replace physicians in the diagnosis of skin cancer lesions as the best diagnostic results can be achieved by using both trained computer classifier and experienced dermatologist diagnosis.

Ferris et al. [74] presented a computer-assisted diagnosis of dermoscopic images for classification of melanoma. A classifier based on random forest is trained on dermoscopic images of benign and malignant skin lesions to generate a severity score for each lesion. Each image is manually segmented and 54 features were computed for all segmented lesions. Fitness measures are sensitivity, specificity and area under the curve (AUC) which is the area under Receiver Operating Characteristics (ROC) curve, showing trade-off between sensitivity and specificity. The classifier results on the same set of images were compared with that of 30 dermatology clinicians. The classifier produced better sensitivity than dermatologists, however, produced lower specificity than dermatologists. The trade-off of a higher sensitivity at the cost of specificity is, to some degree, inevitable and seen not just in devices and tests, but in clinicians as well. In the study, those practitioners who had the highest sensitivity to melanoma generally also had the lowest specificity. This study shows that early identification of melanoma highly depends on the expertise of a dermatologist and there is a need of developing effective CAD system where experienced dermatologists are not available.

For the detection of melanoma, a classification method has been proposed based on ANNs in [22]. The method consists of four stages: preprocessing, pigment network extraction, feature extraction and classification. The thresholding and directional Gabor filter is applied to the blue component of images for the first stage. For pigment network detection, again the Gabor filter is applied with different thresholding values. For feature extraction, mean and standard deviation are computed on the pixel values of the sub-images. Classification is then performed using ANNs fed with the extracted features where the performance is assessed by the commonly used classification accuracy measure and the method achieved 94% accuracy. The limitation of this work is that using balanced classification accuracy on an imbalance dataset leads to bias towards the majority class.

Variation in color of melanoma is a major discriminative aspects for dermatologists that is studied in [36]. This paper evaluates the importance of color in key-points detection steps of the bag-of-features model for the classification of melanoma images based on k-NN. Furthermore, gray scale and color sampling methods using Harris Laplace detector and its color extensions are compared. The performance of scale-invariant feature transform (SIFT) and color-SIFT patch descriptors are also analyzed. The method achieved 85% sensitivity, 87% specificity and 87% accuracy.

A computer system based on image processing and pattern recognition techniques can provide a quantitative evaluation of skin lesions, while keeping good diagnostic ability [216]. Zortea et al. [216] developed a low-cost CAD tool applicable in primary care based on a consumer grade camera with attached dermatoscope and compared its performance to that of experienced dermatologists. The system extracts several new image-derived features computed from automatically segmented images. These are related to the asymmetry, color, border, geometry, and texture of skin lesions. Three well-known statistical methods for classification are compared; linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and classification and regression trees (CART). The diagnostic accuracy of the system is compared with that of three dermatologists. The classifier (QDA, being the best classifier) was able to provide competitive sensitivity (86%) and specificity (52%) scores compared with the sensitivity (85%) and specificity (48%) of the most accurate dermatologist on the dermoscopic images. This method is evaluated on a very small dataset having only 206 images.

In [80], a CAD system is developed for melanoma classification which selects an optimal set of features from different types of features such as texture, border-based, and geometrical shape. To classify melanoma and benign images, four classification algorithms (Naïve Bayes, support vector machine (SVM), random forest, and hidden logistic model tree) are employed. Though this diagnostic system produced very good results (91.26% with 23 features), it utilizes different types of features individually and lacks an appropriate way to combine them.

Kawahara et al. [96] demonstrated how filters from a pre-trained CNN can be used to classify 10 classes of non-dermoscopy images in the Dermofit dataset. However, they reported a standard overall classification accuracy of 81.80% for the highly imbalanced Dermofit dataset, which is not suitable as it may give biased results towards classes with more images. From the confusion matrix shown in [96], the overall accuracy is 81.80%, whereas the balanced accuracy is 60.12% for the 10-class classification problem.

Recently, Alfed et al. [21] 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 and histogram of Lines, instead of the conventional histogram of oriented gradients and histogram of oriented lines, in skin cancer detection. Three classifiers are used: Adaboost, SVM, and ANN. The experiments were performed using dermoscopy and standard skin image datasets.

The robustness of a CAD system is one of the most important characteristics for dermoscopy images [30]. It is difficult to develop a robust system for multi-source images acquired under different conditions, such as varying illumination and different acquisition devices. Hence, it has been suggested to use the color constancy algorithms and the results of SVM have shown increased performance using RGB histograms as features. For effective feature learning from color images, a quaternion-based grassmann average network (QGANet) is developed [178]. The experiment results proved the goodness of the method on three histopathological color image datasets. Since the QGANet algorithm embeds the grassmann average network (GANet) into a principal component analysis network (PCANet), the computational complexity of this method with QGANet is four times more than the baseline GANet.

Identifying the score of the ABCD rule of dermoscopy has been recently studied [94]. In pre-processing, Gabor filters and active contours are utilized to detect lesion boundaries. The extracted features, according to the ABCD rule, are used to compute the total dermoscopy score, which is then used for binary classification. The method has produced good sensitivity and specificity results and revealed the potential of extracted features in building a good classification model.

Adjed et al. [8] developed a binary classification method for melanoma detection through fusion of texture and structure features. The method extracts texture features from different variants of LBP, and structure features from curvelet and wavelet transforms. SVM classifier produced good results in terms of sensitivity (78.93%) and specificity (93.25%). The method concatenates the different features together in a single feature vector for fusion, however, a better way of combining different types of features can help improve the classification performance.

To solve 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-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 when dealing with large image datasets. Shimizu et al. [179] 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, 179] produced a better performance compared to the standard non-hierarchical classification algorithms. However, hierarchical structures are more expensive in terms of computation time than standard non-hierarchical algorithms.

In the recent years, convolutional neural networks (CNNs) have become popular in skin image analysis. Codella et al. [60] used the Caffe architecture to perform feature extraction. Esteva et al. [72] used a huge private dataset which consists of both clinical and dermoscopy images to train an Inception network from scratch, aiming at a performance close to a human expert. However, the deep learning approaches typically required thousands of images to effectively train a model, and due to a "black-box architecture”, the models may not directly provide insights of prominent features. In addition, using a pre-trained CNN generally requires preprocessing a dataset to the same input configurations for which that CNN was originally designed for such as fixed-size images, and RGB or gray scale images, which increases the computation time and decreases flexibility to apply to any size of image.

Xie et al. [208] proposed an ANN-based ensemble model for melanoma detection from skin images. The algorithm works by first extracting the lesion area with a self-generating neural network. Various types of features such as border, texture, and color are extracted, which are then given to a neural network ensemble method for binary classification. The results revealed the goodness of the new border features, which played a vital role in achieving improved accuracy. For melanoma detection from skin images, Yu et al. [210] developed a 2-stage convolutional neural network (CNN) architecture. The first stage performs lesion segmentation using a fully convolutional residual network and the second stage performs classification with a very deep residual network. Their results revealed the potential of very deep CNNs, even with limited training data to solve such a complex task of melanoma detection. These methods [208, 210] are expensive in terms of computation time and require large computing resources.

Identification of suitable data augmentation methods have gained immense importance recently, which can generally cope well with the limited size of datasets [156]. Transfer learning has gained attention, which has been explored with and without fine-tuning [124]. Moreover, other relevant criteria such as size of images and selected architecture in CNNs has recently been studied [201]. Such methods require a lot of extra work such as parameter tuning and identifying suitable data augmentation strategies.

Recently, Brinker et al. [47] proved that automated melanoma image classification using CNN achieved significantly better results than board-certified dermatologists. Barata et al. [32] used pre-trained DenseNet-

161 architecture to perform a hierarchical diagnosis for three skin cancer classes. Additionally, they provided comparative studies on the significance of color normalization, lesion segmentation, and evaluation metrics. Patiño et al. [152] developed a lesion segmentation and classification method using morphological operations to estimate asymmetry, border and color features of the lesions in the PH^2 dataset. The method incorporated SVM, logistic regression and a fully connected neural network where the neural network has shown the best performance achieving 86.5% on average for multi-class classification.

Summary – Most researchers have used overall accuracy until the International Skin Imaging Collaboration (ISIC) 2018[3] challenge started collecting balanced accuracy along with other evaluation measures. Though the above methods have achieved better performance than their baseline methods, they are not effective in terms of computation time (e.g., four times slower than existing approaches). Using a pre-trained CNN generally requires pre-processing a dataset to the same input configurations for which that CNN was originally designed for such as fixed-size images, and RGB or gray scale image, which increases the computation time and decreases the flexibility to use any size of image.

Genetic Programming for Cancer Image Classification

In [177], a method for brain tumor classification on MRI is proposed based on statistical methods for pre-processing, fuzzy c-means (FCM) for brain image segmentation and GP for tumor classification that achieved 97% accuracy. GP here utilizes two fitness functions defined as;

- $\mathit{fitness}_1$ decides to which class a candidate belongs (three tumor classes; menningionma, glioma and medulla blastoma), and
- $\mathit{fitness}_2$ selects the best individual for each class.

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

It is more important in the field of medicine to diagnose a patient with a disease than to diagnose a normal patient with no disease. The method is evaluated on the Fish's Iris dataset; however the details of the class distributions are not provided. The proposed fuzzy logic based GP procedure has enabled the n-class classification problem to be solved as a whole rather than as n two-class problems, hence, reducing the computation time. The authors stated that with the availability of more image data, better results can be achieved.

Early detection of defective nodules in lung computed tomography (CT) images increase the survival rate of the patients by 50%, hence, a GP-based nodule detection method is developed in [57]. After segmentation of lung region from CT image sequence using 18-connectedness voxel labelling and ball rolling algorithm, nodule candidates are detected using adaptive multiple thresholding and rule based classifier. Three-dimensional (3D) geometric based features are extracted from the Region of Interest (ROI) namely volume, elongation factor, compactness and approximated radius. Vessels are distinguished from nodules on basis of elongation factor and compactness having more elongation factor and are not compact. Nodules have more volume and radius as compared to vessels. Hence, simple if-then rule based classifier is used to detect nodules from vessels. Fourteen two-dimensional (2D) features are computed from image matrix made after normalizing the image size of nodule candidates. The nodule candidates are then classified using GP based classifier and achieved 92% detection rate. This work has included expert knowledge to achieve a good detection rate.

Ryan et al. [169] described a fully automated work-flow for performing stage-1 breast cancer detection using GP. The method detects suspicious regions called ROI, which are then examined by more specialized routines either radiologist or a CAD system, which outputs the likelihood of malignancy. It is a seven stage method, where first five stages implement pre-processing, breast segmentation and feature extraction while the last two stages employ multi-objective GP approach for building and testing the classifier. Results have revealed the ability of GP to produce solutions that are in some way human-readable, and capable of examining the GP individuals. This is to ascertain which terminals (features) are most useful, and extract more information related to those from the data. This system accepts raw mammograms and outputs marked ROIs. However, the evolved program has not been analyzed which may help in further investigation of cancer detection.

Summary – It is encouraging to see GP being utilized to solve real-world problems. However, the methods described above need several steps to reach a final classification label. This makes them complex and also consume more computational resources. Since available medical image data is limited, these methods still need to be validated on bigger datasets to check their effectiveness.

2.7 Chapter Summary

Some of the limitations in the existing work are described here which have become the motivation of this thesis.

- The existing GP approaches to feature manipulation in image dataset have mostly used gray level images, which may not produce good results for skin cancer images in which color is a crucial characteristic for distinguishing various types of skin cancer. Having features extracted from skin images, feature selection and feature construction can be used to improve the diagnostic performance.

- The existing approaches have mostly constructed new high-level features by selecting more relevant features from the original set of extracted features. GP has successfully provided very good results using its powerful ability to feature selection. Utilizing feature selection by GP to first select prominent features and then construct new high-level features by using only the selected features can help improve the classification performance which has not been investigated in the past.

- Most of the existing methods extract a single type of features e.g., texture based LBP features, from skin images, and not employ multiple types of features together such as texture as well as color. Using multiple kinds of features can provide more information which helps to train the classification model well or construct useful high-level features from them. How to design GP to automatically combine different types of features to improve classification performance in skin cancer image classification has not been investigated.

- The extracted features can be redundant and may not prove useful in developing good ensemble classifiers. In the existing methods, either complete set of original features or selected features are provided to ensembles [82, 199], however, multiple constructed features have not been provided to an ensemble of classifiers. Most of the existing feature construction approaches that usually generate new features for only a single classifier, remain unable to provide good classification performance [141]. Since the constructed features tend to have more distinguishing ability than the original extracted features, it is expected that newly constructed features evaluated by an ensemble of classifiers will help improve the classification performance.

- Sometimes classifying a particular image to cancer or non-cancer is not enough, here clinicians are more interested to investigate which specific features or clinical properties such as asymmetry, or color variation in the skin lesion are responsible for developing the cancer. In such a scenario, an interpretable classification model is interesting to investigate which not only provides good classification performance but also help identify prominent features.

- With advancements in technology, various optical instruments are in use to capture skin cancer images such as dermatoscope and standard cameras. Images captured from different instruments might have different visual properties such as illumination, scale, and reflection, therefore, which feature extraction methods are suitable for which type of images (captured from different instruments) is still an open question. Most of the existing skin cancer image classification methods are developed for a single image modality, developing a robust skin cancer classification method which can produce good results across multiple image modalities need to be explored.

Chapter 3

Two-Stage GP for Feature Selection and Construction

3.1 Introduction

Feature selection selects a subset of original features while feature construction creates a new feature(s) from the original set of features [209]. Feature construction involves transforming a given set of input features to generate a new set of more powerful features [138]. Feature selection and construction both can help improve performance by selecting relevant features and constructing new high-level features. Hence, these are good tools not only to improve performance, but also to reduce the dimensionality and hence provide features which take less computation time while being processed by the classification algorithm. With the ability of GP in selecting good features that can improve the classification performance and generating classification models that can also be treated as a classifier, it can be used to effectively classify melanoma images. This chapter describes in detail various methods developed to effectively utilize GP's ability to feature selection and feature construction for targeting skin image classification.

Since the skin images are large, a classification method provided with these large-sized images generally requires substantial computing resources to train/generate a good classification model. This also leads to high computation time. Hence, there is a need for dimensionality reduction where these images are effectively converted into feature vectors using suitable feature extraction methods. A classification algorithm can easily handle the resulting feature vectors; using a limited number of resources and computationally fast, can help achieve good performance. Dimensionality reduction aims to reduce the number of features and select only prominent features with good discriminating ability between classes. However, there is limited work done to feature selection and construction in skin cancer image classification.

Moreover, the medical practitioners are interested in finding the cause of a disease, and a system is highly recommended to have such causal information. With the property of GP evolved programs being interpretable, giving information about which features are prominent in constructing new high-level features, the medical practitioners can gain deep understanding of which specific texture patterns and color variations are the cause of the disease.

Existing GP approaches to feature selection and construction aim at improving the generalizability of GP such as in symbolic regression problems [56], improving classification performance in high-dimensional data [196] and effective biomarker identification and classification [10]. These methods have used the complete original set of features to construct new features, which might potentially limit the performance. The proposed methods construct new high-level features from the selected features which are expected to perform better compared to features constructed from complete original set of features.

This chapter initially focuses on developing a classification method for melanoma detection, which is a binary classification task. Early detection of melanoma is crucial, since it helps increase the survival rate of the patient. For melanoma detection, an embedded approach for feature selection and feature construction will be investigated. To extend this method further in order to achieve multi-class classification, a wrapper approach for feature selection and construction will be developed.

3.1.1 Chapter Objectives

Motivated by the intrinsic ability of GP to feature selection and construction, two methods (an embedded approach and a wrapper approach) are developed in this chapter for skin cancer image classification problems. Different from most existing methods, the proposed methods aim at constructing new features only using previously selected features by GP, i.e., a two-stage approach, which can have the ability to construct more informative features as compared to construct features from all of the original features. This chapter address the following research objectives:

- Design a new two-stage GP method in an embedded approach (2SGP-E) for feature selection and feature construction.
- Extend 2SGP-E method to a new two-stage GP method in a wrapper approach (2SGP-W) for feature selection and feature construction to achieve performance gains.
- Compare the two methods by analyzing their effectiveness for feature selection and construction for skin image classification.
- Identify the prominent features selected by GP during the evolutionary process to construct new informative features.
- Compare the discriminating ability of gray scale features with color features for melanoma detection.
- Investigate the efficiency of the two methods by analyzing the average time required to evolve a solution, and average time to evaluate an instance.

- Analyze the pixel-based texture patterns of the selected features.
- Investigate the interpretability of the evolved programs by those two methods.

3.1.2 Chapter Organization

The rest of the chapter is organized as follows. Section 3.2 presents the proposed GP methods in an embedded approach to feature selection and construction, including the two stages, search space, fitness function and describes the evaluation procedure. Section 3.3 describes the proposed GP method in a wrapper approach to feature selection and construction in detail. Section 3.4 describes the experiments performed, GP parameters and benchmark methods for comparison. Section 3.5 presents the results. Section 3.6 provides detailed analysis by examining the GP selected and constructed features, computation time and evolutionary processes. Section 3.7 concludes the chapter with the achievements of the two methods, and their potential limitations.

3.2 The proposed embedded two-stage GP approach

The proposed Two-Stage GP (2SGP-E) method is described in this section. The overall structure is depicted in Figures 3.1 and 3.2. First, the images are converted to feature vectors by using LBP image descriptor as described in Section 2.1.2 (on page 30). In stage-1, these features then are fed into GP. GP utilizes its traditional representation where an individual consists of a single tree. The GP process starts by randomly creating a predefined number of initial solutions via using different combinations of the elements in the function and terminal sets. Adopting a carefully designed fitness measure (presented in Section 3.2.1 on page 87), the performance of

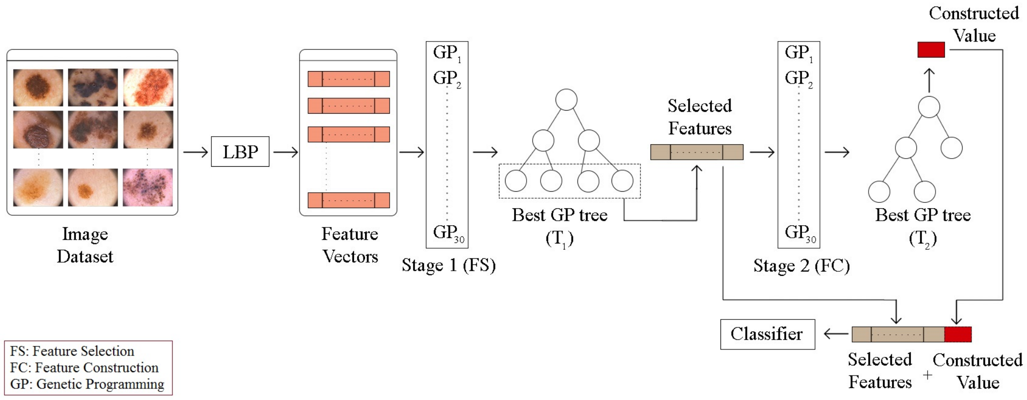


Figure 3.1: The overall process of the two stages in 2SGP-E.

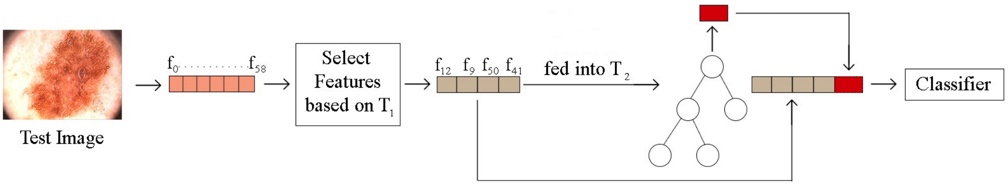


Figure 3.2: The test process of 2SGP-E.

each individual is calculated. GP uses genetic operators such as crossover, mutation, and elitism to produce new individuals for the next generation from those of the current generations. Individuals with better fitness values are more likely to be selected for participating in the next generation using ramped half-and-half selection. The best evolved individual, the one with the highest fitness value, represents the best evolved solution to the problem.

GP, implicitly, performs feature selection during the evolutionary process, since not all the features are used as the leaf nodes in the tree of an evolved GP individual. The leaf nodes, i.e., features, of a GP tree are the selected features. The task of the evolutionary process at stage-1 is to evolve a classifier (GP individual), and the leaf nodes of the best individual at the end of the evolutionary process will be considered as the selected features. These selected features usually have high discriminating ability between classes. In stage-1, after performing GP for multiple runs, i.e., 10, the features appearing in the best individual (evolved tree) giving highest performance on training data are selected. These features are called GP-selected features.

The selected features which are obtained from stage-1 are used as the input to stage-2 for feature construction. Here again after the 30 independent GP runs, the evolved individual having the highest performance on the training data is selected. This individual represents a single constructed feature that will be used along with the GP-selected features (computed after stage-1) for classification. To this end, we have the selected features (outcome of stage-1) and a constructed feature (outcome of stage-2). These GP-selected and GP-constructed features are concatenated to form the final feature vector, which will be given to the classification method.

Figure 3.2 shows an example of how an unseen image is classified. Based on best GP tree (T_1) evolved on training data in stage-1, some of the features are selected (e.g., f_3, f_{17}, f_{32}, f_{47}). These feature values are fed into best GP tree (T_2) evolved on training data in stage-2 to get the GP-constructed feature value for each test image. The GP-selected and GP-constructed features make the final feature vector to be given to a classification algorithm such as a decision tree.

In order to deal with feature selection bias and feature construction bias issues, each image dataset is divided into 10 folds where 9 folds are used for training and 1 fold for testing, such that only training folds are used for feature selection and feature construction and the test fold remain unseen during the learning process. The method used for feature selection and feature construction using the training data to evolve selected features (outcome of stage-1) and to evolve constructed feature (outcome of stage-2) is illustrated in Figure 3.1. For getting the transformed feature vectors for the test instances, the method illustrated in Figure 3.2 has been adopted. Hence, the problems of feature selection bias and feature construction bias have been avoided in this work.

3.2.1 Fitness Function

Having (very) different numbers of instances in different classes is commonly referred as a *class imbalance* problem. In this case, the use of the standard overall classification accuracy, defined as the ratio ($\frac{N_{\text{correct}}}{N_{\text{total}}}$) between the correctly classified instances N_{correct} and the total number of instances N_{total}, is inappropriate, since it may lead to bias towards the majority class. Alternatively, the balanced classification accuracy has been used as a good measure for imbalance classification problems [30, 153], since it gives equal importance to both classes without any bias. Therefore, we adopted it as the fitness function in this study, which is defined as

$$fitness = \frac{1}{m}\sum_{i=1}^{m}\frac{correct_i}{total_i} \tag{3.1}$$

where m refers to the total number of classes, $correct_i$ refers to the correctly classified images of class i, and $total_i$ refers to the total number of images of the class i. The fitness value ranges between 0 and 1, where 1 represents the *ideal* case.

3.2.2 Terminal Set and Function Set

The terminal set consists of uniform LBP features. Gray-level LBP features (referred as LBP_{Gray}) include a total of 59 features and colour LBP features (referred as LBP_{RGB}) include 177 features. For computing LBP_{RGB}, a color image is converted to its red, green and blue channel images and then LBP features are extracted from each of them. These three color channel features are concatenated together to make a total of 177 (= 59 LBP features

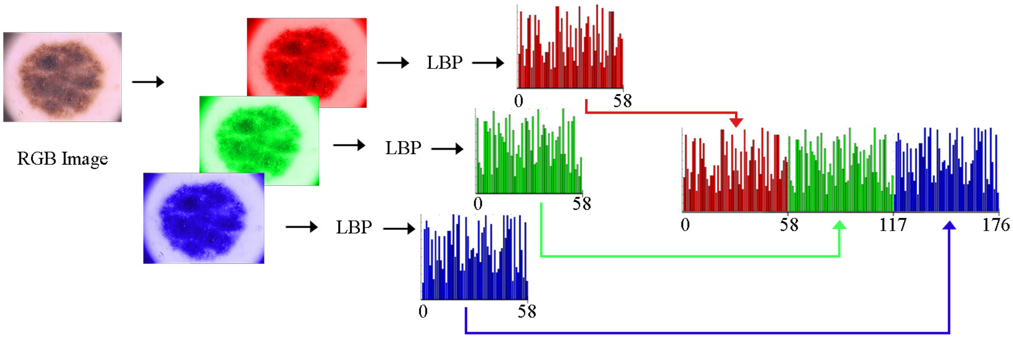


Figure 3.3: Step-by-Step procedure to generate the $\mathrm{LBP}_{\mathrm{RGB}}$ feature vector from a color image.

× 3 channels) $\mathrm{LBP}_{\mathrm{RGB}}$ features. This process is illustrated in Figure 3.3.The value of the i^{th} feature is indicated as Fi. The window size of 3×3 pixels and a radius of 1 pixel ($LBP_{8,1}$) is used, which are the fundamental and widely used settings for extracting LBP features.

The function set consists of four arithmetic operators, two trigonometric functions and one conditional operator, which are $\{add, sub, mul, div, Sin, Cos, if\}$. The first three arithmetic operators and the two trigonometric operators have the same arithmetic and trigonometric meaning. However, division is protected that returns 0 when divided by 0. The if operator takes four inputs and returns the third if the first is greater than the second; otherwise, it returns the fourth [195].

3.3 The proposed wrapper two-stage GP approach

This method (2SGP-W) is similar to 2SGP-E in terms of the program representation, the terminal set, the function set and the fitness function. It differs in terms of using a wrapper approach instead of an embedded approach in 2SGP-E. In a wrapper approach, the classification algorithm such as a decision tree not only provides classification result, but also takes part in feature selection and construction during the evolutionary process, which greatly helps achieve improved performance. Different from 2SGP-E which solves only binary classification problem, 2SGP-W can be applied to multi-class classification as well without changing GP program structure. The 2SGP-W method starts by converting the image datasets to feature vectors similar to 2SGP-E. The stage-1 and stage-2 in 2SGP-W is similar to 2SGP-E, except that 2SGP-W is a wrapper method where a classification method such as decision tree helps to improve feature selection and classification during whole of the evolutionary process. It is important to note here that both the stages are implemented as wrapper approaches for features selection and construction. The main aim of 2SGP-W is to keep improving the feature subset selection (stage-1) and keep improving the goodness of the constructed feature (stage-2) during the evolutionary process while providing good classification performance by a machine learning classification algorithm such as a J48.

The 2SGP-W method has the same terminal set, the function set and the fitness function as the 2SGP-E method.

3.4 Experiment Design

For performing the experiments, *10-fold cross validation* is used using random stratified sampling. This is because 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 such that nine folds are used for training and one fold for testing. In our experiments, features are selected and constructed using nine (training) folds and the last (test) fold remains unseen during this feature selection and feature construction processes in order to avoid feature selection and feature construction biases. This process is repeated ten times where each fold is used for testing and the results are reported as mean of the accuracy values. All the folds are randomly selected but are ensured that the ratio of instances of each class in each fold is the same as in the original dataset.

For our experiments of detecting melanoma in a binary classification setup, Melanocytic Nevus / Mole (ML) and Melanoma (MEL) classes in Dermofit are used to explore a dataset of 407 total images. For multi-class classification, we have used the 10 classes. In PH^2, for the binary classification experiments, atypical nevi class and common nevi class are together considered as one class and denoted as "non-melanoma", and melanoma class are denoted as "melanoma". For the multi-class classification experiments, PH^2 has three classes: atypical nevi, common nevi, and melanoma.

In case of 2SGP-E, for stage-1, the number of individual GP runs is 10. Among these 10 evolved trees, the one having highest performance on the training data is selected and the features appearing in that tree (GP-selected features) are used as input to stage-2 for feature construction. Here in stage-2, GP runs for 30 times and evolves trees. Again, the best performing tree among the 30 evolved trees on the training data is selected as the constructed feature. The above procedure is repeated 30 times to get 30 sets of selected and constructed features. These are provided to the classification algorithm to get 30 accuracy values. The results are reported as the mean and standard deviation of these accuracy values.

Similar to 2SGP-E, 2SGP-W is executed for 10 and 30 times in stage-1 and stage-2, 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, 2SGP-W is executed 30 times in stage-2. Hence, the results are reported in terms of mean and standard deviation of these 30 accuracy values.

Table 3.1: Parameter Settings of the GP method.

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

3.4.1 GP Parameters

The GP parameters are listed in Table 3.1. For generating the initial population, the "Ramped half-and-half" method is used and the population size is set to 1024. Tournament selection with size 7 is applied to pick good individuals for producing new generations while maintaining population diversity. During the evolutionary process, the ratios for producing new individuals through crossover, mutation and elitism are 80%, 19% and 1%, respectively. The depth of the trees ranges between 2 and 8 in order to avoid code bloating [203]. After reaching a maximum of 50 generations, the evolutionary process stops unless a perfect individual with accuracy 100% is found. These parameters are specified empirically as they gave the best training performance amongst other settings in our experiments.

3.4.2 Methods for Classification

To check the performance of the two proposed method (2SGP-E and 2SGP-W), six classification methods are applied: Naïve Bayes (NB), k-Nearest Neighbor (k-NN) where $k = 5$, Support Vector Machines (SVM), Decision Trees (J48), Random Forest (RF), and Multilayer Perceptron (MLP). In a study [97] on kernel functions in SVM, it has been shown that non-linear kernel can achieve similar or better performance than linear kernel. Hence, a Radial basis Function (RBF) kernel is used instead of the default linear kernel in WEKA. For MLP, the learning rate, momentum, training epochs and number of hidden layers are set to 0.1, 0.2, 60, and 20, respectively. These parameters are specified empirically as they gave the best performance amongst other settings in our experiments.

3.4.3 Implementation

The implementations of all the non-GP methods are taken from the most commonly used Waikato Environment for Knowledge Analysis (WEKA) software [85] version 3.8. The implementation of GP method is done using the Evolutionary Computing Java-based (ECJ) package [114] package version 23.

3.5 Results and Discussions

3.5.1 Overall Results

The results of the two methods for binary classification are presented in Tables 3.2 and 3.3 using $\mathrm{LBP}_{\mathrm{Gray}}$ and $\mathrm{LBP}_{\mathrm{RGB}}$ features, respectively. The results are represented in terms of sensitivity, specificity and balanced accuracy. Vertically, each table comprises of three blocks where first block corresponds to the results of using *All* features directly provided to commonly used classification algorithms. The second and the third blocks show the results of the embedded 2SGP-E, and the wrapper 2SGP-W methods, respectively. Horizontally, these tables consist of 7 columns where first lists the classification algorithm, second, third and fourth show, respectively, the test performances in terms of sensitivity, specificity and balanced accuracy on the PH^2 datasets. Similarly, the rest of the columns show these test performances on Dermofit datasets. The values of the results provided by deterministic methods using *All* features is the mean of applying 10-folds cross validation to the dataset. The proposed 2SGP-E, and 2SGP-W methods are repeated 30 times, hence we get 30 accuracies for each classifier which are represented as mean and standard deviation

$(\bar{x} \pm s)$ in Tables 3.2 and 3.3. The results of multi-class classification are presented in Table 3.4 for both $\mathrm{LBP}_{\mathrm{Gray}}$ and $\mathrm{LBP}_{\mathrm{RGB}}$ features.

Table 3.2: Binary Classification Results with LBP_{Gray}: The accuracy (%) on the test set using *All* features, 2SGP-E, and 2SGP-W (results are represented in terms of sensitivity, specificity and accuracy showing their mean and standard deviation $(\bar{x} \pm s)$) along with the statistical significance tests.

	Algorithm	PH^2			Dermofit		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
		(No. of features = 59)			(No. of features = 59)		
All	NB	60.00	66.38	63.44 ↑	55.32	65.67	60.39 ↑
	SVM	66.50	74.38	70.94 ↑	53.68	57.84	55.95 ↑
	k-NN	65.25	75.63	70.62 ↑	58.07	62.79	60.99 ↑
	J48	58.50	64.50	61.56 ↑	60.57	64.27	62.93 ↑
	RF	59.00	65.13	62.81 ↑	56.79	58.70	57.03 ↑
	MLP	65.50 ± 2.80	69.50 ± 1.54	67.34 ± 2.56 +	25.34 ± 1.76	34.35 ± 2.60	59.16 ± 2.31 +
		(No. of features = 29.26)			(No. of features = 32.16)		
2SGP-E	NB	63.73 ± 6.78	67.25 ± 1.79	65.71 ± 2.46 +	61.25 ± 5.64	65.57 ± 4.98	63.75 ± 2.01 +
	SVM	67.25 ± 3.91	73.49 ± 2.35	70.64 ± 2.55 +	52.50 ± 6.92	56.02 ± 5.68	54.23 ± 3.81 +
	k-NN	68.50 ± 6.44	66.90 ± 5.14	67.55 ± 1.99 +	60.96 ± 7.96	62.30 ± 3.47	61.15 ± 3.63 +
	J48	63.75 ± 9.83	65.03 ± 7.98	64.84 ± 3.55 +	55.27 ± 4.79	57.21 ± 7.80	56.62 ± 3.51 +
	RF	64.94 ± 5.68	66.13 ± 3.93	65.97 ± 2.02 +	58.86 ± 5.20	62.40 ± 1.95	60.87 ± 4.33 +
	MLP	64.50 ± 6.89	70.75 ± 3.15	67.33 ± 2.29 +	57.43 ± 1.36	61.71 ± 3.81	59.46 ± 5.18 +
		(No. of features = 25.43)			(No. of features = 34.73)		
2SGP-W	NB	77.42 ± 2.14	81.12 ± 1.35	79.22 ± 2.48	67.60 ± 1.24	77.22 ± 1.72	72.32 ± 0.86
	SVM	85.83 ± 2.27	89.13 ± 2.25	87.03 ± 3.48	74.18 ± 1.65	78.61 ± 2.11	76.22 ± 0.78
	k-NN	78.00 ± 1.68	80.68 ± 1.58	79.53 ± 2.56	65.00 ± 1.39	73.42 ± 2.63	69.67 ± 1.96
	J48	79.08 ± 1.04	83.14 ± 2.64	81.87 ± 3.22	68.11 ± 2.84	74.14 ± 2.52	71.61 ± 1.20
	RF	88.75 ± 1.67	94.98 ± 1.30	90.62 ± 2.21	73.75 ± 1.78	78.96 ± 1.69	75.83 ± 2.53
	MLP	76.94 ± 3.58	82.68 ± 2.46	79.53 ± 1.73	71.58 ± 2.47	77.08 ± 1.46	74.56 ± 1.84

For making a clear comparison between using different methods, the results are also tested using *Wilcoxon signed-rank test* and *One-sample t-test*. *Wilcoxon signed-rank test* (with a significance level of 5%) is applied to compare two stochastic methods e.g., 2SGP-W and 2SGP-E. *One-sample t-test* is applied to compare a stochastic method (2SGP-W) with the deterministic methods such as NB, SVM, k-NN, SVM, and RF. The statistical test has

Table 3.3: Results of *Binary Classification* with LBP_{RGB}: The accuracy (%) on the test set using *All* features, 2SGP-E, and 2SGP-W (results are represented in terms of sensitivity, specificity and accuracy showing their mean and standard deviation ($\bar{x} \pm s$)) along with the statistical significance tests.

	Algorithm	PH^2			Dermofit		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
		(No. of features = 177)			(No. of features = 177)		
All	NB	75.97	77.38	76.25 ↑	65.17	67.85	66.37 ↑
	SVM	72.61	78.07	75.41 ↑	55.42	53.66	54.77 ↑
	k-NN	73.77	75.21	74.02 ↑	62.24	60.49	61.39 ↑
	J48	70.46	76.57	73.13 ↑	57.77	59.50	58.43 ↑
	RF	74.84	76.49	75.94 ↑	52.71	58.78	55.01 ↑
	MLP	75.68 ± 2.95	77.50 ± 3.91	76.06 ± 3.87 +	56.71 ± 0.31	64.78 ± 0.44	60.43 ± 2.81 +
		(No. of features = 35.95)			(No. of features = 40.34)		
2SGP-E	NB	74.08 ± 3.31	78.32 ± 1.33	76.21 ± 1.91 +	59.21 ± 6.78	63.69 ± 0.69	61.27 ± 2.87 +
	SVM	73.25 ± 7.20	77.79 ± 0.57	75.77 ± 2.41 +	52.52 ± 5.33	56.14 ± 0.70	54.26 ± 6.56 +
	k-NN	72.08 ± 4.59	74.44 ± 0.72	73.31 ± 1.88 +	60.00 ± 5.62	64.54 ± 0.71	62.55 ± 2.75 +
	J48	71.67 ± 6.22	73.69 ± 0.63	72.74 ± 2.84 +	58.48 ± 6.58	60.88 ± 0.82	59.33 ± 4.52 +
	RF	74.50 ± 3.12	76.57 ± 2.00	75.53 ± 1.72 +	57.72 ± 1.68	59.46 ± 0.00	58.38 ± 3.31 +
	MLP	75.88 ± 6.43	79.86 ± 1.47	77.54 ± 1.67 +	60.48 ± 5.33	62.51 ± 0.70	61.37 ± 4.62 +
		(No. of features = 36.57)			(No. of features = 42.07)		
2SGP-W	NB	82.42 ± 1.16	86.12 ± 0.70	84.54 ± 0.16	78.60 ± 0.49	84.22 ± 0.32	81.92 ± 0.58
	SVM	86.83 ± 1.27	92.13 ± 0.25	89.84 ± 1.41	81.18 ± 0.60	85.61 ± 0.11	83.17 ± 1.37
	k-NN	80.42 ± 0.68	83.38 ± 0.28	81.88 ± 0.64	72.00 ± 0.45	82.42 ± 0.13	77.81 ± 2.64
	J48	85.08 ± 0.08	89.14 ± 0.04	87.19 ± 1.32	81.11 ± 0.54	89.14 ± 0.51	85.60 ± 1.78
	RF	90.75 ± 0.67	96.98 ± 0.10	**93.65 ± 0.83**	84.75 ± 0.48	88.96 ± 0.10	**86.23 ± 1.59**
	MLP	78.94 ± 5.98	88.68 ± 0.86	83.13 ± 0.63	75.67 ± 0.00	81.23 ± 0.00	78.38 ± 2.78

been applied on the test results to check which method has better ability to discriminate between *benign* and *malignant* classes. *Wilcoxon signed-rank test*, the symbols "+", "−" and "=" are used to represent significantly better, significantly worse and not significantly different performance, respectively, of the 2SGP-W compared to 2SGP-E and MLP. For example on PH^2 dataset, in Table 3.3, the test performance of SVM with 2SGP-E is represented as "75.77 ± 2.41+" where the "+" sign represents that SVM with 2SGP-W producing 89.84 ± 1.41 average accuracy significantly outperformed 2SGP-E. Similarly, on Dermofit dataset, the test performance of RF with 2SGP-E using *All* features is represented as "55.01 ↑" where the "↑" sign represents that RF with 2SGP-W producing 86.23 ± 1.59 average accuracy significantly outperformed RF with *All* features.

For sensitivity and specificity, it is more important to get higher sensitivity which represents the total number of correctly classified melanoma, as compared to specificity which represents the correctly classified benign lesions. Analyzing the results in terms of sensitivity and specificity, it has been observed that for PH^2, the 2SGP-W method is the most effective for identifying melanoma images by achieving the highest sensitivity of 90.75% on average among all the other methods as shown in Table 3.3. On Dermofit, the 2SGP-W achieved the highest sensitivity of 84.75% on average using $\mathrm{LBP}_{\mathrm{RGB}}$ features.

Analyzing the effect of dimensionality reduction in the proposed 2SGP-E method, it has been seen that while using $\mathrm{LBP}_{\mathrm{Gray}}$ features (59 in total) on PH^2 dataset as shown in Table 3.2, GP selects only half of the features (around 28) in its tree having tree depth of 8. Here, the number of features is 28.26 computed as average number of features appeared in 30 evolved GP trees. Adding the one constructed feature to these average number of 28.26 features make as total of 29.26 features as shown in the second block of Table 3.2. In case of $\mathrm{LBP}_{\mathrm{RGB}}$, the reduction in number of features is significant (from 177 to around 35). A similar trend in dimensionality reduction has been observed in 2SGP-W evolved programs. Using $\mathrm{LBP}_{\mathrm{Gray}}$ and $\mathrm{LBP}_{\mathrm{RGB}}$ features, the average number of selected features are 24.43 and 35.57 reduced from a total of 59 and 177 features, respectively. Similarly on Dermofit dataset, the effect of dimensionality reduction can be clearly seen. In 2SGP-E and 2SGP-W methods as shown in Table 3.2, the number of $\mathrm{LBP}_{\mathrm{Gray}}$ features are reduced from 59 to 31.16 and 33.73, respectively. Similarly, a total of 177 $\mathrm{LBP}_{\mathrm{RGB}}$ features are reduced to 39.34 and 41.07 in 2SGP-E and 2SGP-W methods as shown in Table 3.3. In the multi-class classification task, a similar trend in dimensionality reduction has been shown by 2SGP-E and 2SGP-W methods, while achieving better classification performance than using all set of features.

In 2SGP-E, most of the classification algorithms have achieved either better or similar performance compared to other classification algorithms using *All* features. In 2SGP-W, all the classification algorithms have achieved better performance compared to 2SGP-E and the non-GP classification algorithms. This shows that GP with its feature selection ability, has pushed most of the classification algorithms to achieve good performance even with reduced number of features. Moreover, the feature constructed by GP-selected features are more powerful in creating good training models as compared to feature constructed by the full set of features. 2SGP-E and 2SGP-W allow GP to perform both feature selection and feature construction during each stage, which helps improve the performance.

Table 3.4 shows that 2SGP-W is effective in providing much better results compared to the non-GP classification algorithms using *All* set of features. Among the two datasets, 2SGP-W provides good results on the PH^2 dataset with 200 images (relatively easy task). However, for the difficult task of distinguishing between ten types of skin cancers in the full Dermofit dataset with 1300 images, the performance is not very good. Here, RF achieved the highest test performance using $\mathrm{LBP}_{\mathrm{RGB}}$ features producing 69.62% average accuracy. The result of applying the statistical test shows that 2SGP-W (with an "↑" sign in Table 3.4) has significantly outperformed all the commonly used classification algorithms on both datasets.

Variation in color of *malignant* melanoma is a major discriminative aspect for dermatologists [30] which is validated by the results as well. In case of the PH^2 dataset, comparing the results of gray features and color features, color features have shown better performance in almost all cases. According to the overall results on PH^2, RF achieved the highest performances, i.e., 93.65% and 86.97% on the unseen images in binary and multi-class classification, respectively. This binary classification performance is comparatively much better than the state-of-the-art method [30]

Table 3.4: 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 $(\bar{x} \pm s)$).

Algorithm			PH^2	Dermofit
All	LBP_{Gray}		(No. of features = 59)	(No. of features = 59)
		NB	51.00 ↑	24.77 ↑
		SVM	57.50 ↑	38.69 ↑
		k-NN	53.50 ↑	31.62 ↑
		J48	47.00 ↑	22.38 ↑
		RF	54.50 ↑	33.62 ↑
		MLP	49.50 ± 4.57 +	40.65 ± 3.21 +
	LBP_{RGB}		(No. of features = 177)	(No. of features = 177)
		NB	55.00 ↑	25.15 ↑
		SVM	60.00 ↑	42.38 ↑
		k-NN	60.00 ↑	33.08 ↑
		J48	47.50 ↑	27.23 ↑
		RF	59.50 ↑	34.85 ↑
		MLP	60.25 ± 2.68 +	44.89 ± 3.54 +
2SGP-W	LBP_{Gray}		(No. of features = 28.33)	(No. of features = 31.52)
		NB	82.87 ± 4.58 +	55.12 ± 3.16 +
		SVM	77.34 ± 3.42 +	63.33 ± 3.81 +
		k-NN	79.60 ± 2.65 +	48.25 ± 5.36 +
		J48	74.02 ± 1.04 +	58.50 ± 2.64 +
		RF	83.78 ± 2.22 +	59.44 ± 5.12 +
		MLP	75.46 ± 3.92 +	60.20 ± 4.22 +
	LBP_{RGB}		(No. of features = 34.68)	(No. of features = 44.26)
		NB	82.33 ± 1.46	53.27 ± 2.87
		SVM	80.44 ± 2.80	49.33 ± 5.56
		k-NN	81.36 ± 1.25	51.22 ± 4.75
		J48	84.02 ± 2.45	56.71 ± 3.52
		RF	**86.97 ± 2.02**	**69.62 ± 4.62**
		MLP	79.68 ± 3.70	62.30 ± 3.31

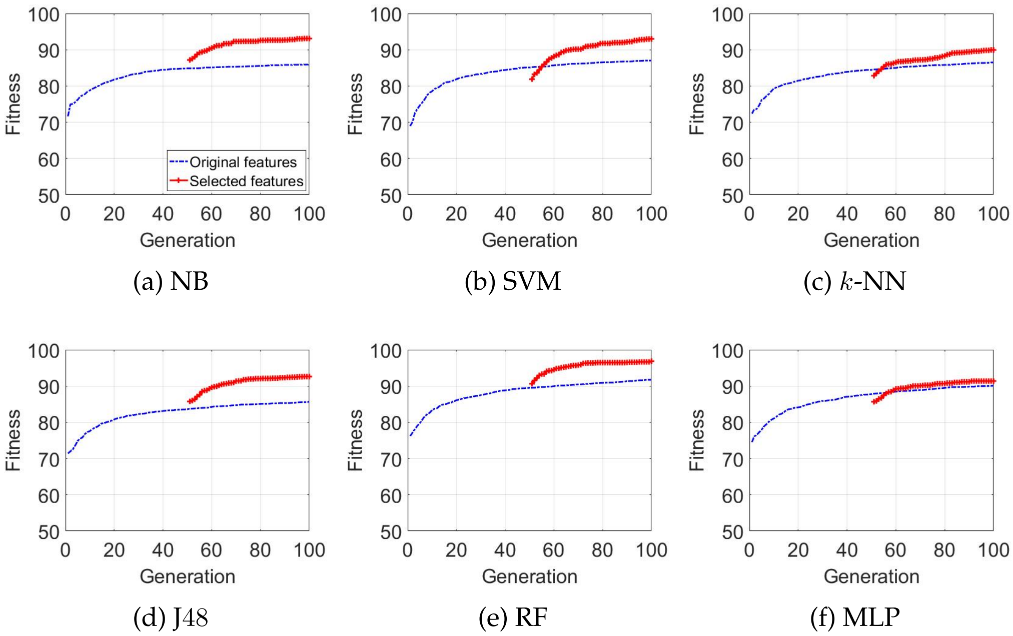
which produced 84.30% balanced accuracy on the same dataset using the same fitness measure. Moreover, this state-of-the-art method employs preprocessing and manual segmentation, which generally requires human expertise [30].

3.6 Further Analysis

3.6.1 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 3.4 to 3.7. These convergence plots have been taken from the experiments using $\mathrm{LBP}_{\mathrm{RGB}}$ features. Though there are 50 generations in both stages but for comparison purposes, here we have shown stage-1 executed till 100 generations. By doing so, we would like to see the difference in training performance among the 51^{st} to 100^{th} generations in stage-1 and the 1^{st} to 50^{th} generations in stage-2. To make this obvious from the graphs, we have plotted stage-2 from 51 generation onwards on x-axis. It is important to note here that stage-1 uses all the original features whereas stage-2 uses only the features selected in 50 generations of stage-1.

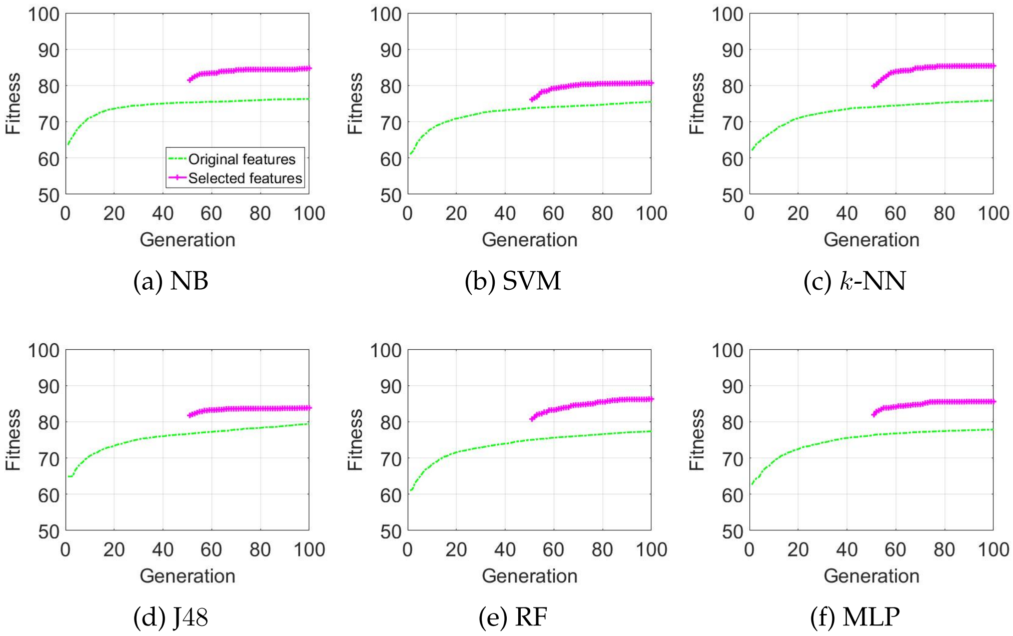
From the plots in Figures 3.4 to 3.7, a general GP trend has been observed; in the start of the evolutionary process, GP tries to explore the search space and makes larger jumps, regardless of whether it has been provided with all the original features or only the selected features. To get a clear understanding of how stage-2 is effective, we observe that the stage-2 (shown by selected features) starts from a higher average accuracy most of the time as compared to the average accuracy of 51^{st} generation in stage-1 (shown by original features). For example, in Figure 3.4(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 51^{st} generation reaches 84.52% average accuracy. This trend is not always true. In a few cases, stage-2 with selected features starts with a lower average accuracy compared to the stage-1. Such an example is given in Figure 3.4(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



(a) NB (b) SVM (c) k-NN

(d) J48 (e) RF (f) MLP

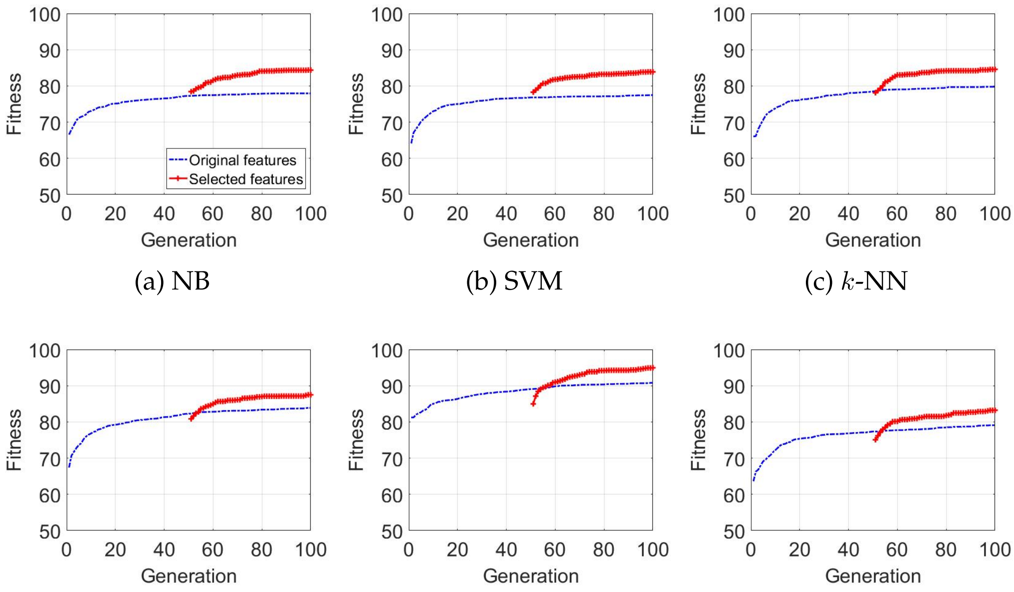
Figure 3.4: Convergence plots for PH^2 dataset in *binary classification*.



(a) NB (b) SVM (c) k-NN

(d) J48 (e) RF (f) MLP

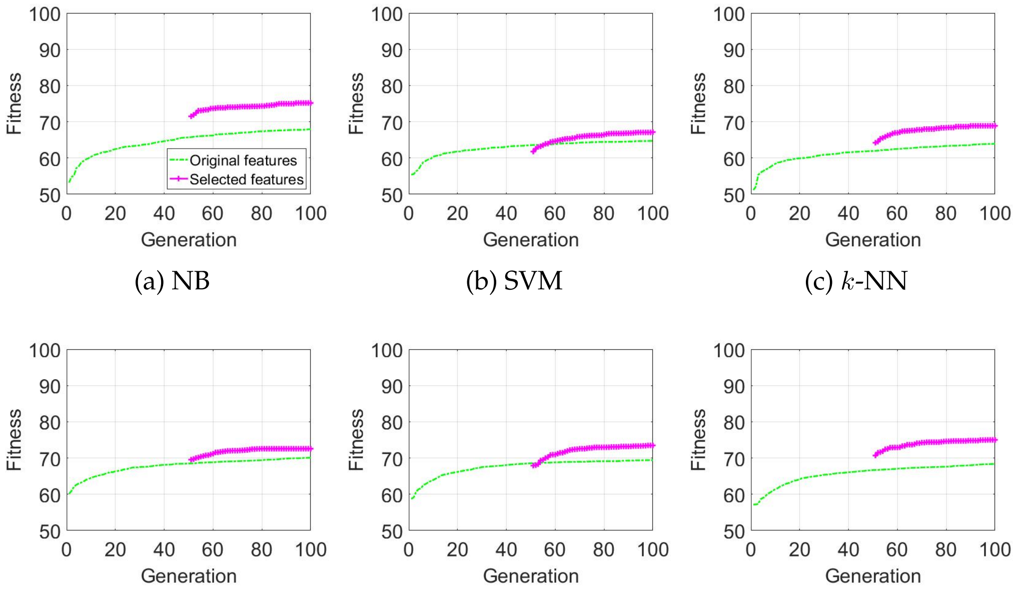
Figure 3.5: Convergence plots for Dermofit dataset in *binary classification*.



(a) NB (b) SVM (c) k-NN

(d) J48 (e) RF (f) MLP

Figure 3.6: Convergence plots for PH^2 dataset in *multi-class classification*.



(a) NB (b) SVM (c) k-NN

(d) J48 (e) RF (f) MLP

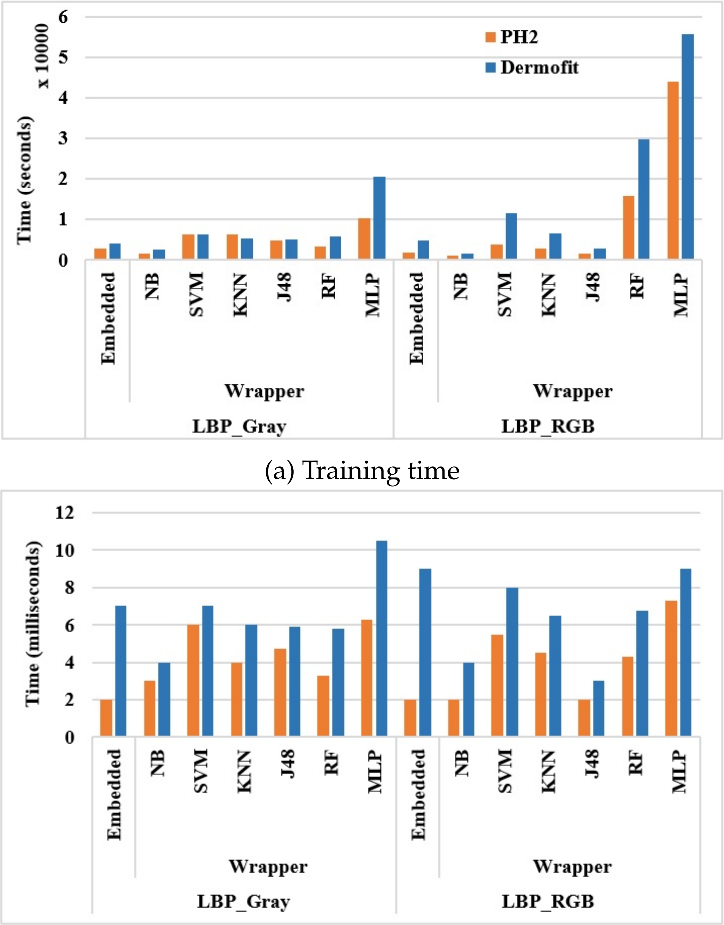
Figure 3.7: Convergence plots for Dermofit dataset in *multi-class classification*.

cycle. We can clearly see this in Figure 3.4(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.

3.6.2 Computation time

The average training time needed for the two methods to execute the two stages and to test their performances on the unseen data for solving binary classification task is presented in Figure 3.8. The average training and test time required by 2SGP-W for multi-class classification task is presented in Figure 3.9. Clearly, the time required to train a classification algorithm is affected by the number of images in a dataset, the number of features used to evolve an individual, and whether a wrapper or an embedded approach is adopted. Although the 2SGP-W method is more expensive than 2SGP-E, it does not take more than 18 minutes on average to evolve a solution.

In Figure 3.8, among the six wrapper binary classification algorithms, NB is the fastest to train a model. Overall, the highest-performing average training time of using the 2SGP-W method is given by RF on both the PH^2 and Dermofit datasets, and takes on average, 4.37 and 8.29 hours, respectively, to differentiate between melanoma and benign images. Similarly, having these trained methods at hand, they take only 0.33 and 0.58 milliseconds on average to test an unseen skin image. Therefore, we can say that our proposed 2SGP-W binary classification method is very effective and efficient for melanoma detection in real-time clinic situations and can help dermatologists to decide whether a biopsy is required or not in diagnosis of skin images.

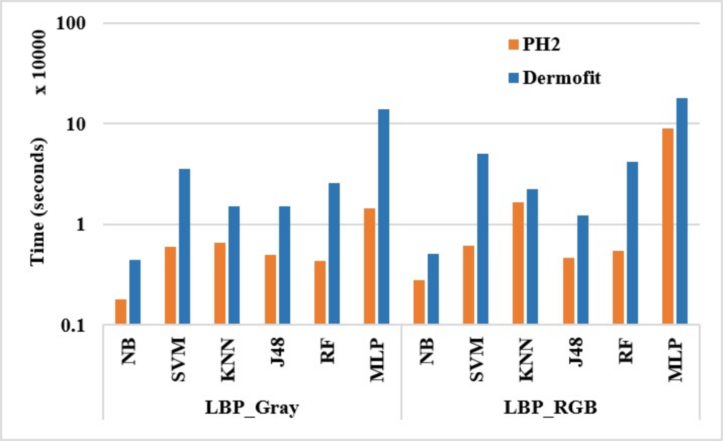


(a) Training time

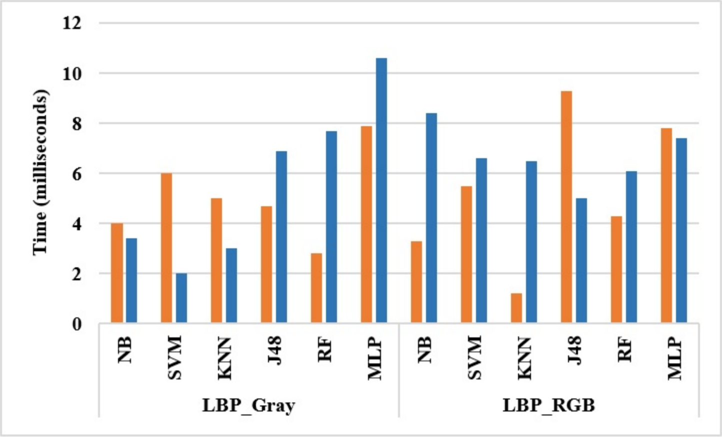
(b) Test time

Figure 3.8: The average computation time for *binary classification* using 2SGP-E method on the two skin image datasets.

For multi-class classification, Figures 3.8(a) and (b) depict that training a dataset with ten classes having 1300 instances (the Dermofit dataset) increases the computation time by many folds as compared to training a dataset with three classes having 200 instances (the PH^2 dataset). Since multi-class classification methods require more training time as compared to binary classification methods, this behavior can easily be observed while comparing Figures 3.8 and 3.9. However, an unseen image can be tested in fractions of a second using these trained models as shown by the



(a) Training time



(b) Test time

Figure 3.9: The average computation time for *binary classification* using 2SGP-W method on the two skin image datasets.

test time depicted in Figure 3.9(b).

Clearly, the wrapper approaches take more time to train a classification method as compared to embedded approach. Similarly, the bigger dataset (Dermofit) takes more time as compared to the smaller dataset (PH^2), regardless of which approach (wrapper or embedded) is used. Overall, the 2SGP-W approaches for multi-class classification are taking more test time as compared to the binary classification.

3.6.3 Example of an Evolved Feature

To see why the GP-selected and constructed features can achieve good performance, we show a good GP tree (Figure 3.10) from the 30 GP runs in 2SGP-E after stage-2 producing 90.63% accuracy on the training set. This tree is taken from $\mathrm{LBP_{RGB}}$ experiments where the total number of features is 177. In the figure, gray nodes represent functions and white nodes represent terminals.

Note that for constructing the tree as shown in Figure 3.10, features selected by a tree in stage-1 are used only and not the whole feature set. This tree is constructed from ten $\mathrm{LBP_{RGB}}$ features appeared in a tree in stage-1, which are F15, F40, F68, F90, F95, F105, F113, F117, F119, and F154. The values of these 10 selected features (after stage-1) and the constructed feature (after stage-2) are plotted in a bar chart shown in Figure 3.11.

For analysis of the selected feature, we take the simple example of features F15 and F154. As an example, we take the values of these features for only two instances from each class. The bar plot shows that the values of F15 (shown in black) and F154 (shown in green) for the *benign* instances (B1 and B2) are high as compared to values for *malignant* instances (M1 and M2). Hence, by combining these GP-selected features, the constructed feature divides instances of the two classes into two completely separate intervals as shown by blue color in Figure 3.11. Therefore, using these powerful GP-selected-constructed features from the selected features, the common classification algorithms become able to achieve better discrimination between the *benign* and *malignant* classes, resulting in improved performance.

We further analyze the LBP texture pattern of these two features F15 and F154 to match skin cancer image properties like streaks and blobs. Figure 3.12(a) shows the extracted 3×3 window for F15, its transformed LBP mask and the histogram showing the given pattern added to the *malignant* class bin represented as C_2. This mask shows that the presence of line ends in the image, which matches the presence of streaks in *malignant* images.

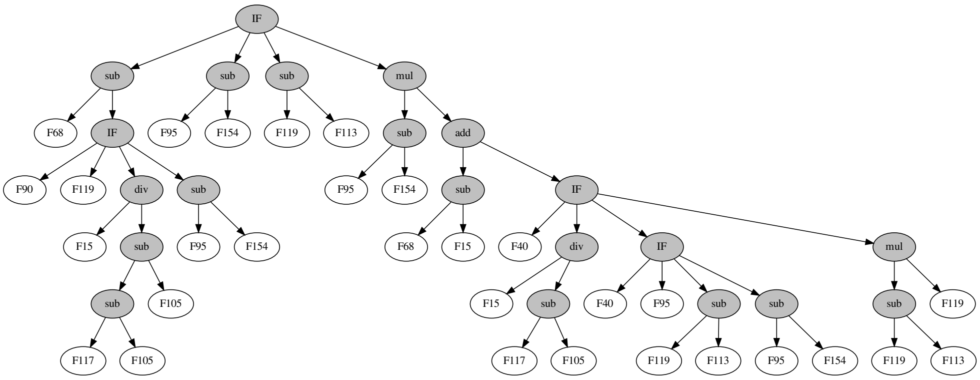


Figure 3.10: A good evolved GP tree in 2SGP-E after stage-2 having 90.63% accuracy on the training data.

According to the bar chart, this value is less for *malignant* images and high for *benign* images, which helps our method to distinguish between the two classes effectively. Similarly, Figure 3.12(b) shows the extracted 3×3 window for F154, its transformed LBP mask and the histogram showing the given pattern added to the *benign* class bin represented as C_1. This mask shows the presence of corners in an image. Its value for the *malignant* class is lower as compared to the *benign* class. This maps to the structure of the *benign* and *malignant* lesions. The *benign* lesions are often a confined dense structure having less variation in color, however, *malignant* lesions have often sparse structure, spreading over a larger region with no defined boundary and varying color (refer to Figures 1.2 and 1.3 for a visual illustration).

3.7 Chapter Summary

Motivated by the powerful ability of GP in feature selection and feature construction, this chapter has described the two GP based methods for solving the skin cancer binary and multi-class image classification. The

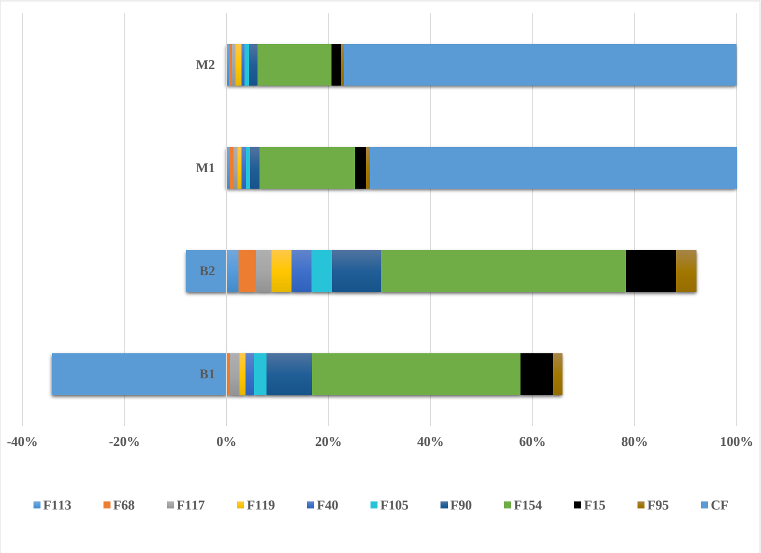
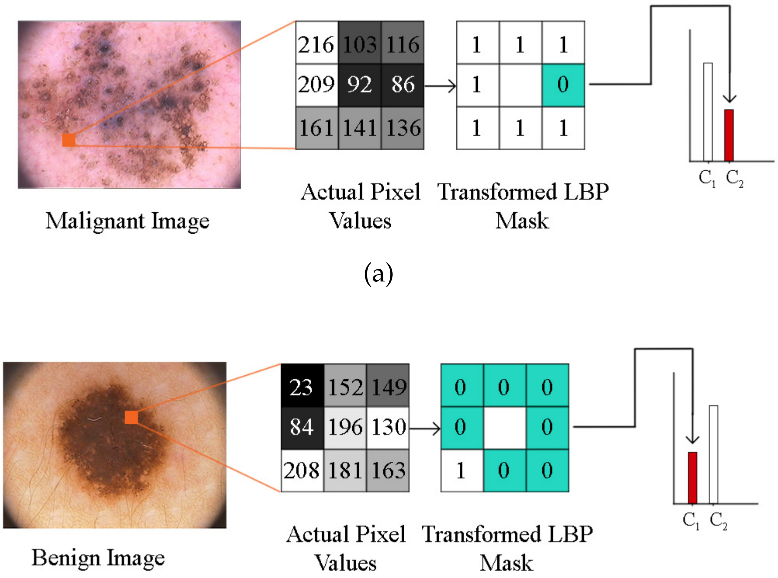


Figure 3.11: Bar chart showing the values of different selected features after stage-1 and the value of constructed feature "CF" after stage-2 in 2SGP-E.

methods aim to achieve feature selection in stage-1 and to achieve feature selection and construction in stage-2. The GP selected and constructed features together have shown powerful ability to help common classification algorithms achieve better performance as compared to using the full set of features. These methods constructed new features from the GP selected features, hence using the feature selection ability of GP twice, resulting in more powerful constructed features. Using these GP selected and constructed features, the classification algorithms have shown to provide effective solutions for the real-world cancer detection problem. The results have also shown that color features have more potential to distinguish between *benign* and *malignant* skin lesions as compared to gray features. We further analyzed the GP selected features and GP constructed features to get into the insights of skin cancer properties. It has been observed that



(a)

(b)

Figure 3.12: Feature analysis (a) *Malignant*, and (b) *Benign*.

the LBP patterns can be mapped to skin cancer properties, explaining the contribution of the selected features.

Though these methods have shown good performance, they remain unable to incorporate various kinds of features such as gray features and color features, simultaneously in a suitable GP approach. These methods have used gray features and color features separately and need to execute GP for multiple times for different features. Therefore, to tackle this problem, the next chapter will develop a suitable GP approach where multiple set of features can be utilized simultaneously to include information from multiple sets of features.

Furthermore, the geometrical shape of skin lesion which includes asymmetry and border features, is an important distinguishing characteristics between different types of skin cancers. The next chapter will show how these domain-specific features can be included in GP and used to help improve classification performance in melanoma detection.

Chapter 4

A Multi-tree GP Approach to Embedded Feature Selection

4.1 Introduction

Several computer-aided diagnostic (CAD) systems [72, 80, 84, 172, 210] have been developed to help dermatologists in diagnosing *benign* and *malignant* skin lesions. Although the existing CNN methods [72, 84, 210] have shown very good performance, however as most of them are implemented as a black-box model, hence, are not interpretable. In assisting a dermatologist, these methods cannot suggest which features are critical in classifying skin cancer images. With advancements in technology, various optical instruments are in use to capture these skin cancer images such as dermatoscope and standard cameras. Images captured from different instruments might have different visual properties such as illumination, scale, and reflection, therefore, which feature extraction methods are suitable for which type of images (captured from different instruments) is still an open question.

Some existing approaches [80, 172] rely on extracting various kinds of features e.g., texture, color and shape features from skin cancer images and compared the classification performances of these features using commonly used machine learning classification algorithms. These methods remain unable to design a way of using all these different types of features simultaneously, in order to get increased performance.

Skin cancer image classification is a complex task which requires enough informative features in order to achieve good classification performance. Hence, using only a single type of features such as texture features, may not provide sufficient information to discriminate between different classes of skin images. Hence, there is a need to employ various kinds of texture and color as well as local and global features to mimic the clinical properties such as asymmetry, border, color and diameter size. Developing a classification method for skin images to incorporate various kinds of features is not a trivial task because concatenating different types of features together in a single feature vector has shown poor classification performance. Therefore, there is a need of a new melanoma detection method, which not only incorporate various types of features, but is also capable of evolving a classification model based on selecting prominent features efficiently and effectively.

GP can evolve an individual having more than one tree to solve a particular problem, which is termed as multi-tree GP (MTGP)[150]. In the literature, MTGP has been explored for multi-class classification where each tree in an individual represents a classifier for a particular class [135]. To efficiently discover a set of patterns necessary for self-assembling swarm robots, a MTGP method is proposed which evolves patterns which are then incorporated into the corresponding robot modules [106]. Recently, it has been employed for constructing new redundant features for supervised and unsupervised problems [111]. MTGP has been used for automatically evolving image descriptors for multi-class texture image classification tasks [18]. Multi-tree approaches on non-image classification datasets have been studied in the literature [106, 111, 135], however they have not been investigated for complex image classification tasks where different kinds of features (based on local and global information as well as color and texture information) are necessary to be incorporated in the evolved solution. MTGP can be used to effectively employ different kinds of features simultaneously for handling specifically the complex skin cancer image classification tasks.

Selecting a suitable fitness function for the proposed MTGP method is important for effective algorithm design. In this chapter, each tree in a MTGP individual is considered as a binary classifier. We have explored two situations; 1) all the trees are considered equally important and improve themselves during the evolutionary process, and 2) more weight is provided to highest performing tree as compared to rest of the trees during the evolutionary process.

4.1.1 Chapter Objectives

This chapter develops a new multi-tree GP method for melanoma detection. Different from most existing methods, the proposed method aims at evolving a GP individual based on different types of texture, color, border shape and geometrical information features for skin cancer images taken from different optical instruments (specialized dermatosocope and standard camera), as compared to evolving models using only one type of feature. This chapter aims to address the following research objectives:

- Design a new MTGP based embedded feature selection method capable of handling different types of features.
- Compare the classification performances of proposed multi-tree GP approach and the traditional single-tree GP approaches across different skin image datasets.
- Compare the proposed GP method with the other commonly used non-GP classification algorithms.
- Identify whether all type of features are contributing equally to classification performance or a specific type of feature has more distin-

guishing ability for an image dataset captured from a specific instrument.

- Further analyze the effectiveness of the proposed MTGP framework by developing a new weighted fitness function considering the importance of specific type of features.

4.1.2 Chapter Organization

The rest of the chapter is organized as follows. Section 4.2 describes the feature extraction methods for extracting different types of features to be used in the proposed method. Section 4.3 presents the proposed embedded multi-tree GP method, its representation, the terminal set, the function set, crossover and mutation operators, and the fitness function. Section 4.4 describes the experiments performed, GP parameters and benchmark methods for comparison. Section 4.5 presents the experimental results and discusses how well they address the chapter goals. Section 4.6 provides detailed analysis by extending the proposed method with a new weighted fitness function. Section 4.7 concludes the chapter with the achievements of the method, and its possible limitations.

4.2 Feature Extraction

4.2.1 Local Binary Patterns (LBP)

LBP is used in this chapter to extract two types of texture features from gray and color skin images, respectively. To extract texture features from gray images, LBP is applied to the entire skin image to get a total of 59 features. To get color information from skin images, LBP features are extracted from red, green and blue channel images. Feature vectors from these three color channels are concatenated together to form a single feature vector with a total of 177 (= 59 LBP features × 3 channels) features.

LBP is described in detail in Section 2.1.2 on page 30.

4.2.2 Lesion Color Variation

Color is an important characteristics often used by dermatologists to classify skin lesions as a significant component of ABCD-rule [193] and 7-point checklist method [24]. Melanoma skin lesions are characterized by variation in color across the lesion area. This color variation induces high variance in the RGB color space. Therefore, features extracted from RGB color channels may have high discriminating ability between classes. To incorporate such global color features, the pixels in the segmented skin lesion of red, green and blue color channels are used. There are a total of 12 lesion color variation features extracted as follows.

The mean (μ) and variance (σ) of each channel is calculated and represented as μR, μG, μB and σR, σG, σB. To capture complex non-uniform color distributions within the skin lesion region, mean ratios of the mean values is calculated, i.e., $\frac{\mu_R}{\mu_G}$, $\frac{\mu_R}{\mu_B}$, $\frac{\mu_G}{\mu_B}$. Variations in color of the skin lesion with respect to the surrounding skin is also considered to show how much the lesion has grown compared to the normal skin of that specimen. These features are calculated as $\frac{\mu_R}{\overline{\mu}_R}$, $\frac{\mu_G}{\overline{\mu}_G}$, $\frac{\mu_B}{\overline{\mu}_B}$, where $\overline{\mu}$ represents the mean value of surrounding/normal skin region. These features are adopted from [172].

4.2.3 Geometry-based Features

Border information and geometrical properties of the shape of a lesion provide significant diagnostic information for detecting melanoma. According to the ABCD-rule of dermoscopy [193], asymmetry is given the highest score among its four characteristics; asymmetry, border irregularity, color, and diameter. A number of studies have been carried out on quantifying asymmetry in skin lesions [59, 140, 192]. Here, we used some standard geometry features (area, perimeter, greatest diameter, circularity index, irregularity index A, irregularity index B, and asymmetry index) adopted

Table 4.1: Geometrical border shape features.

Name	Description
Area (A)	Number of pixels of the lesion.
Perimeter (P)	Number of pixels along the detected boundary.
Greatest Diameter (GD)	The length of the line which connects the two farthest boundary points and passes across the lesion centroid c, which is given by $(x_c, y_c) = \left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n}\right)$ where n shows number of pixels inside the lesion, and $(x_i\ , y_i)$ are the coordinates of the ith lesion pixel.
Shortest Diameter (SD)	The length of the line which connects the two nearest boundary points and passes across the lesion centroid.
Circularity Index (CRC)	It explains the shape uniformity expressed as CRC = 4πA/P^2
Irregularity Index A (IrA)	IrA = P/A
Irregularity Iindex B (IrB)	IrB = P/GD
Irregularity Index C (IrC)	IrC = P $\times$ $\big($(1/SD) - (1/GD)$\big)$
Irregularity Index D (IrD)	IrD = GD - SD
Major and Minor Asymmetry Indices	These indices are defined as the area difference between the two halves of the lesion, taken the principal axes as the major symmetry axis, and its 90° rotation as the minor axes of the symmetry.
Asymmetry Index (AI)	Having known major and minor symmetry axes, the lesion is folded along the axes and the differences between the two halves of the lesion are calculated by applying XOR operation on the binary segmentation plane. The asymmetry index is measured by AI = $(\mathrm{A}_D/A)\times 100$ where $\mathrm{A_D}$ denotes the difference between the two halves.

from [117] complemented by others (shortest diameter, irregularity index C, and irregularity index D) adopted from [80]. To extract these features, we used the segmentation masks provided along with the datasets to get the lesion region. These features are extracted from only the lesion region, and not the entire image. Figure 4.1 shows a sample dermoscopy image, and the process of acquiring its major symmetry axis and calculating the major asymmetry index. Images within each dataset in this study have fairly similar spatial resolution; thus, there has been no scale issue for features such as area and perimeter. We extracted a set of 11 geometry-based features (described in Table 4.1) from each skin lesion image.

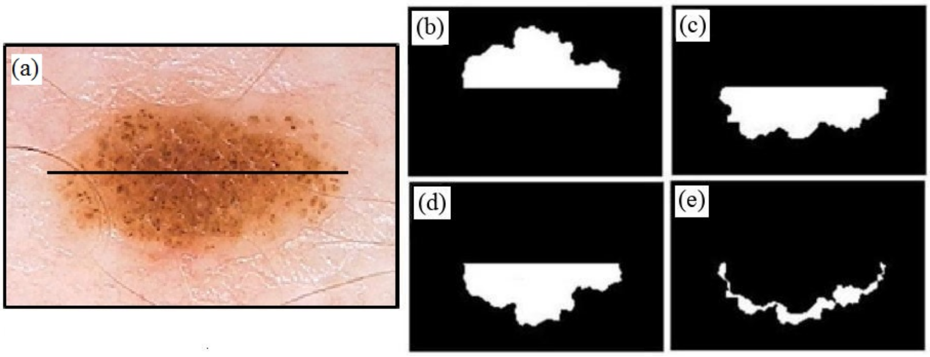


Figure 4.1: Calculating the major symmetry index: (a) major symmetry axis, (b) upper half, (c) lower half, (d) folded upper half, and (e) difference.

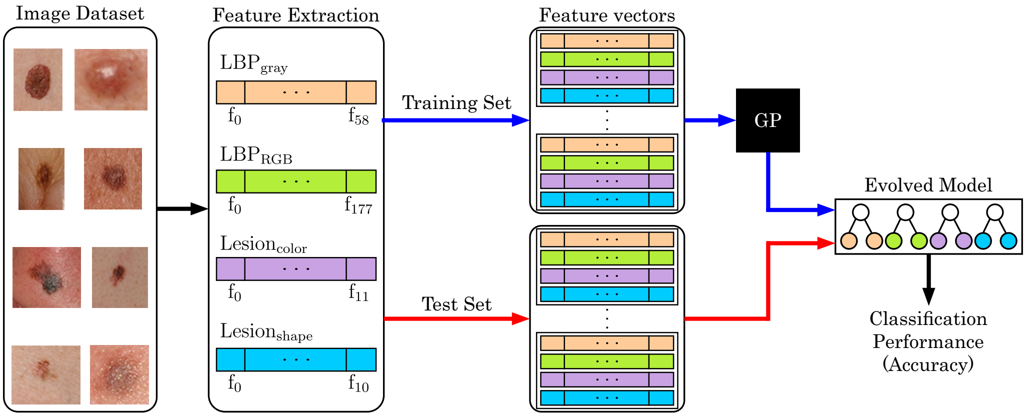


Figure 4.2: The overall algorithm.

4.3 The Proposed Method

The proposed method for melanoma detection from skin cancer images is described in this section. The overall structure of the proposed multi-tree GP based embedded feature selection approach (EGP-4) is presented in Figure 4.2.

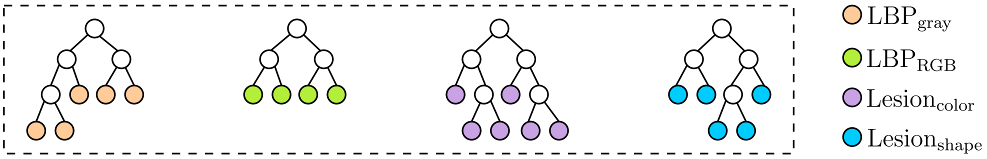


Figure 4.3: An EGP-4 individual with different types of features at terminals of each tree.

4.3.1 Representation

The images are first converted to feature vectors by employing the four feature extraction methods described in Section 4.2. These four types of features ($\mathrm{LBP}_{\mathrm{gray}}$, $\mathrm{LBP}_{\mathrm{RGB}}$, $\mathrm{Lesion}_{\mathrm{color}}$, and $\mathrm{Lesion}_{\mathrm{shape}}$) are fed into multi-tree GP method. Example of an individual in the proposed method is shown in Figure 4.3. During the evolutionary process, the proposed method is designed such that each tree can select from only one type of features. In other words, our multi-tree GP method evolves an individual (model) which consists of four trees; one is evolved using $\mathrm{LBP}_{\mathrm{gray}}$ features, second using $\mathrm{LBP}_{\mathrm{RGB}}$ features, third using $\mathrm{Lesion}_{\mathrm{color}}$ features and fourth using $\mathrm{Lesion}_{\mathrm{shape}}$ features as shown in Figure 4.3.

4.3.2 Terminal Set and function Set

The terminal set consists of four types of features, extracted from the four different feature extraction methods.

1. $\mathrm{LBP}_{\mathrm{Gray}}$: A total of 59 LBP features are extracted from gray-level skin cancer images.
2. $\mathrm{LBP}_{\mathrm{RGB}}$: From each color channel (red, green, blue), 59 LBP features are extracted. These features are concatenated to make a total of 177 ($= 59$ LBP features $\times$ 3 channels) $\mathrm{LBP}_{\mathrm{RGB}}$ features.
3. $\mathrm{Lesion}_{\mathrm{Color}}$: Color variation inside the lesion area, and between the lesion area and skin is calculated by a total of 12 $\mathrm{Lesion}_{\mathrm{Color}}$ features adopted from [172].

4. $\text{Lesion}_{\text{Shape}}$: The geometrical properties and border information of the lesion region are included in our method by extracting 11 $\text{Lesion}_{\text{Shape}}$ features adopted from [80, 117].

The value of the ith feature for the above four feature types is indicated as Gi, Ri, Ci, and Si, respectively, as shown by the GP individual in Figure 4.5.

The function set consists of the most commonly used seven operators; four arithmetic $\{+, -, \times, /\}$, two trigonometric $\{sin, cos\}$, and one conditional $\{if\}$ operator. Among the arithmetic operators, the first three operators have the same arithmetic meaning, however, division is protected that returns zero when divided by zero. The if operator takes four inputs and returns the third input if the first input is greater than the second input; else, it returns the fourth input.

4.3.3 Crossover and Mutation

To meet the objective of having only one type of features in a single GP tree, genetic operators, such as crossover and mutation, are designed accordingly, which is called *same-index-crossover/mutation* [111]. The step-by-step process is given in Algorithms 1 and 2. This crossover/mutation guarantees that the GP individual evolved at the end of the evolutionary process, consists of four trees where each tree evolves from a single type of features. For example, in case of crossover having two parents, the tree generated using LBP_{RGB} features in the first parent can only crossover with the tree generated using the LBP_{RGB} features in the second parent, and it is ensured that it cannot crossover with a tree built using $\text{Lesion}_{\text{Shape}}$, LBP_{Gray} or $\text{Lesion}_{\text{Color}}$ features as described in Algorithm 1. Similarly, for example, in case of mutation having one parent, a newly created tree generated using $\text{Lesion}_{\text{Color}}$ features can only mutate with a previously generated tree in parent from $\text{Lesion}_{\text{Color}}$ features as described in Algorithm 2. A graphical illustration of this crossover/mutation is shown in Figure 4.4.

Algorithm 1 Same-Index Crossover

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

Algorithm 2 Same-Index Mutation

```
function MUTATION($P^1$)                 ▷ One GP Individual (parent)
                                          having $n$ trees
    for i = 1 to $n$ do
        $P^1 \leftarrow init(T_{\mathrm{i})}$              ▷ Generate a new tree with
                                          a single type of features
        MUTATE($P_i, P^1_{)}$               ▷ Mutate the tree from parent
                                  individual with the new generated tree,
                                  both having the same type of features
    end for
    return $C^1$                        ▷ One child obtained after MUTATE
end function
```

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

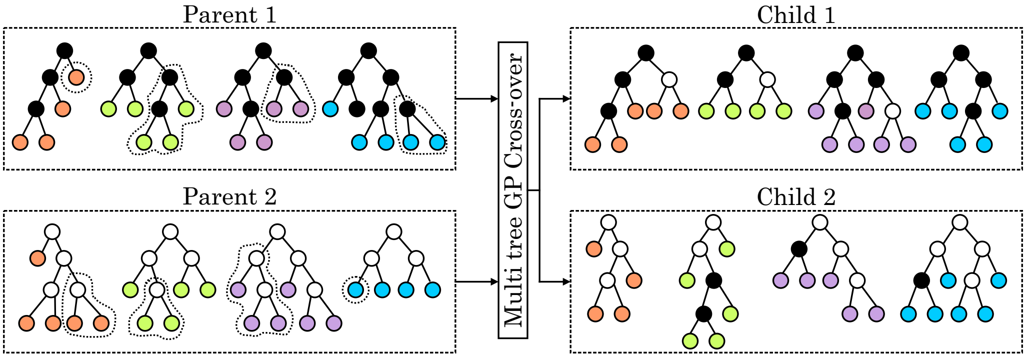


Figure 4.4: The proposed same-index-crossover operator.

4.3.4 Fitness Function

For evaluating each individual in the proposed multi-tree GP approach, we have used a fitness function based on average of the classification accuracy of all the trees in one GP individual. The fitness is defined as

$$fitness = \frac{1}{m}\sum_{i=1}^{m} accuracy(T_i) \tag{4.1}$$

$$accuracy(x) = \frac{1}{2}\left(\frac{TP_x}{TP_x + FN_x} + \frac{TN_x}{TN_x + FP_x}\right) \tag{4.2}$$

where m shows the number of trees and T_i shows the ith tree in a GP individual and accuracy is the balanced accuracy among the two classes given by Equation (4.2). TP refers to true positive, TN refers to true negative, FP refers to false positive, and FN refers to false negative.

Using this fitness function, we allow all the four trees to improve themselves during the evolutionary process, rather maximizing the accuracy of only one tree. When there is a class imbalance problem (different number of instances in different classes), it is more appropriate to use balanced accuracy rather than standard overall accuracy, defined as the ratio between correctly classified instances and total number of instances.

After evolving a GP individual on the training data, we know the different accuracies produced by different tree in that GP individual. Among these trees, we take the highest performing tree on the training set and test it on the test (unseen) data. Since the tree providing the best results on training data, we expect it will perform better on the test data as well compared to the other trees.

4.4 Experiment Design

For carrying out the experiments, each dataset is split by *10-fold cross validation*. The division of instances among the folds is random but it is ensured that the ratio of instances of each class in each fold is the same as in the original dataset. The number of individual GP runs is 30 and the results are reported in terms of the mean and standard deviation of the fitness values. For evolving an individual having four trees on the training data (9 folds), the fitness given in Equation (4.2) is used which computes the average of the accuracies of the four trees. This evolved model is then tested on the test data (1 fold) using only a single-tree having the highest accuracy on the training data. This procedure is repeated 10 times to get the result for 10-*fold cross validation*. Hence for 30 GP runs, the above procedure is repeated 30 times to get 30 fitness values each for training and test sets. In one set of experiments, the random seeds for each of the 30 runs are all different.

4.4.1 GP Parameters

The parameter settings of our proposed multi-tree GP method are listed in Table 4.2. The initial population is generated by "Ramped half-and-half" method and the population size is set to 1024. Tournament selection with size 7 is used to select good individuals for producing new generations while maintaining population diversity. During the evolutionary process, new individuals are produced through crossover, mutation and elitism with percentages of 0.80, 0.19 and 0.01, respectively. The depth of the trees

Table 4.2: Parameter Settings of the GP method.

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

ranges between 2 and 6. The evolutionary process keeps evolving until a maximum of 50 generations is reached or it stops unless an individual with accuracy 100% is found.

4.4.2 Benchmark Methods for Comparison

To check the performance of our proposed multi-tree GP method on the test set, six classification methods are used: Naïve Bayes (NB), k-Nearest Neighbor (k-NN) where $k = 5$, Support Vector Machines (SVM), Decision Trees (J48), Random Forest (RF), and Multilayer Perceptron (MLP).

We have also compared the performance of EGP-4 with the single-tree GP methods. We provide a single set of features e.g., $\text{Lesion}_{\text{Color}}$ features to the traditional GP with one evolved tree in an individual. Here, we will investigate whether a single set of features can effectively distinguish between malignant and benign lesions or remain unable to do so.

4.4.3 Implementation

The implementation of our multi-tree GP method is done using the Evolutionary Computing Java-based (ECJ) package version 23 [114]. The implementations of all the non-GP methods are taken from the most commonly used Waikato Environment for Knowledge Analysis (WEKA) software [85] version 3.8.

4.5 Results and Discussions

This section presents the overall results of the proposed EGP-5 method for the binary classification task of melanoma detection. It also includes analysis of an evolved GP individual with four trees.

4.5.1 Overall Results

The results of the experiments are presented in Table 4.3. Vertically, the table consists of three blocks where the first gives the results of the proposed multi-tree GP method (EGP-4), the second shows the results of the other non-GP based classification methods, and the third shows the results of the single-tree GP methods each using one type of features. Horizontally, the table consists of five columns where the first lists the classification algorithm, the second and the third show respectively the training and the test performances for the PH^2 dataset, and the fourth and the fifth show these performances for the Dermofit dataset. The values of these results are represented as the mean and standard deviation of applying *10-fold cross validation* to the datasets. For all the GP methods (multi-tree and single-tree), the training and test processes are repeated 30 times (as shown in Figure 4.2), hence we get 30 accuracies for each method which are represented as mean and standard deviation ($\bar{x} \pm s$) in Table 4.3.

For making a clear comparison between the proposed method and the other non-GP classification algorithms, and single-tree GP methods, the results are also investigated using *Wilcoxon signed-rank test* with a significance level of 5% and *one-sample t-test*. These statistical test has been applied on the test results to check which classification method has better ability to discriminate between benign and malignant classes. For *Wilcoxon signed-rank test*, the symbols "+", "$-$" and "=" are used to represent significantly better, significantly worse and not significantly different performance, respectively, of the proposed EGP-4 method in comparison with the stochastic single-tree GP classification methods. For *one-sample*

Table 4.3: Comparison between the proposed EGP-4 method, the non-GP and the single-tree GP classification methods: The accuracy (%) on the training and test set of both datasets (results are represented in terms of mean accuracy and standard deviation ($\bar{x} \pm s$)).

	PH^2		Dermofit	
	training	test	training	test
EGP-4	79.69 ± 1.35	**78.87 ± 2.92**	75.63 ± 0.99	**74.57 ± 1.86**
Non-GP Classification Methods				
NB	93.85	77.81 ↑	86.42	72.26 ↑
SVM	89.62	70.00 ↑	95.16	70.02 ↑
k-NN	100.0	74.52 ↑	100.0	72.08 ↑
J48	97.05	71.25 ↑	97.09	73.98 ↑
RF	100.0	76.56 ↑	99.93	71.30 ↑
MLP	77.31 ± 3.64	78.09 ± 3.36 +	80.85 ± 2.48	72.77 ± 2.54 +
Single-tree GP Classification Methods				
$\mathrm{LBP}_{\mathrm{gray}}$	82.84 ± 1.35	65.96 ± 3.96 +	73.41 ± 1.87	59.91 ± 3.57 +
$\mathrm{LBP}_{\mathrm{RGB}}$	84.42 ± 1.43	73.87 ± 2.34 +	75.52 ± 1.62	63.26 ± 3.19 +
$\mathrm{Lesion}_{\mathrm{Color}}$	81.59 ± 2.31	65.70 ± 3.61 +	81.06 ± 1.31	74.13 ± 2.67 +
$\mathrm{Lesion}_{\mathrm{Shape}}$	78.06 ± 1.97	49.89 ± 5.34 +	74.74 ± 2.67	61.74 ± 7.06 +

t-test, the symbols "↑", and "↓" are used to represent significantly better, and significantly worse performance, respectively, of the proposed EGP-4 method in comparison with the deterministic non-GP classification methods. For example, in case of PH^2, the test performance of RF is represented as "76.56" where the "↑" sign represents that EGP-4 has significantly outperformed the RF classification method. Similarly, in case of PH^2, the test performance of $\mathrm{LBP}_{\mathrm{RGB}}$ features in single-tree GP methods is represented as "73.87 ± 2.34" where the "+" sign represents that EGP-4 has significantly outperformed this single-tree GP classification method.

From the results of the statistical tests, it has been observed that the proposed EGP-4 method has not only outperformed all the non-GP classification methods but has also outperformed all the single-tree GP methods.

Comparing the EGP-4 and single-tree GP methods, we have seen that

EGP-4 has potential to evolve good classification models that have more discriminating ability between classes. Moreover, among the two datasets, different type of features are prominent in playing the role of classification. In other words, for PH^2, the $\mathrm{LBP}_{\mathrm{RGB}}$ features have shown highest performance (73.87 ± 2.34) among the four single-tree GP methods. This shows that for images captured from specialized instruments (such as in PH^2 dataset), $\mathrm{LBP}_{\mathrm{RGB}}$ has the most potential to discriminate between *"benign"* and *"malignant"* classes. Whereas, for images captured from standard camera (such as in Dermofit dataset), the $\mathrm{Lesion}_{\mathrm{Color}}$ feature has produced the best results (74.13 ± 2.67) among the four type of features. Hence, we can say that for images captured from different instruments, different feature extraction methods play a vital role in distinguishing between classes.

Comparing the two embedded methods: EGP-4 (developed in this chapter) and the 2SGP-E (developed in chapter 3), we found that the EGP-4 has outperformed 2SGP-E on both datasets. This is due to the fact that in addition to domain independent information, domain specific information in the form of $\mathrm{Lesion}_{\mathrm{Color}}$, and $\mathrm{Lesion}_{\mathrm{Shape}}$ features helps improve the performance of EGP-4 method. On the other hand, 2SGP-E method utilizes only domain independent features either $\mathrm{LBP}_{\mathrm{Gray}}$ or $\mathrm{LBP}_{\mathrm{RGB}}$. Moreover, designing a suitable way of utilizing different types of feature in an effective way has helped EGP-4 achieve performance gains.

We have also seen such trend while evolving an individual using our multi-tree approach. Among all the four trees, on the PH^2 dataset, $\mathrm{LBP}_{\mathrm{RGB}}$ features gave highest accuracy most of the time and in case of evolving an individual on the Dermofit dataset, the tree representing $\mathrm{Lesion}_{\mathrm{Color}}$ features has the highest accuracy among the four trees. Therefore, we decided to use the highest performing tree to check the performance on the unseen (test) data which has produced better results as compared to using average of the accuracies of the four trees on the test data. It is evident from the results of single-tree GP methods for both datasets that selecting an appropriate feature extraction method is important in evolving good classification models.

The existing approaches to skin cancer image classification using GP [12, 13] have used single-tree GP methods and employed only a single dataset (PH^2) to test their performance, however, they have produced good results. Our proposed EGP-4 method has outperformed both the existing methods in terms of classification performance giving 78.87% average accuracy as compared to 70.49% and 78.17% average accuracies of these existing methods on PH^2 dataset, respectively.

4.5.2 Analysis of an Evolved Individual

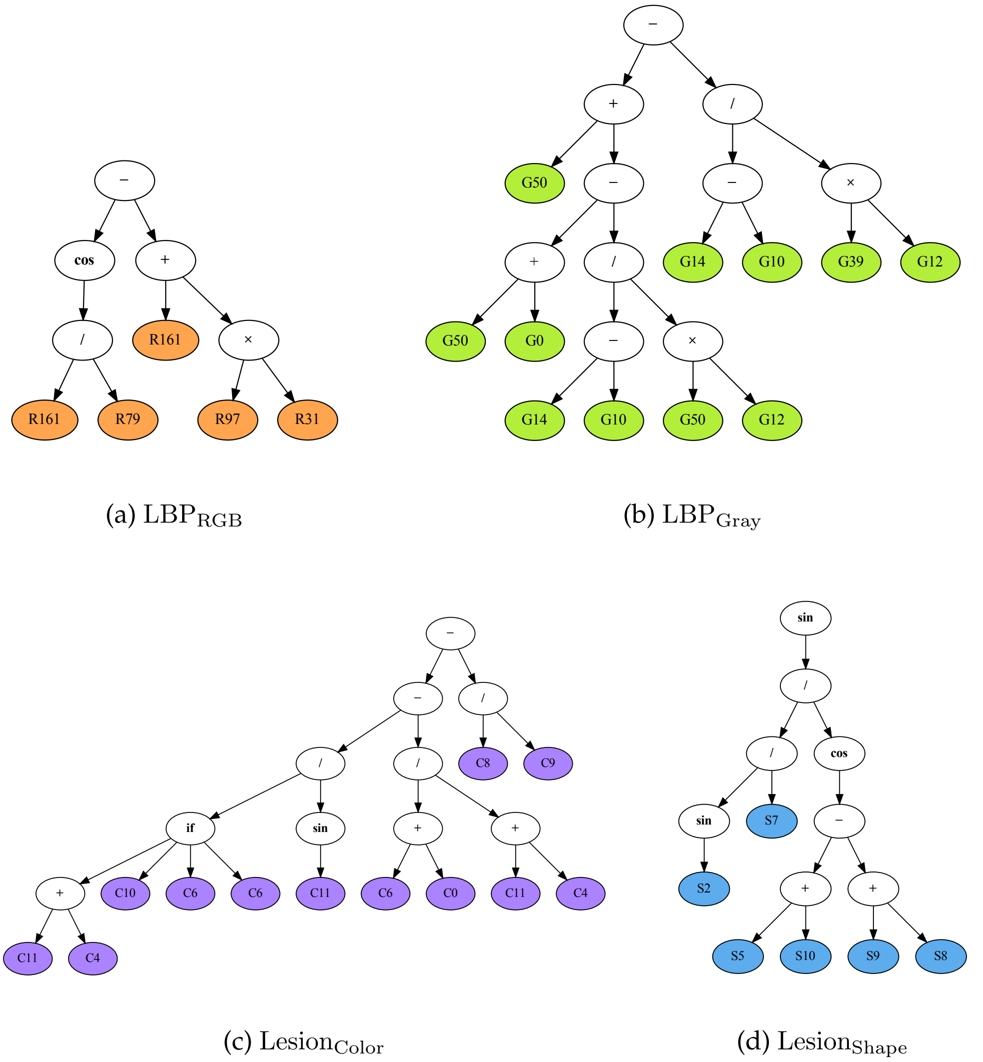
To understand why our proposed method can achieve good performance, we show a good evolved GP individual (Figure 4.5) with four trees evolved using the four types of features, namely a) LBP_{Gray}, b) LBP_{RGB}, c) $\text{Lesion}_{\text{Color}}$, and d) $\text{Lesion}_{\text{Shape}}$, having 80.32% accuracy on the test data. This individual is taken from the PH^2 experiments. In the figure, white nodes represent functions and colored nodes represent terminals. While evolving this model on the training data, the individual accuracy values for LBP_{Gray} tree, LBP_{RGB} tree, $\text{Lesion}_{\text{Color}}$ tree, and $\text{Lesion}_{\text{Shape}}$ tree are 77.08%, 76.74%, 70.49% and 65.63%, respectively. As discussed earlier in Section 4.5, for the PH^2 dataset LBP_{RGB} features have played the most prominent role in classification as compared to the other feature types. This shows that for this dataset, local pixel-based features with color information can extract good information from images about the presence/absence of melanoma. Also the two feature types ($\text{Lesion}_{\text{Color}}$ and $\text{Lesion}_{\text{Shape}}$) which cover the global properties like color variation between the lesion area and the skin region, and border shape are not as good as LBP feature types which have the local pixel-based information.

From Figure 4.5(b) in the LBP_{Gray} tree, the features $G50$ and $G12$ get selected 3 and 2 times, respectively, whereas the expression $G14 - G10$

appears 2 times, which shows that these features have high discriminating ability. Among a total of 177 $\mathrm{LBP}_{\mathrm{RGB}}$ features, a tree (Figure 4.5(a)) constructed from only four dominant features ($R161, R79, R97, R31$) has shown 77.08% accuracy on the training data. This is the highest performing tree among the four trees in this individual, hence applied on the test data and achieved an accuracy of 80.32%. In $\mathrm{Lesion}_{\mathrm{color}}$ tree as shown in Figure 4.5(c), $C6$ and $C11$ (corresponding to $\frac{\mu_R}{\mu_G}$ and $\frac{\mu_B}{\bar{\mu}_B}$) showing the two ratios between the mean of 1) the red channel lesion area and the green channel lesion area, and 2) the blue channel lesion area and the blue channel skin area, are significant. In $\mathrm{Lesion}_{\mathrm{shape}}$ tree as shown in Figure 4.5(d), $S2$, $S5$, $S7$, $S8$, $S9$, and $S10$ are selected which corresponds to the greatest diameter, irregularity indices A, C and D, minor and major asymmetry indices, and Asymmetry Index. These border shape features can provide significant knowledge to the dermatologist in making a diagnosis.

4.5.3 Convergence Plots

The average fitness value per generation of the 30 independent runs (each having 10 independent runs for the 10 folds in $10-$*fold cross validation*) using different seed values on the training data of the two datasets is depicted in Figure 4.6 for EGP-4. These graphs show that on average the programs make larger jumps in the first few generations than in the later generations. This trend has been observed in both datasets. However, in case of PH^2, the improvement in average accuracy is more as compared to the Dermofit dataset. On PH^2, as shown in Fig. 4.6(a), the fitness value has increased from 62.11% to 73.46% in the first 10 generations compared to the increase in fitness from 73.46% to 78.22% over the remainder 40 generations. Similarly on Dermofit as shown in Figure 4.6(b), the highest jump is made from 60.62% to 68.36% only in the first 10 generations compared to the increase from 68.37% to 74.50% over the remainder 40 generations. It can be clearly seen that PH^2 made larger jumps in the start of the evolu-



(a) LBP_{RGB}

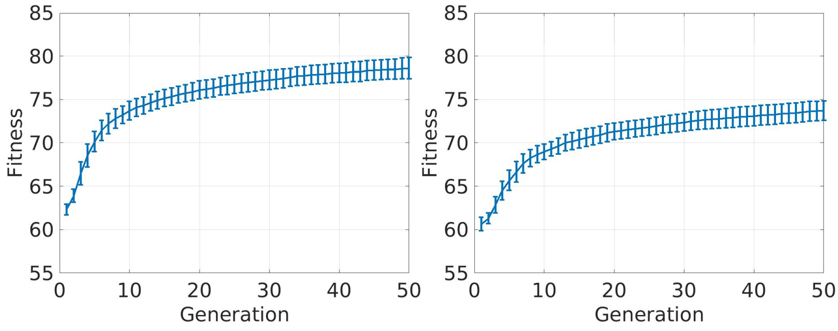
(b) LBP_{Gray}

(c) $\text{Lesion}_{\text{Color}}$

(d) $\text{Lesion}_{\text{Shape}}$

Figure 4.5: The best evolved GP individual for PH^2 dataset having 77.08%, 76.74%, 70.49% and 65.63% accuracy for the four evolved trees on training data and 80.32% accuracy on the test data.

tionary process as compared to Dermofit (11.35 vs 7.76 average improvement in first 10 generations). The standard deviation bars of these 30 independent runs show a different behavior where the earlier generations



(a) PH^2 (b) Dermofit

Figure 4.6: Convergence plots for both datasets in EGP-4.

have less variations than the later ones.

4.6 Further Analysis

We have also analyzed the potential of our EGP-4 method by using a different weighted fitness measure to achieve better classification performance. For evaluating each individual in the proposed MTGP approach, a new weighted fitness function is developed, where the weights are assigned based on the classification accuracy of each tree in one GP individual. Each tree in an individual also works as a simple classifier that can classify binary problem: if an instance x has a negative value on the constructed high-level feature, GP will classify x to "*benign*" class; otherwise to "*malignant*" class. This embedded method with weighted fitness function is termed as "EGP-4_{weighted}". The weighted fitness is defined as

$$fitness = \sum_{i=1}^{k} (W_i \times accuracy(t_i)) \tag{4.3}$$

$$W_i = \frac{accuracy(t_j)}{\sum_{j=1}^{k} accuracy(t_i)} \tag{4.4}$$

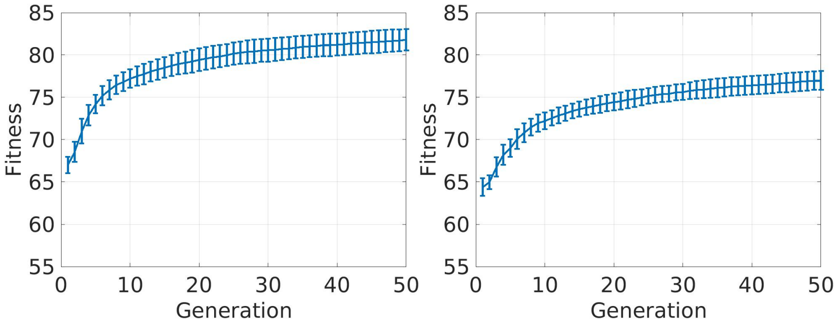
where k is the number of trees and t_i is the ith tree in a GP individual, W_i is the weight assigned to the ith tree, and $accuracy\,(\cdot)$ is the balanced accuracy among the two classes given by Equation (4.2).

Using the weighted fitness function, we allow all the trees to be able to evolve during the evolutionary process and the tree having higher accuracy would contribute more towards the fitness of that individual, via being allocated a higher weight. When the average accuracy of the trees is used as a fitness function using Equation (4.2) in the multi-tree representation in EGP-4, it allows all the trees to grow while giving equal importance to all the four trees. However, the performance of one tree has no influence on the performance of other trees. In other words, the interaction between trees during the evolutionary process was quite limited. Therefore, we designed a new fitness function in this work to evolve GP individuals, where trees influence each other's performance and interacts during the evolutionary process. It is important to note here that the interaction between trees is not in terms of genetic operators (crossover and mutation), but via the weighted fitness function, which encourages the GP method to search for an individual with all the four trees having high classification accuracy.

Furthermore, after getting an evolved model on the training data in EGP-4_{weighted}, each tree in a GP individual often produces a different accuracy on the training data. Among these trees, we take the top two highest performing trees on the training data and use them to classify unseen test data. This is to use the power of two classification models (two trees) to increase the confidence of the prediction. Hence, there are four possible situations: 1) both trees predict an image as benign, 2) both trees predict an image as malignant, 3) first tree predicts an image as benign whereas second tree predicts the same image as malignant, and 4) first tree predicts an image as malignant whereas second tree predicts the same image as benign. For the first two situations, the final prediction is easy to determine as both trees predicts the same class label. However, when the two trees have different predictions, we allocate the final prediction as malignant. This is due to the fact that incorrectly diagnosing a malignant image is too much worse than not diagnosing it at all. For illustration, if either of the trees predicts an image as malignant, there is a possibility that melanoma might be present in that image and hence, this prompts the medical practitioner to get alert and immediately take further medical procedures.

Using this weighted fitness function, the accuracies obtained on the PH^2 and Dermofit datasets are $81.21 \pm 2.23\%$ and $77.14 \pm 1.96\%$, respectively. To highlight the impact of incorporating the new weighted fitness function into the multi-tree representation on finding better solutions, we apply the Wilcoxon signed-rank test to compare its performance with the EGP-4 method without the weights as the fitness measure. We found that the EGP-4$_{\mathrm{weighted}}$ method with the weighted fitness has significantly outperformed the EGP-4 method with balanced accuracy as fitness measure.

4.6.1 Convergence Plots



(a) PH^2

(b) Dermofit

Figure 4.7: Convergence plots on both datasets using the weighted fitness function in the MTGP approach.

To further analyze the effectiveness of using the weighted fitness function, we plot convergence graphs of the EGP-4$_{\mathrm{weighted}}$ method as shown in Figure 4.7(a) and (b) on the PH^2 and the Dermofit datasets, respectively. These graphs of EGP-4$_{\mathrm{weighted}}$ method show quite similar behavior with the graphs shown in Figure 4.6 for the EGP-4 method on both datasets.

However, these graphs start and end at a higher average accuracy compared to the graphs in EGP-4. In case of PH^2, the average accuracy after generation 1 in EGP-4 is 62.11% as shown in Figure 4.6(a), however, it is 66.68% in EGP-4$_{\mathrm{weighted}}$ as shown in Figure 4.7(a). Similarly, in case of dermofit, the average accuracy after generation 1 in EGP-4 is 60.62% as shown in Figure 4.6(b), however, it is 66.68% in EGP-4$_{\mathrm{weighted}}$ as shown in Figure 4.7(b). In addition, the highest training accuracies on average on PH^2 dataset achieved by EGP-4 and EGP-4$_{\mathrm{weighted}}$ are 78.22% and 82.81%, respectively. Similarly, the highest training accuracies on average on dermofit dataset achieved by EGP-4 and EGP-4$_{\mathrm{weighted}}$ are 74.50% and 77.23%, respectively. The standard deviation bars of these 30 independent runs show a different behavior where the variation almost remains the same in the 50 generations.

4.7 Chapter Summary

This chapter has developed a novel embedded feature selection method for skin cancer image classification using multi-tree GP. The method works by incorporating four different types of local and global features extracted from skin cancer images that have information regarding pixel-based gray-level and RGB characteristics, variation in color across the image (inside and between the lesion and skin regions) and geometrical border shape properties. These four type of features are provided to MTGP with newly designed *same-index-crossover/mutation* such that during the evolutionary process, same type of features undergo crossover/mutation in order to avoid mixing of different features in one tree. The proposed EGP-4 method has outperformed all the most commonly used classification algorithms and all the single-tree GP methods showing evidence of powerful discriminating ability between *"malignant"* and *"benign"* skin lesions. We have also found an interesting behavior for selecting suitable feature extraction method for particular type of images captured from a specific instrument. The local pixel-based features have more potential for classifying dermoscopy images, however, global color variation and geometrical shape features provide good discriminating ability between classes for skin cancer images captured from standard camera.

The proposed method has also been extended by developing a new weighted fitness function. This weighted fitness function allowed interaction between the trees in a single MTGP individual during the evolutionary process, which was not found in the EGP-4 method without the weights as the fitness measure. Using the weighted fitness function, the new MTGP method achieved significantly better accuracies on both datasets than the fitness function without the weights.

Though these methods have provided very good results for the complex task of melanoma detection, the binary classification case was only considered in this chapter. Motivated by the promising results, extending the method for multi-class classification will be investigated in the next chapter. Moreover, these methods relied on using only the implicit feature selection ability of GP and did not explore feature construction which can also help improve the classification performance. As we have seen in Chapter 3 that new high-level features constructed from the selected features help improve the classification performance, we will explore feature construction using a MTGP approach in the next chapter.

Chapter 5

Multi-tree GP for Wrapper based Feature Construction

5.1 Introduction

GP has implicit feature selection ability to automatically select important features as its terminals. Feature selection only selects the prominent features and cannot improve the quality of the original features by generating new features. Feature construction can be utilized to generate new informative features from the original set of features. The evolved GP tree(s) can be considered as a new constructed feature(s). Since GP keeps improving the fitness of these new constructed features (CFs) during the evolutionary cycle by measuring their goodness against a fitness function, the evolved CFs most probably have high discriminating ability between classes, which can greatly help in achieving good classification performance [197, 198].

In the previous chapter, an embedded feature selection method using multi-tree GP has been developed which performs feature selection as part of the model construction process. In order to generate new high level features, a wrapper feature construction method can be utilized which uses a predictive model to score feature subsets which are used to train a classification model such as a decision tree. To evolve multiple trees in a single GP individual, the multi-tree GP (MTGP) representation is used as discussed in Chapter 4. Each tree in a MTGP individual can be considered as a new constructed feature. The method proposed in Chapter 4 considers each tree as a classifier, however, each tree can also be considered as a constructed feature. Without changing the design of the MTGP individual, using a wrapper method can generate new informative features. The wrapper feature construction method can be utilized to solve both the binary and multi-class classification tasks without the need of changing the algorithm design.

Since, we are interested to encompass the different local, global, texture, and color image properties of the skin lesion images, we have employed MTGP to effectively evolve multiple trees (constructed features) each based on a specific property, e.g., one tree for gray-scale features, one for pixel-based color features, and another for border shape features. On the other hand, in a MTGP approach evolving multiple trees based on all different type of features may not result in meaningful constructed features. This is because the interactions between different kinds of features may ruin effectiveness of each feature type. Moreover, the number of redundant features increases when combining together all the sets of features which may hinder in evolving good constructed features. Similarly, evolving these CFs individually in a single-tree GP approach (such as in Chapter 3) will use only one specific property of skin images (e.g. based on either local features or global features) and, hence, may not provide sufficient information necessary for classification. Moreover, using all these different features together to evolve a single-tree GP-constructed feature has resulted in poor performance in our preliminary experiments. Therefore, MTGP is suitable where different image properties (local, global, texture, and color information) encompassed in different sets of features are necessary to have sufficient informative features in terminal set.

Since wavelet features extracted from skin images encompass detailed internal structure and global border shape characteristics, this chapter develops a method to construct new high-level features using three-level pyramid structured wavelet decomposition as well as using the four types of features used in Chapter 4.

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

5.1.1 Chapter Objectives

Unlike existing approaches, this chapter develops two feature construction methods where one method constructs four features in a wrapper based feature construction approach (WGP-4) using the four types of features as described in Chapter 4 (on page 112), and the second method (WGP-5) constructs five features by adding a new set of wavelet features to the first method. The CFs are provided to a machine learning classification algorithm (such as $k-$nearest neighbor or decision trees) for classification. This feature construction ability of the MTGP methods generate knowledge-guided features which help the classification algorithm to produce good results. These methods aim at automatically generating new features from a variety of local and global features to discriminate images of different classes. The following objectives will be explored in this chapter:

- Developing two new multi-tree based GP methods both with a wrapper feature construction approach with different types of features for binary and multi-class skin image classification problems.

- Assessing the performance of the proposed classification methods quantitatively and comparing it to six commonly used classification algorithms and twelve single-tree GP methods.

- Investigating the efficiency of the proposed methods in terms of analyzing computation time to train the proposed method and to test its performance on the test images.
- Analyzing the interpretability of the constructed features from WGP-4 and WGP-5.
- Investigating the different types of prominent features for the diagnosis of skin images based on the frequency of appearance in CFs .

5.1.2 Chapter Organization

The rest of the chapter is organized as follows. Section 5.2 describes the feature extraction methods used in this chapter. Section 5.3 presents the proposed multi-tree GP wrapper based feature construction methods, their representation, the terminal sets, the function sets, crossover and mutation operators, and the fitness functions. Section 5.4 describes the experiments performed, GP parameters and benchmark methods for comparison. Section 5.5 presents the results and discusses how well they address the chapter objectives. Section 5.6 provides detailed analysis in terms of evolved CFs, computation time and frequency of features appearing in the CFs. Section 5.7 concludes the chapter with the achievements of the method, and its possible limitations.

5.2 Feature Extraction

In this chapter, we capture texture information from images using three-level pyramid-structured wavelet decomposition [54], local information using LBP image descriptor [145], global information using lesion color variation [172], and border shape features [80, 117]. The details of these methods are presented in Section 2.1.2 on page 30, Section 4.2.2 on page

113, and Section 4.2.3 on page 113, respectively. These different types of features are incorporated to: 1) provide necessary discriminative information to GP for effective feature construction, 2) analyze which type of features are more prominent to classify which type of images (dermoscopy and standard camera).

5.2.1 Wavelet-based Features

The visual characteristics of a skin lesion, which formulates the basis of clinical diagnosis (e.g. the asymmetry, border, color and diameter (ABCD) rule of dermoscopy), can be represented through texture analysis [80]. The pyramid-structured wavelet analysis [54] 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:

$$luminance = (0.3 \times R) + (0.59 \times G) + (0.11 \times B) \tag{5.1}$$

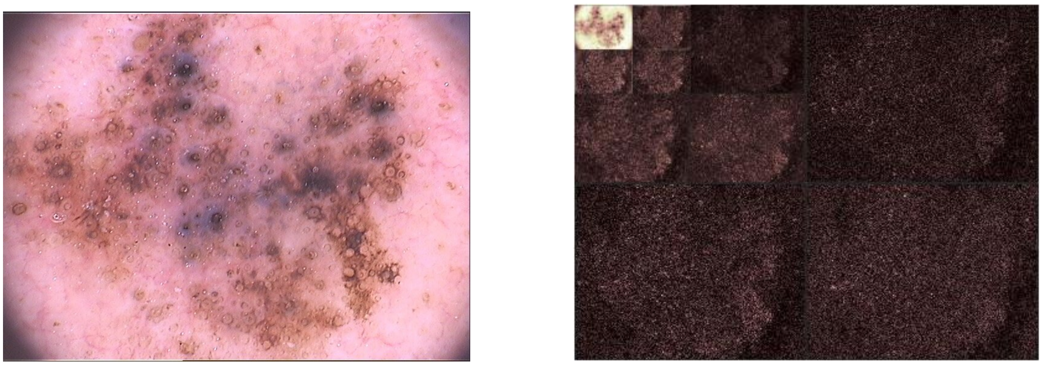
where R, G and B are, respectively, the red, green, and blue color channels.

Eight statistical measures and ratios are extracted from the wavelet coefficients. These measures are mathematically represented in Table 5.1 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 as shown in Figure 5.1.

Figure 5.1(a) shows a skin lesion image and Fig, 5.1(b) shows its pyramid-structured wavelet decomposition. Unlike [80], three-level pyramid-structured wavelet decomposition extracted from four color channels has been reported for the first time in this work. Figure 5.2

Table 5.1: Statistical measures applied to the wavelet coefficients [80].

Measure	Formula	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	$H_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} \left(x_{jk}^2 \times log(x_{jk}^2)\right)}{J \times K}$
Average Energy	$AvgE_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} \lvert x_{jk} \rvert}{J \times K}$	Skewness	$S_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} \left(\frac{x_{jk}-M(n_i)}{Std(n_i)}\right)^3}{J \times K}$



(a)

(b)

Figure 5.1: A skin image, shown in (a), with a three-level pyramid-structured wavelet decomposition, shown in (b).

displays a symbolic representation of wavelet tree where ovals represent nodes. There are 13 nodes in the wavelet tree (1 parent node which is the original image, and 4 nodes in each of the three subsequent levels ($4 \times 3 = 12$)). The eight measures computed on each tree node yield a total of 8×13 features, for each color channel. Hence, there are a total of 416 (= 8 measures $\times$ 13 nodes $\times$ 4 color channels) wavelet features extracted.

5.3 The Proposed Methods

This section provides a detailed description of the proposed MTGP wrapper methods: 1) MTGP in a wrapper feature construction approach using

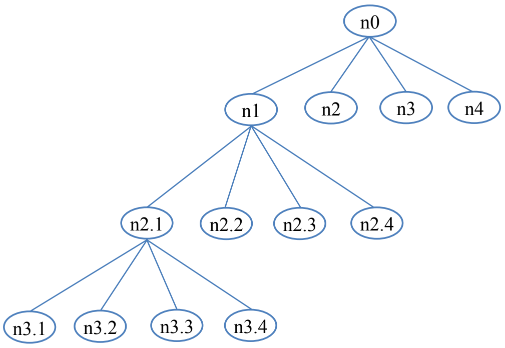


Figure 5.2: A schematic three-level wavelet tree with nodes in oval.

four sets of features (WGP-4), and 2) MTGP in a wrapper feature construction approach using five sets of features (WGP-5). The detail starts by presenting an overview of the algorithm to evolve a GP individual in order to highlight the key components of the proposed methods, and how the constructed features from the evolved individuals are used for classification. Then the program structure, i.e., the terminal and the function sets, the crossover and mutation operators, and the fitness function, are discussed.

The proposed methods operates on a set of predefined/extracted features which include local and global information about the skin images. The local features are extracted with the help of LBP descriptor which works with the pixel values and can significantly capture informative features about various skin properties such as lines/streaks, blobs, homogeneous regions, and irregular border patterns. The global features are extracted by focusing on shape and color variation characteristics of skin lesions. These features are defined in [172] and [80]. These features are of utmost importance because without using these human crafted features, it is difficult to achieve good performance for such a difficult task as skin image classification. These global features capture the properties of the ABCD rule of dermoscopy, which plays a vital role for the dermatologist in distinguishing malignant from benign images. The pyramid-

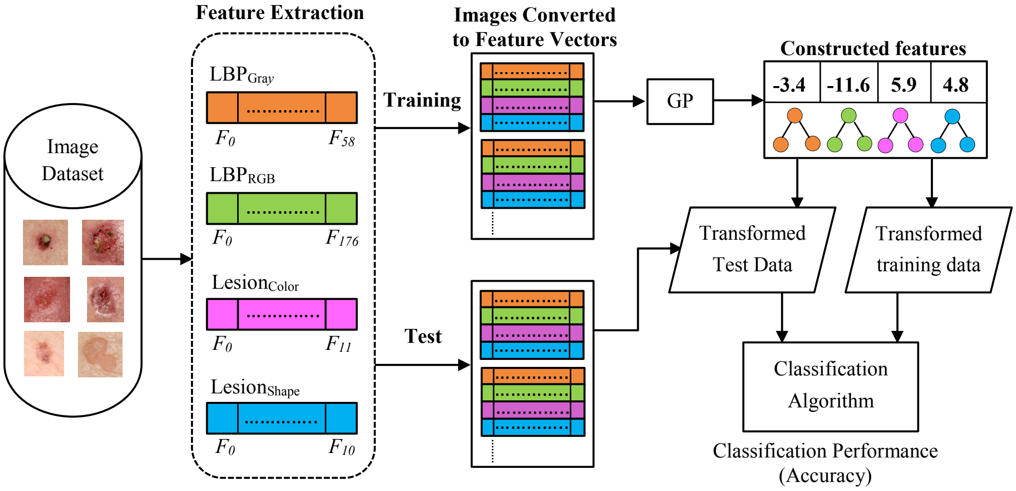


Figure 5.3: Overview of the proposed WGP-4 method.

structured wavelet analysis [54] provides detailed texture properties (local features), as well as overall characteristics (global features) of the skin lesions. Hence, incorporating these informative features help the classifier learn better and produce an effective model.

5.3.1 The Overall Algorithm of WGP-4

The overall structure of the MTGP in a wrapper approach using four different sets of image features for skin image classification is shown in Figure 5.3. First, the four types of features are extracted from each image of a dataset. Hence, one image is represented by four feature vectors namely LBP_{Gray}, LBP_{RGB} $\text{Lesion}_{\text{Color}}$, and $\text{Lesion}_{\text{Shape}}$. Then the dataset is divided into training and test sets. The MTGP algorithm runs on the training set of the dataset to select a subset of relevant features for each type of features among the four feature types. It constructs four features from these selected features. In other words, GP evolves four trees in a single individual based on the four types of features, which is the evolutionary feature construction process. Then using these four trees (constructed features) the training set and the test set are transformed to a new training set and a new test set by constructing new features from the four trees evolved during the evolutionary process. A classification algorithm (such as a decision tree) is then trained on the transformed training set. The learned classifier is then applied to the transformed test set to obtain the final test classification performance.

GP Program Representation of WGP-4

Each tree in a GP individual is generated using a single type of features. A GP individual consists of four trees in WGP-4. For illustration (as shown in Figure 4.3), the terminal set of the first tree consists of $\mathrm{LBP}_{\mathrm{Gray}}$ features only. Similarly, the terminal sets of the second, third and fourth trees consist of $\mathrm{LBP}_{\mathrm{RGB}}$, $\mathrm{Lesion}_{\mathrm{Color}}$, and $\mathrm{Lesion}_{\mathrm{Shape}}$ features, respectively. However, all the trees share the same function set that consists of seven operators as described in Section 5.3.4.

5.3.2 The Overall Algorithm of WGP-5

WGP-5 is an extension of WGP-4 with a set of new wavelet features added to the method. With the multi-scale and multi-channel properties of these wavelet features as described in Section 5.2.1, WGP-5 aims at providing more robust CFs to a classification algorithm to achieve performance gains. The overall structure of the WGP-5 method is shown in Figure 5.4.

5.3.3 Terminal Set

The terminal set of WGP-4 comprises of the four types of features, similar to EGP-4 in Chapter 4. The terminal set of WGP-5 consists of the four types of features from EGP-4 and a fifth set of wavelet features has been added. All the five types of features are listed in Table 5.2. The value of the ith feature for the above five types of features is indicated by Gi, Ri, Ci, Si,

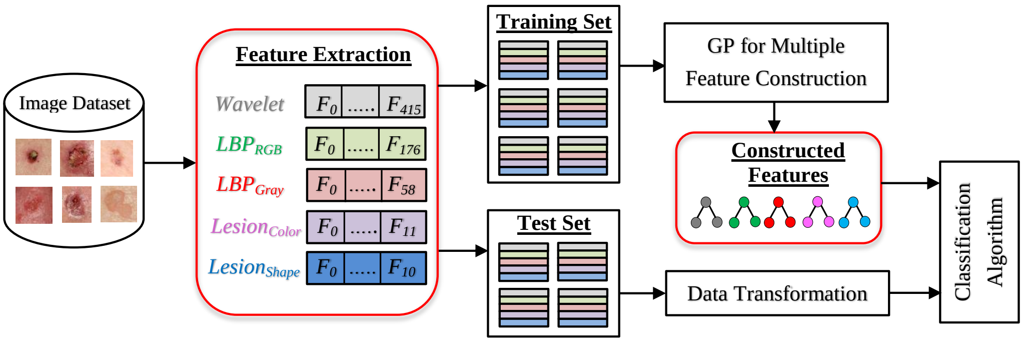


Figure 5.4: Overview of the proposed WGP-5 method.

Table 5.2: Various types of features in the terminal set.

Features	Number	Description
LBP$_{\text{RGB}}$ (C_i)	177	LBP from RGB channels (Section 4.2.1).
LBP$_{\text{Gray}}$ (G_i)	59	LBP from gray images (Section 4.2.1).
Lesion$_{\text{Color}}$ (L_i)	12	Color variation from RGB channels (Section 4.2.2).
Lesion$_{\text{Shape}}$ (S_i)	11	Domain-specific geometry-based shape features extracted from lesion binary masks (Section 4.2.3).
Wavelet (W_i)	416	Wavelet decomposition from RGB and luminance channels (Section 5.2.1).

and Wi respectively, as shown by the GP individual in Figure 5.13.

5.3.4 Function Set

The function set of WGP-4 and WGP-5 is the same as EGP-4 method described in Chapter 4. The details of function set can be found in Section 4.3.2 on page 116.

5.3.5 Crossover and Mutation

Both the proposed methods: WGP-4 and WGP-5 use same-index-crossover/mutation, similar to EGP-4. The details including description, and algorithms of this type of crossover and mutation operators can be found in Section 4.3.3 on page 117. A graphical representation is shown in Figure 5.5.

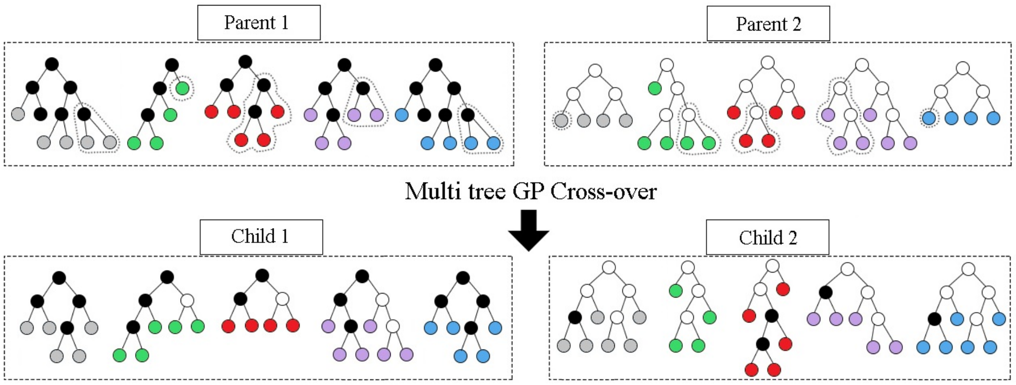


Figure 5.5: The same-index-crossover operator in WGP-5.

5.3.6 Fitness Function

The balanced classification accuracy is used as the fitness function in WGP-4 and WGP-5, which is defined and explained in detail in Section 3.2.1 on page 87.

5.4 Experiment Design

The aim and design of the experiments are discussed in this section. The discussions also include the datasets, the benchmark methods used for comparison, the experiments and the parameter settings.

5.4.1 Methods for Comparison

To evaluate the performance of the proposed WGP-4 and WGP-5 methods, six classification methods are used: Naïve Bayes (NB), k-Nearest Neighbor (k-NN) where $k = 5$, Support Vector Machines (SVMs) with a Radial Basis Function (RBF) kernel, Decision Trees (J48) where the minimum number of instances per leaf equals 2, Random Forest (RF), and Multilayer Perceptron (MLP). These methods are implemented through the commonly used

Waikato Environment for Knowledge Analysis package [85]. The parameters for these classification methods are kept the same as in Chapters 3 and 4 where they are specified empirically as they have shown the best performance amongst other settings.

Single-tree GP Methods

We compare the two methods with twelve standard single-tree GP feature construction methods. WGP-1 means wrapper based single-tree GP method using each set of features and the combination of all features. RF classification algorithm is used for the WGP-1 methods as it has mostly produced the best results among the six classification algorithms (NB, SVM, k-NN, J48, RF, and MLP) in case of MTGP methods. Similarly, EGP-1 is the embedded based single-tree GP method. WGP-4 is the wrapper based 4-tree GP method, run with different classification algorithms. Similarly, WGP-5 is the wrapper based 5-tree GP method experimented with different classifiers such as NB, SVM, k-NN, J48, RF, and MLP. EGP-4 and EGP-5 are the embedded based 4-tree and 5-tree GP methods. We have also experimented the performance of all types of features by combining them together in one vector and providing them to GP to evolve a single CF. The results are represented by 'All' in Tables 5.3 and 5.4 (block 2 and 3). As a GP individual in both WGP-1 and EGP-1 methods has only one tree, they use only one CF for classification.

5.4.2 Experiments

Two sets of experiments are conducted in this study. The first set of experiments are intended for the purpose of binary classification specifically detecting melanoma, which attempts to differentiate melanoma images from all the collection of images provided. The second set of experiments investigate the effectiveness of the two feature construction methods for multi-class classification. The binary classification is relatively easy (two classes) compared to the multi-class classification (three classes in PH^2 and ten classes in Dermofit). The results are compared with the other classification methods as described in Section 5.4.1. The binary classification results are also compared with the two embedded methods 1) EGP-4 developed in Chapter 4, and 2) EGP-4 extended by adding fifth set of wavelet features denoted by EGP-5. Both the EGP-4 and EGP-5 are binary classification methods for the diagnosis of melanoma. In WGP-4 and WGP-5, six classification methods, namely NB, SVM, $k-$NN, J48, RF, and MLP (each individually executed) are used as a wrapper feature construction algorithm to search which classification algorithm performs better for binary and multi-class classification methods.

For carrying out the experiments, the *10-fold cross validation* approach is adopted in this chapter. This is because PH^2 is very small (200 images) and some classes in Dermofit have very small number of images (Pyogenic Granuloma with 24 images). To segment the data into 10 folds, stratified random sampling is applied. The number of GP runs is 30. The results are represented in terms of the mean and standard deviation of the test performance values. For the task of binary classification, average sensitivity and average specificity values have also been reported.

5.4.3 Parameter Settings

The GP parameter settings of WGP-4 and WGP-5 are the same, except that an individual in WGP-4 has four trees and an individual in WGP-5 has five trees. The details of these parameter settings can be found in Section 3.4.1 on page 91.

5.5 Results and Discussions

This section presents and describes the findings of the experiments. The results are expressed as the mean and standard deviation of the 30 runs of GP each of which is calculated is the average of the 10-fold cross validation. For binary classification, the results are represented in terms of sensitivity, specificity and balanced accuracy where mean and standard deviation ($\bar{x} \pm s$) are shown in Table 5.3. For the task of multi-class classification, the balanced accuracy ($\bar{x} \pm s$) is shown in Table 5.4. The deterministic methods' results are given in terms of the mean of implementing *10-fold cross validation* on the datasets.

Wilcoxon signed-rank test (with a significance level of 5%) and *one sample t-test* are used to compare the performance of different methods. The former is performed between the highest performing FC stochastic method (WGP-5) and the other stochastic GP methods, whereas the latter is performed between WGP-5 and the non-GP deterministic methods. In case of one sample t-test, two symbols "↑" and "↓" are shown (in Tables 5.3 and 5.4) in front of a deterministic method which show that WGP-5 has significantly better and worse performance, respectively, compared to the deterministic method. In case of Wilcoxon signed-rank test, "+", "−" and "=" represent that the WGP-5 method has outperformed, has been outperformed, or has equal performance compared to the corresponding GP method. In each block of the Tables 5.3 and 5.4, the highest classification performance on each dataset in terms of balanced accuracy is made **bold** to clearly see which method has provided the best results among the methods listed in one block.

5.5.1 Binary Classification

The binary classification results of the two datasets are presented in Table 5.3. The first column shows the wrapper or embedded based single or multi-tree GP and non-GP methods. The second, third, and fourth columns show the sensitivity, specificity and balanced accuracy on average for PH^2 dataset, respectively. The fifth, sixth, and seventh columns show these values for the Dermofit dataset, respectively.

Table 5.3: Results of Non-GP, single-tree GP and Multi-tree GP methods for **Binary Classification**: Sensitivity, Specificity, and Balanced Accuracy (%) on the two real-world skin cancer datasets, along with the statistical significance tests.

Algorithm		PH^2			Dermofit		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Non-GP Methods	NB	60.00	94.38	77.19 ↑	97.32	96.67	**96.99** ↑
	SVM	25.00	99.38	62.19 ↑	27.68	100.0	63.84 ↑
	k-NN	57.50	90.63	74.06 ↑	76.07	98.79	87.43 ↑
	J48	57.50	87.50	72.50 ↑	93.57	97.27	95.42 ↑
	RF	55.00	98.13	76.56 ↑	61.79	99.70	80.74 ↑
	MLP	59.00 ± 2.80	98.50 ± 1.54	**78.93 ± 2.47** ↑	92.34 ± 1.76	98.35 ± 2.60	94.48 ± 1.97 ↑
WGP-1	LBP_{Gray}	80.61 ± 4.88	96.38 ± 2.12	85.25 ± 3.10 +	83.17 ± 4.52	90.85 ± 3.89	85.58 ± 2.06 +
	LBP_{RGB}	85.61 ± 3.09	99.07 ± 3.41	88.25 ± 3.02 +	83.42 ± 4.22	92.66 ± 2.15	86.75 ± 2.03 +
	$\text{Lesion}_{\text{Color}}$	78.77 ± 4.23	94.21 ± 2.19	82.00 ± 2.03 +	88.24 ± 4.69	98.49 ± 1.52	92.31 ± 0.92 +
	$\text{Lesion}_{\text{Shape}}$	74.46 ± 5.87	90.57 ± 4.56	84.00 ± 1.22 +	86.77 ± 5.44	97.50 ± 2.31	90.29 ± 1.32 +
	Wavelet	95.73 ± 1.11	100.0 ± 0.00	**90.75 ± 2.45** +	99.71 ± 0.32	99.78 ± 0.44	**99.33 ± 0.69** +
	All	95.68 ± 1.33	100.0 ± 0.00	90.50 ± 3.02 +	99.71 ± 0.31	99.78 ± 0.44	99.31 ± 0.81 +
EGP-1	LBP_{Gray}	57.00 ± 6.78	75.25 ± 1.79	65.96 ± 3.96 +	49.25 ± 5.64	70.57 ± 4.98	59.91 ± 3.57 +
	LBP_{RGB}	61.25 ± 3.91	86.49 ± 2.35	73.87 ± 2.34 +	61.50 ± 6.92	65.02 ± 5.68	63.26 ± 3.19 +
	$\text{Lesion}_{\text{Color}}$	53.50 ± 6.44	77.90 ± 5.14	65.70 ± 3.61 +	72.96 ± 7.96	75.30 ± 3.47	74.13 ± 2.67 +
	$\text{Lesion}_{\text{Shape}}$	43.75 ± 9.83	56.03 ± 7.98	49.89 ± 5.34 +	61.27 ± 4.79	62.21 ± 7.80	61.74 ± 7.06 +
	Wavelet	59.50 ± 5.68	85.13 ± 3.93	**72.31 ± 2.75** +	82.86 ± 5.20	93.40 ± 1.95	**88.13 ± 3.58** +
	All	57.50 ± 6.89	85.75 ± 3.15	71.63 ± 3.93 +	86.43 ± 1.36	89.71 ± 3.81	88.07 ± 2.26 +
WGP-4	NB	80.08 ± 3.31	91.32 ± 1.33	85.70 ± 2.65 +	77.21 ± 6.78	83.69 ± 0.69	80.45 ± 2.18 +
	SVM	67.25 ± 7.20	95.79 ± 0.57	81.52 ± 3.58 +	62.52 ± 5.33	98.14 ± 0.70	80.33 ± 2.71 +
	k-NN	29.08 ± 11.59	93.44 ± 0.72	61.26 ± 4.05 +	41.00 ± 5.62	97.54 ± 0.71	69.27 ± 2.89 +
	J48	74.67 ± 6.22	95.69 ± 0.63	85.18 ± 3.72 +	71.48 ± 6.58	96.88 ± 0.82	84.18 ± 4.11 +
	RF	79.50 ± 3.12	100.0 ± 0.00	**89.75 ± 1.55** +	90.72 ± 1.68	100.0 ± 0.00	**95.35 ± 0.83** +
	MLP	56.88 ± 6.43	90.86 ± 1.47	73.87 ± 1.52 +	63.48 ± 5.33	96.51 ± 0.70	80.47 ± 2.02 +
WGP-5	NB	86.42 ± 1.16	93.12 ± 0.70	89.77 ± 1.84	95.60 ± 0.49	97.22 ± 0.32	96.21 ± 1.09
	SVM	73.83 ± 1.27	99.13 ± 0.25	86.48 ± 2.35	95.18 ± 0.60	99.61 ± 0.11	97.26 ± 1.25
	k-NN	30.00 ± 0.68	96.68 ± 0.28	63.34 ± 2.67	73.00 ± 0.45	99.42 ± 0.13	86.04 ± 2.52
	J48	77.08 ± 0.08	98.14 ± 0.04	87.61 ± 3.08	95.11 ± 0.54	99.14 ± 0.51	96.99 ± 0.70
	RF	99.75 ± 0.67	99.98 ± 0.10	**99.93 ± 0.24**	99.75 ± 0.48	99.96 ± 0.10	99.87 ± 0.23
	MLP	69.94 ± 5.98	89.68 ± 0.86	74.29 ± 1.94	100.0 ± 0.00	100.0 ± 0.00	**100.0 ± 0.00**
EGP-4		73.65 ± 4.92	84.09 ± 5.10	78.87 ± 2.92 +	75.82 ± 3.08	73.32 ± 3.45	74.57 ± 1.86 +
EGP-5		75.25 ± 4.47	87.31 ± 3.48	81.28 ± 1.13 +	83.75 ± 3.48	81.43 ± 4.14	82.59 ± 3.85 +

Among the six classification algorithms in the WGP-4 method (Table 5.3), it has been found that RF achieved the highest accuracy with 89.75% and 95.35% on average on the PH^2 and Dermofit datasets, respectively. However, there is a substantial performance improvement by using the WGP-5 method, where the wavelet-based texture features help boost the classifier's distinguishing capability. In case of WGP-5, RF and MLP achieved the highest average accuracies with 99.93% and 100.0% on PH^2 and Dermofit, respectively. It is clearly seen from the results in Table 5.3 that the wrapper approaches (WGP-4 and WGP-5) have outperformed the embedded approaches (EGP-4 and EGP-5) with around 18% increase on both PH^2 and Dermofit. For sensitivity and specificity, it is more important to get higher sensitivity which represents the total number of correctly classified melanoma, as compared to specificity which represents the correctly classified benign lesions. Analyzing the results in terms of sensitivity and specificity, it has been observed that for PH^2, the WGP-5 method is the most effective for identifying melanoma images by achieving the highest sensitivity of 99.75% on average among all the other methods. On Dermofit, the WGP-5 achieved the highest sensitivity of 100.0%, classifying all the melanoma images correctly.

The outcomes of the statistical significance test shown in Table 5.3 have revealed that the WGP-5 method not only dominated all single-tree GP methods (EGP-1 and WGP-1) but also consistently outperformed all multi-tree GP methods such as WGP-4, EGP-4 and EGP-5 which has demonstrated the efficacy and validity of this approach for identifying melanoma in skin images.

5.5.2 Multi-class Classification

The multi-class classification results for the two datasets are given in Table 5.4. We can clearly observe that RF has achieved the best classification performance on the test data among the six classification algorithms in the WGP-4 and WGP-5 methods. In WGP-4, it has provided 96.42% and

80.74% average accuracy which improves to around 1% and 6% on PH^2 and Dermofit, respectively, by using WGP-5 demonstrating the potential of wavelet-based texture features. It is worth mentioning here that PH^2 has three classes (relatively easy task) and Dermofit has ten classes (more challenging task). Most of the classifiers in both WGP-4 and WGP-5 have performed well for a 3-class problem like SVM that has produced 77.17% and 84.92% average accuracy, respectively, but only RF achieved well enough for the complicated 10-class problem with an average accuracy as high as 86.77%. To the best of our knowledge, the state-of-the-art result on Dermofit for this 10-class skin image classification problem is presented by [96]. The authors have reported an *overall accuracy* of 80.80% using 5-fold cross validation which comes out to be 60.12% *balanced accuracy* (as calculated from the confusion matrix provided in [96]). The WGP-5 method has outperformed this state-of-the-art method by achieving an increase of nearly 26% on the Dermofit dataset.

The outcomes of the statistical significance test shown in Table 5.4 have revealed that the WGP-4 and WGP-5 methods significantly performed better than all the non-GP methods and the WGP-1 methods on the simple (PH^2) and challenging (Dermofit) datasets indicating their usefulness for skin image classification problems.

5.5.3 Comparisons with Other Classification Methods

From the results in Tables 5.3 and 5.4, it has been observed that the FC methods have more potential to classify melanoma images than the non-GP classification methods. MLP has dominated other non-GP classification methods by providing 78.75% average accuracy on PH^2 for melanoma detection in binary classification. The embedded MTGP methods (EGP-4 and EGP-5) remain unable to provide good results as compared to the non-GP methods. WGP-4 has four sets of features (not using wavelet features), whereas, non-GP methods have five sets of features, which is the possi-

Table 5.4: Results of non-GP, single-tree GP and multi-tree GP methods **Multi-Class Classification**: Balanced Accuracy (%) on the two real-world skin cancer datasets, along with the statistical significance tests.

Algorithm		PH^2	Dermofit
Non-GP Methods	NB	71.00 ↑	45.92 ↑
	SVM	59.50 ↑	51.08 ↑
	k-NN	65.50 ↑	43.54 ↑
	J48	58.00 ↑	50.08 ↑
	RF	**71.50** ↑	47.92 ↑
	MLP	68.23 ± 2.67 +	**65.78** ± 3.42 +
WGP-1	LBP_{Gray}	52.00 ± 6.34 +	35.27 ± 1.02 +
	LBP_{RGB}	62.42 ± 4.84 +	41.80 ± 1.94 +
	$\text{Lesion}_{\text{Color}}$	52.17 ± 3.23 +	43.41 ± 0.00 +
	$\text{Lesion}_{\text{Shape}}$	51.33 ± 4.37 +	41.28 ± 0.00 +
	Wavelet	67.17 ± 4.78 +	43.48 ± 1.12 +
	All	**92.08 ± 2.08** +	**65.98 ± 4.67** +
WGP-4	NB	75.01 ± 1.76 +	49.23 ± 1.51 +
	SVM	77.17 ± 2.00 +	38.69 ± 1.34 +
	k-NN	57.43 ± 2.40 +	41.13 ± 0.91 +
	J48	80.64 ± 2.24 +	69.25 ± 1.41 +
	RF	**96.42 ± 1.45** +	**80.74 ± 1.24** +
	MLP	47.47 ± 2.42 +	31.88 ± 1.18 +
WGP-5	NB	80.31 ± 2.03	58.99 ± 1.25
	SVM	84.92 ± 2.31	53.05 ± 1.57
	k-NN	63.46 ± 2.55	47.46 ± 1.85
	J48	85.82 ± 1.60	74.05 ± 1.52
	RF	**97.39 ± 1.00**	**86.77 ± 0.88**
	MLP	65.64 ± 1.92	41.84 ± 1.30

ble reason of non-GP methods outperforming WGP-4. However, WGP-5 dominated all the six classification methods by producing 99.93% average accuracy on the test data. Similarly, among the six non-GP methods, NB has provided the best accuracy (96.99%) on Dermofit, which has been significantly outperformed by the WGP-5 by producing an accuracy of 100.0% on the unseen data.

On the difficult multi-class classification problem, RF with 71.50% and MLP with 65.78% remain most prominent among non-GP methods on PH^2 and Dermofit, respectively. However, the WGP-5 has significantly outperformed both these classification methods by achieving 97.39% and 86.77% average accuracy on PH^2 and Dermofit, respectively. To have a fair comparison, the non-GP methods are given the five sets of features concatenated together in a single vector with a total of 675 (= 416 Wavelet + 177 $\mathrm{LBP}_{\mathrm{RGB}}$ + 59 $\mathrm{LBP}_{\mathrm{Gray}}$ + 12 $\mathrm{Lesion}_{\mathrm{Color}}$ + 11 $\mathrm{Lesion}_{\mathrm{Shape}}$) features. Despite this, several of these methods are still incapable of producing good results. That is the key reason why various sets of features are unable to produce quality results without designing an appropriate way to integrate them. Hence, we conclude that the WGP-5 method effectively and automatically evolves powerful CFs which help the wrapper classification algorithm achieve good performance.

5.5.4 Comparisons with Single-tree GP Methods

In comparing the MTGP and single-tree GP methods for binary classification, we have observed that the MTGP method is the most capable of generating good classification models that can achieve good sensitivity as well as specificity values. The WGP-5 method constructs five features, each from Wavelet, $\mathrm{LBP}_{\mathrm{RGB}}$, $\mathrm{LBP}_{\mathrm{Gray}}$, $\mathrm{Lesion}_{\mathrm{Color}}$ and $\mathrm{Lesion}_{\mathrm{Shape}}$ feature sets. These five CFs are provided to the wrapper algorithm for classification, e.g., J48. On the other hand, in case of the single-tree GP method with a wrapper approach, GP constructs only one feature (based on either of the five sets of features), which is given to the same classification algorithm. In this work, we have calculated the performance using all sets of features collectively provided to GP to generate a single tree (CF), represented by 'All' in Tables 5.3 and 5.4. RF has been selected as the wrapper algorithm for the WGP-1 methods as it has shown the best performance most of the time on the MTGP (WGP-4 and WGP-5) approaches. However, with only one CF in WGP-1, we observe from the results in Table 5.3 that it is difficult for RF to generate a good classification model.

Among the two datasets, different types of features are prominent to produce good classification performance. The $\mathrm{LBP}_{\mathrm{RGB}}$ and wavelet features have produced better results compared to other feature sets among WGP-1 and EGP-1 in case of PH^2 as shown in Table 5.3, blocks 2 and 3. On the other hand, the $\mathrm{Lesion}_{\mathrm{Color}}$ and wavelet features are most prominent in producing good results among WGP-1 and EGP-1 on Dermofit. In multi-class classification (Table 5.4, block 2), such a trend is not seen for any of the datasets. This can be explained by the complexity of switching from binary classification to multi-class classification which is less on PH^2 (2 classes to 3 classes) and quite high on Dermofit (2 classes to 10 classes).

The experimental results have shown that images taken from different instruments require different feature extraction methods to get useful information to distinguish between images of different classes. Such a pattern has been seen when multiple features are generated in the WGP-4 and WGP-5 methods. On both datasets, wavelet features produced the best performance in the majority of cases. Moreover, $\mathrm{LBP}_{\mathrm{RGB}}$ and $\mathrm{Lesion}_{\mathrm{Color}}$ features still stay significant on the PH^2 and Dermofit datasets, respectively. From the EGP-1 and WGP-1 results on both datasets, it is inferred that choosing an appropriate feature extraction method is crucial in achieving performance gains.

5.5.5 Comparisons between WGP-4 and WGP-5

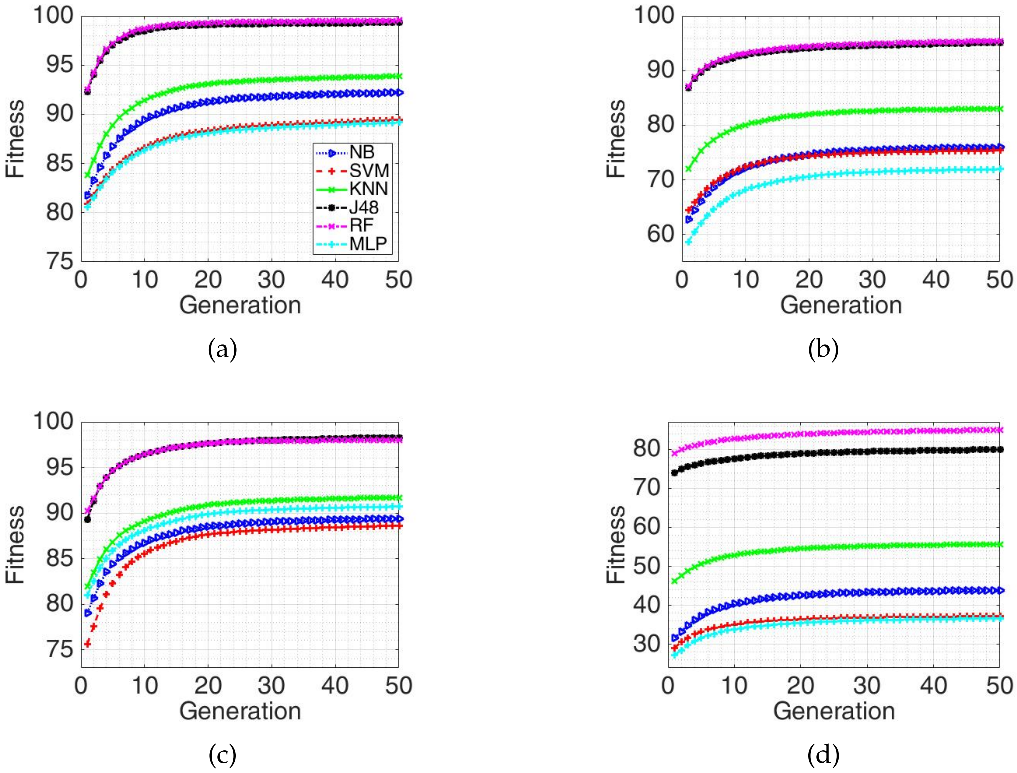
It can be observed that the generation of the wavelet-based CF in the WGP-5 approach allows the classification algorithm to train a classifier more effectively compared to the previous methods. The WGP-5 method has shown around 10% and 4.5% improvement on WGP-4 in the binary classification task on PH^2 and Dermofit, respectively. Although, these wavelet features greatly support the non-GP and WGP-1 methods to produce good results compared to the WGP-5 method on binary classification, these methods remain incapable of producing quality results in the multi-class classification tasks. For instance, on Dermofit, WGP-1 with wavelet features give 99.33% average accuracy (Table 5.3) for binary classification. On the other hand, the same method has achieved only 43.48% average accuracy (Table 5.4) for multi-class classification. This concludes that the multi-scale properties of frequency-based features great enhance the potential of WGP-5 method to achieve good results for both the easy (binary classification) and the challenging (multi-class classification) problems leaving the earlier GP and the other non-GP methods far behind.

5.5.6 Comparisons between Wrapper MTGP and Embedded MTGP Approaches

Embedded MTGP approaches are limited to binary classification only due to their design, whereas wrapper MTGP approaches are also utilized for multi-class classification in addition to binary classification. The wrapper approaches have proved to be more powerful where GP builds multiple features suitable for a specific classifier such as NB or RF. The highest performances achieved by EGP-5 are 81.28% and 82.59% on the PH^2 and Dermofit, respectively. However, WGP-5 outruns EGP-5 by many folds achieving 99.93% and 100% on the PH^2 and Dermofit, respectively.

5.6 Further Analysis

This section provides detailed overall analysis of the WGP-4 and WGP-5 methods by examining their convergence plots, and showing how efficient these methods are by observing their computation time. It further shows evolved GP individuals in WGP-4 and WGP-5, and identifies prominent features by plotting frequency of occurrence in CFs of various types of features.



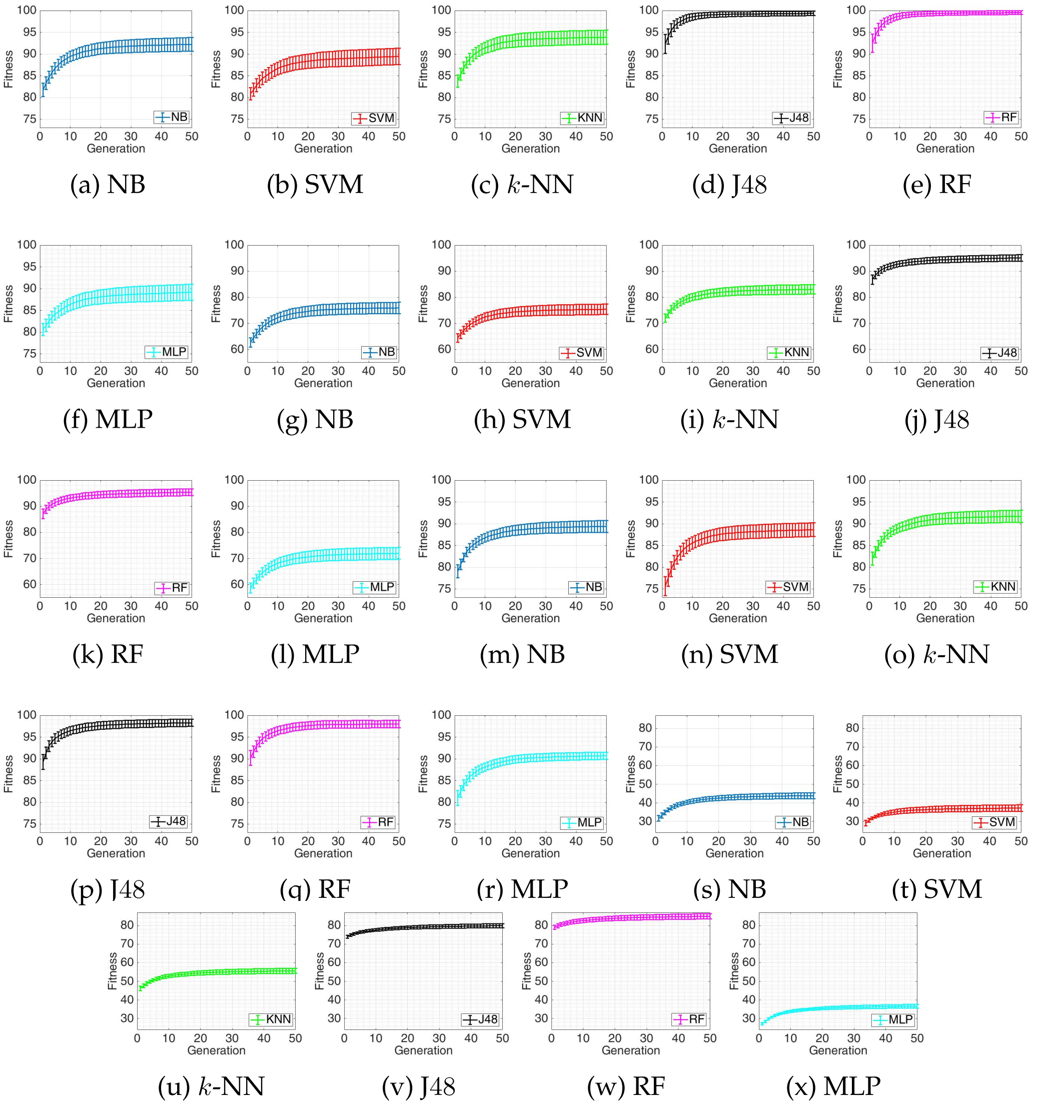
(a) (b)

(c) (d)

Figure 5.6: The average fitness value per generation in WGP-4 on PH^2 dataset for (a) binary classification, and (b) multi-class classification, and on Dermofit dataset for (c) binary classification, and (d) multi-class classification.

5.6.1 Overall Analysis

The average fitness value per generation of the 30 independent runs (each having 10 independent runs for the 10 folds in $10-$*fold cross validation*) using different seed values on the training data of the two datasets is depicted in Figure 5.6 for WGP-4, and in Figure 5.8 for WGP-5. Figures 5.6 and 5.8(a) and (b) show these plots for binary and multi-class classification on the PH^2 dataset, respectively, and Figures 5.6 and 5.8(c) and (d) show these plots on the Dermofit dataset.



(a) NB (b) SVM (c) k-NN (d) J48 (e) RF

(f) MLP (g) NB (h) SVM (i) k-NN (j) J48

(k) RF (l) MLP (m) NB (n) SVM (o) k-NN

(p) J48 (q) RF (r) MLP (s) NB (t) SVM

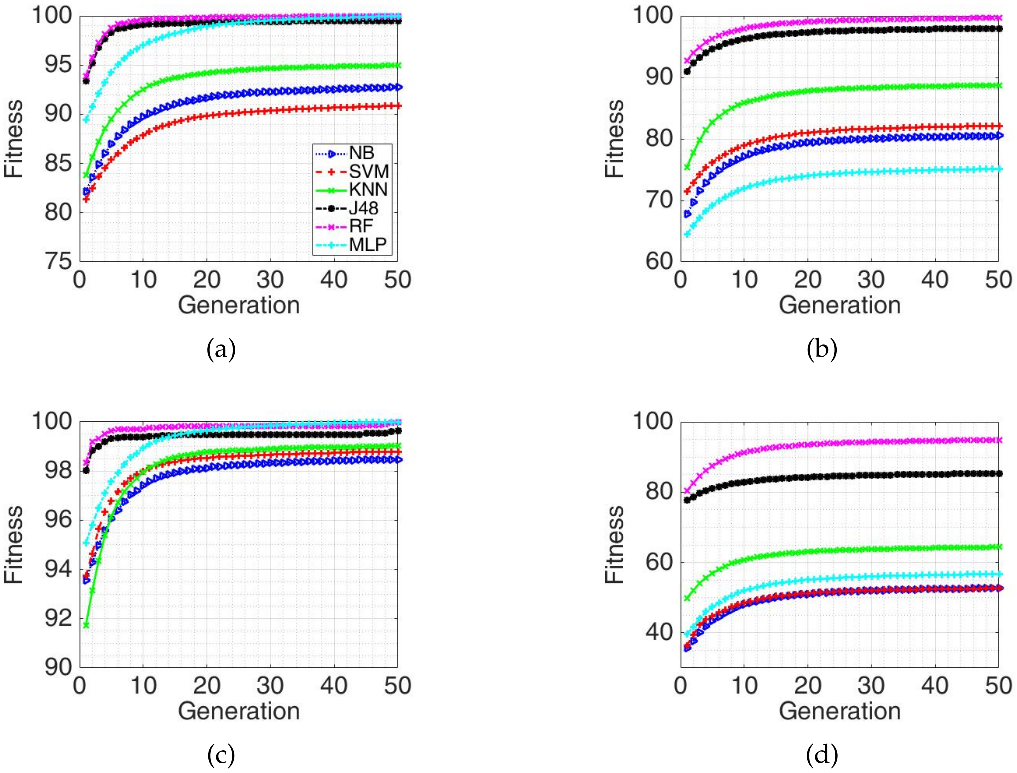
(u) k-NN (v) J48 (w) RF (x) MLP

Figure 5.7: The average fitness value per generation in *WGP-4* on PH^2 dataset for binary classification: (a) to (f), multi-class classification: (g) to (l), and on Dermofit dataset for binary classification: (m) to (r), and multi-class classification: (s) to (x).

For binary classification tasks (Figure 5.6 and 5.8(a) and (c)), these graphs show that on average the programs make larger jumps in the first few generations than in the later generations. In WGP-4, this trend has been observed in all the six wrapped classifiers (NB, SVM, k-NN, J48, RF, and MLP). In case of the PH^2 dataset using RF as the classifier, as shown in Figure 5.6(a), the fitness value has increased from 92.41% to 99.08% in the first 20 generations compared to the increase in fitness from 99.08% to 99.34% over the remainder 30 generations. Similarly in WGP-5, the highest jump is made by k-NN in case of Dermofit, as shown in Figure 5.8(c), from 91.67% to 98.12% only in the first 10 generations compared to the increase from 98.12% to 99.03% over the remainder 40 generations. In case of PH^2 using RF classifier, as shown in Figure 5.8(a), the fitness value has increased from 93.9% to 99.6% in the first 10 generations compared to the increase in fitness from 99.90% to 100% over the remainder 40 generations.

The plots for multi-class classification tasks, as given in Figure 5.6 and 5.8(b) and (d), show different behavior compared to the binary classification tasks as shown in Figure 5.6(a) and (b). There is an abrupt increase in the first few generations (around 10) which becomes slightly insignificant in later generations (last 40 generations). In WGP-4, this trend is more visible in case of the PH^2 dataset as compared to the Dermofit dataset where a significant increase in fitness is only seen among the first 5 generations. In comparing RF and J48, we have seen that both these classifiers have shown similar training curves except in the case of multi-class classification on the Dermofit dataset where RF outperforms J48 by a relevant margin as can be clearly seen in Figure 5.6(d). Similarly in WGP-5, from Figure 5.8(d), it has been seen that RF and J48 start around 80% for the complex task of 10-class classification, however, with increase in the number of generations, RF gets much bigger jumps as compared to J48 providing far better accuracy than J48.

Figures 5.7 and 5.9 further expand Figures 5.6 and 5.8 to highlight the standard deviation along the evolutionary process. For illustration, Figure 5.7(a) to (f) corresponds to Figure 5.6(a), where these six individual plots show standard deviation as well of the WGP-4 method run with NB, SVM, k-NN, J48, RF, and MLP (shown in same colors across both Figures



(a)

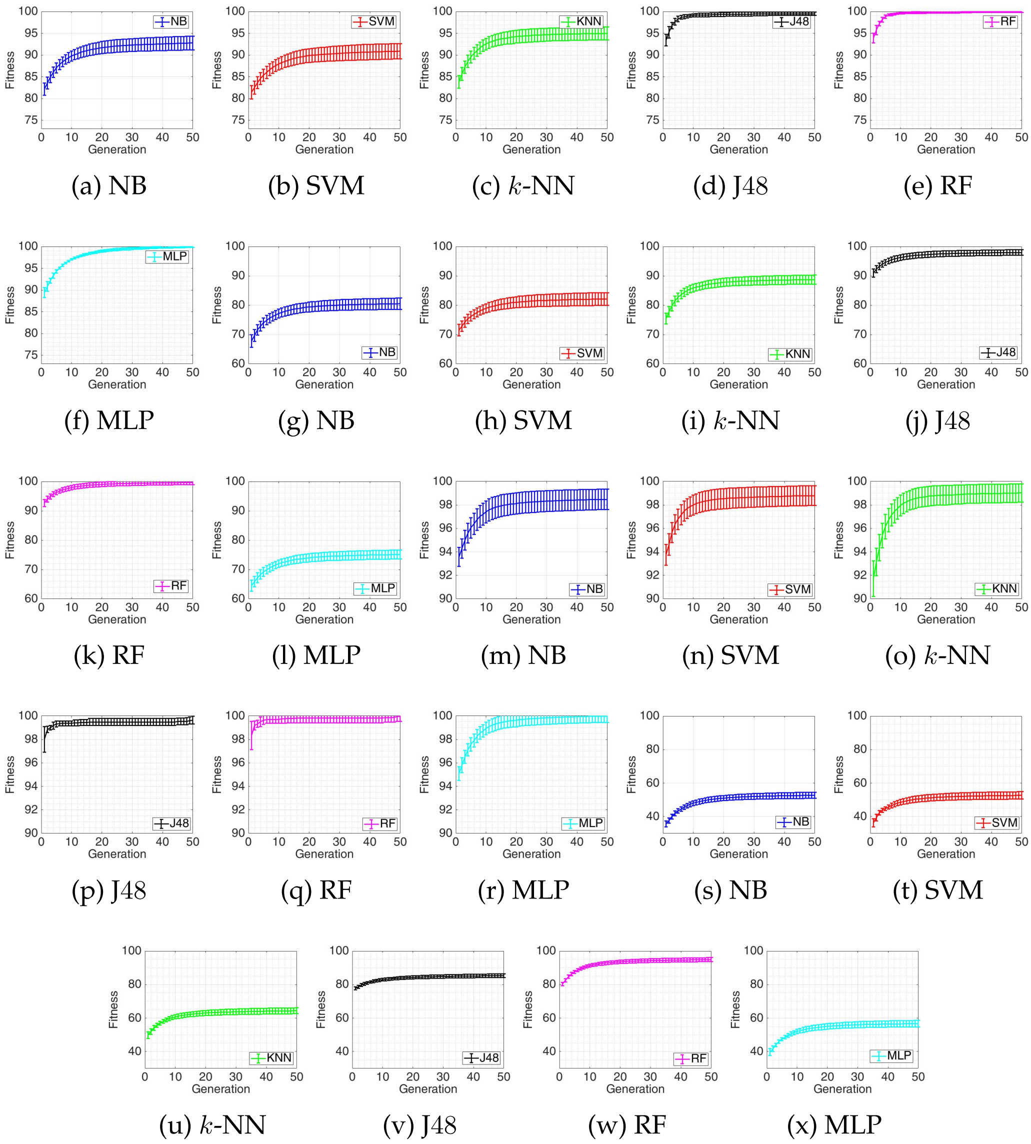
(b)

(c)

(d)

Figure 5.8: The average fitness value per generation in WGP-5 on PH^2 dataset for (a) binary classification, and (b) multi-class classification, and on Dermofit dataset for (c) binary classification, and (d) multi-class classification.

Figure 5.6 and Figure 5.7). The standard deviation bars of these 30 independent runs also show a similar behavior in case of J48 and RF where the earlier generations have more variations than the later ones. However, this trend is opposite in case of NB, SVM, $k-$NN and MLP where the later generations have more variations than the earlier generations. Overall the variation in standard deviation is high on the PH^2 dataset as compared to the Dermofit dataset.



(a) NB (b) SVM (c) k-NN (d) J48 (e) RF

(f) MLP (g) NB (h) SVM (i) k-NN (j) J48

(k) RF (l) MLP (m) NB (n) SVM (o) k-NN

(p) J48 (q) RF (r) MLP (s) NB (t) SVM

(u) k-NN (v) J48 (w) RF (x) MLP

Figure 5.9: The average fitness value per generation in *WGP-5* on PH^2 dataset for binary classification: (a) to (f), multi-class classification: (g) to (l), and on Dermofit dataset for binary classification: (m) to (r), and multi-class classification: (s) to (x).

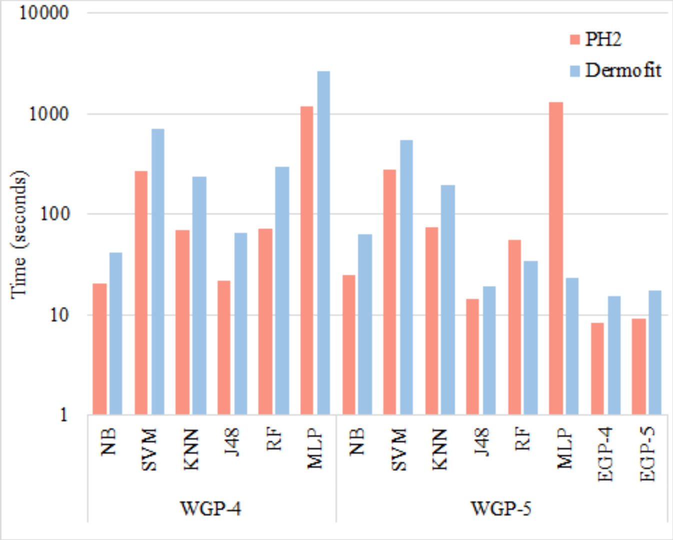
5.6.2 Computation Time

The average training time required for the WGP-4, WGP-5, EGP-4, and EGP-5 methods and to evaluate their performance on the test data in the binary and multi-class classification is plotted in Figures 5.10 and 5.11, respectively. While observing these plots, we can see that the time taken to evolve a model is typically influenced by the number of trees in a GP individual, the number of classes and the number of images in a dataset, the number of input features used to generate a tree(s) in a GP individual, and whether an embedded or a wrapper algorithm is utilized.

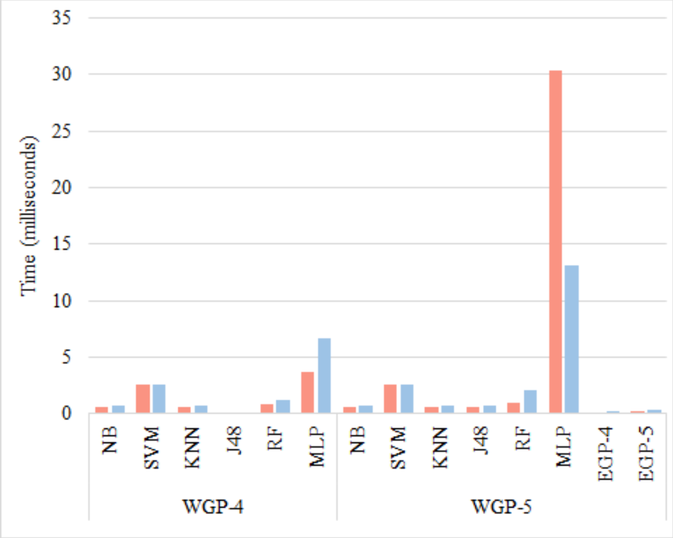
The average training time required for the WGP-4, WGP-5, EGP-4, and EGP-5 methods and to evaluate their performance on the test data in the binary and multi-class classification is plotted in Figures 5.10 and 5.11, respectively. While observing these plots, we can see that the time taken to evolve a model is typically influenced by the number of trees in a GP individual, the number of classes and the number of images in a dataset, the number of input features used to generate a tree(s) in a GP individual, and whether an embedded or a wrapper algorithm is utilized.

This is due to the fact that fitness evaluation of an individual having a single tree requires less time on average as compared to fitness evaluation of an individual with multiple trees. Similarly, the genetic operators are applied on five trees during the evolutionary process which also executes longer than the single-tree methods. Moreover, the wrapper approaches are computationally more expensive than embedded approaches. For instance, in WGP-5, the original training and test datasets are converted to new training and test datasets (by applying the evolved CFs). These new datasets are then utilized to train (and test) a classification algorithm, e.g., J48. Therefore, generating a classification algorithm using the new CFs leads to an increased computation time.

For binary classification in Figures 5.10(a) and (b), J48 requires the least computation time to generate a model among the six wrapper algorithms in WGP-5. Overall, RF and MLP are the fastest and highest performing

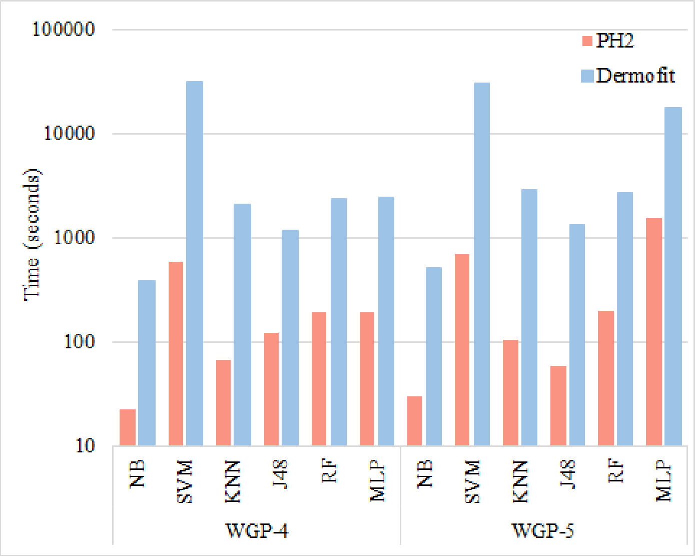


(a) Training time

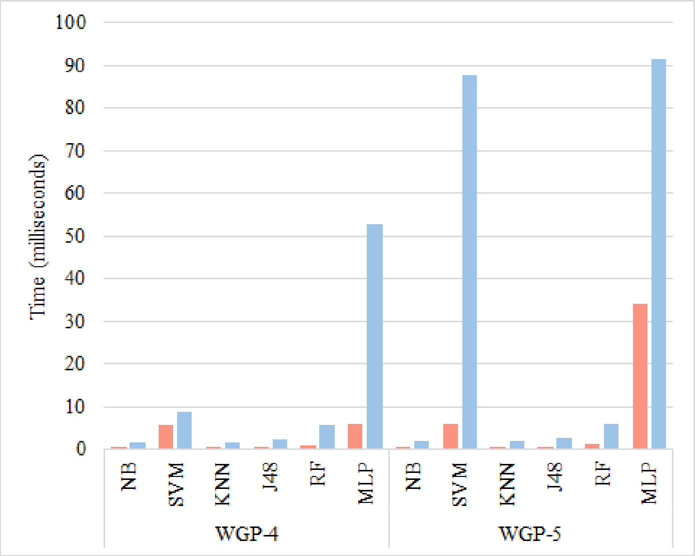


(b) Test time

Figure 5.10: The average computation time for *binary classification* using the WGP-4, WGP-5, EGP-4, and EGP-5 approaches on the two skin cancer image datasets.



(a) Training time



(b) Test time

Figure 5.11: The average computation time for *multi-class classification* using the WGP-4 and WGP-5 approaches on the two skin cancer image datasets.

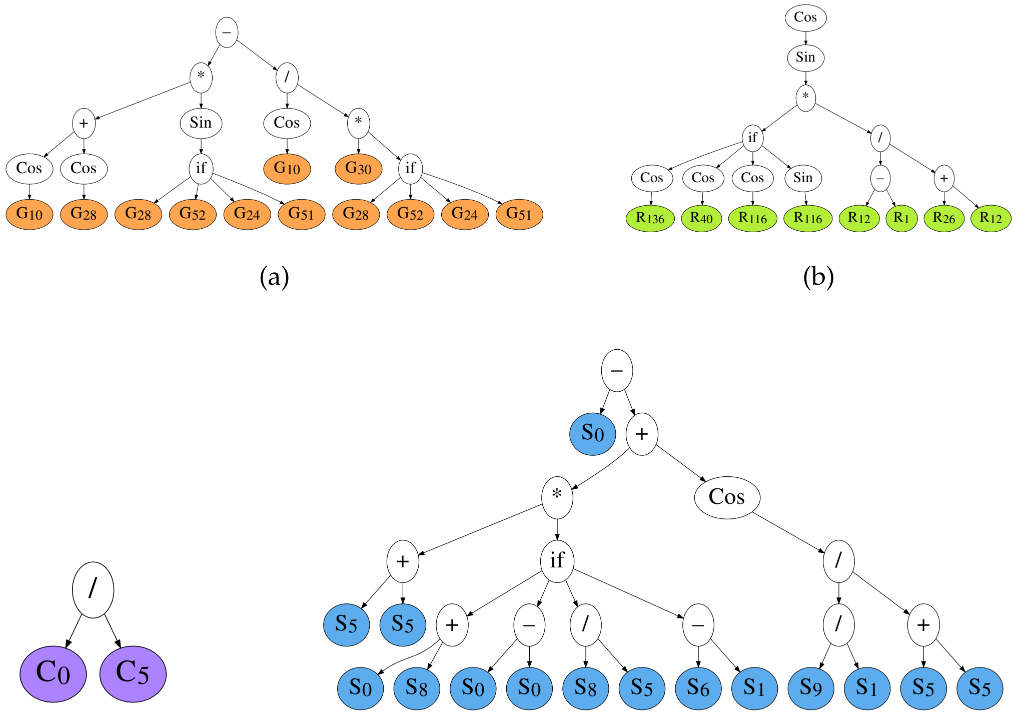
wrapper algorithms in the WGP-5 approach by using an average evolutionary time of only 55.76 and 23.31 seconds on PH^2 and Dermofit, respectively. Moreover, with these trained methods at hand, they spend only 0.96 and 13.1 milliseconds on average to test an unseen skin image.

For multi-class classification, Figures 5.11(a) shows that a dataset with large number of images (Dermofit) generally requires more time to evolve a model compared to a dataset with small number of images (PH^2). Such a behavior has been clearly seen while comparing Figures 5.10 and 5.11. On the contrary, a test image can be checked in fractions of a second having these trained models as given by the test time shown in Figure 5.11(b). The highest performing RF takes only 1.11 and 6.09 milliseconds on PH^2 and Dermofit, respectively, to test an unseen image.

In comparison to the embedded methods, the wrapper methods require more time to train a classifier as can be seen in Figure 5.10. Similarly, the large dataset (Dermofit) needs more time in most cases than the smaller dataset (PH^2), regardless of the approach (wrapper or embedded) used.

5.6.3 An Evolved GP Individual of WGP-4

GP evolves models that can be interpretable. To see why our proposed WGP-4 method can achieve good classification results, we have analyzed a good GP individual with four trees in Figure 5.12 from the PH^2 binary classification experiments. The four constructed features have given 87.5% accuracy on the test data. GP found this perfect solution giving 100% accuracy on the training data, just after 24 generations. In Figure 5.12, colored nodes show terminals, whereas white nodes show functions. As discussed previously in Chapter 4 (Section 4.5 on page 4.5, $\mathrm{LBP}_{\mathrm{RGB}}$ features have the most potential compared to other feature types to classify images in PH^2 dataset. Since LBP captures local pixel-based properties of an image, these features with gray and color information can incorporate good discrimina-



(a) (b)

(c) (d)

Figure 5.12: A good evolved GP individual for PH^2 dataset using a) $\mathrm{LBP}_{\mathrm{Gray}}$, b) $\mathrm{LBP}_{\mathrm{RGB}}$, c) $\mathrm{Lesion}_{\mathrm{Color}}$, and d) $\mathrm{Lesion}_{\mathrm{Shape}}$ features producing 87.5% accuracy on the test data in the binary classification task.

tive information regarding the presence or absence of melanoma in a skin image. Furthermore, $\mathrm{Lesion}_{\mathrm{Shape}}$ and $\mathrm{Lesion}_{\mathrm{Color}}$ features, which capture the global properties such as geometrical border shape and color variation between the lesion region and the skin region, respectively cannot provide as good performance as LBP feature.

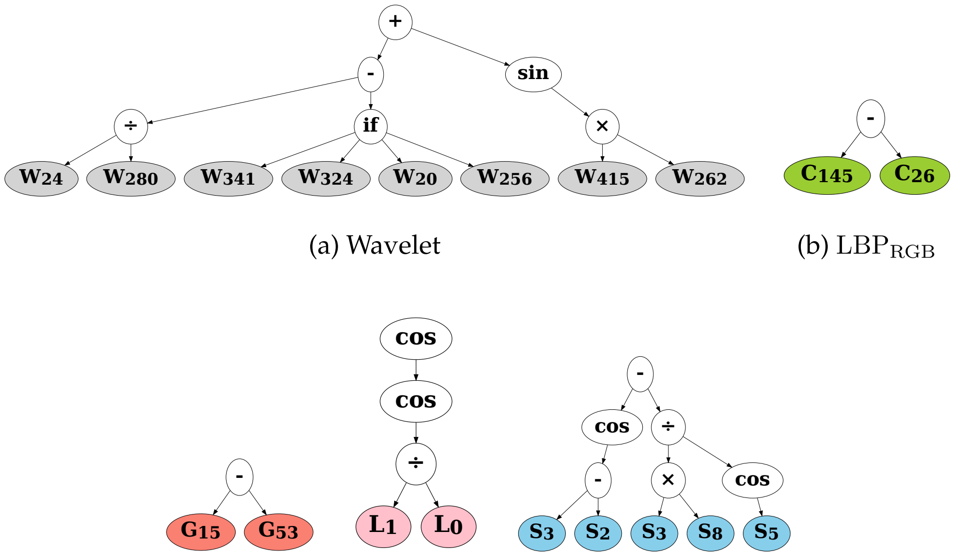
In the $\mathrm{LBP}_{\mathrm{Gray}}$ tree from Figure 5.12(a), the features G_{10} and G_{28} are selected two times and three times, respectively. In addition, the expression $\mathit{if}(G_{28}, G_{52}, G_{24}, G_{51})$ is selected twice. This illustrates that these features possess good distinguishing ability between classes. Among the 177 $\mathrm{LBP}_{\mathrm{RGB}}$ features, only six prominent features (R_1, R_{12}, R_{26}, R_{40}, R_{116}, and R_{136}) are used to construct a tree (Figure 5.12(b)). Similarly, only six features (G_{10}, G_{24}, G_{28}, G_{30}, G_{51}, and G_{52}) among the 59 LBP_{Gray} features have been selected to build the tree in Figure 5.12(a). The $\text{Lesion}_{\text{Color}}$ tree in Figure 5.12(c) has been built from only two features among the 12 $\text{Lesion}_{\text{Color}}$ features, and the $\text{Lesion}_{\text{Shape}}$ tree in Figure 5.12(d) has been constructed from six features among the 11 $\text{Lesion}_{\text{Shape}}$ features. Hence, the feature selection and construction ability of GP has provided discriminative constructed features as input to the decision tree classification algorithm, which helps RF achieve promising results.

In the $\text{Lesion}_{\text{Color}}$ tree from Figure 5.12(c), C_0 and C_5 representing the mean of the red color channel (μR) and the variance of the blue color channel (σB), are combined to produce a significant constructed feature. In the $\text{Lesion}_{\text{Shape}}$ tree from Figure 5.12(d), S_0, S_1, S_5, S_6, S_8, and S_9 are selected which correspond to the geometrical shape features: area, perimeter, irregularity indices A, C and D, and the major asymmetry index. These shape features can assist the dermatologist in real-time situations by providing significant knowledge about the lesion geometrical properties and hence, making a diagnosis much easier.

5.6.4 An Evolved GP Program in WGP-5

GP has the ability to evolve models that can be interpretable. To analyze why the WGP-5 method can achieve good performance, we show a good evolved GP individual in Figure 5.13. This individual is taken from the Dermofit experiments for the binary classification task using RF. It has five trees evolved using the five types of features. These CFs when used by J48 have given 98.48% accuracy on the test data. This individual is the perfect solution for the training data where GP has evolved it just after 21 generations.

With the ability of feature selection and feature construction, GP plays a vital role in dimensionality reduction. From the evolved GP individual



(a) Wavelet (b) LBP_{RGB}

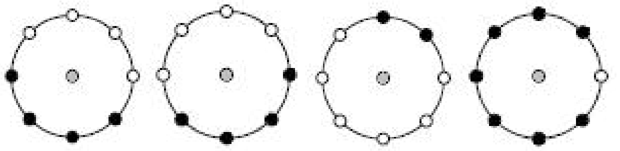
(c) LBP_{Gray} (d) $\text{Lesion}_{\text{Color}}$ (e) $\text{Lesion}_{\text{Shape}}$

Figure 5.13: A good WGP-5 individual on the Dermofit dataset with five trees (CFs) evolved using the five sets of features providing 98.48% accuracy on the test data in the binary classification.

shown in Figure 5.13, GP has selected only 8 features among a total of 416 wavelet features, only 2 features among 177 LBP_{RGB} features, only 2 features among 59 LBP_{Gray} features, only 2 features among 12 $\text{Lesion}_{\text{Color}}$ features, and 4 features among 11 $\text{Lesion}_{\text{Shape}}$ features. The wavelet texture-based features appearing in a tree of the GP individual shown in Figure 5.13(a) are listed in Table 5.5. The following conclusions are derived from this table: 1) only one out of eight features belong to the nodes from the third level, which indicates our use of three-level wavelet decomposition as further decomposition may not obtain informative features for the purpose of classification, 2) texture features extracted from the blue and the luminance channels have more significant information than red and green channels, 3) the selected features are derived from both the low and middle frequency channels as shown by the node column in Table 5.5, 4) par-

Table 5.5: Wavelet features appearing in the GP individual shown in Figure 5.13(a).

Feature	Measure	Channel	Level	node
W_{24}	Energy	Luminance	0	0
W_{280}	Energy	Blue	2	2.1
W_{341}	Norm	Luminance	1	1.1
W_{324}	Kurtosis	Luminance	1	1.2
W_{20}	Kurtosis	Blue	0	0
W_{256}	Energy	Blue	2	2.2
W_{415}	Average Energy	Luminance	3	3.1
W_{262}	Entropy	Blue	2	2.2



(a) C_{145} (b) C_{26} (c) G_{15} (d) G_{53}

Figure 5.14: LBP patterns of the features appearing in $\mathrm{LBP_{RGB}}$ and $\mathrm{LBP_{Gray}}$ trees in the GP individual shown in Figure 5.13 (b) and (c), respectively.

ticular measures such as energy and kurtosis are prominent selected features.

Figure 5.14 shows the LBP patterns of the features appeared in the $\mathrm{LBP_{RGB}}$ and $\mathrm{LBP_{Gray}}$ trees in the GP individual shown in Fig 5.13 (b) and (c), respectively. C_{145} and C_{26} patterns in Figure 5.14(a) and (b), respectively, show the presence of edges in the skin cancer images. Similarly, G_{15} and G_{53} show the presence of corners in these images. Edges and corners identify various visual patterns such as streaks, dots, blobs and pigment network inside a lesion area in the skin cancer image. Therefore, these GP trees have selected prominent LBP features corresponding to signifi-

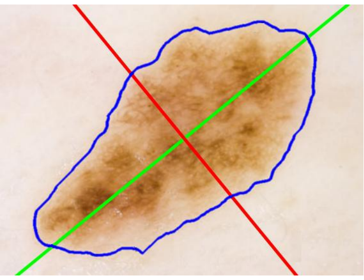
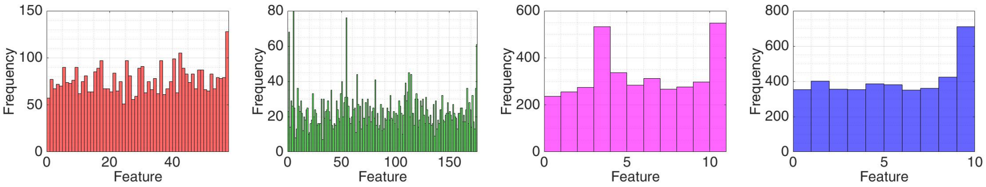


Figure 5.15: The minor symmetry axis (shown by red line) and the major symmetry axis (shown by green line) for a dermoscopy image.

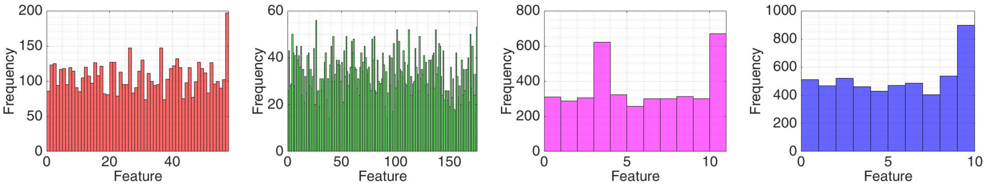
cant visual characteristics necessary to accurately identify a type of skin cancer.

The $\text{Lesion}_{\text{Color}}$ tree in Figure 5.13(d) is generated by selecting only two features L_0 and L_1 among the 12 $\text{Lesion}_{\text{Color}}$ features, and the $\text{Lesion}_{\text{Shape}}$ tree in Figure 5.13(e) is built by choosing four features among the 11 $\text{Lesion}_{\text{Shape}}$ features. In $\text{Lesion}_{\text{Color}}$ tree as shown in Figure 5.13(d), L_0 and L_1 (corresponding to μR and μG) showing the mean of the red channel and the mean of the blue channel, combine to generate an informative CF. In $\text{Lesion}_{\text{Shape}}$ tree as shown in Figure 5.13(d), $\text{S}_2, \text{S}_3, \text{S}_5$, and S_8, are selected which correspond to GD, SD, IrA and IrD. The greatest and the shortest diameter of a lesion is shown in Figure 5.15 which seems important in capturing the shape of the lesion. The value of irregularity index D (IrD) depends on GD and SD as shown by its mathematical expression in Table 4.1 on page 114. Similarly, rather selecting area and perimeter as individual features, GP has selected IrA which is the ratio between perimeter and area of a lesion. Hence, GP has incorporated important hand-crafted features effectively in evolving $\text{Lesion}_{\text{Shape}}$ tree. These border shape features include significant knowledge regarding the lesion geometrical characteristics which can greatly help the dermatologist in actual clinic settings making a diagnosis much easier. Hence, we can conclude here that the automatic feature selection and feature construction abilities of GP produce discriminative CFs as input to the classification algorithm contributing sig-



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

Figure 5.16: The average frequency of features in trees, each evolved with a single type of features on the PH^2 dataset in the *binary classification* task using RF as a classifier.



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

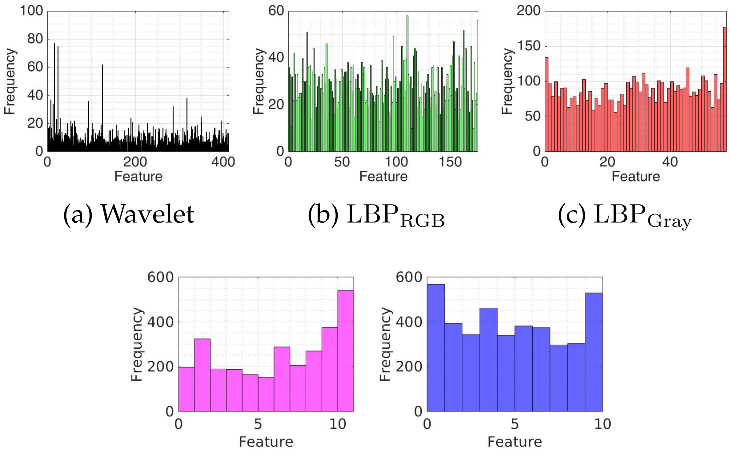
Figure 5.17: The average frequency of features in trees, each evolved with a single type of features on the PH^2 dataset in the *multi-class classification* task using RF as a classifier.

nificantly to achieve promising results.

5.6.5 Feature Appearance in CFs of WGP-4

GP automatically constructs new features by selecting more relevant and discriminative features among the whole set of original features. We have also explored and analyzed this intrinsic ability of GP to feature selection. Figures 5.16 and 5.17 show the bars for the average number of times each feature appears in the constructed features among the 30 GP runs (in all the 10 folds) in the PH^2 experiments for binary and multi-class classification using RF as a classifier, respectively. It is evident from these plots that there are some features which are selected more frequently as compared to other features, e.g., G_{58}, the last feature among $\mathrm{LBP}_{\mathrm{Gray}}$ features in Figure 5.16(a) appears almost twice as frequently as the other 58 $\mathrm{LBP}_{\mathrm{Gray}}$ features. Similarly, R_5, C_{11} and S_{10} have the highest frequency of occurrence among the $\mathrm{LBP}_{\mathrm{RGB}}$, $\mathrm{Lesion}_{\mathrm{Color}}$, and $\mathrm{Lesion}_{\mathrm{Shape}}$ features as shown in Figures 5.16(b), (c) and (d), respectively. We have seen a similar pattern while having a closer look at Figure 5.17 for the multi-class classification task in the PH^2 experiments, where these features have the highest frequency again except R_5. This shows that these features have significant discriminative ability between classes, not only for the binary classification task but also for the multi-class classification task. R_{26} which also represents the same structural properties as R_5, has the highest frequency among $\mathrm{LBP}_{\mathrm{RGB}}$ in the multi-class classification task.

For a deep analysis of these significant features (G_{58}, C_{11} and S_{10} for both tasks, R_5 for binary classification task alone, and R_{26} for multi-class classification task), digging further into the local and global properties of these features, we see that G_{58} are the non-uniform LBP features combined in one bin for gray-scale images. Though non-uniform features are not considered to have discriminative properties for texture analysis (that is why they are binned together in one bin), however, in our dataset, the number of times these non-uniform features appear in one class of images is quite different from their appearance in other classes, which makes them highly significant. For $\mathrm{LBP}_{\mathrm{RGB}}$ features, R_5 and R_{26} represent presence of edges in an image. Inside the skin lesion, the different structures such as dots, streaks and regression areas with varying colors are highlighted by these pixel-level edge properties. Among the different classes, these structures vary and hence, these edge detecting $\mathrm{LBP}_{\mathrm{RGB}}$ features become prominent in distinguishing between classes. Among $\mathrm{Lesion}_{\mathrm{Color}}$ features, C_{11} corresponds to $\frac{\mu_B}{\bar{\mu}_B}$, which shows the ratio between mean of blue color channel of the lesion region and its surrounding skin region. Among $\mathrm{Lesion}_{\mathrm{Shape}}$ features, S_{10} is the most significant as its frequency is almost double as compared to the other 11 $\mathrm{Lesion}_{\mathrm{Shape}}$ features (Figures 5.16(d)



(a) Wavelet (b) LBP_{RGB} (c) LBP_{Gray}

(d) $\text{Lesion}_{\text{Color}}$ (e) $\text{Lesion}_{\text{Shape}}$

Figure 5.18: The average frequency of features in trees (the CFs) of WGP-5, each generated with one set of features on the Dermofit dataset using J48 classifier in the multi-class classification.

and 5.17(d)). It corresponds to asymmetry index, which provides the necessary information about the shape, particularly being computed from the asymmetry axes and area of the lesion. As described earlier in Section 4.2.3 on page 113, our analysis also confirms that asymmetry plays an essential role in making a diagnosis for the binary and multi-class classification of skin cancer images.

5.6.6 Feature Appearance in CFs of WGP-5

GP selects more relevant features from the original set of features to automatically construct new informative features. In WGP-5, this built-in ability of GP to FS has also been investigated by giving an example from the Dermofit experiments for the multi-class classification. In Figure 5.18, the bars show the frequency of occurrence of each feature in the CFs among the 30 GP runs. From these plots, it is obvious that certain features are more frequently selected than the other features, W_{16}, and W_{24} features appear almost four times more than the other 414 wavelet features. Similarly, G_{58}, the last feature among the $\mathrm{LBP}_{\mathrm{Gray}}$ features in Fig 5.18(c) appear almost three times more than the other 58 $\mathrm{LBP}_{\mathrm{Gray}}$ features. Similarly, C_{110}, L_{11} and S_0 appeared the most among the $\mathrm{LBP}_{\mathrm{RGB}}$, $\mathrm{Lesion}_{\mathrm{Color}}$, and $\mathrm{Lesion}_{\mathrm{Shape}}$ features as given in Figure 5.18(b), (d) and (e), respectively.

We have further digged deeper into the texture, color, local and global properties of these important features; W_{16}, W_{24}, C_{110}, G_{58}, L_{11} and S_0. We have found that both W_{16}, and W_{24} are the energy measures of the blue and the luminance channels of the original image, respectively. W_{24} has also been selected by the evolved GP individual shown in Figure 5.13(a). Moreover, we see that G_{58} are the non-uniform LBP features integrated together in the last bin for gray-scale skin images. Although the non-uniform LBP features are not known to provide useful information for texture analysis (that's the reason, they are combined together in one bin), on the contrary, in our dataset, their frequency of occurrence in one class of images greatly differs from their frequency of occurrence in other classes, which renders them most significant. Among the $\mathrm{LBP}_{\mathrm{RGB}}$ features, C_{110} corresponds to a 3×3 LBP window from the green channel image which shows presence of edges in an image. Within the skin lesion, these pixel-level edge properties mainly highlight the various structures such as streaks, blobs, and corners. The presence of these structures differ among the different types of skin cancers making these edge detecting $\mathrm{LBP}_{\mathrm{RGB}}$ features highly significant in discriminating between different classes. Among the $\mathrm{Lesion}_{\mathrm{Color}}$ features, L_{11} corresponds to $\frac{\mu_B}{\bar{\mu}_B}$, which is the ratio between the mean of the blue channel of the lesion region and the skin region around the lesion. Among the $\mathrm{Lesion}_{\mathrm{Shape}}$ features, S_0 corresponds to the area of the lesion and remains most prominent with almost twice frequency of occurrence compared to the other $\mathrm{Lesion}_{\mathrm{Shape}}$ features as shown in Figure 5.18(e).

5.7 Chapter Summary

This chapter has developed and analyzed two multi-tree GP based wrapper methods developed for multiple feature construction in skin cancer image classification. These wrapper methods utilize the ability of GP to automatically construct multiple features in a single evolved GP individual. Chapter 4 described an embedded approach which solves the binary classification problem, whereas the wrapper approaches developed in this chapter are employed to solve both the binary and the multi-class classification tasks. For effective classifier generation and feature construction using multi-tree GP, these methods incorporate various types of local, global, color, texture and frequency-based features extracted from skin images.

Experiments revealed the effectiveness of these GP methods for both the tasks of binary and multi-class classification of skin cancer images. The higher performing method among these two methods is WGP-5, which has outperformed the state-of-the-art method achieving 86.77% balanced accuracy for the difficult task of 10-class skin cancer image classification. It has also outperformed all the six commonly used classification algorithms (NB, k-NN, SVM, J48, RF, and MLP), the MTGP embedded approaches and the single-tree GP methods, demonstrating its efficacy to discriminate effectively between classes. Similar to the previous chapters, it has been observed that the local pixel-based features have strong ability for classifying dermoscopy images, whereas the global color variation and the domain-specific shape features are prominent for distinguishing different classes of images obtained from standard camera. However, the wavelet-based texture features introduced in this chapter with multi-scale image properties have shown the best potential for both dermoscopy and standard camera images.

Though these wrapper approaches have provided far better results than embedded approaches for skin image classification tasks, they evolve potentially good CFs which are fit to only a single classification algorithm.

These method generate CFs suitable to the classification algorithm used as a wrapper approach. For example, the GP individual with five CFs shown in Figure 5.13 is evolved with RF as a classifier, so the CFs are generated to fit RF classification method. In other words, performance can be improved by generating generic CFs which when provided to multiple classifiers can help achieve better results. This will be investigated in the next chapter.

Chapter 6

GP based Feature Construction for Ensemble Learning

6.1 Introduction

Ensembles of classifiers have shown to be more effective than a single classification algorithm in skin image classification problems [86, 115, 208]. Generally, the ensembles are created using the whole set of extracted features. However, some extracted features can be redundant and may not provide useful information in building good ensemble classifiers. To deal with this, existing feature construction methods that usually generate new features for only a single classifier have been developed but they remain unable to provide good classification performance [141]. In the past, either a complete set of originally extracted features or selected features are provided to ensembles [82, 199], however, using multiple constructed features for learning an ensemble of classifiers has not been investigated. Since constructed features tend to have more distinguishing ability than original features, it is expected that constructed features will help improve the classification performance in an ensemble.

This chapter develops a new classification method that combines the effectiveness of feature construction and ensemble learning using GP to address the above limitations. GP has been used successfully for feature selection and feature construction. However, generating new features specific to a single classifier might have limited performance. An ensemble of classifiers combining the predictions of multiple classifiers can improve the performance, where the assumption here is that those classifiers are different from each other, i.e., they work together and can cover more data points. Hence, an ensemble of multiple classifiers can produce more accurate results [82]. Moreover, since GP has demonstrated good potential for feature construction as presented in Chapters 3 and 5, utilizing GP for feature construction for an ensemble of classifier can help achieve good classification performance.

6.1.1 Chapter Objectives

In order to deal with the limitations of existing approaches and utilize the effectiveness of ensemble learning, we combine the benefits of feature construction and ensemble classification in a GP framework to construct informative features for the tasks of both binary and multi-class skin cancer image classification. This chapter aims at investigating the following objectives:

- Design a new feature construction method using GP to generate new features for an ensemble of classifiers.

- Assess the performance of the proposed classification method in comparison to bagging, boosting, random forests, other commonly used machine learning classification algorithms, and the existing deep learning and GP methods on two real-world skin cancer image datasets.

- Visualize the multiple constructed features evolved by the proposed ensemble method and identify prominent image features.

6.1.2 Chapter Organization

The rest of the chapter is organized as follows. Section 6.2 presents the proposed ensemble method, its program representation, the fitness function, the terminal sets, the function sets, crossover and mutation operators used. Section 6.3 describes the experiments performed, GP parameters and benchmark methods for comparison. Section 6.4 presents the results of these experiments and discusses how well they address the chapter goals. Section 6.5 provides detailed analysis in terms of evolved constructed features, and convergence graphs. Section 6.6 concludes the chapter with the achievements of the method, and its possible limitations.

6.2 The Proposed Method

This section presents our proposed algorithm, i.e., multiple feature construction with ensemble classification (MFCEC) using GP for skin cancer image classification.

6.2.1 Program Representation and Fitness Function

MFCEC utilizes the multi-tree GP approach similar to the WGP-5 method in the previous chapter where GP constructs five trees in one individual during the evolutionary process. Each constructed feature is built from one and only one type of features as described in Chapter 5. These constructed features are utilized to transform the original training and test sets to new training and test sets. The transformed training set with new five constructed features is provided as input to the ensemble classification algorithm, which is formed by SVM, J48, and RF. These three classification algorithms have been selected since they have shown promising results in the previous multi-tree GP method in Chapter 5. These classifiers are trained on the training data during the evolutionary process. MFCEC uses the accuracy produced on the training data by the ensemble classifier as its fitness function, where each image is classified based on the majority voting. The balanced accuracy is used in order to avoid bias towards the majority class, since the datasets are highly imbalanced.

6.2.2 Terminal Set and Function Set

The terminal set consists of five feature sets, similar to WGP-5 listed in Table 5.2 on page 142. The function set of MFCEC is the same as previous multi-tree GP methods and the details can be found in Section 4.3.2 on page 116.

6.2.3 Crossover and Mutation

MFCEC uses same-index crossover/mutation. A graphical representation can be found in Figure 5.5 on page 143. Algorithm 1 and Algorithm 2 on page 118 describes the step-by-step procedure of same-index crossover and mutation in detail, respectively.

6.2.4 The Overall Algorithm

The overall structure of the proposed MFCEC method is presented in Figure 6.1. First the images are transformed to feature vectors using the feature extraction methods described in Chapters 4 and 5. The *LBP image descriptor*, as described in Section 2.1.2 on page 30, is used to extract texture features from gray skin images. LBP is also used to extract both texture and color information from the red, green, and blue color channels of the skin images as presented in Section 3.2.2 on page 87. The graphical illustration of extracting color LBP features is shown in Figure 3.3 on page 88. The color variation inside the lesion area and the surrounding skin region is calculated by extracting *lesion color variation* features as described in Section 4.2.2 on page 113. The domain specific information such as the border shape, asymmetry and size of the lesion are included by ex-

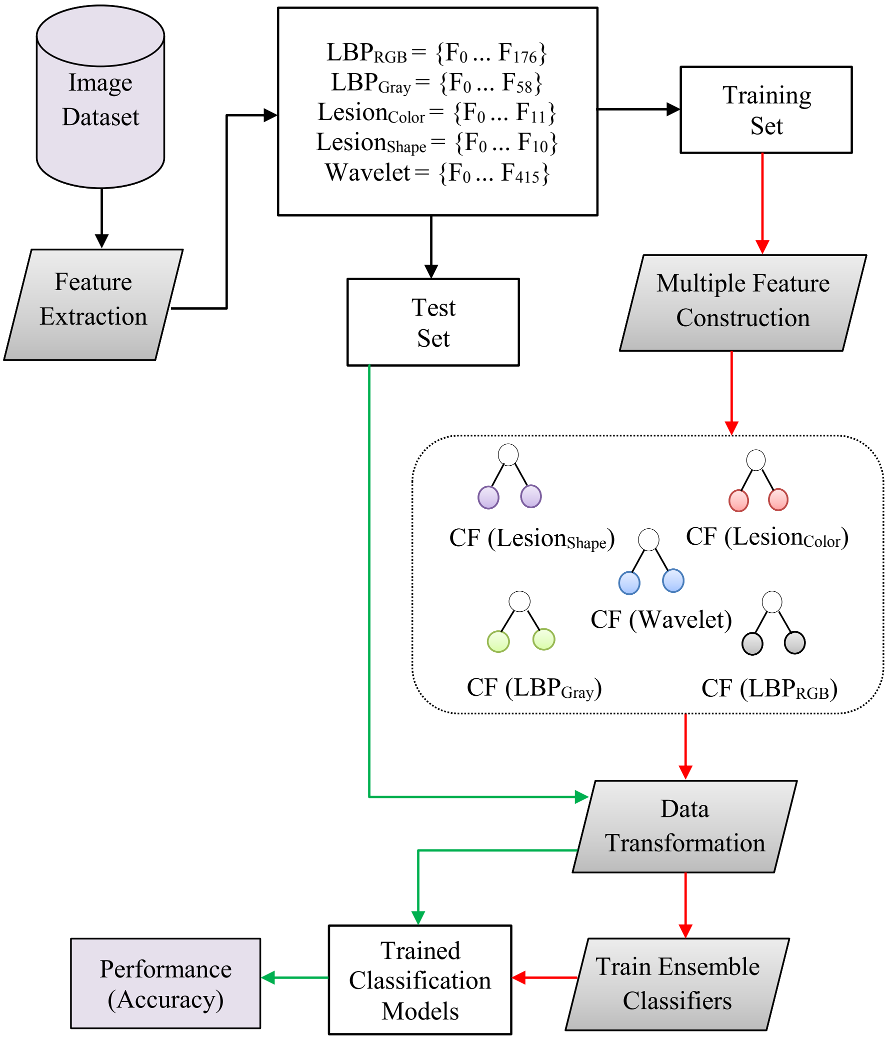


Figure 6.1: The workflow of the proposed algorithm.

tracting *geometry-based features* as described in Section 4.2.3 on page 4.2.3. To include both global and local as well as color and texture properties of skin images, *wavelet-based features* are extracted using three-level pyramid structured wavelet decomposition as described in Section 5.2.1 on page 137. These different methods are used to extract five sets of features as outlined in Table 5.2 on page 142.

MFCEC uses the same five sets of features as the WGP-5 and EGP-5 methods described in Chapter 5. Note that the feature extraction method is performed before the training and test data split because features are extracted image by image, so test images are not used for extracting features from training images, i.e., no bias produced. For each image, we get five feature vectors, namely LBP_{Gray}, LBP_{RGB}, $\text{Lesion}_{\text{Color}}$, $\text{Lesion}_{\text{Shape}}$, and Wavelet features. The dataset is then divided into training and test sets. The training process is shown by red line and test process is shown by green line in Figure 6.1. GP utilizes the training set to construct multiple features in one GP individual. Each tree in a GP individual is considered as one constructed feature. The constructed features in WGP-5 have shown to have more discriminating ability between classes as compared to the original sets of features. Similarly, we expect the same behavior of evolved constructed features, here, in MFCEC.

The three classification algorithms: SVM, J48, and RF are used in the ensemble. Since they have shown effective as a single classifier in WGP-5, we expect that combining them together in an ensemble will increase the classification performance. Multiple classification algorithms (SVM, J48, and RF) are incorporated as an ensemble to use the constructed features as input, hence, the constructed features evolved are generic to all these three classification algorithms. The constructed features are not tailored to one specific classifier, rather generated regardless of which classifier is used to classify them. After finishing the evolutionary process, we get the three (SVM, J48, and RF) trained classification models on the training data. The original test set is transformed by utilizing the same constructed features to a new test set. This new test set is used to evaluate the test performances. The trained classification model providing highest accuracy among the SVM, J48, and RF models on the training data is selected. The test set is provided to this selected classification model to get the accuracy on the unseen data.

6.3 Experiment Design

The aim and design of the experiments are discussed in this section. The discussion includes the benchmark methods used for comparisons, the parameter settings, and the experiment settings.

6.3.1 Benchmark Methods

In this chapter, we compare the performance of our proposed method with ten commonly used machine learning algorithms: Naïve Bayes (NB), k-Nearest Neighbor (k-NN), Support Vector Machines (SVMs), Decision Trees (J48), and Multi-layer Perceptron (MLP). We have also compared our proposed MFCEC method with the common used ensemble methods: Random Forest (RF), Bagging (Bgg), AdaBoost (AB), LogitBoost (LB), and Random Committee (RC). The settings of these methods are adopted from previous chapters, where they have been empirically searched via experiments. All other settings are set to default as in the Waikato Environment for Knowledge Analysis (WEKA) package [85]. The ten classification algorithms are trained one time on the five sets of features, appended to make a single feature vector. The trained classifiers are then tested to obtain their test performances.

For GP implementation, the Evolutionary Computation in Java (ECJ) package is used [114]. We also compare MFCEC with the two existing GP approaches for skin cancer image classification:

- Embedded-GP (EGP-5), described in Chapter 5, uses five types of features (LBP_{Gray}, LBP_{RGB}, $\text{Lesion}_{\text{Color}}$, and $\text{Lesion}_{\text{Shape}}$, and wavelet) to evolve five trees in its GP individual. Since this is an embedded approach where GP also performs classification, each tree acts as a binary classifier. The best tree with highest accuracy on the training data is used to test the performance on the test data.

- Wrapper-GP (WGP-5), developed in Chapter 5, uses five types of features listed in Table 5.2 on page 142 to evolve five trees in a single GP individual. These trees act as constructed features to be classified by a machine learning algorithm such as decision tree. The trained model is applied on the test set to check the performance of this method.

In addition, we compare MFCEC with the state-of-the-art CNN methods recently developed for the PH^2 and Dermofit datasets:

- Patino et al. [152] developed a lesion segmentation and classification method using morphological operations to estimate asymmetry, border and color features of the lesions in the PH^2 dataset. The method incorporated SVM, logistic regression and a fully connected neural network where the neural network has shown the best performance achieving 86.5% on average for multi-class classification.

- Kawahara et al. [96] trained a logistic regression classifier with deep features extracted from a convolutional neural network, pre-trained on natural images, to classify ten classes of skin lesions in the Dermofit dataset. They reported a standard overall accuracy of 81.80%, whereas the balanced accuracy computed from the confusion matrix provided in [96] is 60.12%.

- Fisher et al. [75] 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, they have reported a standard overall accuracy of 78.1% and a balanced accuracy of 70.5%.

6.3.2 Parameter Settings

The parameters settings for GP are the same as those listed in Table 4.2 on page 121.

6.3.3 Experiments

For carrying out the experiments, the *10-fold cross validation* approach is adopted. This is because PH^2 is very small (200 images) and some classes in Dermofit have a very small number of images (Pyogenic Granuloma with 24 images). To segment the data into 10 folds, stratified random sampling is applied. The number of GP runs is 30. The results are represented in terms of the mean and standard deviation of the test performance values. For the task of binary classification, average sensitivity and average specificity values have also been reported.

Two sets of experiments are conducted. The first set of experiments are intended for the purpose of binary classification, which attempts to differentiate melanoma images from all the images provided. The second set of experiments investigate the effectiveness of MFCEC for multi-class classification. The results are compared with the other classification methods as described in Section 6.3.1. The binary classification results are also compared with the two embedded methods: EGP-5, and WGP-5 developed in Chapter 5. In WGP-5, six classification methods, namely NB, SVM, $k-$NN, J48, RF, and MLP (each individually executed) are used as a wrapper feature construction algorithm to search which classification algorithm performs better for binary and multi-class classification methods.

6.4 Experiment Results

The results are represented as the mean and standard deviation ($\bar{x} \pm s$) of the 30 GP runs, and are listed in Table 6.1. Since *10-fold cross validation* is used, the result of one GP run is the mean of the accuracies of the 10-folds. *Wilcoxon signed-rank test* (with a significance level of 5%) is applied to compare MFCEC to the other stochastic methods. *One-sample t-test* is applied to compare MFCEC to the other deterministic methods. For Wilcoxon signed-rank test, "+", "$-$" or "=" represents that MFCEC is significantly better, worse, or similar to the other algorithms. For one-sample t-test, ↑ or ↓ represents that MFCEC is significantly better or worse to the other algorithm. In each block of the Tables 6.1 and 6.2, the highest classification performance on each dataset in terms of balanced accuracy is made **bold** to clearly see which method has provided the best results among the methods listed in one block.

6.4.1 Binary Classification

The binary classification results are presented in Table 6.1. Among the non-GP methods, MLP has shown the highest average accuracy of 78.93% on PH^2, whereas NB produced the best accuracy 96.99% on Dermofit. Among the four ensemble methods, Bagging outperformed the other three methods giving 76.56% accuracy on the dermoscopic (PH^2) dataset. AdaBoost showed the highest accuracy 98.30% on the standard camera (Dermofit) dataset. Although the Embedded-GP method provided good accuracy on dermoscopic datasets outperforming all the non-GP and ensemble methods, it remain unable to achieve good results for standard camera images, where ensemble methods dominated all the non-GP and Embedded-GP methods. Similarly, among the EGP-5, WGP-5, non-GP, and ensemble methods, WGP-5 produced the best results on both the dermoscopic and Dermofit images. However, MFCEC produced the best results among all the methods, achieving 100.0% accuracy on both dermoscopic and standard camera images, respectively. This shows that feature construction in ensemble learning has huge potential to solve complex real-world problems like melanoma detection. The main reason of dominance of MFCEC over WGP-5 is that MFCEC constructs features for an ensemble of classifiers which are expected to be more general as compared to features constructed for a single classifier in WGP-5.

Table 6.1: Results of binary classification on the two real-world skin cancer datasets.

Algorithm		PH^2			Dermofit		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Non-GP Methods	NB	60.00	94.38	77.19 +	97.32	96.67	**96.99** +
	SVM	25.00	99.38	62.19 +	27.68	100.0	63.84 +
	k-NN	57.50	90.63	74.06 +	76.07	98.79	87.43 +
	J48	57.50	87.50	72.50 +	93.57	97.27	95.42 +
	MLP	59.00 ± 2.80	98.50 ± 1.54	**78.93 ± 2.47** ↑	92.34 ± 1.76	98.35 ± 2.60	94.48 ± 1.97 ↑
Ensemble Methods	RF	55.00	98.13	76.56 +	61.79	99.70	80.74 +
	Bgg	80.71	90.35	**76.56** +	90.15	97.64	93.46 +
	AB	59.05	87.41	68.44 +	96.75	99.41	**98.30** +
	LB	70.12	87.82	70.00 +	97.78	99.12	97.82 +
	RC	85.17	89.25	74.69 +	94.92	96.54	91.54 +
EGP-5	–	73.65 ± 4.92	84.09 ± 5.10	78.87 ± 2.92 ↑	75.82 ± 3.08	73.32 ± 3.45	74.57 ± 1.86 ↑
WGP-5	NB	86.42 ± 1.16	93.12 ± 0.70	**89.77 ± 1.84** ↑	95.60 ± 0.49	97.22 ± 0.32	96.21 ± 1.09 ↑
	SVM	73.83 ± 1.27	99.13 ± 0.25	86.48 ± 2.35 ↑	95.18 ± 0.60	99.61 ± 0.11	**97.26 ± 1.25** ↑
	k-NN	30.00 ± 0.68	96.68 ± 0.28	63.34 ± 2.67 ↑	73.00 ± 0.45	99.42 ± 0.13	86.04 ± 2.52 ↑
	J48	77.08 ± 0.08	98.14 ± 0.04	87.61 ± 3.08 ↑	95.11 ± 0.54	99.14 ± 0.51	96.99 ± 0.70 ↑
	RF	99.75 ± 0.67	99.98 ± 0.10	**99.93 ± 0.24** =	99.75 ± 0.48	99.96 ± 0.10	99.87 ± 0.23 =
	MLP	69.94 ± 5.98	89.68 ± 0.86	74.29 ± 1.94 ↑	100.0 ± 0.00	100.0 ± 0.00	**100.0 ± 0.00** =
MFCEC	–	100.0 ± 0.00	100.0 ± 0.00	**100.0 ± 0.00**	100.0 ± 0.00	100.0 ± 0.00	**100.0 ± 0.00**

6.4.2 Multi-class Classification

The multi-class classification results are presented in Table 6.2. Among the five non-GP algorithms, RF achieved the best accuracy 71.50% on PH^2, whereas MLP achieved the best accuracy 66.85% on Dermofit. Similar to binary classification results, among the ensemble methods, bagging provided the best results for PH^2 whereas Boosting (LogitBoost) provided highest accuracy for Dermofit. WGP-5 outperformed all the non-GP and ensemble methods providing an increase in accuracy by around 14% and 7% on average on the PH^2 and Dermofit datasets, respectively. It is worthwhile to note here that PH^2 has 3 classes and Dermofit has 10 classes (more difficult). For WGP-5, most of the single classifiers are performing well for a 3-class problem such as SVM, J48 and RF producing 84.92%, 85.82% and

Table 6.2: Results of multi-class classification on the two real-world skin cancer datasets.

Algorithm		PH^2	Dermofit
Non-GP Methods	NB	**71.00** ↑	45.92 ↑
	SVM	59.50 ↑	51.08 ↑
	k-NN	65.50 ↑	43.54 ↑
	J48	58.00 ↑	50.08 ↑
	MLP	67.50 ± 3.47 +	**64.92 ± 4.31** +
Ensemble Methods	RF	**71.50** ↑	47.92 ↑
	Bgg	**71.50** ↑	62.38 ↑
	AB	56.50 ↑	29.46 ↑
	LB	66.50 ↑	**62.62** ↑
	RC	70.00 ↑	58.38 ↑
WGP-5	NB	80.31 ± 2.03 +	58.99 ± 1.25 +
	SVM	84.92 ± 2.31 +	53.05 ± 1.57 +
	k-NN	63.46 ± 2.55 +	47.46 ± 1.85 +
	J48	85.82 ± 1.60 +	74.05 ± 1.52 +
	RF	**97.39 ± 1.00** +	**86.77 ± 0.88** =
	MLP	65.64 ± 1.92 +	41.84 ± 1.30 +
MFCEC	–	**98.03 ± 0.85**	**86.23 ± 1.30**

97.39% average accuracy, respectively, however, only RF performed well enough for the complex 10-class problem reaching 86.77% average accuracy. MFCEC remained prominent among all the methods in multi-class classification as well achieving 98.03% and 86.23% on average on the PH^2 and Dermofit datasets, respectively.

From the results of the statistical tests presented in Table 6.2, MFCEC outperformed WGP-5 on PH^2 and shown comparable performance on Dermofit dataset. Moreover, MFCEC outperformed all the non-GP, and ensemble methods on the easy (PH^2) and difficult (Dermofit) datasets, which shows its effectiveness for these complex skin cancer image classification problems.

Table 6.3: State-of-the-art Methods and their results.

Dataset	Method	Overall Accuracy	Balanced Accuracy
PH^2	Patino et al. [152]	-	86.50
Dermofit	Kawahara et al. [96]	81.80	60.12
	Fisher et al. [75]	78.10	70.50

6.4.3 Comparison to the State-of-the-Art Methods

For PH^2, the most recent state-of-the-art reported by Patino et al. [152] achieved 86.5% balanced accuracy using 10-fold cross validation. Since the experimental setup is the same as MFCEC, we can make a direct comparison. MFCEC outperformed this method by providing an increase of nearly 11% accuracy.

To the best of our knowledge, the state-of-the-art result on Dermofit for this 10-class skin image classification problem is presented by CNNs [96]. The authors reported an overall accuracy of 81.80% using 5-fold cross validation which came out to be 60.12% balanced accuracy (as calculated from the confusion matrix provided in the study). Recently, Fisher et al. [75] used 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 (5-folds vs 10-folds, and leave-one out vs 10-folds), we have provided a general idea what accuracy has been achieved by the current state-of the arts on Dermofit dataset.

6.5 Further Analysis

This section provides detailed overall analysis of the MFCEC method by examining their convergence plots, and showing how efficient these methods are by observing their computation time. It further shows evolved GP individual, and identifies prominent features by plotting frequency of occurrence in CFs of various types of features.

6.5.1 Overall Analysis

The average of best-of-generation fitness value of the 30 independent GP runs using different seed values on the training data of the PH^2 dataset in multi-class classification experiments is depicted in Figure 6.2. The plot shows how the accuracies of individual classifiers (SVM, J48, and RF) progress with the increase in generations and how much each of them contribute to the ensemble classification curve. Since elitism is applied to the ensemble classification and not on the individual classifiers, the individual classifiers' accuracies show behaviors of increase and decrease during the evolutionary process. However, they ensure that the collective performance increases as the number of generations increase. The benefit of using ensemble of classifiers is evident from this plot which clearly illustrates that if one classifier cannot produce good results, the ensemble can still rely on other classifiers to maintain good performance. From this plot, we observe that RF and J48 are producing far better results individually than SVM. However, when there is a decrease in the performance of RF and J48 in the subsequent generation, SVM makes larger jumps to maintain or even improve the performance of the ensemble classifier. This behavior is seen in the third and fourteenth generations.

We also compare the evolutionary process of MFCEC with the previous WGP-5 (Chapter 5) method as shown in Figure 6.3. The MFCEC curve shows an abrupt increase in the first five generations, being more powerful it achieves good performance in a very few earlier generations. However, the existing WGP-5 individual classifiers (NB, SVM, k-NN, and J48) start with lower average accuracy than MFCEC, thereby get the chance of making larger jumps as shown in first twenty generations. It is evident that MFCEC remained prominent and outperformed all the WGP-5 methods.

6.5.2 Computation Time

The average training time needed for the proposed MFCEC method and to test its performance on the unseen images for solving binary and multi-

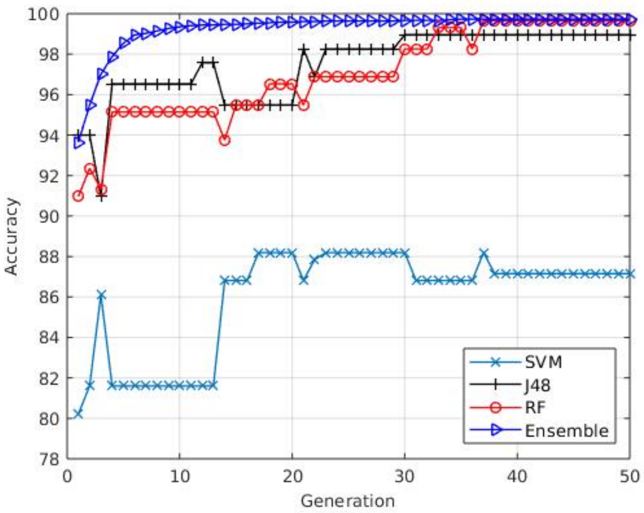


Figure 6.2: The convergence plot for SVM, J48, RF, and ensemble of these three classifiers.

class classification tasks is presented in Figure 6.4. Here, Figure 6.4(a) shows the training time in seconds to evolve the constructed features and train the three classification models in the ensemble to provide training performance. Figure 6.4(b) shows the time taken in milliseconds to apply the trained classifier on an unseen image. Clearly, the time required to train a classification algorithm is affected by the number of images and classes in a dataset. The binary classification task spend less time to train an ensemble of classifiers as compared to the multi-class classification task as shown in Figure 6.4(a). During the training process, the time is needs not only on evolving five constructed features but also on the transformation of the original training dataset to a new dataset (with the help of the constructed features) which are then used to train an ensemble of three classifiers (SVM, J48, and RF). Hence, training an ensemble of classification algorithms with the new constructed features results in increased computation time.

It is evident that evaluating a population of individuals with five trees

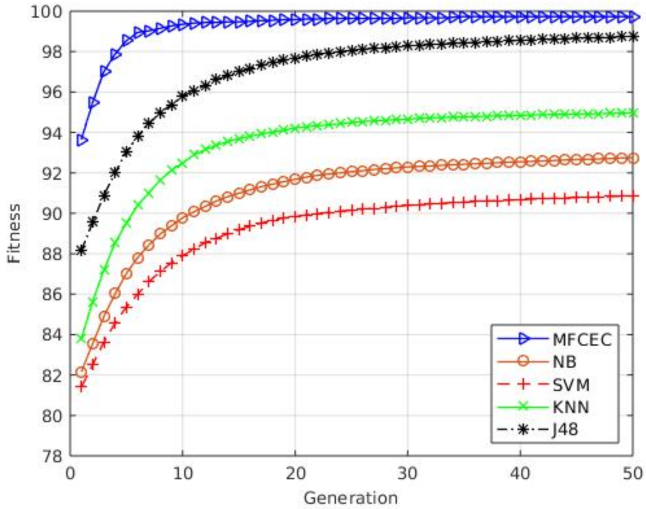
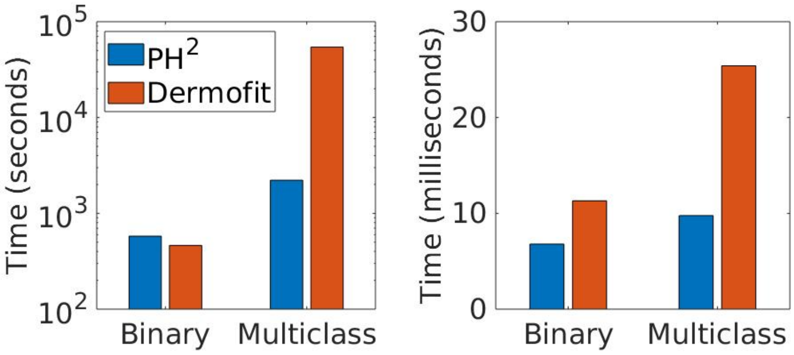


Figure 6.3: The convergence plot to compare MFCEC and the four existing WGP-5 with NB, SVM, k-NN, and J48, respectively.

requires more time as compared to evaluating a population of individuals having one tree. Moreover, similar to WGP-5, during the evolutionary process, the *same-index crossover/ mutation* is applied on four trees which has more computational complexity, thereby takes more time as compared to the simple crossover in case of the single-tree approaches. Although MFCEC is more expensive as compared to WGP-5, it takes only 9.6 and 7.7 minutes on average to evolve a binary class solution on PH^2 and Dermofit dataset, respectively. With these trained methods at hand, they take only 6.8 and 11.3 milliseconds on average to provide a label for an unseen skin image as *melanoma* or *benign*.

For multi-class classification, similar to WGP-5, Figure 6.4(a) depict that training a dataset with ten classes (Dermofit dataset) increases the computation time by many folds as compared to training a dataset with three classes (PH^2 dataset). Since multi-class classification methods require more training time as compared to binary classification methods, this behavior can easily be observed while comparing Figure 6.4(a). For



(a) Training

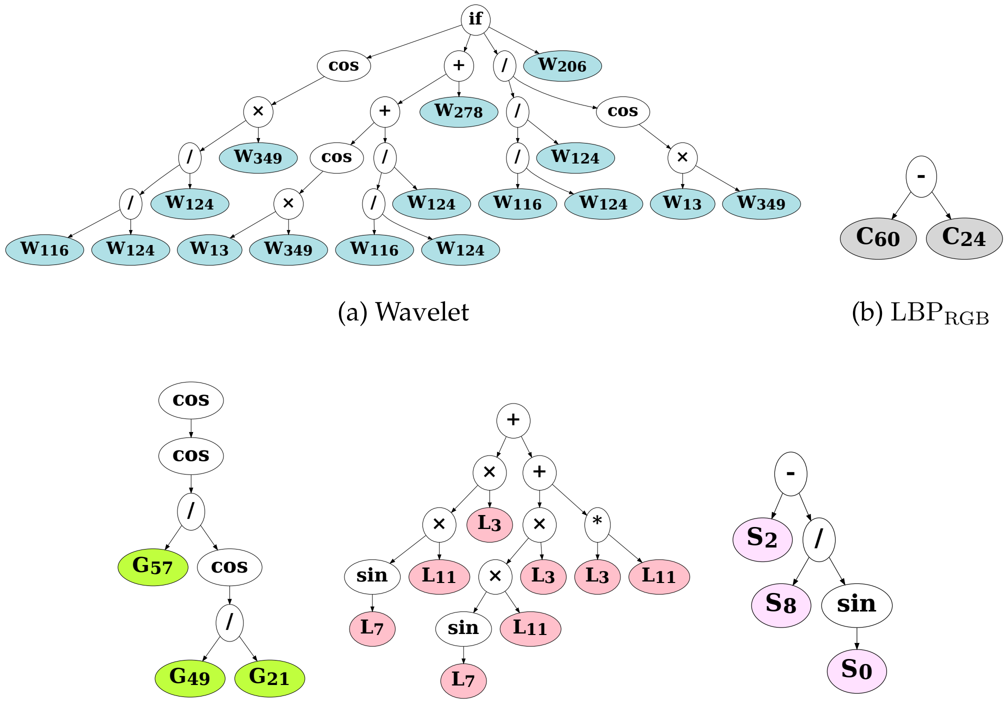
(b) Test

Figure 6.4: The average computation time for *binary classification* using WGP-4, WGP-5, EGP-4, and EGP-5 approaches on the two skin cancer image datasets.

illustration, PH^2 and Dermofit require 36.95 seconds and 15.1 hours on average to train an ensemble of classifiers in MFCEC, respectively. However, a label can be provided to a test image in fractions of a second using these trained models. For example, PH^2 and Dermofit require only 9.8 and 25.4 seconds on average to test an unseen image in a multi-class classification task as shown by the test time depicted in Figure 6.4(b).

6.5.3 Analysis of an Evolved GP Program

GP has the ability to evolve models that can be interpretable. To analyse why MFCEC can achieve good performance, we show a good evolved GP individual in Figure 6.5. This individual is taken from the PH^2 experiments for the binary classification task producing highest training performance. It has five trees evolved using the five types of features: a) Wavelet, b) $\mathrm{LBP}_{\mathrm{RGB}}$, c) $\mathrm{LBP}_{\mathrm{Gray}}$, d) $\mathrm{Lesion}_{\mathrm{Color}}$, and e) $\mathrm{Lesion}_{\mathrm{Shape}}$. These constructed features achieved 100.0% fitness produced by the ensemble classifier, where SVM produced 99.33%, J48 produced 99.83%, and RF produced 100% accuracy on the training data. Hence, selecting the highest



(a) Wavelet

(b) LBP_{RGB}

(c) LBP_{Gray}

(d) $\text{Lesion}_{\text{Color}}$

(e) $\text{Lesion}_{\text{Shape}}$

Figure 6.5: A good MFCEC evolved individual on the PH^2 dataset in the binary classification task producing 100% accuracy on the test data.

performing RF model when applied to the test data, produced 100% accuracy on the test data. In Figure 6.5, colored nodes represent terminals (each color represents one type of features) and white nodes represent functions.

With the ability of feature construction, GP plays a vital role in dimensionality reduction. From the evolved GP individual shown in Figure 6.5, GP has selected only 6 features from a total of 416 wavelet features, only 2 features from the 177 LBP_{RGB} features, only 3 features from the 59 LBP_{Gray} features, only 3 features from the 12 $\text{Lesion}_{\text{Color}}$ features, and 3 features from the 11 $\text{Lesion}_{\text{Shape}}$ features. The wavelet texture-based features appearing in a tree of the GP individual shown in Figure 6.5(a) are listed in Table 6.4. The following conclusions can be derived from this table:

- three out of the six features belong to the nodes from the third level, which indicates our use of three-level wavelet decomposition as further decomposition may not obtain informative features for the purpose of classification,

- texture features extracted from all the four color channels are selected to construct this informative constructed feature,

- the selected features are derived from both the low and the middle frequency channels as shown by the node column in Table 6.4, and

- among the eight statistical measures, norm, kurtosis, and entropy are prominent selected features.

Moreover, the sub-trees "$\cos(\mathrm{W}_{13} \times \mathrm{W}_{349})$" and "$(\mathrm{W}_{116}$ / $\mathrm{W}_{124})$ / W_{124}" appear twice and thrice, which shows the potential of these sub-trees getting selected multiple times to construct this informative wavelet-based constructed feature.

Among the $\mathrm{LBP}_{\mathrm{RGB}}$ and $\mathrm{LBP}_{\mathrm{Gray}}$ features, the constructed features shown in Figure 6.5(a) and (b) selected prominent LBP patterns corresponding to corners, edges and flat areas in these skin lesion images. Edges and corners identify various visual patterns such as streaks, blobs and pigment network inside a lesion area, whereas flat areas identify blue whitish veil and regions inside the blobs in the skin images. Therefore, these GP trees have selected prominent LBP patterns corresponding to significant visual characteristics of the skin lesions to build even more informative constructed features.

The $\mathrm{Lesion}_{\mathrm{Color}}$ tree in Figure 6.5(c) is built from three features L_3, L_7, and L_{11} which correspond to variance of red color channel ($\sigma\mathrm{R}$), ratio between mean of red and mean of blue color channels ($\frac{\mu_{\mathrm{R}}}{\mu_{\mathrm{B}}}$), and ratio between mean of blue color channel of lesion area and mean of blue color channel of skin area $\frac{\mu_{\mathrm{B}}}{\bar{\mu}_{\mathrm{B}}}$. They are combined in simple arithmetic operators to produce a significant constructed feature. In addition, the mathematical expression "$\mathrm{L}_3 \times (\mathrm{L}_{11} \times \sin(\mathrm{L}_7))$" appears twice which shows that this

Table 6.4: Wavelet features appearing in the GP individual shown in Figure 6.5(e).

Feature	Measure	Channel	Level	node
W_{13}	Norm	Green	0	–
W_{116}	Kurtosis	Red	3	3.4
W_{124}	Kurtosis	Red	3	3.1
W_{278}	Entropy	blue	2	2.4
W_{206}	Entropy	Green	3	3.3
W_{349}	Norm	Luminance	1	1.1

sub-tree captures significant information. In the $\mathrm{Lesion}_{\mathrm{Shape}}$ tree as shown in Figure 6.5(d), S_0, S_2, snd S_8 correspond to area of the lesion, greatest diameter, and the difference between greatest and shortest diameter of the lesion region, respectively. The lesion area, greatest and shortest diameter are vital in capturing the shape of the lesion. Here, rather selecting shortest diameter as individual feature, GP selected the difference of the greatest and shortest diameter, thereby incorporating important hand-crafted features effectively in evolving the $\mathrm{Lesion}_{\mathrm{Shape}}$ tree. These border shape features can hugely assist the dermatologist in real-time situations by providing significant knowledge about the lesion geometrical properties and hence, making a diagnosis much easier. Hence, the feature selection and construction ability of GP has provided discriminative constructed features as input to the ensemble classification algorithm contributing significantly to achieve promising results.

6.5.4 Comparisons between WGP-5 and MFCEC on the selected wavelet features

Since wavelet features have shown promising ability in improving the performance of skin cancer image classification, we are also interested to analyze their evolved constructed features in WGP-5 and MFCEC methods.

A good evolved individual by WGP-5 has been discussed in Section 5.6.3 on page 162. In Chapter 5, Figure 5.13(a) on page 165 represents the constructed feature using the wavelet features in the WGP-5 method. A good constructed feature evolved from the wavelet features in the MFCEC ensemble method has been shown in Figure 6.5(a).

While comparing the two features constructed from the wavelet features in Figure 5.13(a) and Figure 6.5(a), we see that MFCEC evolved bigger tree as compared to WGP-5. For illustration, the MFCEC wavelet feature, as shown in Figure 6.5(a), has 16 terminal nodes and 15 function nodes, whereas the WGP-5 wavelet tree, as shown in Figure 5.13(a), has 8 terminal nodes and 5 function nodes. However, the number of selected features in the WGP-5 wavelet tree is more than the number of selected features in MFCEC. The WGP-5 wavelet tree has selected eight features, whereas the MFCEC wavelet tree has selected six features. The MFCEC wavelet tree, with a smaller number of selected features while evolving a bigger tree than WGP-5, has many features and sub-trees appearing multiple times in this tree. As discussed in Section 6.5.3, MFCEC has used the effective sub-trees multiple times to make it more useful in distinguishing different types of images.

Furthermore, the details of these selected wavelet features by WGP-5 and MFCEC have been listed in Tables 5.5 and 6.4. These tables have shown the statistical measure each feature represents, the color channel from which each feature has been extracted, the level of pyramid structured wavelet decomposition, and the node in a level. We have reached the following conclusions by comparing these two tables: 1) the MFCEC wavelet tree has selected features from all the four color channels, whereas the WGP-5 wavelet tree has selected features from only two color channels (blue and luminance), 2) both the trees have selected Kurtosis, norm, and Entropy measures which demonstrates their higher discriminating ability than other five statistical measures for skin image classification, and 3) both the trees have selected wavelet features from all the three levels which proves that decomposing wavelets till third level is necessary to incorporate informative features. In conclusion, MFCEC tries to evolve general features (not tailored to a single classification algorithm) and, hence, searches the search space more to include features from all the four channels and all the three nodes of wavelet coefficients.

6.5.5 Feature Appearance in CFs of MFCEC

We have also explored and analyzed the intrinsic ability of GP to feature selection in the evolved GP trees of the MFCEC method. Figures 6.6 and 6.7 show the bars for the average number of times each feature appears in the constructed features among the 30 GP runs in the Dermofit experiments for binary and multi-class classification with ensemble classifiers, respectively.

It is evident from these plots that some features are selected more frequently as compared to the other features, e.g., W_8 and W_{216}, the two features among the wavelet features in Figure 6.6(a) appears almost four times and three times as frequently as the other 414 wavelet features. Similarly, C_{154}, G_2, L_{11} and S_{10} have the highest frequency of occurrence among the $\mathrm{LBP_{RGB}}$, $\mathrm{LBP_{Gray}}$, $\mathrm{Lesion_{Color}}$, and $\mathrm{Lesion_{Shape}}$ features as shown in Figures 6.6(b), (c), (d) and (e), respectively. We have seen a similar pattern while having a closer look at Figure 6.6 for the multi-class classification task in the Dermofit experiments, where these features have also been selected more frequently than other features. This shows that these features have significant discriminative ability between classes, not only for binary classification task but also for multi-class classification.

Among the wavelet features, Figure 6.6(a) and 6.7(a) show that W_{13}, and W_{216} have the most frequency of occurrence. W_{13}, which represents the norm statistical measure calculated from the green color channel of the original image, has also been selected by the evolved program shown in Figure 6.5(a) where it has appeared two times. W_{216} represents the kurtosis statistical measure calculated from the green color channel of the third level wavelet coefficients. This demonstrates that even at the third level of wavelet coefficients, the features have good discriminating ability between different types of skin images. Tables 5.5 and 6.4 list the details of the selected wavelet features in good evolved GP individuals of WGP-5 and MFCEC, respectively. A closer look to these tables provide evidence of *norm* and *kurtosis* statistical measures being more prominent.

Among the $\mathrm{LBP}_{\mathrm{RGB}}$ features, 6.6(b) and 6.7(b) show that C_{76} and C_{154} are selected more frequently, around four times more, than most of the other 175 $\mathrm{LBP}_{\mathrm{RGB}}$ features. These features represent edge texture pattern in LBP which is effective in identifying streaks, blobs, pigmented networks, and regions with abrupt color change within the lesion region. We have found that C_{154} has also been selected by a good GP tree shown in Figure 3.10 on page 105 in 2SGP-E, the two stage GP embedded feature selection and construction method in Chapter 3, where C_{154} has been selected four times in the GP tree. A detailed analysis of C_{154} $\mathrm{LBP}_{\mathrm{RGB}}$ feature can be found in Section 3.6.3 on page 104 with graphical illustrations earlier provided in Figures 3.11 and 3.12.

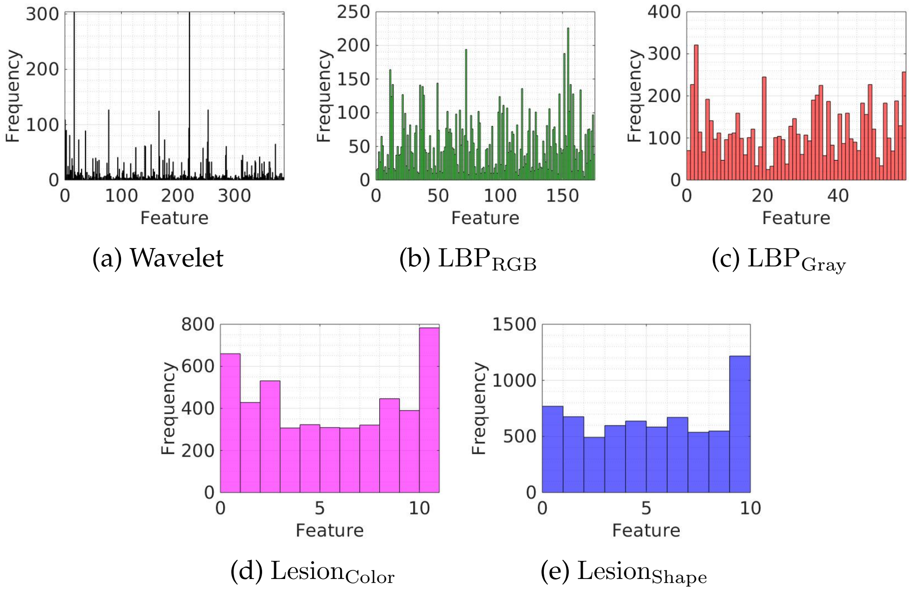
Among the $\mathrm{LBP}_{\mathrm{Gray}}$ features, G_{58} which is the last feature in the $\mathrm{LBP}_{\mathrm{Gray}}$ feature set has shown good discriminating ability between classes by getting selected most frequently in both the binary and multi-class classification tasks of WGP-4 on the PH^2 dataset as shown in Figures 5.16 and 5.17, respectively. Similarly, on the Dermofit dataset, G_{58} occurred most frequently in WGP-5 for multi-class classification task as shown in Figure 5.18. However, this pattern is not shown in MFCEC by $\mathrm{LBP}_{\mathrm{Gray}}$ features. But still G_{58} has been selected roughly two times compared to almost half of the $\mathrm{LBP}_{\mathrm{Gray}}$ features in both the binary and multi-class classification tasks on the Dermofit dataset as shown in Figures 6.6(c) and 6.7(c), respectively.

Among the $\mathrm{Lesion}_{\mathrm{Color}}$ features, L_{10} corresponds to $\frac{\mu_B}{\bar{\mu}_B}$, which shows the ratio between mean of blue color channel of the lesion region and its surrounding skin region. This feature has shown the highest frequency in both the binary and multi-class classification tasks as shown by Figures 6.6(d) and 6.7(d), respectively. This pattern is similar to WGP-4 and WGP-5 as shown in Figures 5.16(c), 5.17(c), and 5.18(c).

Among the $\mathrm{Lesion}_{\mathrm{Shape}}$ features, S_{11} which corresponds to asymmetry index, is the most prominent and appeared twice as compared to the rest of the $\mathrm{Lesion}_{\mathrm{Shape}}$ features. This trend has been observed in both the binary and multi-class classification tasks as shown in Figures 6.6(e) and 6.7(e). Asymmetry index of lesion is computed based on the major and minor asymmetry indices, hence, asymmetry index itself is a handcrafted constructed feature which include useful border shape geometrical information to differentiate different types of skin cancer images. It is the most important clinical property in the ABCD rule [125], and our evolved program in MFCEC confirms this as well. We have observed that this feature has been selected by all our methods described in previous chapters. It has been picked by EGP-4 as shown in Figure 4.5(d) on page 127. In Chapter 5, this is again the most selected feature in WGP-4 and WGP-5 as discussed in Section 5.6.5 on page 168, and Section 5.6.5 on page 170. We can conclude here that asymmetry index is the most prominent feature in all our methods including EGP-4, WGP-4, WGP-5, and MFCEC.

6.6 Chapter Summary

This chapter has developed an ensemble classification method based on GP for feature construction to solve the complex task of skin cancer image classification. The method constructs new powerful features from the pre-extracted texture, color, frequency-based, local and global features. These new constructed features when being provided to an ensemble of classifiers in a GP framework result in generating good trained models. The results have revealed that the constructed features generated for building the ensemble of classifiers have more distinguishing ability between

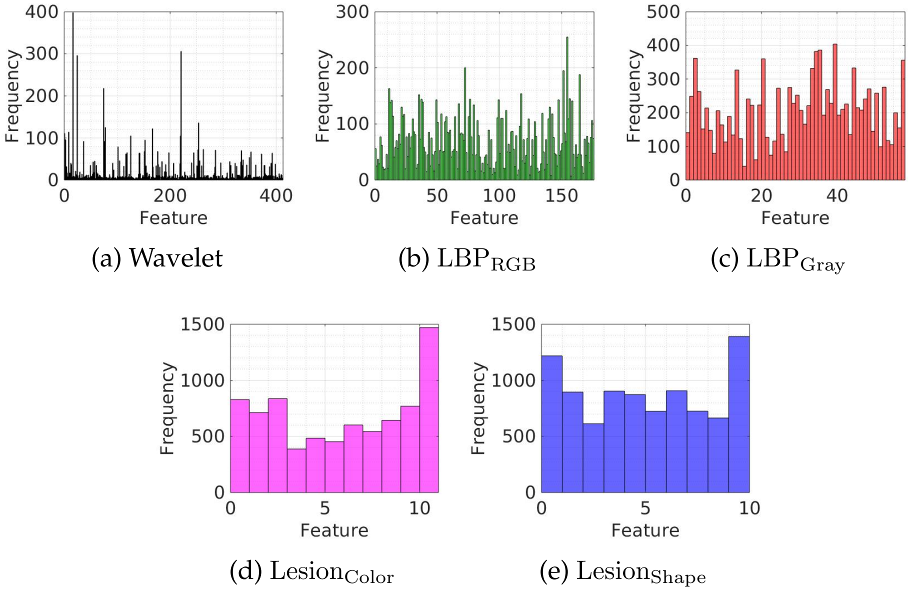


(a) Wavelet (b) LBP_{RGB} (c) LBP_{Gray}

(d) $\text{Lesion}_{\text{Color}}$ (e) $\text{Lesion}_{\text{Shape}}$

Figure 6.6: The average frequency of features in trees (the CFs) of MFCEC, each generated with one set of features on the Dermofit dataset in the *binary classification* task.

classes as compared to constructed features generated for a single classifier. The proposed method is evaluated on two benchmark real-world skin image datasets. The experimental results have revealed that the proposed algorithm has shown better or comparable performance than two existing GP approaches. Moreover, the proposed algorithm has significantly outperformed three state-of-the-art convolutional neural network methods, and ten commonly used classification algorithms. The proposed method has significantly outperformed all the existing GP approaches developed in Chapters 3 and 4. In comparison to the state-of-the-art CNN methods for the two datasets, the proposed method has produced significantly better results. Moreover, the proposed method significantly outperformed the commonly used classification algorithms (NB, SVM, k-NN, J48, and MLP) and ensemble methods (RF, Bagging, AdaBoost, LogitBoost, and RandomCommittee). Since the evolved individual that is considered as a



(a) Wavelet (b) LBP$_{\mathrm{RGB}}$ (c) LBP$_{\mathrm{Gray}}$

(d) Lesion$_{\mathrm{Color}}$ (e) Lesion$_{\mathrm{Shape}}$

Figure 6.7: The average frequency of features in trees (the CFs) of MFCEC, each generated with one set of features on the Dermofit dataset in the *multi-class classification* task.

set of constructed features are interpretable, the insights of good evolved constructed features have identified important features selected from the original set of features. This information can be helpful to dermatologists in making a diagnosis.

Although the proposed method has achieved very good results, its performance can be increased by generating more constructed features and investigating a suitable number of constructed features. Selecting only prominent constructed features, e.g. measuring their information gain, and providing those selected constructed features to the ensemble classifiers may improve results and will be investigated in the future. The computation time for the Dermofit dataset in the multi-class classification task is very high. An alternative method which can provide good results while using less training time is still needed to be explored.

Chapter 7

Conclusions

This chapter concludes the discussions of this thesis, highlights the main contributions and outlines directions for future work.

The overall goal of this thesis was to develop a new genetic programming (GP) based approach to skin cancer image classification by utilizing GP to evolve programs that are capable of automatically selecting prominent features, constructing new high-level features, interpreting useful features which can help dermatologist to diagnose cancer, and are robust to skin images captured from specialized instruments and standard cameras. This goal has been successfully achieved by developing a number of new GP methods incorporating various types of texture, color, local, global, and frequency based features to automatically select and construct new high-level features that have increased discriminating ability compared to the full set of features extracted by different methods. The proposed methods have been evaluated on two real-world skin cancer image datasets and compared with existing state-of-the-art methods. The experimental results show that the newly proposed methods in this thesis have achieved significantly better performance than the previous state-of-the-art methods. The interpretability of the evolved models has validated the importance of particular clinical features in skin cancer diagnosis.

The rest of this chapter provides conclusions for the individual objectives and outlines the key findings from each contribution chapter and outlines some potential research directions for future work.

7.1 Achieved Objectives

This thesis has achieved the following objectives:

- Proposes new two-stage GP methods for *feature selection and construction* in skin cancer image classification. The first method is an embedded based feature selection and construction approach, whereas the second method is a wrapper based feature selection and construction approach. These methods employed feature selection in stage-1 to select features from the original set of features, and employed feature construction in stage-2 to construct new high-level features from the selected features. These selected and constructed features are used to form a single feature vector which is given to the classification method for classification. The second method achieves better classification performance than the first method but uses longer computational time. The second method can solve binary and multi-class classification problems, whereas the first method can only solve binary classification problem. The selected and constructed features together have shown ability of GP to help common classification methods achieve better performance compared to using the full set of features. These methods constructed new features from the GP selected features, using the feature selection ability twice, resulting in more useful constructed features.

- Proposes two novel embedded *feature selection* methods using multi-tree GP (MTGP) for skin cancer image classification. The first method uses balanced accuracy as a fitness function, whereas the second method adopts a weighted fitness function and provides significantly better results than the first method on both skin cancer datasets. These methods can identify a suitable way of incorporating various local and global features extracted from skin cancer images. These different types of features are provided to MTGP by using suitable *same-index-crossover/mutation*. These methods have successfully demonstrated the ability to utilize the MTGP representation for melanoma detection by significantly outperforming all the single-tree GP methods. The methods evolve classification models to effectively and efficiently discriminate *"malignant"* from *"benign"* images. These methods have identified an interesting behavior for selecting a suitable feature extraction method for a particular type of images captured from a specific instrument. Both methods show that the local pixel-based features have a strong ability to classify dermoscopy images, whereas the global color variation and the domain-specific shape features are prominent for classifying standard camera images.

- Proposes two MTGP based wrapper methods for multiple *feature construction* in skin cancer image classification. The two methods can construct multiple features where each feature is generated from a single type of features. The first method utilizes only texture and color features to generate new features, whereas the second method also utilizes multi-scale properties of frequency based wavelet features in addition to texture and color features. Both the methods have provided good results for both the melanoma detection and multi-class classification tasks, where the second method significantly outperforms the first method. The second method demonstrates that wavelet based texture features have the best potential for both the dermoscopy and the standard camera images.

- Proposes a new GP method to *construct multiple features by an ensem-*

ble approach for skin cancer image classification. The method uses a multi-tree GP representation to construct new features to train an ensemble of classifiers. The experimental results demonstrate that the constructed features generated for an ensemble of classifiers have better discriminating ability between the different classes of skin cancer images than the constructed features generated for a single classifier. The method, evaluated on two benchmark real-world skin image datasets, shows its effectiveness by significantly outperforming existing GP approaches, state-of-the-art convolutional neural network methods, and commonly used classification and ensemble methods. The method provides interpretable constructed features that show the selection of critical clinical features to provide improved classification performance. With these trained classification models at hand, a label is predicted for an unseen image in fractions of a second.

7.2 Main Conclusions

Overall, this thesis finds that GP has good potential to address the problem of skin cancer images for classification by automatically selecting, from a number of extracted features, the prominent features and constructing new, more informative features. Most of the newly proposed methods in this thesis have successfully provided better classification performance than the prior state-of-the-art algorithms. The main conclusions are drawn from each of the four contribution chapters (Chapter 3 through Chapter 6) for the four research objectives are presented and discussed in this section.

7.2.1 Two-stage Feature Selection to Improve Classification Performance

Chapter 3 proposes two new GP based feature selection and construction algorithms, where one method is an embedded feature selection and construction approach and the second method is a wrapper approach. The feature selection in the first stage reduces the number of features in the second stage, hence, reducing the search space of feature construction yielding better constructed features which help improve the performance of skin cancer image classification.

Two-Stage GP

The two-stage GP based embedded feature selection and construction method is proposed in this thesis. It is found that employing multiple stages to select prominent features and construct new high-level features results in better classification performance compared to the traditional approach of feature selection in a single stage.

Since the full set of features includes redundant or irrelevant features, feature selection helps eliminate them. The second stage picks features from the selected features picked in the first stage has provided better classification performance compared to using all features. Moreover, new features constructed from the selected features and not from the complete set of features have more potential to discriminate between classes. This is mainly because the search space is smaller, hence, it is easier to find optimal solutions.

Embedded versus Wrapper Algorithms

This thesis finds that two-stage wrapper based feature selection and construction achieves better skin cancer image classification performance than embedded approaches. However, wrappers are computationally more expensive than embedded approaches. The performance of wrappers depends on the classification algorithm used during the feature selection and construction process. Simple classification algorithms such as NB and k-NN, take less time but cannot achieve better results than complex algorithms such as SVM and MLP.

7.2.2 Multi-tree GP for Embedded Feature Selection

This thesis designs a multi-tree GP based embedded feature selection approach to effectively incorporate different sets of features with color and texture, as well as local and global characteristics of skin images. In this embedded approach, each tree is considered as a binary classifier. Mixing different sets of features with different image properties results in poor classification performance. This thesis develops a multi-tree GP approach to effectively evolve each tree based on one set of features to improve the classification performance. Therefore, each tree picks the most prominent features in one set of features and discards the irrelevant or redundant features in that set of features, thereby improving the classification performance.

Single-tree versus Multi-tree GP

It is found in Chapter 4 that single-tree GP based embedded feature selection methods cannot provide good classification results. Using either all the sets of features with texture, color, local and global features, or a single set of features, in single tree GP cannot achieve good classification performance. This thesis proposes using multiple trees to handle each type of feature in a single tree which helps improve classification accuracy.

Fitness Function

This thesis proposes a new weighted fitness function, where the weights are assigned based on the classification accuracy of each tree in one GP individual. Using this fitness function, trees inside a GP individual influence each other's performance and interact during the evolutionary process, increasing their classification performance. This thesis finds that using a weighted fitness function provides better classification accuracy than using the average accuracy of the GP trees as a fitness function.

Domain Independent and Domain Specific Skin Image Features

GP can use different types of features concurrently. Chapter 4 uses four sets of features: two domain independent and two domain dependent. LBP features extracted from gray and color channels are the two domain independent feature sets. Lesion color variation and geometrical border shape features are the two domain dependent feature sets. This thesis has identified that using a single type of features does not provide sufficient information to discriminate against different skin cancer images. Therefore, GP has successfully included useful information from both the domain specific and domain independent features in a suitable way that helps improve classification performance.

Specialized versus Standard Cameras to Capture Skin Images

Though single tree GP did not achieve good performance, they identified an interesting behavior to classify images captured from the standard camera and specialized instruments. LBP color features (namely $\mathrm{LBP}_{\mathrm{RGB}}$) provide the highest discriminating ability between images of different cancer types captured from specialized instruments compared to other types of features. On the other hand, lesion color variation features (namely $\mathrm{Lesion}_{\mathrm{Color}}$) have the most potential to discriminate between images captured from a standard camera.

7.2.3 Multi-tree GP for Wrapper Feature Construction

This thesis proposes two multi-tree GP based multiple feature construction (WGP-4 and WGP-5) methods in a wrapper approach. WGP-4 utilizes four set of texture and color as well as local and global features whereas WGP-5, in addition to the four sets of features in WGP-4, uses a fifth set of frequency based wavelet features extracted from three-level pyramid structured wavelet decomposition to encompass detailed internal structure and global properties of skin cancer images. The experimental results show that the proposed methods have outperformed existing multi-tree GP embedded methods (proposed in Chapter 4), and commonly used machine learning classification algorithms. In addition, the proposed methods in Chapter 5 have provided improved classification performance on both datasets compared to the state-of-the-art skin cancer image classification methods.

Wavelet based Texture and Color Features

The goodness of utilizing wavelet features is evident while comparing the results of WGP-4 and WGP-5 methods. WGP-5 adds wavelet features to the previous WGP-4 method and constructs five new high-level features that have shown increased skin image classification performance. WGP-5 shows an increase of around 6% on the difficult task of classifying ten types of skin cancer, i.e., a 10-class classification problem.

Interpretability of Constructed Features in WGP-4 and WGP-5

This thesis has explored the interpretability of constructed features to identify prominent features for skin cancer image classification. Some LBP patterns showing corners, edges, and line ends are selected in the evolved features which correspond to presence of streaks and blobs in skin images. Moreover, this thesis analyzes the frequency of features in the constructed features which shows that particular features occur more frequently (twice, thrice or even four times) as compared to the other features. This thesis digged deeper into the details of these prominent features to analyze how well they try to include information of a clinical feature such as streaks, border edges, etc.

7.2.4 Ensemble Classification with Multiple Feature Construction in Multi-tree GP

This thesis proposes the *first* multi-tree GP based multiple feature construction method for an ensemble of classifiers (MFCEC). When provided to an ensemble of classifiers in a GP framework, the newly constructed features result in generating well-trained models during the evolutionary process. The results in Chapter 6 demonstrate that the constructed features generated for an ensemble of classifiers have more potential to discriminate between classes than the constructed features generated for a single classifier. In other words, more informative features can be constructed when the accuracy of an ensemble of classifiers is used to evaluate their goodness as compared to the accuracy of a single classifier.

Interpretability of Constructed Features in MFCEC

By analyzing the evolved constructed features, this thesis validates the importance of asymmetry property of skin lesions. The domain specific feature, *"asymmetry index"*, is found to be selected almost twice as compared to other handcrafted domain specific features, which shows its potential for generating useful evolved features. With the interpretability of GP programs, this thesis confirms that asymmetry is a crucial handcrafted feature in diagnosing skin cancers.

In summary, this thesis develops several methods to provide effective and efficient feature selection and feature construction strategies to improve the classification accuracy for the complex task of skin cancer image classification. The exiting approaches to skin cancer image classification have not studied feature selection and construction to improve the classification performance of their methods. This thesis validates the importance of utilizing domain specific knowledge as well as domain independent knowledge in producing good results for skin cancer image classification. Moreover, features extracted from both the gray-scale images and color images have been shown to have good information necessary to distinguish between images of different types of cancers. This thesis has further explored the use of frequency based features along with other types of features, which helped improve the classification accuracy. With the help of interpretable GP programs, this thesis has identified important features and associate their textural patterns with skin cancer characteristics such as streaks, blobs and globules. This is a vital contribution to the field of skin cancer image classification. This thesis has produced the best results for the binary classification task, mainly melanoma detection, by achieving 100% average accuracy on both datasets. For the multi-class classification task, this thesis significantly outperforms existing GP methods, state-of-the-art methods, and the commonly used classification and ensemble methods.

7.3 Future Work

Finally, this section provides some possible research directions for future work.

7.3.1 Feature Extraction Using GP for Skin Cancer Image Classification

This thesis has extensively investigated the feature selection and feature construction abilities of GP for providing effective and efficient solutions to skin cancer image classification. However, this thesis is based on existing feature extraction methods to extract features from the raw pixel values in the images. It will be interesting to investigate GP to directly extract features from skin images for feature selection and construction. Since these skin images are large in size, proposing feature extraction methods using GP, which are not computationally expensive, needs careful consideration.

1. It would be interesting to incorporate skin cancer detection operators in the function set of GP to extract informative features. New operators for cancer detection can be proposed aiming at identifying essential characteristics of cancer type present in the image. An appropriate set of operators may include directional Gabor filters, Harris Laplace, Differential of Gaussian and Laplacian functions, which will help in evolving discriminative features. Such spatial operators are good at highlighting rapid intensity change. Moreover, frequency and orientation representations of Gabor filters are found to be particularly appropriate for texture representation and discrimination, which targets at identifying the vital characteristics of skin cancer images. This thesis has demonstrated the potential of wavelet features for generating useful constructed features. Hence, using Gabor filters in the GP function set will be interesting to investigate in the future.

2. Skin images come with various artifacts. GP can be designed as a pre-processing step to deal with noise (such as hair, illumination, and gel) in skin cancer images. In the previous systems, a pre-processing step for hair removal and illumination correction is essential [170]. A GP based denoising method where GP applies an automatically generated filter to the lesion region may help extract informative image features. For real-world images, how to reduce noise without losing discriminative features is a challenging task and still requires research.

7.3.2 GP for Region Detection in Skin Cancer Image Classification

An image often cannot be fed directly into a classifier because of the amount of data in each image; therefore, a feature or a set of features is extracted from full images [131]. An image consists of millions of pixels, so reducing the enormous amount of data is an integral part of image analysis; this can be achieved by region detection and extracting features from the detected regions. In dermoscopy images, region detection plays a fundamental role in extracting important clinical features such as atypical pigment networks, globules, and blue-whitish veils present in selected regions of the lesion and do not cover the whole lesion area. Identifying such features can potentially enhance the classification performance.

1. Lesion segmentation [132] and border detection [4, 66] have been extensively used in skin cancer image classification; however, region detection for extracting prominent features have not been investigated. A GP based region detection system that selects important regions from dermoscopy images and then extracts features from those regions will be interesting to explore in the future.

2. A strongly typed GP system can be helpful in defining the appropriate shape and size of the region from which features are to be extracted. Using various shapes for region detection that correspond to the shapes of clinical features, are required for proper feature extraction. Moreover, the size of the features varies across the lesion area; for example, some blobs have bigger sizes than others, while the pigment network is present in a small diagonal area in one corner of the lesion. With GP's ability to dynamically select various kinds of shapes with different sizes, this system is expected to extract better features than human crafted features.

3. This thesis has explored both color and gray image spaces for ef-

fective feature selection and construction. These GP methods can be extended where different image spaces target to extract different properties of cancerous lesions to achieve better performance. According to the 7-point checklist method [24], the color characteristic is vital in distinguishing various classes of skin cancer. Without using color spaces, the regions prominent for blue-whitish veil and presence/absence of various colors (white, red, light-brown, dark-brown, blue-grey, black) cannot be properly extracted.

7.3.3 Multi-objective (MO) GP for Skin Cancer Image Classification

In skin cancer images, cancer may or may not be spread across the whole area of the mole. Moreover, it may not lie in one part or corner of the image and might be found in different regions. Therefore, there might be multiple important regions that only have cancer present in them. Therefore, detecting those regions helps to classify the image as cancer or non-cancer. Since a different number of regions may achieve the same classification accuracy, the minimum number of regions is desired. Therefore, it is a multi-objective problem where multiple objectives are to minimize the size of the region(s) and maximize the accuracy.

1. A multi-objective GP (MOGP) method for image classification can be explored with the objectives of maximizing the classification accuracy, minimizing the size of the detected regions, and minimizing the complexity of the evolved programs. In a multi-objective approach, a single MOGP experiment can evolve multiple solutions that show a trade-off between different objectives, allowing doctors to choose between these solutions depending on their preferences.

2. For multi-objective diagnostic classification, the members of the Pareto-optimal set correspond to operating points on an optimal receiver operating characteristic curve, whose performances describe

the limiting sensitivity-specificity trade-offs that the classifier can provide for the given training dataset [61]. Binary classifiers consider two conflicting objectives: 1) Sensitivity describing how well they classify the abnormal/diseased cases, and 2) Specificity describing how well they classify the normal/non-diseased cases. There is a trade-off between these two objective functions, and it is not always possible to simultaneously improve both the sensitivity and specificity.

3. With the promising results shown by the multi-tree approach in this thesis, it can be extended for MO approach where each tree in an evolved program targets two or more objectives mentioned above while performing multi-class classification. A multi-objective approach to multi-class classification using multi-tree representation can be explored simultaneously to achieve better performance on multiple objectives. More specifically, for each type of disease, each tree targets to maximize classification accuracy, minimize the size of the evolved program, and minimize the number of detected cancer regions.

7.3.4 GP using knowledge transfer in Skin Cancer Image Classification

Transfer learning or knowledge transfer is a promising approach to solving complex image classification tasks such as dermoscopic images, by utilizing the knowledge learned from more straightforward tasks such as Imagenet dataset [124]. In image analysis, GP with knowledge transfer has shown improved performance [90], where two crucial aspects of transfer learning in GP have been studied: "what to transfer", "how to transfer", and "when to transfer". These mainly address whether transfer subtrees from the evolved program or the whole tree is crucial in the learning process. Moreover, how much knowledge transfer (where features are based on local, global, color and gray scale information) is essential while maintaining good classification performance.

1. A feature representation using GP for the target domain can be investigated where knowledge used to transfer across domains is encoded into the feature representation. With this new feature representation having knowledge learned from the source domain, the performance of the target task is expected to improve significantly.

2. To investigate knowledge transfer between different cancer domains such as using the whole trained model or subtrees generated for one cancer domain to classify the instances of another cancer domain under the assumption that both of the domains, e.g., dermoscopic images (skin cancer) and mammograms or breast tissue images (breast cancer) are related may achieve improved classification performance.

Bibliography

[1] Cancer facts & figures. *American Cancer Society* (2008).

[2] WHO Disease and injury country estimates. *World Health Organization* (2012).

[3] WHO Disease and injury country estimates. *World Health Organization* (2018).

[4] ABBAS, Q., CELEBI, M. E., F. GARCÍA, I., AND RASHID, M. Lesion border detection in dermoscopy images using dynamic programming. *Skin Research and Technology 17*, 1 (2011), 91–100.

[5] ABBAS, Q., EMRE C., M., AND FONDÓN, I. Computer-aided pattern classification system for dermoscopy images. *Skin Research and Technology 18*, 3 (2012), 278–289.

[6] ABBASI, N. R., SHAW, H. M., RIGEL, D. S., FRIEDMAN, R. J., MCCARTHY, W. H., OSMAN, I., KOPF, A. W., AND POLSKY, D. Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. *Journal of the American Medical Association 292*, 22 (2004), 2771–2776.

[7] ABEDINI, M., CHEN, Q., CODELLA, N. C., GARNAVI, R., AND SUN, X. Accurate and scalable system for automatic detection of malignant melanoma. *Dermoscopy Image Analysis* (2015), 293–343.

[8] ADJED, F., GARDEZI, S. J. S., ABABSA, F., FAYE, I., AND DASS, S. C. Fusion of structural and textural features for melanoma recognition. *IET Computer Vision 12*, 2 (2017), 185–195.

[9] AFZALI, S., AL-SAHAF, H., XUE, B., HOLLITT, C., AND ZHANG, M. Genetic programming for feature selection and feature combination in salient object detection. In *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)* (2019), Springer, pp. 308–324.

[10] AHMED, S., ZHANG, M., AND PENG, L. Enhanced feature selection for biomarker discovery in LC-MS data using GP. In *Proceedings of the 2013 IEEE Congress on Evolutionary Computation* (2013), IEEE, pp. 584–591.

[11] AHMED, S., ZHANG, M., PENG, L., AND XUE, B. Multiple feature construction for effective biomarker identification and classification using genetic programming. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation* (2014), ACM, pp. 249–256.

[12] AIN, Q. U., XUE, B., AL-SAHAF, H., AND ZHANG, M. Genetic programming for skin cancer detection in dermoscopic images. In *Proceedings of the 2017 Congress on Evolutionary Computation* (2017), pp. 2420–2427.

[13] AIN, Q. U., XUE, B., AL-SAHAF, H., AND ZHANG, M. Genetic programming for feature selection and feature construction in skin cancer image classification. In *The 15th Pacific Rim International Conference on Artificial Intelligence (PRICAI). Lecture Notes in Computer Science.* (2018), Springer, p. (to appear).

[14] AL-SAHAF, H., AL-SAHAF, A., XUE, B., JOHNSTON, M., AND ZHANG, M. Automatically evolving rotation-invariant texture im-

age descriptors by genetic programming. *IEEE Transactions on Evolutionary Computation 21*, 1 (2016), 83–101.

[15] AL-SAHAF, H., NESHATIAN, K., AND ZHANG, M. Automatic feature extraction and image classification using genetic programming. In *Proceedings of the 5th International Conference on Automation, Robotics and Applications (ICARA), 2011* (2011), IEEE, pp. 157–162.

[16] AL-SAHAF, H., SONG, A., NESHATIAN, K., AND ZHANG, M. Two-tier genetic programming: towards raw pixel-based image classification. *Expert Systems with Applications 39*, 16 (2012), 12291–12301.

[17] AL-SAHAF, H., SONG, A., AND ZHANG, M. Hybridisation of genetic programming and nearest neighbour for classification. In *Proceedings of the 2013 IEEE Congress on Evolutionary Computation* (2013), IEEE, pp. 2650–2657.

[18] AL-SAHAF, H., XUE, B., AND ZHANG, M. Evolving texture image descriptors using a multitree genetic programming representation. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (2017), ACM, pp. 219–220.

[19] AL-SAHAF, H., XUE, B., AND ZHANG, M. A multitree genetic programming representation for automatically evolving texture image descriptors. In *Asia-Pacific Conference on Simulated Evolution and Learning* (2017), Lecture Notes in Computer Science, Springer, pp. 499–511.

[20] AL-SAHAF, H., ZHANG, M., AND JOHNSTON, M. Binary image classification using genetic programming based on local binary patterns. In *Proceedings of the 28th International Conference on Image and Vision Computing New Zealand* (2013), IEEE, pp. 220–225.

[21] ALFED, N., AND KHELIFI, F. Bagged textural and color features for melanoma skin cancer detection in dermoscopic and standard images. *Expert Systems with Applications 90* (2017), 101–110.

[22] ALFED, N., KHELIFI, F., BOURIDANE, A., AND SEKER, H. Pigment network-based skin cancer detection. In *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2015), IEEE, pp. 7214–7217.

[23] ALPAYDIN, E. *Introduction to Machine Learning*, 2nd ed. The MIT Press, 2010.

[24] ARGENZIANO, G., FABBROCINI, G., CARLI, P., DE GIORGI, V., SAMMARCO, E., AND DELFINO, M. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of Dermatology 134*, 12 (1998), 1563–1570.

[25] ASWIN, R. B., JALEEL, J. A., AND SALIM, S. Hybrid genetic algorithm–artificial neural network classifier for skin cancer detection. In *Proceedings of the 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies* (2014), IEEE, pp. 1304–1309.

[26] ATKINS, D., NESHATIAN, K., AND ZHANG, M. A domain independent genetic programming approach to automatic feature extraction for image classification. In *Proceedings of the 2011 IEEE Congress on Evolutionary Computation* (2011), IEEE, pp. 238–245.

[27] AWATE, S. P., AND WHITAKER, R. T. Unsupervised, information-theoretic, adaptive image filtering for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 3 (2006), 364–376.

[28] Ballard, D. H., and Brown, C. M. Computer vision. *Prenice-Hall, Englewood Cliffs* (1982).

[29] Ballerini, L., Fisher, R. B., Aldridge, B., and Rees, J. 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] Barata, C., Celebi, M. E., and Marques, J. S. Improving dermoscopy image classification using color constancy. *IEEE Journal of Biomedical and Health Informatics 19*, 3 (2015), 1146–1152.

[31] Barata, C., Celebi, M. E., and Marques, J. S. Development of a clinically oriented system for melanoma diagnosis. *Pattern Recognition 69* (2017), 270–285.

[32] Barata, C., and Marques, J. S. Deep learning for skin cancer diagnosis with hierarchical architectures. In *Proceedings of the International Symposium on Biomedical Imaging* (2019), vol. 2, IEEE.

[33] Barata, C., Marques, J. S., and Mendonça, T. Bag-of-features classification model for the diagnose of melanoma in dermoscopy images using color and texture descriptors. In *Proceedings of the International Conference Image Analysis and Recognition* (2013), Springer, pp. 547–555.

[34] Barata, C., Marques, J. S., and Rozeira, J. A system for the detection of pigment network in dermoscopy images using directional filters. *IEEE Transactions on Biomedical Engineering 59*, 10 (2012), 2744–2754.

[35] Barata, C., Marques, J. S., and Rozeira, J. A system for the detection of pigment network in dermoscopy images using directional filters. *IEEE Transactions on Biomedical Engineering 59*, 10 (2012), 2744–2754.

[36] Barata, C., Marques, J. S., and Rozeira, J. Evaluation of color based keypoints and features for the classification of melanomas using the bag-of-features model. In *Proceedings of the International Symposium on Visual Computing* (2013), Springer, pp. 40–49.

[37] Barata, C., Ruela, M., Mendonça, T., and Marques, J. S. A bag-of-features approach for the classification of melanomas in dermoscopy images: The role of color and texture descriptors. In *Computer vision techniques for the diagnosis of skin cancer*. Springer, 2014, pp. 49–69.

[38] Bhanu, B., and Lin, Y. Object detection in multi-modal images using genetic programming. *Applied Soft Computing 4*, 2 (2004), 175–201.

[39] Bhatia, S. K., and Deogun, J. S. Data mining tools: Formal concept analysis and rough sets. In *Encyclopedia of Business Analytics and Optimization*. IGI Global, 2014, pp. 655–663.

[40] Bi, Y., Xue, B., and Zhang, M. An automatic feature extraction approach to image classification using genetic programming. In *International Conference on the Applications of Evolutionary Computation* (2018), Springer, pp. 421–438.

[41] Bi, Y., Xue, B., and Zhang, M. An automated ensemble learning framework using genetic programming for image classification. In *Proceedings of the Genetic and Evolutionary Computation Conference* (2019), pp. 365–373.

[42] Bi, Y., Xue, B., and Zhang, M. Genetic programming with a new representation to automatically learn features and evolve ensembles for image classification. *IEEE Transactions on Cybernetics* (2020).

[43] Bi, Y., Xue, B., and Zhang, M. Genetic programming with image-related operators and a flexible program structure for feature learn-

ing in image classification. *IEEE Transactions on Evolutionary Computation* (2020).

[44] BI, Y., ZHANG, M., AND XUE, B. Genetic programming for automatic global and local feature extraction to image classification. In *2018 IEEE Congress on Evolutionary Computation (CEC)* (2018), IEEE, pp. 1–8.

[45] BINDER, M., SCHWARZ, M., WINKLER, A., STEINER, A., KAIDER, A., WOLFF, K., AND PEHAMBERGER, H. Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. *Archives of Dermatology 131*, 3 (1995), 286–291.

[46] BISHOP, C. M. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., 1995.

[47] BRINKER, T. J., HEKLER, A., ENK, A. H., BERKING, C., HAFERKAMP, S., HAUSCHILD, A., WEICHENTHAL, M., KLODE, J., SCHADENDORF, D., HOLLAND-LETZ, T., ET AL. Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer 119* (2019), 11–17.

[48] BROWNLEE, J. *Clever algorithms: nature-inspired programming recipes*. 2011.

[49] BULLER, D. B., COKKINIDES, V., HALL, H. I., HARTMAN, A. M., SARAIYA, M., MILLER, E., PADDOCK, L., AND GLANZ, K. Prevalence of sunburn, sun protection, and indoor tanning behaviors among americans: review from national surveys and case studies of 3 states. *Journal of the American Academy of Dermatology 65*, 5 (2011), S114–e1.

[50] Burks, A. R., and Punch, W. F. Genetic programming for tuberculosis screening from raw x-ray images. In *Proceedings of the Genetic and Evolutionary Computation Conference* (2018), pp. 1214–1221.

[51] Burling-Claridge, F., Iqbal, M., and Zhang, M. Evolutionary algorithms for classification of mammographie densities using local binary patterns and statistical features. In *Proceedings of the 2016 IEEE Congress on Evolutionary Computation* (2016), pp. 3847–3854.

[52] Cagnoni, S., Dobrzeniecki, A. B., Poli, R., and Yanch, J. C. Genetic algorithm-based interactive segmentation of 3d medical images. *Image and Vision Computing 17*, 12 (1999), 881–895.

[53] Cagnoni, S., Poli, R., Smith, G. D., Corne, D., Oates, M., Hart, E., Lanzi, P. L., Willem, E. J., Li, Y., Paechter, B., and Fogarty, T. C. *Real-world applications of evolutionary computing*. Springer Science & Business Media, 2000.

[54] Chang, T., and Kuo, C.-C. J. Texture analysis and classification with tree-structured wavelet transform. *IEEE Transactions on Image Processing 2*, 4 (1993), 429–441.

[55] Chen, Q., Xue, B., and Zhang, M. Generalisation and domain adaptation in GP with gradient descent for symbolic regression. In *Proceedings of the 2015 IEEE Congress on Evolutionary Computation* (2015), IEEE, pp. 1137–1144.

[56] Chen, Q., Xue, B., and Zhang, M. Improving generalization of genetic programming for symbolic regression with angle-driven geometric semantic operators. *IEEE Transactions on Evolutionary Computation 23*, 3 (2018), 488–502.

[57] Choi, W., and Choi, T. Computer-aided detection of pulmonary nodules using genetic programming. In *Proceedings of the 2010 IEEE*

International Conference on Image Processing (2010), IEEE, pp. 4353–4356.

[58] CHOI, W., AND CHOI, T. Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images. *Information Sciences 212* (2012), 57–78.

[59] CLAWSON, K. M., MORROW, P. J., SCOTNEY, B. W., MCKENNA, D. J., AND DOLAN, O. M. Determination of optimal axes for skin lesion asymmetry quantification. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on* (2007), vol. 2, IEEE, pp. II–453.

[60] CODELLA, N., CAI, J., ABEDINI, M., GARNAVI, R., HALPERN, A., AND SMITH, J. R. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In *Proceedings of the International Workshop on Machine Learning in Medical Imaging* (2015), Springer, pp. 118–126.

[61] COELLO, C. A. C., AND LAMONT, G. B. *Applications of multi-objective evolutionary algorithms*, vol. 1. World Scientific, 2004.

[62] COLLOBERT, R., AND BENGIO, S. Links between perceptrons, MLPs and SVMs. In *Proceedings of the Twenty-first International Conference on Machine learning* (2004), ACM, p. 23.

[63] COTTRELL, G. W. Extracting features from faces using compression networks: Face, identity, emotion and gender recognition using holons. In *proceedings of the 1990 summer school on Connectionist models* (1990), pp. 328–337.

[64] DEMYANOV, S., CHAKRAVORTY, R., ABEDINI, M., HALPERN, A., AND GARNAVI, R. Classification of dermoscopy patterns using deep convolutional neural networks. In *Proccedings of the 2016 IEEE 13th*

International Symposium on Biomedical Imaging (2016), IEEE, pp. 364–368.

[65] DENG, H., RUNGER, G., AND TUV, E. Bias of importance measures for multi-valued attributes and solutions. *Artificial Neural Networks and Machine Learning* (2011), 293–300.

[66] E. CELEBI, M., KINGRAVI, H. A., IYATOMI, H., A. ASLANDOGAN, Y., STOECKER, W. V., MOSS, R. H., MALTERS, J. M., GRICHNIK, J. M., MARGHOOB, A. A., RABINOVITZ, H. S., ET AL. Border detection in dermoscopy images using statistical region merging. *Skin Research and Technology 14*, 3 (2008), 347–353.

[67] EBERHART, R., AND KENNEDY, J. A new optimizer using particle swarm theory. In *Proceedings of the Sixth International Symposium on Micro Machine and Human Science* (1995), IEEE, pp. 39–43.

[68] EBNER, M., AND ZELL, A. *Evolving a Task Specific Image Operator*. Springer Berlin Heidelberg, 1999, pp. 74–89.

[69] ENGLISH, D. R., ARMSTRONG, B. K., KRICKER, A., WINTER, M. G., HEENAN, P. J., AND RANDELL, P. L. Case-control study of sun exposure and squamous cell carcinoma of the skin. *International Journal of Cancer 77*, 3 (1998), 347–353.

[70] EROL, R. *Skin Cancer Malignancy Classification with Transfer Learning*. University of Central Arkansas, 2018.

[71] ESPEJO, P. G., VENTURA, S., AND HERRERA, F. A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40*, 2 (2010), 121–144.

[72] ESTEVA, A., KUPREL, B., NOVOA, R. A., KO, J., SWETTER, S. M., BLAU, H. M., AND THRUN, S. Dermatologist-level classification

of skin cancer with deep neural networks. *Nature 542*, 7639 (2017), 115–118.

[73] EVANS, B., AL-SAHAF, H., XUE, B., AND ZHANG, M. Evolutionary deep learning: A genetic programming approach to image classification. In *IEEE Congress on Evolutionary Computation (CEC)* (2018), IEEE, pp. 1–6.

[74] FERRIS, L. K., HARKES, J. A., GILBERT, B., WINGER, D. G., GOLUBETS, K., AKILOV, O., AND SATYANARAYANAN, M. Computer-aided classification of melanocytic lesions using dermoscopic images. *Journal of the American Academy of Dermatology 73*, 5 (2015), 769–776.

[75] FISHER, R. B., REES, J., AND BERTRAND, A. Classification of ten skin lesion classes: Hierarchical knn versus deep net. In *Annual Conference on Medical Image Understanding and Analysis* (2019), Springer, pp. 86–98.

[76] FOGELBERG, C., AND ZHANG, M. Linear genetic programming for multi-class object classification. In *Australasian Joint Conference on Artificial Intelligence* (2005), Springer, pp. 369–379.

[77] FRIEDMAN, R. J., RIGEL, D. S., AND KOPF, A. W. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *A Cancer Journal for Clinicians 35*, 3 (1985), 130–151.

[78] FU, W., JOHNSTON, M., AND ZHANG, M. Genetic programming for edge detection: a gaussian-based approach. *Soft Computing 20*, 3 (2016), 1231–1248.

[79] GANSTER, H., PINZ, P., ROHRER, R., WILDLING, E., BINDER, M., AND KITTLER, H. Automated melanoma recognition. *IEEE transactions on medical imaging 20*, 3 (2001), 233–239.

[80] GARNAVI, R., ALDEEN, M., AND BAILEY, J. Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis. *IEEE Transactions on Information Technology in Biomedicine 16*, 6 (2012), 1239–1252.

[81] GREEN, R. D., GUAN, L., AND BURNE, J. A. Video analysis of gait for diagnosing movement disorders. *Journal of Electronic Imaging 9*, 1 (2000), 16–21.

[82] GUERRA-SALCEDO, C., AND WHITLEY, D. Genetic approach to feature selection for ensemble creation. In *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 1* (1999), pp. 236–243.

[83] GUSTAFSON, S., BURKE, E. K., AND KRASNOGOR, N. On improving genetic programming for symbolic regression. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation* (2005), vol. 1, IEEE, pp. 912–919.

[84] HAENSSLE, H., FINK, C., ET AL. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology* (2018).

[85] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. The WEKA data mining software: an update. *Special Interest Group on Knowledge Discovery and Data Mining Explorations Newsletter 11*, 1 (2009), 10–18.

[86] HARANGI, B., BARAN, A., AND HAJDU, A. Classification of skin lesions using an ensemble of deep neural networks. In *Proceeddings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2018), IEEE, pp. 2575–2578.

[87] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *The elements of statistical learning*. Springer series in statistics, 2009.

[88] HE, X., PAN, J., JIN, O., XU, T., LIU, B., XU, T., SHI, Y., ATALLAH, A., HERBRICH, R., BOWERS, S., AND Q., C. J. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising* (2014), ACM, pp. 1–9.

[89] HSU, C., CHANG, C., AND LIN, C. A practical guide to support vector classification, 2003.

[90] IQBAL, M., XUE, B., AL-SAHAF, H., AND ZHANG, M. Cross-domain reuse of extracted knowledge in genetic programming for image classification. *IEEE Transactions on Evolutionary Computation 21*, 4 (2017), 569–587.

[91] JEREZ-ARAGONÉS, J. M., GÓMEZ-RUIZ, J. A., RAMOS-JIMÉNEZ, G., MUÑOZ PÉREZ, J., AND ALBA-CONEJO, E. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence in Medicine 27*, 1 (2003), 45–63.

[92] JOACHIMS, T. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning* (1999), Morgan Kaufmann Publishers Inc., pp. 200–209.

[93] JONES, W. O., HARMAN, C. R., NG, A. K., AND SHAW, J. H. Incidence of malignant melanoma in Auckland, New Zealand: highest rates in the world. *World Journal of Surgery 23*, 7 (1999), 732–735.

[94] KASMI, R., AND MOKRANI, K. Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule. *IET Image Processing 10*, 6 (2016), 448–455.

[95] KAUR, R., ALBANO, P. P., COLE, J. G., HAGERTY, J., LEANDER, R. W., MOSS, R. H., AND STOECKER, W. V. Real-time supervised detection of pink areas in dermoscopic images of melanoma: importance of color shades, texture and location. *Skin Research and Technology 21*, 4 (2015), 466–473.

[96] KAWAHARA, J., BENTAIEB, A., AND HAMARNEH, G. Deep features to classify skin lesions. In *Proceedings of the 13th International Symposium on Biomedical Imaging* (2016), IEEE, pp. 1397–1400.

[97] KEERTHI, S. S., AND LIN, C.-J. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation 15*, 7 (2003), 1667–1689.

[98] KENNEDY, J. Swarm intelligence. In *Handbook of Nature-Inspired and Innovative Computing*. Springer, 2006, pp. 187–219.

[99] KENNEDY, J. Particle swarm optimization. In *Encyclopedia of Machine Learning*. Springer, 2011, pp. 760–766.

[100] KIANI, K., AND SHARAFAT, A. R. E-shaver: An improved dullrazor® for digitally removing dark and light-colored hairs in dermoscopic images. *Computers in Biology and Medicine 41*, 3 (2011), 139–145.

[101] KOLMOGOROV, V., AND ZABIH, R. Multi-camera scene reconstruction via graph cuts. *Computer Vision* (2002), 8–40.

[102] KOROTKOV, K., AND GARCIA, R. Computerized analysis of pigmented skin lesions: a review. *Artificial Intelligence in Medicine 56*, 2 (2012), 69–90.

[103] KOZA, J. R. *Genetic programming: on the programming of computers by means of natural selection*, vol. 1. MIT press, 1992.

[104] LANG, K. J., WAIBEL, A. H., AND HINTON, G. E. A time-delay neural network architecture for isolated word recognition. *Neural Networks 3*, 1 (1990), 23–43.

[105] LECUN, Y., BOSER, B., DENKER, J. S., HENDERSON, D., HOWARD, R. E., HUBBARD, W., AND JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation 1*, 4 (1989), 541–551.

[106] LEE, J., AHN, C. W., AND AN, J. An approach to self-assembling swarm robots using multitree genetic programming. *The Scientific World Journal* (2013).

[107] LEE, J., ANARAKI, J. R., AHN, C. W., AND AN, J. Efficient classification system based on fuzzy–rough feature selection and multitree genetic programming for intension pattern recognition using brain signal. *Expert Systems with Applications 42*, 3 (2015), 1644–1651.

[108] LEE, T., NG, V., GALLAGHER, R., COLDMAN, A., AND MCLEAN, D. Dullrazor®: A software approach to hair removal from images. *Computers in Biology and Medicine 27*, 6 (1997), 533–543.

[109] LENSEN, A., AL-SAHAF, H., ZHANG, M., AND VERMA, B. Genetic programming for algae detection in river images. In *proceedings of the 2015 IEEE Congress on Evolutionary Computation* (2015), IEEE, pp. 2468–2475.

[110] LENSEN, A., AL-SAHAF, H., ZHANG, M., AND XUE, B. Genetic programming for region detection, feature extraction, feature construction and classification in image data. In *Proceedings of the European Conference on Genetic Programming* (2016), Springer, pp. 51–67.

[111] LENSEN, A., XUE, B., AND ZHANG, M. Generating redundant features with unsupervised multi-tree genetic programming. In *European Conference on Genetic Programming* (2018), Springer, pp. 84–100.

[112] LIU, H., AND MOTODA, H. *Feature extraction, construction and selection: A data mining perspective*, vol. 453. Springer Science & Business Media, 1998.

[113] LOWE, D. G. Object recognition from local scale-invariant features. In *Proceedings of the 7th International Conference on Computer Vision* (1999), vol. 2, IEEE, pp. 1150–1157.

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

[115] LYNN, N. C., AND WAR, N. Melanoma classification on dermoscopy skin images using bag tree ensemble classifier. In *2019 International Conference on Advanced Information Technologies (ICAIT)* (2019), IEEE, pp. 120–125.

[116] MACKIE, R. M., AND DOHERTY, V. R. Seven-point checklist for melanoma. *Clinical and Experimental Dermatology 16*, 2 (1991), 151–152.

[117] MAGLOGIANNIS, I., AND DOUKAS, C. N. Overview of advanced computer vision systems for skin lesions characterization. *IEEE Transactions on Information Technology in Biomedicine 13*, 5 (2009), 721–733.

[118] MANIEZZO, A. Distributed optimization by ant colonies. In *Proceedings of the First European Conference on Artificial Life: Toward a Practice of Autonomous Systems* (1992), Mit Press, p. 134.

[119] MANIKAS, T. W., ASHENAYI, K., AND WAINWRIGHT, R. L. Genetic algorithms for autonomous robot navigation. *IEEE Instrumentation Measurement Magazine 10*, 6 (2007), 26–31.

[120] MASOOD, A., AND A. AL-JUMAILY, A. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. *International Journal of Biomedical Imaging* (2013).

[121] MATTHEWS, N. H., LI, W. Q., QURESHI, A. A., WEINSTOCK, M. A., AND CHO, E. Epidemiology of melanoma. In *Cutaneous Melanoma: Etiology and Therapy [Internet]*. Codon Publications, 2017.

[122] MCPHEE, N., OHS, B., AND HUTCHISON, T. Semantic building blocks in genetic programming. In *Proceedings of the 11th European conference on Genetic programming* (2008), Springer, pp. 134–145.

[123] MENDONÇA, T., FERREIRA, P. M., MARQUES, J. S., MARCAL, A. R. S., AND ROZEIRA, J. Ph2–a dermoscopic image database for research and benchmarking. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2013), IEEE, pp. 5437–5440.

[124] MENEGOLA, A., FORNACIALI, M., PIRES, R., BITTENCOURT, F. V., AVILA, S., AND VALLE, E. Knowledge transfer for melanoma screening with deep learning. In *Proceedings of the 14th International Symposium on Biomedical Imaging* (2017), IEEE, pp. 297–300.

[125] MENZIES, S. W., GUTENEV, A., AVRAMIDIS, M., BATRAC, A., AND MCCARTHY, W. H. Short-term digital surface microscopic monitoring of atypical or changing melanocytic lesions. *Archives of Dermatology 137*, 12 (2001), 1583–1589.

[126] MICHALSKI, R. S., CARBONELL, J. G., AND MITCHELL, T. M. *Machine learning: an artificial intelligence approach*. Springer Science & Business Media, 2013.

[127] MILLER, J. F., AND SMITH, S. L. Redundancy and computational efficiency in cartesian genetic programming. *IEEE Transactions on Evolutionary Computation 10*, 2 (2006), 167–174.

[128] MISHRA, N. K., AND CELEBI, M. E. An overview of melanoma detection in dermoscopy images using image processing and machine learning. *arXiv preprint arXiv:1601.07843* (2016).

[129] MITCHELL, T. M. *Machine Learning*, 1st ed. McGraw-Hill, Inc., 1997.

[130] MOLINARO, A. M., SIMON, R., AND PFEIFFER, R. M. Prediction error estimation: A comparison of resampling methods. *Bioinformatics 21*, 15 (2005), 3301–3307.

[131] MØLLERSEN, K. *Melanoma Detection: Colour, clustering and classification*. PhD thesis, Faculty of Science and Technology, Department of Mathematics and Statistics, The Arctic University of Norway, 2015.

[132] MØLLERSEN, K., KIRCHESCH, H. M., SCHOPF, T. G., AND GODTLIEBSEN, F. Unsupervised segmentation for digital dermoscopic images. *Skin Research and Technology 16*, 4 (2010), 401–407.

[133] MONTANA, D. J. Strongly typed genetic programming. *Evolutionary Computation 3*, 2 (1995), 199–230.

[134] MORRIS, T. *Computer vision and image processing*. Palgrave Macmillan, 2004.

[135] MUNI, D. P., PAL, N. R., AND DAS, J. A novel approach to design classifiers using genetic programming. *IEEE Transactions on Evolutionary Computation 8*, 2 (2004), 183–196.

[136] MUNI, D. P., PAL, N. R., AND DAS, J. Genetic programming for simultaneous feature selection and classifier design. *IEEE Transactions on Systems, Man, and Cybernetics, Part B 36*, 1 (2006), 106–117.

[137] MURPHY, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[138] NESHATIAN, K., ZHANG, M., AND ANDREAE, P. A filter approach to multiple feature construction for symbolic learning classifiers using genetic programming. *IEEE Transactions on Evolutionary Computation 16*, 5 (2012), 645–661.

[139] NESHATIAN, K., ZHANG, M., AND JOHNSTON, M. Feature construction and dimension reduction using genetic programming. In *Proceedings of the Australasian Joint Conference on Artificial Intelligence* (2007), Springer, pp. 160–170.

[140] NG, V. T., FUNG, B. Y., AND LEE, T. K. Determining the asymmetry of skin lesion with fuzzy borders. *Computers in Biology And Medicine 35*, 2 (2005), 103–120.

[141] NGUYEN, B. H., XUE, B., ZHANG, M., AND ANDREAE, P. Population-based ensemble classifier induction for domain adaptation. In *Proceedings of the Genetic and Evolutionary Computation Conference* (2019), pp. 437–445.

[142] NIXON, M., AND AGUADO, A. S. *Feature Extraction & Image Processing for Computer Vision*, 3rd ed. Academic Press, 2012.

[143] OECHSLE, O., AND CLARK, A. F. Feature extraction and classification by genetic programming. In *Proceedings of the International Conference on Computer Vision Systems* (2008), Springer, pp. 131–140.

[144] OF HEALTH, M. Melanoma facts. Tech. Rep. 1, Ministry of Health and the New Zealand Guidelines Group, Wellington: Ministry of Health, March 2017.

[145] OJALA, T., PIETIKÄINEN, M., AND HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition 29*, 1 (1996), 51–59.

[146] OLAGUE, G., AND TRUJILLO, L. A genetic programming approach to the design of interest point operators. In *Bio-inspired Hybrid Intelligent Systems for Image Analysis and Pattern Recognition*. Springer, 2009, pp. 49–65.

[147] OLAGUE, G., AND TRUJILLO, L. Evolutionary-computer-assisted design of image operators that detect interest points using genetic programming. *Image and Vision Computing 29*, 7 (2011), 484–498.

[148] OLAGUE, G., AND TRUJILLO, L. Interest point detection through multiobjective genetic programming. *Applied Soft Computing 12*, 8 (2012), 2566–2582.

[149] OLTEAN, G., IVANCIU, L., AND BALEA, H. Pedestrian detection and behaviour characterization for video surveillance systems. In *2019 IEEE 25th International Symposium for Design and Technology in Electronic Packaging (SIITME)* (2019), IEEE, pp. 256–259.

[150] OLTEAN, M., AND DUMITRESCU, D. Multi expression programming. Tech. rep., Department of Computer Science, Babes-Bolyai University, 2006.

[151] PAN, S. J., AND YANG, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering 22*, 10 (2010), 1345–1359.

[152] PATIÑO, D., CEBALLOS-ARROYO, A. M., RODRIGUEZ-RODRIGUEZ, J. A., SANCHEZ-TORRES, G., AND BRANCH-BEDOYA, J. W. 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* (2020), vol. 11330, International Society for Optics and Photonics, p. 1133018.

[153] PATTERSON, G., AND ZHANG, M. Fitness functions in genetic programming for classification with unbalanced data. In *Proceedings of the 2007 Australasian Joint Conference on Artificial Intelligence* (2007), Springer, pp. 769–775.

[154] PENNISI, A., BLOISI, D. D., NARDI, D., GIAMPETRUZZI, A. R., MONDINO, C., AND FACCHIANO, A. Melanoma detection using delaunay triangulation. In *Proceedings of the 27th International Conference on Tools with Artificial Intelligence* (2015), IEEE, pp. 791–798.

[155] PEREZ, C. B., AND OLAGUE, G. Evolutionary learning of local descriptor operators for object recognition. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation* (2009), ACM, pp. 1051–1058.

[156] PEREZ, F., VASCONCELOS, C., AVILA, S., AND VALLE, E. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018, pp. 303–311.

[157] PICCOLO, D., FERRARI, A., PERIS, K., DAIDONE, R., RUGGERI, B., AND CHIMENTI, S. Dermoscopic diagnosis by a trained clinician vs. a clinician with minimal dermoscopy training vs. computer-aided diagnosis of 341 pigmented skin lesions: a comparative study. *British Journal of Dermatology 147*, 3 (2002), 481–486.

[158] POLI, R. Genetic programming for feature detection and image segmentation. In *Artificial Intelligence and Simulation of Behaviour Workshop on Evolutionary Computing* (1996), Springer, pp. 110–125.

[159] POLI, R. Genetic programming for image analysis. In *Proceedings of the 1st Annual Conference on Genetic Programming* (1996), MIT Press, pp. 363–368.

[160] POLI, R., LANGDON, W. B., MCPHEE, N. F., AND KOZA, J. R. *A field guide to genetic programming*. Lulu, 2008.

[161] PRADHAN, B., JEBUR, M. N., SHAFRI, H. Z. M., AND TEHRANY, M. S. Data fusion technique using wavelet transform and taguchi methods for automatic landslide detection from airborne laser scanning data and quickbird satellite imagery. *IEEE Transactions on Geoscience and remote sensing 54*, 3 (2015), 1610–1622.

[162] RAJPAR, S., AND MARSDEN, J. *ABC of skin cancer*, vol. 94. John Wiley & Sons, 2009.

[163] RAJPARA, S. M., BOTELLO, A. P., TOWNEND, J., AND ORMEROD, A. D. Systematic review of dermoscopy and digital dermoscopy/artificial intelligence for the diagnosis of melanoma. *British Journal of Dermatology 161*, 3 (2009), 591–604.

[164] RAO, D. H., AND PANDURANGA, P. P. A survey on image enhancement techniques: classical spatial filter, neural network, cellular neural network, and fuzzy filter. In *Proceedings of the International Conference on Industrial Technology* (2006), IEEE, pp. 2821–2826.

[165] RATLE, A., AND SEBAG, M. Grammar-guided genetic programming and dimensional consistency: application to non-parametric identification in mechanics. *Applied Soft Computing 1*, 1 (2001), 105–118.

[166] RUBERTO, S., TERRAGNI, V., AND MOORE, J. H. Image feature learning with a genetic programming autoencoder. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion* (2020), pp. 245–246.

[167] RUELA, M., BARATA, C., MENDONÇA, T., AND MARQUES, J. S. What is the role of color in dermoscopy analysis? In *Iberian Conference on Pattern Recognition and Image Analysis* (2013), Springer, pp. 819–826.

[168] RUSSELL, S. J., AND NORVIG, P. *Artificial intelligence: a modern approach.* Prentice-Hall, Inc., 1995.

[169] RYAN, C., KRAWIEC, K., O'REILLY, U., FITZGERALD, J., AND MEDERNACH, D. Building a stage 1 computer aided detector for breast cancer using genetic programming. In *Proceedings of the European Conference on Genetic Programming* (2014), Springer, pp. 162–173.

[170] SABOURI, P. *Melanoma Detection Using Image Processing and Computer Vision Algorithms*. PhD thesis, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, 2016.

[171] SABOURI, P., AND GHOLAMHOSSEINI, H. Lesion border detection using deep learning. In *Proceedings of the 2016 IEEE Congress on Evolutionary Computation* (July 2016), pp. 1416–1421.

[172] SATHEESHA, T., SATYANARAYANA, D., PRASAD, M. G., AND DHRUVE, K. D. Melanoma is skin deep: A 3d reconstruction technique for computerized dermoscopic skin lesion classification. *IEEE Journal of Translational Engineering in Health and Medicine 5* (2017), 1–17.

[173] SAUSVILLE, E. A., AND LONGO, D. L. Principles of cancer treatment: surgery, chemotherapy, and biologic therapy. *Harrisons Principles of Internal Medicine 16*, 1 (2005), 464.

[174] SEGARAN, T. *Programming collective intelligence.* O'Reilly Media, Inc., 2007.

[175] SHAO, L., LIU, L., AND LI, X. Feature learning for image classification via multiobjective genetic programming. *IEEE Transactions on Neural Networks and Learning Systems 25*, 7 (2014), 1359–1371.

[176] SHAPIRO, L. G., AND STOCKMAN, G. C. Computer vision: Theory and applications, 2001.

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

[178] Shi, J., Zheng, X., Wu, J., Gong, B., Zhang, Q., and Ying, S. Quaternion grassmann average network for learning representation of histopathological image. *Pattern Recognition 89* (2019), 67–76.

[179] Shimizu, K., Iyatomi, H., Celebi, M. E., Norton, K.-A., and Tanaka, M. Four-class classification of skin lesions with task decomposition strategy. *IEEE Transactions on Biomedical Engineering 62*, 1 (2015), 274–283.

[180] Siegel, R., Naishadham, D., and Jemal, A. Cancer statistics, 2013. *A Cancer Journal for Clinicians 63*, 1 (2013), 11–30.

[181] Siegel, R. L., Miller, K. D., and Jemal, A. Cancer statistics, 2016. *A Cancer Journal for Clinicians 66*, 1 (2016), 7–30.

[182] Siegel, R. L., Miller, K. D., and Jemal, A. Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians 69*, 1 (2019), 7–34.

[183] Singh, T., Kharma, N., Daoud, M., and Ward, R. Genetic programming based image segmentation with applications to biomedical object detection. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation* (2009), ACM, pp. 1123–1130.

[184] Singh, T., Kharma, N., Daoud, M., and Ward, R. Genetic programming based image segmentation with applications to biomedical object detection. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation* (2009), pp. 1123–1130.

[185] SMART, W., AND ZHANG, M. Classification strategies for image classification in genetic programming. In *Proceedings of the International Conference on Image and Vision Computing New Zealand* (2003), IEEE, pp. 402–407.

[186] SMART, W., AND ZHANG, M. Probability based genetic programming for multiclass object classification. In *PRICAI* (2004), Springer, pp. 251–261.

[187] SOLOMON, C., AND BRECKON, T. *Fundamentals of Digital image processing: A practical approach with examples in Matlab*. John Wiley & Sons, 2011.

[188] SONG, A., AND CIESIELSKI, V. Texture analysis by genetic programming. In *2004 IEEE Congress on Evolutionary Computation* (2004), vol. 2, IEEE, pp. 2092–2099.

[189] SPANHOL, F. A., OLIVEIRA, L. S., PETITJEAN, C., AND HEUTTE, L. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering 63*, 7 (2016), 1455–1462.

[190] STEWART, B. W., AND WILD, C. P. World cancer report 2014. *Health* (2017).

[191] STOECKER, W. V. *Computer applications in dermatology*. Igaku-Shoin New York, 1993.

[192] STOECKER, W. V., LI, W. W., AND MOSS, R. H. Automatic detection of asymmetry in skin tumors. *Computerized Medical Imaging and Graphics 16*, 3 (1992), 191–197.

[193] STOLZ, W., RIEMANN, A., COGNETTA, A. B., PILLET, L., ABMAYR, W., HOLZEL, D., BILEK, P., NACHBAR, F., AND LANDTHALER, M. ABCD rule of dermatoscopy: a new practical method for early

recognition of malignant-melanoma. *European Journal of Dermatology 4*, 7 (1994), 521–527.

[194] SUTTON, R. S. Learning to predict by the methods of temporal differences. *Machine learning 3*, 1 (1988), 9–44.

[195] TACKETT, W. A. Genetic programming for feature discovery and image discrimination. In *Proceedings of the 5th International Conference on Genetic Algorithms* (1993), pp. 303–311.

[196] TRAN, B., XUE, B., AND ZHANG, M. Genetic programming for feature construction and selection in classification on high-dimensional data. *Memetic Computing 8*, 1 (2015), 3–15.

[197] TRAN, B., XUE, B., AND ZHANG, M. Class dependent multiple feature construction using genetic programming for high-dimensional data. In *Australasian Joint Conference on Artificial Intelligence* (2017), Springer, pp. 182–194.

[198] TRAN, B., ZHANG, M., AND XUE, B. Multiple feature construction in classification on high-dimensional data using GP. In *Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence* (Dec 2016), pp. 1–8.

[199] TRAN, C. T., ZHANG, M., XUE, B., AND ANDREAE, P. Genetic programming with interval functions and ensemble learning for classification with incomplete data. In *Australasian Joint Conference on Artificial Intelligence* (2018), Springer, pp. 577–589.

[200] ULUSOY, I., AND BISHOP, C. M. Comparison of generative and discriminative techniques for object detection and classification. In *Toward Category-Level Object Recognition*. Springer, 2006, pp. 173–195.

[201] VALLE, E., FORNACIALI, M., MENEGOLA, A., TAVARES, J., BITTENCOURT, F. V., LI, L. T., AND AVILA, S. Data, depth, and design:

Learning reliable models for melanoma screening. *arXiv preprint arXiv:1711.00441* (2017).

[202] WASSERMAN, P. D., AND SCHWARTZ, T. Neural networks, II. what are they and why is everybody so interested in them now? *IEEE Expert 3*, 1 (1988), 10–15.

[203] WHIGHAM, P. A., AND DICK, G. Implicitly controlling bloat in genetic programming. *IEEE Transactions on Evolutionary Computation 14*, 2 (2009), 173–190.

[204] WHITEMAN, D. C., GREEN, A. C., AND OLSEN, C. M. The growing burden of invasive melanoma: projections of incidence rates and numbers of new cases in six susceptible populations through 2031. *Journal of Investigative Dermatology 136*, 6 (2016), 1161–1171.

[205] WILSON, G., AND BANZHAF, W. A comparison of cartesian genetic programming and linear genetic programming. In *European Conference on Genetic Programming* (2008), Springer, pp. 182–193.

[206] WITTEN, I. H., FRANK, E., HALL, M. A., AND PAL, C. J. *Data Mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, 2005.

[207] WORZEL, W. P., YU, J., ALMAL, A. A., AND CHINNAIYAN, A. M. Applications of genetic programming in cancer research. *The international journal of biochemistry & cell biology 41*, 2 (2009), 405–413.

[208] XIE, F., FAN, H., LI, Y., JIANG, Z., MENG, R., AND BOVIK, A. Melanoma classification on dermoscopy images using a neural network ensemble model. *IEEE Transactions on Medical Imaging 36*, 3 (2017), 849–858.

[209] XUE, B., ZHANG, M., BROWNE, W. N., AND YAO, X. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation 20*, 4 (2016), 606–626.

[210] YU, L., CHEN, H., DOU, Q., QIN, J., AND HENG, P.-A. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging 36*, 4 (2017), 994–1004.

[211] ZHANG, M., AND BHOWAN, U. Pixel statistics and program size in genetic programming for object detection. *Technical Report CS-TR-04-3, Computer Science, Victoria University of Wellington, New Zealand* (2004).

[212] ZHANG, M., AND CIESIELSKI, V. Genetic programming for multiple class object detection. In *Proceedings of the 12th Australasian Joint Conference on Artificial Intelligence* (1999), Springer, pp. 180–192.

[213] ZHANG, M., CIESIELSKI, V. B., AND ANDREAE, P. A domain-independent window approach to multiclass object detection using genetic programming. *EURASIP Journal on Advances in Signal Processing 2003*, 8 (2003), 206791.

[214] ZHANG, Y., AND ZHANG, M. A multiple-output program tree structure in genetic programming. In *Proceedings of* (2004).

[215] ZHENG, S., YUILLE, A., AND TU, Z. Detecting object boundaries using low-, mid-, and high-level information. *Computer Vision and Image Understanding 114*, 10 (2010), 1055–1067.

[216] ZORTEA, M., SCHOPF, T. R., THON, K., GEILHUFE, M., HINDBERG, K., KIRCHESCH, H., MØLLERSEN, K., SCHULZ, J., SKRØVSETH, S. O., AND GODTLIEBSEN, F. Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists. *Artificial Intelligence in Medicine 60*, 1 (2014), 13–26.