Kennesaw State University

DigitalCommons@Kennesaw State University

Master of Science in Computer Science Theses

Department of Computer Science

Summer 7-22-2020

DATA MINING AND IMAGE CLASSIFICATION USING GENETIC PROGRAMMING

Mahsa Shokri varniab

Follow this and additional works at: https://digitalcommons.kennesaw.edu/cs_etd

Part of the Data Science Commons

Recommended Citation

Shokri varniab, Mahsa, "DATA MINING AND IMAGE CLASSIFICATION USING GENETIC PROGRAMMING" (2020). *Master of Science in Computer Science Theses*. 45. https://digitalcommons.kennesaw.edu/cs_etd/45

This Thesis is brought to you for free and open access by the Department of Computer Science at DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Master of Science in Computer Science Theses by an authorized administrator of DigitalCommons@Kennesaw State University. For more information, please contact digitalcommons@kennesaw.edu.

DATA MINING AND IMAGE CLASSIFICATION USING GENETIC PROGRAMMING

A Thesis Presented to The Faculty of the Computer Science Department

by

Mahsa Shokri Varniab

In Partial Fulfillment

of Requirements for the Degree

Master of Science, Computer Science

Kennesaw State University

July 2020

Approved:

Dr. Chih-Cheng Hung – Thesis Committee Chair

Dr. Edward Jung - Committee Member

Dr. Zhiling Long – Committee Member

Dr. Coskun Cetinkaya - Department Chair

Dr. Jon Preston - Dean

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Kennesaw State University, I agree that the university library shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish, this thesis may be granted by the professor under whose direction it was written, or, in his absence, by the dean of the appropriate school when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from or publication of, this thesis which involves potential financial gain will not be allowed without written permission.

Mahsa Shokri Varniab

Notice To Borrowers

Unpublished theses deposited in the Library of Kennesaw State University must be used only in accordance with the stipulations prescribed by the author in the preceding statement.

The author of this thesis is:

Mahsa Shokri Varniab 1100 S Marietta PKWY, Marietta, GA 30060

The director of this thesis is:

Dr. Chih-Cheng Hung 1100 S Marietta PKWY, Marietta, GA 30060

Users of this thesis not regularly enrolled as students at Kennesaw State University are required to attest acceptance of the preceding stipulations by signing below. Libraries borrowing this thesis for the use of their patrons are required to see that each user records here the information requested.

DATA MINING AND IMAGE CLASSIFICATION USING GENETIC PROGRAMMING

A Thesis Presented to The Faculty of the Computer Science Department

by

Mahsa Shokri Varniab

In Partial Fulfillment

of Requirements for the Degree

Master of Science, Computer Science

Kennesaw State University

July 2020

ABSTRACT

Genetic programming (GP), a capable machine learning and search method, motivated by Darwinian-evolution, is an evolutionary learning algorithm which automatically evolves computer programs in the form of trees to solve problems. This thesis studies the application of GP for data mining and image processing. Knowledge discovery and data mining have been widely used in business, healthcare, and scientific fields. In data mining, classification is supervised learning that identifies new patterns and maps the data to predefined targets. A GP based classifier is developed in order to perform these mappings. GP has been investigated in a series of studies to classify data; however, there are certain aspects which have not formerly been studied.

We propose an optimized GP classifier based on a combination of pruning subtrees and a new fitness function. An orthogonal least squares algorithm is also applied in the training phase to create a robust GP classifier. The proposed GP classifier is validated by 10-fold cross validation. Three areas were studied in this thesis. The first investigation resulted in an optimized genetic-programming-based classifier that directly solves multiclass classification problems. Instead of defining static thresholds as boundaries to differentiate between multiple labels, our work presents a method of classification where a GP system learns the relationships among experiential data and models them mathematically during the evolutionary process. Our approach has been assessed on six multiclass datasets. The second investigation was to develop a GP classifier to segment and detect brain tumors on magnetic resonance imaging (MRI) images. The findings indicated the high accuracy of brain tumor classification provided by our GP classifier. The results confirm the strong ability of the developed technique for complicated image classification problems. The third was to develop a hybrid system for multiclass imbalanced data classification using GP and SMOTE which was tested on satellite images. The finding showed that the proposed approach improves both training and test results when the SMOTE technique is incorporated. We compared our approach in terms of speed with previous GP algorithms as well. The analyzed results illustrate that the developed classifier produces a productive and rapid method for classification tasks that outperforms the previous methods for more challenging multiclass classification problems. We tested the approaches presented in this thesis on publicly available datasets, and images. The findings were statistically tested to conclude the robustness of the developed approaches.

DEDICATION

This thesis is dedicated to my parents, my husband, my brother, and my two sisters who have been my inspiration and strong supports. I love you all so very much and thank you for always being with me!

I thank everyone who supported me throughout my journey.

PREFACE

The thesis is submitted in partial fulfillment of the requirements for the Master of Science degree program at Kennesaw State University (KSU), Kennesaw.

The research project has been conducted under the supervision of Dr. Chih-Cheng Hung from the Center for Machine Vision and Security Research (CMVSR) in the College of Computing and Software Engineering in KSU, during the years 2018-2020. Financial support for the work has been provided in the form of scholarship from the Graduate College of Kennesaw State University through the Graduate Research Assistantship.

ACKNOWLEDGEMENTS

I would like to express thanks to Dr. Chih-Cheng Hung for his supervision and all his valuable advice, guidance, expertise, and support which helped me complete my thesis work through the entire process.

A special thank you to Dr. Vahid Khalilzad Sharghi for his ideas, technical assistance, thoughtful comments, encouragement, motivation, and support through this entire process to improve my work.

I would also like to thank Dr. Coskun Cetinkaya; Department Chair, for his support and help through my master's degree.

In addition, I would also like to thank Katherine Taylor and David Schmidt for their help to review my thesis and papers through my master's degree.

Thanks to my master supervisory committee members, Dr. Edward Jung, and Dr. Zhiling Long, for their their insightful comments and valuable suggestions.

I would like to thank all faculty, and staff in the Department of Computer Science who helped to make my experience here wonderful especially Dr. Hisham Haddad, Dr. Hassan Pournaghshband, Dr. Kun Suo, and Melinda Ross.

I would also like to acknowledge my labmates; Srivatsa Mallapragada, Wajira Priyadarshani, and Michael Wong for their assistance during my master program.

Publication Produced

1. M. Shokri, C. C. Hung, and V. Khalilzad "Classification of Multiclass Datasets Using Genetic Programming", RACS'19, September 24-27, 2019, Chongqing China.

2. M. Shokri, C. C. Hung, and V. Khalilzad "Data Mining and Image Analysis Using Genetic Programming", ACM-Applied Computing Review, winter issue (Vol. 19, No. 4).

3. M. Shokri, C. C. Hung, and V. Khalilzad "Classification of Imbalanced Data Using Genetic Programming with Application in Satellite Images", under preparation.

Table of Contents

CHAPT	CHAPTER 1 19	
Introduc	Introduction	
CHAPT	'ER 2	21
Literatu	re Survey	21
2.1.	Classification in machine learning	21
2.2.	Supervised learning	22
2.1.2.	Unsupervised learning	23
2.2. Cl	assification	23
2.3.	Evolutionary Algorithms	24
2.3.	Genetic Programming	25
2.3.	1. Initializing Population	28
2.3.2.	Fitness function	29
2.3.3.	Selection for Reproduction	29
2.3.4.	Genetic operators	29
2.3.5. Termination criteria		
2.4.	GP Application	31
2.4. 2.5.	GP Application GP related work	31 31
2.4. 2.5. 2.6.	GP Application GP related work Tumor detection on MRI image	31 31 33
2.4. 2.5. 2.6. 2.7.	GP Application GP related work Tumor detection on MRI image Imbalanced data problem	31 31 33 34
 2.4. 2.5. 2.6. 2.7. 2.8. 	GP Application GP related work Tumor detection on MRI image Imbalanced data problem Resampling techniques	31 31 33 34 35
 2.4. 2.5. 2.6. 2.7. 2.8. 2.8.1. 	GP Application GP related work Tumor detection on MRI image Imbalanced data problem Resampling techniques Under-sampling technique	31 31 33 34 35 36
 2.4. 2.5. 2.6. 2.7. 2.8. 2.8.1. 2.8.2. 	GP Application GP related work Tumor detection on MRI image Imbalanced data problem Resampling techniques Under-sampling technique Random under-sampling technique	 31 31 33 34 35 36 37
 2.4. 2.5. 2.6. 2.7. 2.8. 2.8.1. 2.8.2. 2.8.2. 	GP Application GP related work Tumor detection on MRI image Imbalanced data problem Resampling techniques Under-sampling technique Random under-sampling technique Over-sampling technique	 31 31 33 34 35 36 37 38
 2.4. 2.5. 2.6. 2.7. 2.8. 2.8.1. 2.8.2. 2.8.2. 2.8.2. 	GP Application GP related work Tumor detection on MRI image Imbalanced data problem Resampling techniques Under-sampling technique Random under-sampling technique Over-sampling technique Random over-sampling technique	 31 31 33 34 35 36 37 38 38
 2.4. 2.5. 2.6. 2.7. 2.8. 2.8.1. 2.8.2. 2.8.2. 2.8.2. 2.8.3. 	GP Application GP related work Tumor detection on MRI image Imbalanced data problem Resampling techniques Under-sampling technique Random under-sampling technique Over-sampling technique Random over-sampling technique SMOTE technique	 31 31 33 34 35 36 37 38 38 39
 2.4. 2.5. 2.6. 2.7. 2.8. 2.8.1. 2.8.2. 2.8.2. 2.8.2. 2.8.3. 2.9. A 	GP Application GP related work Tumor detection on MRI image Imbalanced data problem Resampling techniques Under-sampling technique Random under-sampling technique Over-sampling technique Random over-sampling technique SMOTE technique ccuracy measurements	 31 31 33 34 35 36 37 38 38 39 40
 2.4. 2.5. 2.6. 2.7. 2.8. 2.8.1. 2.8.2. 2.8.2. 2.8.2. 2.8.3. 2.9. A 2.10. I 	GP Application GP related work Tumor detection on MRI image Imbalanced data problem Resampling techniques Under-sampling technique Random under-sampling technique Over-sampling technique Random over-sampling technique SMOTE technique SMOTE technique ccuracy measurements mbalanced data problem related work	 31 31 33 34 35 36 37 38 39 40 43
 2.4. 2.5. 2.6. 2.7. 2.8. 2.8.1. 2.8.2. 2.8.2. 2.8.3. 2.9. A 2.10. I 2.11. F 	GP Application GP related work Tumor detection on MRI image Imbalanced data problem Resampling techniques Under-sampling technique Random under-sampling technique Over-sampling technique Random over-sampling technique SMOTE technique sMOTE technique ccuracy measurements mbalanced data problem related work Remote Sensing Images	 31 31 34 35 36 37 38 39 40 43 44
 2.4. 2.5. 2.6. 2.7. 2.8. 2.8.1. 2.8.2. 2.8.2. 2.8.3. 2.9. A 2.10. I 2.11. F 2.12. I 	GP Application GP related work Tumor detection on MRI image Imbalanced data problem Resampling techniques Under-sampling technique Random under-sampling technique Over-sampling technique Random over-sampling technique SMOTE technique SMOTE technique ccuracy measurements mbalanced data problem related work Remote Sensing Images Discussion and Analysis	 31 31 33 34 35 36 37 38 39 40 43 44 45

Proposed Algorithm	47
3.1. Classification	49
3.2. Genetic Programming Classifier	50
3.2.1. Fitness Function	51
3.2.2. Genetic programming with pruning subtrees and OLS	52
3.2.3 Genetic programming classifier structure using 10-fold cross validation	53
3.3. Multiclass dataset classification using GP classifier	54
3.4. Tumor detection using the proposed GP classifier	55
3.5. Proposed hybrid system for classifying imbalanced data	59
3.5.1. SMOTE resampling technique	60
3.5.2. Evaluation of the hybrid system on imbalanced satellite images	61
CHAPTER 4	62
Experimental Results and Analysis	62
3.1 Multiclass Datasets	62
3.1.1 Results of multiclass dataset classification	64
3.1.2 Evaluation and Comparison	67
3.2 MRI brain image with a tumor	69
3.2.1 Results of Tumor Detection	71
3.2.3 Evaluation of GP classifier's performance on MRI brain image compared with SVM classifier	73
3.3 Satellite Images	75
3.3.1 Results of plain GP classifier on imbalanced satellite images	77
3.3.2 Hybrid GP classifier equipped with SMOTE technique for resampling	81
3.3.3 Results of the hybrid GP classifier	84
3.3.4 Improvement made by the hybrid GP classifier versus the plain GP classifier	87
3.4 Evaluation of SVM classifier performance on satellite images before and after SMOTE technique	88
3.5 Comparison between the hybrid GP system and SVM-SMOTE classifier	91
CHAPTER 5	97
Conclusion and Future Work	
CHAPTER 6	, , , , , , , , , , , , , , , , , ,

References

List of Multimedia Objects

FIGURES

Figure 2-1. Two major machine learning categories.

Figure 2-2. Types of Supervised learning.

Figure 2-3. An example computer program for the numerical expression (F1+F2) / (F3*F4).

Figure 1-4. Five major steps in a GP algorithm.

Figure 2-5. A basic cycle of GP algorithms.

Figure 2-6. Operations of genetic operators in GP. (a) Mutation; (b) Crossover.

Figure 2-7. Distribution of an imbalanced dataset including minority and majority class.

Figure 2-8. Major resampling techniques discussed in this project.

Figure 2-9. Under-sampling technique.

Figure 2-10. Over-sampling technique.

Figure 2-11. Synthetic Minority Oversampling Technique.

Figure 2-12. Remote sensing.

Figure 3-1. The individual structure for GP classifier.

Figure 3-2. Tree pruning step is performed before the evaluation of fitness value.

Figure 3-3. The structure of GP classifier with 10- fold Cross-validation used to estimate error rates.

Figure 3-4. The structure of proposed tumor detection approach using the improved GP classifier.

Figure 3-5. Transforming an RGB image to a 2D matrix including features and the associated label.

Figure 3-6. Proposed methodology for a hybrid system using the combination of SMOTE and GP.

Figure 3-7. a) The distribution of imbalanced data with minority and majority class. b) Synthetic minority samples produced using SMOTE technique.

Figure 4-1. Scatter plot of Iris dataset classification using the GP classifier.

Figure 4-2. Scatter plot of Wine dataset classification using the GP Classifier.

Figure 4-3. Scatter plot of Glass dataset classification using the GP classifier.

Figure 4-4. Scatter plot of Pima dataset classification using the GP classifier.

Figure 4-5. The evaluation of GP classifier for 10 experiments.

Figure 4-6. The comparison of classification accuracies for each dataset using GP, DT, RF, and RF-SOM.

Figure 4-7. The comparison of classification accuracies for BUPA Liver and Balance Scale datasets using GP, DT, and SVM.

Figure 4-8. a) MRI brain image with a tumor (374×456). b) Mapped image using a custom colormap.

Figure 4-9. a) Cropped MRI image. b) Ground truth image used for training the GP classifier.

Figure 4-10. a) Labelled MRI brain image with a tumor (374×456) before applying a threshold. b) Labelled MRI brain image with a tumor (374×456) after applying a threshold.

Figure 4-11. The evaluation of GP classifier for the brain MRI image dataset classified in 10 experiments.

Figure 4-12. a) The original MRI brain image with a tumor (374×456). b) Labelled MRI brain image with a tumor (374×456) using the SVM classifier.

Figure 4-13. a) The original MRI brain image with a tumor (374×456). b) Labelled image using GP classifier. c) Labelled image using the SVM classifier.

Figure 4-14. (a), (b), (c), and (d) are the original satellite images with their corresponding ground truth images below them (e), (f), (g) and (h). Lyft, green and blue pixels represent village, forest, and river areas respectively.

Figure 4-15. (a), (b), (c), and (d) are the original satellite images. (e), (f), (g) and (h) depicted below them are their corresponding classified images using our plain GP classifier without implementation of SMOTE technique.

Figure 4-16. Evaluation of plain GP classifier on imbalanced satellite images in 10 experiments using 10-fold cross validation.

Figure 4-17. Distribution of the dataset before and after applying SMOTE technique for the satellite image (a).

Figure 4-18. Distribution of the dataset before and after applying SMOTE technique for the satellite image (b).

Figure 4-19. Distribution of the dataset before and after applying SMOTE technique for the satellite image (c).

Figure 4-20. Distribution of the dataset before and after applying SMOTE technique for the satellite image (d).

Figure 4-21. (a), (b), (c), and (d) are the original satellite images, and (e), (f), (g) and (h) are their corresponding classified images using our hybrid GP classifier. The results are significantly improved compared with the plain GP classifier.

Figure 4-22. Evaluation of the hybrid GP classifier on the satellite images in 10 experiments using 10-fold cross validation.

Figure 4-23. (a), (b), (c), and (d) are the original satellite images, and (e), (f), (g) and (h) are their corresponding classified images using SVM before balancing.

Figure 4-24. (a), (b), (c), and (d) are the original satellite images, and (e), (f), (g) and (h) are their corresponding classified images using SVM after SMOTE resampling technique.

Figure 4-25. Accuracy evaluation of SVM classifier on imbalanced satellite images in 10 experiments using 10-fold cross validation.

Figure 4-26. Accuracy evaluation of SVM classifier on balanced satellite images using SMOTE in 10 experiments using 10-fold cross validation.

Figure 4-27. The comparison between classification performance of the hybrid GP and SVM-SMOTE classifiers applied on image (a).

Figure 4-28. The comparison between classification performance of the hybrid GP and SVM-SMOTE classifiers applied on image (b).

Figure 4-29. The comparison between classification performance of the hybrid GP and SVM-SMOTE classifiers applied on image (c).

Figure 4-30. The comparison between classification performance of the hybrid GP and SVM-SMOTE classifiers applied on image (d).

Figure 4-31. The classification comparison for the hybrid GP, and SVM-SMOTE performed in 10 experiments and with 10-fold cross validation.

Figure 4-32. The evaluation of the hybrid GP and SVM-SMOTE classifiers in terms of G-mean measurement for each satellite image.

Figure 4-33. The evaluation of GP-hybrid and SVM-SMOTE classifier in terms of F1 score measurement for each satellite image.

TABLES

 Table 2-1. Confusion matrix.

Table 4-1. Six datasets used to evaluate the GP classifier.

 Table 4-2. Parameters used in the GP algorithm.

Table 4-3. Classification accuracies for Iris, Wine, Glass, Pima, BUPA Liver, and Balance Scale with 10- fold cross-validation.

Table 4-4. The table evaluation of GP Classifier for each dataset based on accuracy and standard deviation in 10 experiments.

Table 4-5. The accuracies of GP brain tumor classifier in 10 experiments.

Table 4-6. The average accuracy of GP brain tumor classifier calculated for 10-fold cross validation in 10 experiments.

Table 4-7. The comparison of classification accuracies for MRI brain image using GP, and SVM classifier using 10-fold cross validation in 10 experiments.

Table 4-8. Experimental satellite image datasets.

Table 4-9. Distribution and ratio of each class in satellite image datasets.

Table 4-10. Confusion matrix for imbalanced satellite image (a) classified by the plain GP classifier.

Table 4-11. Confusion matrix for imbalanced satellite image (b) classified by the plain GP classifier.

Table 4-12. Confusion matrix for imbalanced satellite image (c) classified by the plain GP classifier.

Table 4-13. Confusion matrix for imbalanced satellite image (b) classified by the plain GP classifier.

Table 4-14. Classification evaluation for the plain GP classifier including accuracy, G-mean, and F1score values calculated for the imbalanced satellite images.

Table 4-15. Confusion matrix for imbalanced satellite image (a) classified by the hybrid GP classifier.

Table 4-16. Confusion matrix for imbalanced satellite image (b) classified by the hybrid GP classifier.

Table 4-17. Confusion matrix for imbalanced satellite image (c) classified by the hybrid GP classifier.

Table 4-18. Confusion matrix for imbalanced satellite image (d) classified by the hybrid GP classifier.

Table 4-19. Classification evaluation for the hybrid GP classifier including accuracy, G-mean, and F1score values calculated for the imbalanced satellite images.

Table 4-20. Average accuracies, G-means, and F1scores calculated for both hybrid and plain GP classifiers.

Table 4-21. Evaluation of SVM performance by accuracy, G-mean, and F1score.

Table 4-22. Comparison between hybrid GP and SVM classifiers for the satellite images using mean accuracy, G-mean, and F1scor values.

CHAPTER 1

Introduction

In data mining, classification is supervised learning that labels the data based on predefined targets. The goal of the classification is to create a classifier for a set of instances with some features to predict their class membership using their properties [1].

Given the great quantity of data now being collected and stored in databases and clouds, there is a fast-growing demand for systems that can autonomously do the analysis and find valuable patterns in data for classification without operator intervention. On the other hand, modeling the data and building predictive models that can consistently and accurately classify the input data is challenging. In real world classification scenarios required tackling a tremendous number of learning instances with high dimensions and complicated relationships [2, 3].

Over the years, a series of methods have been introduced to solve data classification problems, comprising statistical and machine learning algorithms such as linear regression [4], logistics regression [5], decision tree[6], Bayesian [7], random forest [8], neural networks [9], KNN [10], SVM [11], FCM [12, 13], CNN [14], and RNN [15] to name a few. Evolutionary algorithms [16] such as genetic algorithms [17] and genetic programming algorithms [18] inspired by nature are also widely used.

The main objective of this thesis is to develop a genetic programming-based classification algorithm, and validate its performance in the domain of three types of data, including multiclass datasets, medical images, and satellite images and to investigate a hybrid system of GP and Synthetic Minority Over-sampling Technique (SMOTE) for multi-dimensional class-imbalanced data. Tied in with the primary objective previously stated, this thesis will provide a complete analysis of the related literature on GP and multiclass data classification. Three objectives were framed for this thesis and are as follows:

Objective 1: Incorporating genetic programming for multiclass classification. To advance genetic programming with a novel fitness function for multiclass dataset classification and employing a pruning subtree technique for improving the training phase. An orthogonal least squares algorithm is also applied in the training phase to create a robust GP classifier. The proposed approach will be applied on six multiclass datasets and compared against existing methods.

Objective 2: Identifying genetic programming representations for medical image analysis.

To identify the performance of the improved genetic programming in classification medical images, and to evaluate the role of the developed algorithm in brain tumor detection using magnetic resonance imaging scans.

Objective 3: Creating a hybrid system based on genetic programming and SMOTE. To propose and study and implement a hybrid method which will classify imbalanced multiclass datasets. The goal is to determine how the SMOTE technique can be employed in the training phase for creating a robust multiclass imbalanced data classifier. To evaluate the implemented hybrid genetic programming algorithm combined with SMOTE, an analysis will be conducted on multiclass imbalanced satellite images in which the features are extracted from the red, blue and green intensities of the pixels. The functionality of the hybrid system will be compared with other techniques.

CHAPTER 2

Literature Survey

This chapter provides the directly relevant works which preceding this project that are related to subcategories of our research and analyzes the details of the methods used in the proposed algorithms and their application. Each subcategory presents an overview of the related work and the concepts of algorithms and techniques used in our research. A detailed analysis of classification, evolutionary algorithms, genetic programming (GP), imbalanced datasets, synthetic minority oversampling technique (SMOTE), and evaluation measurements are provided in this chapter.

The novel system presented in our research is associated with the techniques and concepts introduced regrading improving GP, constructing novel features that are able to be incorporated by GP and SMOTE resampling approaches. The developed system is used for the classification of various multiclass datasets, as well as medical and remote sensing satellite images.

2.1. Classification in machine learning

Machine learning is a method of involving computers to perform tasks without being programmed in an explicit manner. In the development of a machine learning technique, datasets are observed to learn what patterns in datasets are to have better decision making in future. In another words, the major goal of machine learning is to enable computers to learn automatically without requiring human assistance or explicit programming instructions and to develop the knowledge to identify unknown patterns and generate predictive models from data. Among various machine learning methods, two major points of discussion in this work include supervised learning, and unsupervised learning as shown in Figure 2-1.



Figure 2-1. Two major machine learning categories.

2.2. Supervised learning

Supervised learning algorithms construct a mathematical model of a dataset comprising both the input data and the required outputs [19]. In supervised learning, a set of training data with well-labeled classes is used to indicate the correct answers, which is why we refer to this category as "learning with a teacher". To perform the learning phase, a training dataset with input features and output labels is provided to conduct the learning process. Algorithms used in classification and regression (Figure 2-2) [20] are categorized as supervised learning. The output in classification is discontinuous while in regression, the output is continuous.



Figure 2-2. Types of Supervised learning

2.1.2. Unsupervised learning

Unsupervised learning algorithms use a collection of data that only includes inputs and finds patterns among the instances. In unsupervised learning, the types of the variables of the dataset are similar. Therefore, we do not have a set of data with a recognized output and there is no teacher for the training. Unsupervised learning leads to discovering the inherent configuration, relations, or patterns existing in data.

Clustering and association discovery are examples of unsupervised learning tasks [21]. Clustering tasks categorize data into distinctive groups, singles out sets of data that are different from each other, and finds which groups' members are similar to one another. Association discovery is the identification of data values that frequently occur together in a given event or record. Association discovery rules are related to occurrence counts of the number of times items take place alone and in combination in the dataset [22].

2.2. Classification

Classification, supervised learning, is known as one of many effective data modeling and machine learning techniques [23]. An extensive range of problems in various domains can be solved by classification algorithms. For example, disease diagnosis [24], pattern recognition [25], document categorization [26], credit scoring [27], bankruptcy prediction [28], and software quality assessment [28], to name a few. A classification method uses a training set, including properly labeled data instances and a search algorithm, to create a classifier from the training set. To determine the excellence of the resulting classifier, a testing set, including a set of properly labeled data instances, is used. Different kinds of models such as decision trees [29] and random forest [30] have been used by researchers to represent classifiers.

2.3. Evolutionary Algorithms

There is a series of computational techniques for designing new classifiers such as linear classifiers, quadratic classifiers, k-nearest neighbor, K-means, Decision trees and Random Forest. K-means is a widely used unsupervised learning technique, which helps to divide n observations into k clusters; however, the weakness of the K-means algorithm is its need for knowing the number of groups or clusters [31]. This is a big challenge for data mining tasks because in practice, it is difficult to guess the number of clusters properly. In addition, most traditional machine learning algorithms perform a locally greedy search for data classification, and it is difficult to change or to extend their representations. Therefore, the need to develop an algorithm capable of determining answers to problems that are hard to solve without the help of an intelligent machine, results in emerging evolutionary algorithms (EA). EA enables a machine to generate solutions, free of human prejudices or biases, which are equivalent to, and often stronger than a solution developed by human beings [32].

2.3. Genetic Programming

GP is an evolutionary algorithm that utilizes concepts learned from biological evolution and finds answers to problems human beings may not know how to solve directly [33]. Each program in a GP algorithm is expressed as a chromosome in a population, and each chromosome contests for resources and existence, analogous to natural species contending for resources such as nutrition and dwelling. In a GP algorithm, only the most acceptable or near acceptable individuals remain, and they generate newborns in the hope that these newborns can survive [34]. The tree structure of an example computer program is shown in Figure 2-3. Five preliminary steps are taken by an analyst to link the human-level description of the problem to the GP algorithm. These well-defined steps are shown in Figure 2-4. The result of the GP algorithm is the best computer program that appears in the process of generations.

Different control parameters are used for running the GP system. For example, how large the population is, what the probabilities of crossover and mutation are, and how complex the generated programs are. Among them, the population size is the most significant control parameter and needs to be chosen in a way that generates a considerable number of generations within the acceptable processing time and complexity.



Figure 2-3. An example computer program for the numerical expression (F1+F2) / (F3*F4).



Figure 2-4. Five major steps in a GP algorithm.

The following steps describe the complete process of the GP system:

- i. First, a population is initialized.
- ii. The following steps are repeated until an end condition is fulfilled:
 - a. Individual programs are evaluated in the present population and a fitness is calculated for them.

- b. The successive tasks are performed in a loop until the next population is completely produced:
 - Select programs and run crossover and mutation operators on them in the current generation.
 - Place the product of the crossover and mutation operators into the new generation.
- iii. The most viable chromosome of the population is provided as the result of the GP system.

Figure 2-5 illustrates the basic cycle of GP algorithms.



Figure 2-5. A basic cycle of GP algorithms.

2.3.1. Initializing Population

GP provides solutions using programs or functions displayed as a tree consisting of primitive functions (internal nodes) and terminals (leaf nodes). The terminals include independent variables and constants, which are the inputs to the problem. These functions and terminals create a randomly initial population for GP. The user is assigned maximum depth for the initial individuals. Three major techniques are used for individual initialization including grow method, full method, ramped half and half.

- Grow Method

In this method, initial individuals are created by trees with various sizes and shapes. This method selects nodes from the entire primitive set including functions and terminals to reach a limited depth.

- Full Method

In this method, nodes are selected randomly from the function set to reach the maximum tree depth. In this method, the resultant tree is balanced because every branch of the tree continues to reach the full maximum tree depth.

- Ramped Half and Half

Since grow and full method do not create an extensive array of size and shape, in order to improve diversity, the ramped half and half technique is proposed. One half of the initial individuals are built up using the full method and the other half using grow. This method makes diverse individuals including balanced and unbalanced tree.

2.3.2. Fitness function

GP uses a fitness value which is the basic measure for associating the human-level description of the designer's goals to the GP algorithm and determines a desired target. The fitness value is used to compare one individual to another and to determine how fit an individual is [35, 36].

2.3.3. Selection for Reproduction

A selection mechanism is employed in GP to select an appropriate evolved program that will be utilized for crossover and mutation operators. The selected programs are employed to create new individuals for the following generation in the period of the evolutionary steps. There are many selection methods including Roulette Wheel Selection, Tournament Selection, Rank Selection, Elitism, etc. However, in this project, we used roulette wheel selection (fitness proportionate selection), which is the most commonly used selection method. The roulette method works similarly to a simple roulette, randomly rotating and stopping at a point. Every single individual possesses a sector of the roulette that links to its foreseen number of offspring.

2.3.4. Genetic operators

Diversification in the form of mutation and crossover are used for GP systems. Mutation analogous to biological mutation (Figure 2-6.a) is utilized to keep genetic diversity in the population. It can also adjust an evolved program by choosing the appropriate constants. Furthermore, mutation prevents the population of individuals from becoming very similar to each other and therefore creating local minima. Mutation is commonly performed in the form of swap, insert, delete, alter, point, uniform, non-uniform, etc. On the other hand, crossover, which is similar to sexual reproduction, happens between two parents as shown in Figure 2-6.b. Crossover recombines the selected parents to generate one, two or more children. Crossover is performed in the form of one-point, twopoint, n-point, uniform, and cut-and-splice.



Figure 2-6. Operations of genetic operators in GP. (a) Mutation; (b) Crossover.

2.3.5. Termination criteria

Termination criteria need to be defined to terminate the GP process when the result is satisfactory. Specific value of fitness function and how many generations the algorithm can proceed are examples of termination criteria. During the GP process, if the value of the fitness does not improve for a specific number of generations, the GP algorithm will stop the process and will pick the individual with the highest associated fitness value as the result.

2.4. GP Application

GP is being used as an automated development platform, a computer learning tool, and an advanced problem-solving engine with effectiveness. GP is particularly helpful in environments where the precise form of the approach is not planned in advance or an approximate solution is appropriate (maybe since it is so hard to locate the actual solution). Several GP 's applications include curve fitting, data processing, symbolic regression, collection of functions, and classification. John R. Koza [37] lists 76 cases where genetic engineering has worked successfully that are comparable with the effects created by humans (so-called human-based outcomes).

2.5. GP related work

GP has been extensively used to tackle classification problems due to its ability to determine primary data associations. Liu and Xu described GP as a reliable solution to detect and score top-ranked genes as the feature of the experimental data for classification purposes [38-41]. In previous studies, researchers applied GP-based techniques to analyze two-class microarray datasets. The traditional GP system involves evolving tree-based individuals. A tree can generate a binary solution for a classification; therefore, GP is an appropriate method for classifying two-class microarray datasets. Later, this technique was improved to classify multi-class microarray datasets. Liu and Xu showed that multiple-

class datasets can be treated as multiple two-class data instances, and a set of sub-group classifiers were utilized to tackle associated two-class data instances. By combining these groups, an individual is generated leading to solve a multiclass problem without the need for a new algorithm. However, this technique can be time consuming and was not tested on a wide range of challenging datasets to be completely verified. GP is also used in other applications such as feature construction [41].

Tahmasebei et al. have used a GP model to classify high activity regions in the limbic system of the fMRI data. The high dimensionality of fMRI data makes the classification task challenging. In their GP model, a crossover operator was used to select and replace the winner of the tournament with a stochastic subtree. Additionally, their algorithm used mutation to maintain the diversity of subtrees. The authors concluded that accuracy of their algorithm is better than typical machine learning algorithms due to the power of the GP method [42]. Despite the authors' preliminary success, this method was designed for a two-class dataset while GP is previously shown to be much more capable for multi-class problems.

In 2015, Al-Sahaf et al. employed GP for multiclass texture classification. In their method, a combination of raw pixel values as inputs and simple mathematical operators was used. The programs generated were used for initialization of a feature vector that was then grown into a nearest neighbor classifier to predict class labels. The performance of their proposed method was evaluated using multiclass datasets. Then, the results were compared with the performances of two GP-based and nine non-GP methods. The authors

reported high accuracies for their work. However, their algorithm was not tested on instances with rotation or with different dimensions [43].

2.6. Tumor detection on MRI image

The brain is truly the most important and complex organ in the human body; however, development of a brain tumor in the shape of abnormal brain cells could be the origin of numerus brain malfunctions. Neurologists categorize brain tumors into normal, malignant or benign types. Additionally, tumors can be studied in two categories of primary and secondary tumors. If an abnormal growth of brain cells is the origin of the tumor, the abnormal tissue is called a primary tumor. On the other hand, a tumor is called secondary if it originated from abnormal cells spreading from other tissues in the human body. Medical imaging techniques such as the Computed Tomography (CT) scan, Magnetic Resonance Imaging (MRI) [44-47] and Positron Emission Tomography (PET) are used for the early diagnosis of any brain tumor which is very important for successful treatment. Among them, MRI [45, 48] is a noninvasive technique that does not use the damaging ionizing radiation of X-rays or gamma-rays. Although MRI is very reliable to provide the location and size of tumors, there is still a need for a powerful and automated system to accurately diagnose and classify these tumors using MRI. The implementation of such a system will result in fewer human errors and lower medical expenses in poor remote areas [49-52].

2.7. Imbalanced data problem

Imbalanced datasets are a specific condition for classification problems where the class distribution among the classes is not uniform. In recent years, classification problems with imbalanced datasets have attracted attention. There are two types of classes in an imbalanced dataset - majority classes and minority classes. The distribution of imbalanced datasets is visualized in Figure 2-7. The classes with fewer samples, are called the minority, and the others are called majority classes. The small number of minority class instances cannot provide sufficient details to successfully classify both minority and majority classes. In real-world problems, machine learning algorithms have substantial challenges in the classification of datasets with imbalanced distribution because it is difficult to achieve high accuracy in the prediction of minority class due to this lack of information. Indeed, the effect of minority class in classification is not avoidable because it is results from the nature of the problem.

A solution to improve the classification performance of imbalanced datasets is to combine balancing methods with classification algorithms to achieve to a higher accuracy and efficient classification of the minority class along with the majority class [53]. One technique to deal with this issue is to use resampling methods by adding new samples to the dataset, removing existing samples, or a combination of two methods. There are various resampling techniques, so that choosing the appropriate method to deal with the problem is a key factor in solving an imbalanced dataset's classification problem.



Figure 2-7. Distribution of an imbalanced dataset including minority and majority class.

2.8. Resampling techniques

When thinking about Machine Learning and Data Science, we also consider a concept called Imbalanced Class Distribution, which typically occurs when the number of samples are either significantly higher or lower in one of the classes than the other one. Resampling is the simplest strategy to deal with class imbalances by changing class frequencies in a pre-processing phase to balance training data class distribution. This approach, therefore, does not require any change in an original learning algorithm [21].
There may be under-sampling, over-sampling, or both. The sample number can be selected empirically or in conjunction with its misclassification costs. The problem is that undersampling can exclude any useful data, and over-sampling can even contribute to overestimation. Most algorithms also mix under-sampling and over-sampling to benefit from all of them [22]. We will discuss the following resampling techniques as shown in Figure 2-8.



Figure 2-8. Major resampling techniques discussed in this project

2.8.1. Under-sampling technique

Under-sampling remove some of the instances of the majority class to match the number of the minority class. Therefore, the sample sizes of both classes become equal or in the same range. However, the major drawback of this method is that it can remove instances with valuable information which are useful for the learning process of the algorithm. Figure 3-9 shows the under-sampling method.



Imbalanced dataset

Balanced dataset after Under-sampling

Figure 2-9. Under-sampling technique.

2.8.2. Random under-sampling technique

There are different under-sampling methods available but random under-sampling is the simplest one. This under-sampling technique can manage unequal class distribution by random elimination of instances of the dominant class until the optimal equilibrium between the minority and majority classes is reached. This technique has two benefits: it is computationally inexpensive and it reduces in the classification model's learning time by eliminating the size of the training data. A limitation of under-sampling is that examples from the majority class are deleted that may be useful, important, or perhaps critical to fitting a robust decision boundary.

2.8.2. Over-sampling technique

Over-sampling technique increase the size of minority class by replicating some of the samples to match the size of majority class. Figure 2-10 shows the over-sampling technique.



Imbalanced dataset

Balanced dataset after Over-sampling

Figure 2-10. Over-sampling technique

2.8.2. Random over-sampling technique

Random oversampling method is the simplest and most common technique of oversampling which balance the class distribution by replicating randomly selected samples. The main drawback of this method is that it can cause overfitting because it replicates the original samples.

2.8.3. SMOTE technique

The implementation of resampling methods in imbalanced datasets consists of adjusting class data quantities to ensure a balanced class distribution. Chawla has suggested an efficient SMOTE over-sampling technique, a process called Synthetic Minority Oversampling Technique [54]. SMOTE is a method for oversampling the minority class to generate synthetic samples in the line segments which link k nearest minority class neighbours. Figure 2-11 shows the process of the SMOTE technique in which S₀ is one of minority samples considered to generate new artificial samples under it, S₁ to S₄ are the 4 nearest neighbours, and d₁ to d₄ are the synthetic samples created. Neighbours from the k nearest neighbourhood are randomly selected according to the sum of the over-sampling required. It is important to predefine parameter N that is the number of synthetic samples produced by the original minority case and parameter k for the nearest neighbour.

There are several steps to generate the synthetic new instances. First, the difference between minority instances is considered and its nearest neighbour is calculated. Then, the multiplication of this difference by a randomly selected number between 0 and 1 is added to the original instance considered to generate a random instance in the line segment between two different samples.



Figure 2-11. Synthetic Minority Oversampling Technique

Formula 2.1 shows the process of creating new synthetic instances d₁ based on the process of SMOTE technique shown in Figure 2-11.

$$d_{I} = S_{0} + (S_{0} - S_{I}) * \alpha \quad , \alpha \in [0, 1]$$
(2.1)

2.9. Accuracy measurements

Typically, the performance of machine learning algorithms is analyzed with confusion matrix. In the confusion matrix, TN is the right labeled number of negative examples (True Negative), FP is the number of incorrectly labeled negative examples (False Positives), FP is the number of incorrect examples classified as negative, and TP is the number of correctly categorized positive examples (True Positives). A confusion matrix is provided in Table 2-1.

	Predicted		
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FP)
	Negative	False Positive (FN)	True Negative (TN)

Table 2-2. Confusion matrix.

In order to evaluate the effectiveness of an algorithm, overall accuracy is used to calculate the classification accuracy. Formula 2.2 shows the overall accuracy. Since the classification accuracy of the majority class dominates the minority class accuracy in imbalanced datasets, overall accuracy is not an acceptable measurement to evaluate the algorithm. However, overall accuracy can be used to check the performance of an algorithm in the training phase and its general performance. Also, precision and recall measurements are used for the accuracy of information detection, and classification in a computer program. Precision is the fraction of related samples among the whole extracted samples shown in Formula 2.3. Precision measurement shows number of samples correctly classified as a minority. Recall is the fraction of related samples extracted over the total amount of related samples shown in Formula 2.4. Recall shows the number of samples of minority correctly classified.

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$
(2.2)

$$Precision = \frac{TP}{TP + FP}$$
(2.3)

$$Recall = \frac{TP}{TP + FN}$$
(2.4)

Traditional measures, such as overall accuracy, precision, and recall, do not include a valid measure to compare the efficiency of combinations of multiple sampling methods and classifiers. This is not suitable due to the natural imbalance problem [55]. Therefore, F1 measure and G-mean are used for evaluating the classification of imbalanced datasets.

Since overall accuracy is not enough measurement for evaluation of imbalanced data problems, the F1 score is used for assessing the classification algorithm. The F1 score is represented in Formula 2.5.

$$F1 - score = \frac{(1+B2).recall.precision}{B2.recall+precision}$$
(2.5)

Like precision and recall, a poor F1- score is 0.0 and a best or perfect F1- score is 1.0.

In G-mean value, the proportion of positive accuracy and negative accuracy is utilized. G-mean is an efficient measurement for imbalanced dataset problems because it evaluates the balance between classification effectiveness on the majority and minority classes. The best value is 1 and the worst value is 0. If a classifier has a high accuracy for all classes, it is considered as an efficient classifier. Therefore, a high G-mean shows a strong performance for a classifier and low G-mean represents a weak performance for a classifier. In imbalanced data problems, G-mean is considered as the most accepted attitude for evaluating the performance of a classifier. G-means uses the ratio of positive accuracy and negative accuracy. G-mean formula is represented in Formula 2.6.

$$G - mean = \sqrt{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}$$
(2.6)

2.10. Imbalanced data problem related work

Ling and Li [55] offered a particularly valuable reference to our work. They mixed the minority class over-sampling with the dominant class under-sampling. They used lift analysis to measure a classifier's performance, rather than precision. The majority class is under-sampled and the best lift measure collected, if the classes are evenly distributed, then the positive (marginalized) examples have been over-sampled to balance the number of negative (majority) examples to the number of positive ones. The combination of oversampling and under-sampling did not improve the lifted index significantly. Nevertheless, our over-sampling method varies from theirs.

Solberg [56] brought the matter of imbalanced data collections into consideration in the classification of SAR imagery oil slicks. Over-sampling and under-sampling methods were used to improve the detection of oil slicks. The study analyzed 42 oil slicks, and circulated 2'471 look-alikes, with an earlier chance of 0.98 for look-alikes.

The solution for Domingos [57] is close to our research as well. He applies the "meta cost" solution to every sub-sample majority and excess over-sample minority. He noticed the increases in meta prices, and sub-sampling is better than minority oversampling. Cost-sensitive classifiers built on mistakes. With each case, the likelihood for each class was determined, and the cases were relegated optimally with cost of misclassification. Reappointing examples increased the room for judgment, as new examples were generated to benefit from the classifier.

2.11. Remote Sensing Images

Remote sensing is a process for measuring emitted radiation at a remote distance to detect and monitor the physical characteristics of an area. Remote sensing images (RSI) are gathered by finding the energy reflected from earth's surface without physical contact. RSI are analyzed for pulling out the information related to the object. Remote sensors located either on satellite or aircraft are categorized into active and passive remote sensing sends the radiation to an object, then detects the radiation emitted from the object. The process of remote sensing of images is shown in Figure 2-13. Extracting useful information from RSI is a big challenge for image processing in different applications such as agriculture, military, geology, and atmospheric science. Image classification plays an important role in remote sensing images. Classification of images is performed based on certain features using different kinds of machine learning algorithms. Machine learning algorithms teach machines to make them intelligent. Then, the learned machine can automatically classify images.

Satellite images are significant means to extract useful information from remote sensing images for image processing in different applications such as agriculture, military,

geology, and atmospheric science. Image classification plays an important role in remote sensing images due to areas with a few numbers of pixels named minority class. Classification of images is performed based on certain features using different kinds of machine learning algorithms. Many machine learning algorithms are unable to classify RSI effectively. New technologies along with huge interest in collecting data in a rapid and extensive way attract companies and institutions to develop remote sensing further [58, 59].



Figure 2-12. Remote sensing [60]

2.12. Discussion and Analysis

There are many different classifiers available in classification tasks; among them GP proposes many advantages compared to other classifiers in classification applications. GP is a novel method to tackle a broad range of problems due to its flexibility and the fluency of computer program representation as well as the strong proficiencies of its evolutionary search. GP applications have shown a trend of success in recent years [61-66]. The main advantage of the GP algorithm is that it performs a global search for a model allowing evaluation of that model as a whole in the fitness function without focusing on the impact of each possible condition. Additionally, GP allows us to easily change or extend a representation.

The focus of this study is mainly to develop a technique based on GP for classifying datasets precisely without the previous knowledge of numbers of clusters. The developed technique uses a pruning algorithm to promote the accuracy and speed of classification. The resulting classifier is first applied on multiclass datasets then it is tailored for detecting brain tumors based on MRI images. Also, a hybrid system is proposed consisting of the combination of GP and SMOTE technique to enable GP to dominate some of its restrictions and to allow GP to handle classification problem more effectively for imbalanced multiclass datasets. Since GP demonstrates a bias toward the majority class instances, the hybrid system proposed in this study is designed to neutralise that bias. Our experiments on imbalanced remote sensing satellite images using the hybrid system confirm its strength in classification of imbalanced multiclass datasets compared with other techniques.

In this project, we use overall accuracy, mean, and standard deviation for evaluating the performance of our proposed GP classifier applied on multiclass datasets and MRI image data along with comparing the results with existing algorithms. Also, we use overall accuracy, G-mean and F1-score, standard deviation, and mean measurements for evaluating the proposed hybrid GP system for classification of imbalanced RSIs and compare them with the SVM-SMOTE classifier.

CHAPTER 3

Proposed Algorithm

In recent years, classification has become increasingly significant and is used in various aspects of applications including disease diagnosis, image processing, target recognition, and document categorization. There are various algorithms for classifying data into different categories according to some attributes including k-nearest neighbour classifier, SVM, ANN, Naive Bayes, and evolutionary algorithms [67, 68]. GP has also been employed well as a subcategory of evolutionary algorithms for classification of different types of datasets.

We present an optimized genetic-programming-based classifier that directly solves the multi-class classification problems in data mining and image analysis. A new fitness function is proposed for multiclass classification and brain tumor detection, which is validated by 10-fold cross validation. Instead of defining static thresholds as boundaries to differentiate between multiple labels, our work presents a method of classification in which a GP system learns the relationships among samples and models them mathematically during the evolutionary process. We propose an optimized GP classifier based on a combination of pruning subtrees and a new fitness function. An orthogonal least squares algorithm is also applied in the training phase to create a robust GP classifier.

In this research, three types of real-world classification scenarios are used to evaluate the performance of our proposed GP classifier in different applications. First, multiclass datasets collected from various sources in the real-world such as diverse kinds of plants, wines, and diseases were used to validate our developed GP classifier. The GP classifier was tested on Iris, Wine, Glass, Pima, BUPA Liver, and Balance Scale datasets. The results of the six classification problems demonstrated that this method performed very well even when applied on multiclass datasets with very small sample sizes.

Furthermore, brain tumor has been observed as a prevalent malignant disease among human beings, so it is significant to study this area. An MRI is commonly used by physicians to recognize a brain tumor. The correct detection of a tumor area on the MRI images is considered a critical task; therefore, machine learning algorithms assist to recognize tumors in MRI brain images. Therefore, the proposed GP classifier was applied on an MRI brain image for tumor detection. This preliminary experiment demonstrates that by using the features extracted from a mapped image, the GP classifier can provide a robust tumor detection performance. The results of data classification and tumor detection are compared with existing algorithms. The proposed method shows a promising capability in detecting the location of a tumor or a lesion and successfully segments the tumor from the brain tissue. The high accuracy of our GP approach for the classification of multiclass datasets and the brain tumor image confirms the strong ability of the developed technique for assessing complicated classification problems.

Finally, the developed GP classifier was applied on imbalanced remote sensing satellite images to investigate its capability in tackling imbalanced data problems. However, the developed technique shows a bias during performance toward the majority class in imbalanced remote sensing satellite images. Imbalanced data classification is a big challenge in classifying and analyzing remote sensing images (RSI), which aim to receive and process information from earth and its environment remotely. Remote sensing using

strong cameras installed on satellites or aircrafts helps to acquire valuable data about the Earth's surface. Such data is of significance for agriculture, military, geology, and atmospheric science, to name a few. This illustrates the significance of RSI classification, which is a big challenge due to the existence of minority classes such as rivers and roads in which we are interested. In this work, we investigated whether a SMOTE algorithm can be combined with the developed GP approach to successfully deal with imbalanced class distribution in RSI, which is a common drawback of most classification algorithms. The SMOTE resampling approach is combined with the GP algorithm to handle this problem by balancing the training datasets and therefore allow GP algorithm to evolve toward a stronger model. The final classifier is a hybrid system capable of multiclass imbalanced data classification using the combination of GP and the SMOTE technique. We evaluated our system by classifying four imbalanced remote sensing satellite images. For each of these RSIs using 10-fold cross validation, 10 models were developed, and the best one was selected as the outstanding hybrid GP classifier. The results of the satellite image classification were compared with the SVM algorithm. In addition, G-Mean and F-Score values were calculated for the hybrid classifier and SVM before and after SMOTE balancing method in order to compare the performance of both systems.

3.1. Classification

In data mining and machine learning, classification is a common method of creating a predictive model for experiential data. The concept of classification involves creating a model that partitions data into different classes. The model is created by determining a subset of data as the training part by which an algorithm is trained to label the classes. Then, the model is applied on a different dataset, called a test set, to predict the class of each member of the dataset using the model learned from training [69]. In the most cases, the problem uses supervised training in which a portion of a dataset labeled with the type of the class it belongs to is provided to the system.

3.2. Genetic Programming Classifier

In the past few years, researchers have presented a series of computational techniques for designing new classifiers, such as linear classifiers, quadratic classifiers, knearest neighbor, and decision trees. GP has also been employed because it can discover underlying data relationships [70, 71]. We propose an optimized GP classifier based on a combination of pruning subtrees and a new fitness function. An orthogonal least squares algorithm is also applied in the training phase to create a robust GP classifier. GP has several advantages compared with other algorithms for classification applications. First, GP can handle the raw form of the input data without the need for a preprocessing function in most situations while most classifiers require preprocessing of training data. The other advantage is the flexibility of GP. In other words, in a GP algorithm, a solution could be a combination of various functions including arithmetic, conditional, non-linear, and many other functions. Interpretation of the result is another factor that makes GP important. In addition, GP allows us to easily choose to change or extend a representation. This means in redesigning a GP classifier, all we need is a description of what a tree should look like and how to evaluate it.

A GP classifier, which uses a set of arithmetic and mathematical operators as well as conditional/logic operators, provides a mathematical equation as the solution to a classification problem. The individual structure for a GP classifier is shown in Figure 3-1.



Figure 3-1. The individual structure for a GP classifier.

3.2.1. Fitness Function

The fitness function in a GP algorithm represents the evolving quality of a possible solution that depends on the selection probability of the individual. Therefore, we designed a fitness function to guide the GP system to evolve towards a high performing classifier. The new fitness value designed for the n-th individual is shown in equation 3.1.

$$fitness = \frac{a.TCN}{N+b.FCN}$$
(3.1)

Where TCN is the true classification number, FCN is the false classification number, and N is the number of instances in the training set. The factors a and b allow the fitness measure to be adjusted to affect the individuals' sensitivities or specificities. In our algorithm, an individual could be evaluated using the fitness function to measure its fitness in evolving the programs toward the best model that forms the GP classifier. Then, if the value of fitness for an individual is high, it will be chosen. Also, if more than one individual has the same fitness value, the individual with fewer features will be the first one to be chosen.

3.2.2. Genetic programming with pruning subtrees and OLS

In this study, an improved GP algorithm that uses a pruning mechanism is used to perform the classification with higher speed and accuracy. In the process of the GP operation, the algorithm produces multiple possible tree-based solutions, which are the individual parts of the population. These trees are composed of subtrees with good or bad effects on the accuracy of the model. To improve the GP system, the tree structure is disintegrated to subtrees, and the errors of these subtrees are measured. Then the terms with the least importance are removed [72]. This tree pruning step is performed before the calculation of the fitness value of the tree as illustrated in Figure 3-2. The main purpose of the pruning approach is to simplify the trees and still maintain accuracies as close as possible to their original trees. An orthogonal least squares (OLS) algorithm is utilized to monitor the decomposition of the trees to keep the original structure of the trees as much as possible [73]. First, errors of the branches of the tree are calculated and the subtrees with

errors less than a threshold are eliminated with respect to the OLS algorithm. By using this technique, it is not necessary to rearrange the structure of the tree after pruning. Fitness is calculated in the next step, and if it is in the defined range, the associated individual is selected as the final model.



Figure 3-2. The tree pruning step is performed before the evaluation of fitness value.

3.2.3 Genetic programming classifier structure using 10-fold cross validation

Error rates were estimated using 10-fold cross validation as described in Figure 3-3. To estimate how accurately the GP classifier will perform in practice, each dataset is randomly partitioned into 10 folds of equal size subsets. The data in 9 folds are treated as the training set, and the remaining fold is used to estimate the error rate. The crossvalidation step is then redone 10 times, with each of the 10 subsets used once as the test dataset. The 10 results are then averaged to calculate the mean accuracy.



Figure 3-3. The structure of GP classifier with 10- fold cross-validation used to estimate error rates

3.3. Multiclass dataset classification using GP classifier

In the current study, six real world classification problems, including Iris, Wine, Glass, Pima, BUPA Liver, and Balance Scale datasets, are used to evaluate the performance of the GP classifier for multiclass datasets. The results of classification for these datasets are compared with other algorithms, including Decision Tree (DT), Random Forest (RF), and Random Forest with Self Organizing Map (RF-SOM) and Support Vector Machine (SVM) [74, 75].

Our GP classifier works in two steps, including training and testing. In this study, we used the 10- fold cross-validation method for the training and testing phases of classifying the datasets. The GP classifier developed in this work provides potential solutions to a classification problem in terms of computer programs consisting of terminal and function parts that evolve recursively. The function set used in our GP algorithm consists of the primary arithmetic operations (+, - , × , /), and the terminal set consists of the features of each dataset including F_1 , F_2 , ..., and F_k .

A custom-designed fitness function was used to select the best program in the training phase. Then, the best program created during the training phase is applied to classify the test dataset in order to analyze the accuracy of the GP. Furthermore, the pruning mechanism is applied in the training phase to remove insignificant terms of a generated program, which leads to increasing the speed of the GP algorithm and reducing the complexity of programs. The analyzed results illustrate that the developed classifier produces a productive and rapid method for classification tasks that outperforms the previous methods for more challenging multiclass classification problems.

3.4. Tumor detection using the proposed GP classifier

A typical anatomical MRI image is a 2D matrix of pixels with a range of possible values from 0 to 255 representing the brightness of each pixel. Generally, in such a grayscale image, 0 is assumed to be black, and 255 is taken to be white. As a preprocessing step, the grayscale MRI image is transformed to a colored image using a custom colormap. This preprocessing step is required to create red, blue, and green attributes for each pixel that will be used in the training phase of the GP system. The colormap used in this study is

a 2D matrix with 256 rows and 3 columns. In this matrix, each row includes red, green, and blue values in the range of [0, 1] allowing transformation of each gray value to an RGB color. This RGB mapping step creates three features, including red, blue, and green, and improves the pictorial contrast of MRI images. Then the mapped image is transformed to a two-dimensional dataset with four columns in which each row consists of R, G, B, and the ground truth label for the associated pixel. Then, our improved GP classifier is trained and validated against the same dataset using 10-fold cross-validation.

The MRI image is cropped into a smaller window of pixels around the tumor (177×177) used to form the dataset and to train the model. Cross-validation is performed by partitioning the cropped image into a training set to train the model and a test set to evaluate its accuracy. In our 10-fold cross-validation, pixels of the cropped image are randomly partitioned into 10 equally sized subsets. Of the 10 subsets, a single subset is held as the validation data for testing the model, and the remaining 9 subsets are used as training data. The cross-validation step is then redone 10 times. After creating an n-th GP model, it is validated using the n-th training subsets. The 10 results from the folds are used to judge whether a model is an acceptable model or not. The block diagram of our proposed approach for tumor detection using the improved GP classifier is illustrated in Figure 3-4. The high accuracy of brain tumor classification provided by our GP classifier confirms the strong ability of the developed technique for complicated classification problems.



Figure 3-4. The structure of the proposed tumor detection approach using the improved GP classifier.

3.5. Transforming RGB images into 2D datasets

An RGB image includes three 2-dimensional (2D) matrixes (Red, Green, and Blue). Figure 3-5 illustrates how to transform a 2D matrix into a one-dimensional (1D) matrix (a vector) and use it as a feature. For the mapped brain MRI image, the transformed dataset will have four columns. The first column includes red pixel intensities; the second column is comprised of green pixel intensities; the third column lists the blue color. Additionally, the fourth column is added, which includes the class of each pixel extracted from ground truth images.



Figure 3-5. Transforming an RGB image to a 2D matrix including features and the associated label.

3.5. Proposed hybrid system for classifying imbalanced data

Imbalanced class distribution in multiclass datasets makes solving classification problems very challenging. Most standard classifiers are not able to successfully deal with classifying imbalanced data; therefore, the minority class remains undetected. In another words, most classifiers have a bias to the majority group and overlook the minority group. Such a bias could be responsible for a poor minority classification accuracy rate while an outstanding majority classification is observed.

While the developed GP classifier shows a remarkable classification accuracy for balanced multiclass dataset and medical images, it needs improvement to perform efficiently for imbalanced data. To address the classification of imbalanced data, we combined the GP classifier designed in this work with a robust balancing technique named SMOTE (Synthetic Minority Oversampling Technique). The solution is a hybrid system that involves two parts. First, SMOTE is applied to the dataset in order to improve the minority class samples. The SMOTE technique produces new synthetic samples and adds them to the minority classes to make the balanced distribution of all classes in the training dataset. Finally, the balanced training dataset produced by SMOTE and partitioned by 10-fold cross validation is used in the training phase to generate a predictive model for classifying the dataset. The resultant hybrid system using the combination of GP and SMOTE techniques is our proposed classifier to handle imbalanced data. The structure of our proposed hybrid system is shown in Figure 3-6.



Figure 3-6. Proposed methodology for a hybrid system using the combination of SMOTE and GP

3.5.1. SMOTE resampling technique

The Synthetic Minority Oversampling Technique (SMOTE) is proposed to balance the dataset while avoiding the overfitting problem in the random oversampling technique. The SMOTE technique has been represented to be robust and widely used for handling imbalanced data problems in classification [76]. In the SMOTE technique, each minority class sample is taken to be oversampled from the k nearest neighbors of the sample, which are joined by a line ignoring nearby majority samples. This leads to enhancing the number of minority samples to be comparable with majority samples. Figure 3-7.a denotes the distribution of imbalanced data, including minority and majority instances. Figure 3-7.b shows how the SMOTE technique oversamples the minority class in the imbalanced dataset. The number of k nearest neighbours is randomly selected depending on the number of required overdamped instances. New oversampled data become like the original minority class because they are produced based on the features of the original dataset.



Figure 3-7. a) The distribution of imbalanced data with minority and majority classes.b) Synthetic minority samples produced using the SMOTE technique.

3.5.2. Evaluation of the hybrid system on imbalanced satellite images

Imbalanced data is a prevalent problem in remote sensing satellite images (RSI) because classification functionality is affected by imbalanced data. This thesis aims to deal with this problem by using a hybrid system that is implemented in two steps. The proposed solution is applied to the four imbalanced remote sensing satellite images. The proposed hybrid system is compared with SVM classifier and evaluated by calculation of both G-Mean and F-Score before and after incorporating the SMOTE method. The experimental results prove that the proposed hybrid system can efficiently solve the problem of imbalanced satellite images and improve classification proficiency.

CHAPTER 4

Experimental Results and Analysis

To evaluate the functionality of our proposed GP classifier, we conducted test with six multiclass datasets including Iris, Wine, Glass, Pima, Bupa Liver, and Balance Scale [77]. In addition, to illustrate the performance of the developed classifier in medical image analysis, we applied the GP classifier for tumor detection on an MRI brain image [78]. Also, we extended our experiments by applying the developed hybrid GP on imbalanced satellite images [79] in order to evaluate the effectiveness of the proposed technique on imbalanced data. MATLAB (MathWorks, Natick, MA) software is used to implement our algorithm, as it is one of the most recognized platforms for numerical and symbolic computing as well as simulation and model-based design.

3.1 Multiclass Datasets

In the current study, we carry out test with six datasets including Iris, Wine, Glass, Pima, Bupa Liver, and Balance Scale datasets as listed in Table 4-1. We used the 10- Fold cross validation method for the training and testing phases of the multiclass datasets experiments. Key parameters used in GP developed in this work are shown in Table 4-2. As it is mentioned before, a GP system produces a model during an evolutionary process in terms of computer programs consisting of two elements: terminals and functions. The primary arithmetic operations $(+, -, \times, /)$ are employed as the function set and the attributes of each class serve as the terminal set.

Datasets Name	No. Class	No. Features	Dataset Size	No. Each Class
Iris	3	4	150	50+50+50
Wine	3	13	178	59+71+48
Glass	6	9	214	70+76+17+ 13+9+29
Pima	2	8	768	500+268
BUPA Liver	2	6	350	145+200
Balance Scale	3	4	625	49+288+288

Table 4-1. Six datasets used to evaluate the GP classifier

 Table 4-2.
 Parameters used in the GP algorithm

Parameter	Value
Population Size	100
Selection Method	Roulette-wheel
Mutation Operator	Point Mutation
Crossover Operator	One-point Crossover
Proportion of Crossover	70
Proportion of Mutation	30

3.1.1 Results of multiclass dataset classification

For each dataset, the developed GP method was trained with 10-fold cross validation and 10 models were developed. Then the models were tested on the test datasets and the model with the highest accuracy was selected as the best model to form the GP classifier for that dataset. Table 4-3 shows the accuracies analysis of classification results with our GP system on Iris, Wine, Glass, Pima, BUPA Liver, and Balance Scale where the columns "Max", "Min" and "Mean" represent the maximum, minimum and average of the overall accuracies of 10 experiments for each dataset.

Dataset	Max Accuracy	Min Accuracy	Mean Accuracy
Iris	100	95.55	98.44±1.50
Wine	98.11	94.33	97.54±1.27
Glass	98.43	89.06	93.27±3.21
Pima	83.47	75.65	80.34±2.97
Bupa Liver	91.14	80	85.42±4.55
Balance Scale	98.38	93.54	96.124 ±2.176

Table 4-3. Classification accuracies for Iris, Wine, Glass, Pima, BUPA Liver, and Balance Scale with 10- fold cross-validation.

The classification results performed by our developed GP classifier for Iris, Wine, Glass, and Pima datasets are depicted in Figures 4-(1-4) respectively.



Figure 4-1. Scatter plot of Iris dataset classification using the GP classifier.



Figure 4-2. Scatter plot of Wine dataset classification using the GP Classifier.



Figure 4-3. Scatter plot of Glass dataset classification using the GP classifier.



Figure 4-4. Scatter plot of Pima dataset classification using the GP classifier.

3.1.2 Evaluation and Comparison

Average accuracies for 10 experiments are calculated for 6 multiclass datasets and shown in Figure 4-5. Table 4-4 lists the average accuracies and standard deviations of 10 experiments for all datasets. The accuracy of the GP classifier developed in this research was compared with those of Decision Tree, Random Forest and Random Forest with Self Organizing Map methods and the results are illustrated in Figure 4-6. The accuracy performance of our GP classifier on BUPA Liver and Balance Scale datasets are compared with GP, DT, and SVM methods based on the 10-fold cross validation method as shown in Figure 4-7 [36, 37].



Figure 4-5. The evaluation of GP classifier for 10 experiments.

Datasets	Accuracy (%)	Standard deviation
Iris	98.44	1.5
Wine	97.54	1.27
Glass	93.27	3.21
Pima	80.34	2.97
Bupa Liver	85.42	3.82
Balance Scale	96.124	2.17

Table 4-4. The table evaluation of GP Classifier for each dataset based on accuracy and standard deviation in 10 experiments.



Figure 4-6. The comparison of classification accuracies for each dataset using GP, DT, RF, and RF-SOM.



Figure 4-7. The comparison of classification accuracies for BUPA Liver and Balance Scale datasets using GP, DT, and SVM.

3.2 MRI brain image with a tumor

The developed GP classifier is used for automatic detection of tumors on MRI brain images. In the proposed approach, a grayscale MRI brain image is mapped into an RGB color image and then the RGB feature vectors are combined with ground truth labels to form the dataset used for training the GP classifier. We used an MRI brain image (374×456) with a defective area as shown in Figure 4-8.a. to illustrate the proposed tumor detection process. The mapped image using the custom colormap is shown in Figure 4-8.b. We used a cropped version of the original MRI image, including the pathological area shown in Figure 4-9.a, to form the training dataset. The ground truth image for training is shown in Figure 4-9.b.



Figure 4-8. a) MRI brain image with a tumor (374×456). b) Mapped image using a custom colormap.



Figure 4-9. a) Cropped MRI image. b) Ground truth image used for training the GP classifier.

3.2.1 Results of Tumor Detection

To illustrate the performance of the developed classifier in medical image analysis, we applied the GP classifier for tumor detection on an MRI brain image. Table 4-5 shows the accuracies of GP brain tumor classifier in 10 experiments. Table 4-6 lists maximum, minimum and average of the overall accuracies for the GP brain tumor classifier analyzed for 10-fold cross-validation in 10 experiments. Figure 4-10.a indicates the raw MRI brain image labeled from the GP classification process. Then a threshold value is used to categorize the classified data into two categories: the tumor and the remaining section (Figure 4-10. b). Using index labels, we can separate objects in the brain image dataset classified in 10 experiments is shown in Fig 4-11.



Figure 4-10. a) Labelled MRI brain image with a tumor (374×456) before applying a threshold. b) Labelled MRI brain image with a tumor (374×456) after applying a threshold.
Experiments	Accuracy
Experiment 1	93.12%
Experiment 2	94.49%
Experiment 3	95.53%
Experiment 4	94.49%
Experiment 5	95.21 %
Experiment 6	89.85%
Experiment 7	95.45 %
Experiment 8	95.7 %
Experiment 9	95.37%
Experiment 10	95.05%

Table 4-5. The accuracies of GP brain tumor classifier in 10experiments.

Table 4-6. The average accuracy of GP brain tumor classifier calculated for 10-fold cross validation in 10 experiments.

Dataset	Max Accuracy	Min Accuracy	Mean Accuracy
Brain Image	95.70%	89.85%	94.42% ± 1.77



Figure 4-11. The evaluation of GP classifier for the brain MRI image dataset classified in 10 experiments.

3.2.3 Evaluation of GP classifier's performance on MRI brain image compared with SVM classifier

We applied the SVM (Support Vector Machine) classifier [80, 81] on the MRI brain image in order to compare the performance of SVM algorithm in classification of MRI brain image with our GP classifier (Figure 4-12). The evaluation results of classification interns od Max, Min, and Mean accuracy in 10 experiments using 10-fold cross validation for both GP, and SVM classifier are illustrated in Table 4-7. Finally, the classified MRI brain image using the GP classifier compared with the SVM classifier are shown in Figure 4-13.



Figure 4-12. a) The original MRI brain image with a tumor (374×456). b) Labelled MRI brain image with a tumor (374×456) using the SVM classifier.

Table 4-7. The comparison of classification accuracies for MRI brain image using GP, and SVM classifier using 10-fold cross validation in 10 experiments.

Dataset	Max Accuracy	Min Accuracy	Mean Accuracy
Labelled MRI image using GP 95.7%		89.85%	94.42 ± 1.77
Labelled MRI image using SVM	91.32%	85.27%	89.57±1.91



Figure 4-13. a) The original MRI brain image with a tumor (374×456). b) Labelled image using GP classifier. c) Labelled image using the SVM classifier.

3.3 Satellite Images

The capability of effectively classifying imbalanced data is a critical role that a robust classifier should play. Therefore, to validate the effectiveness of the proposed hybrid GP system, imbalanced satellite images are used with their corresponding ground truth image as shown in in Figure 4-14. These satellite images include three regions consisting of forest, river, and village. In ground truth images, lyft pixels represent the village area, green pixels show the forest area and blue pixels are indication of the river area. A quick visual survey on these satellite images reveals that forest and river areas are considered as the majority classes while the village region is considered as the minority class. Table 4-8 describes the details of the satellite images in terms of dimension and the ratio of minority class (Formula 3.1) in the dataset. The class distribution and the number of pixels for both minority and majority classes in the images are shown in the Table 4-9.



Figure 4-14. (a), (b), (c), and (d) are the original satellite images with their corresponding ground truth images below them (e), (f), (g) and (h). Lyft, green and blue pixels represent village, forest and river areas respectively.

$$Imbalanced Ratio = \frac{\text{minority samples}}{\text{minority samples} + \text{majority samples}}$$
(3.1)

Table 4-8. Experi	nental satellite image datasets.

Images	Height	Width	Imbalanced Ratio
Image (a)	200	200	0.03
Image (b)	100	100	0.06
Image (c)	200	200	0.02
Image (d)	412	412	0.01

 Table 4-9. Distribution and ratio of each class in satellite image datasets.

		Forest	Village	River
Image(a)	Population	25166	1515	13319
	Ratio	0.63	0.03	0.34
Image(b)	Population	5427	584	3989
	Ratio	0.54	0.06	0.40
Image(c)	Population	13765	901	25334
	Ratio	0.35	0.02	0.63
Image(d)	Population	112433	2486	54825
	Ratio	0.67	0.01	0.32

3.3.1 Results of plain GP classifier on imbalanced satellite images

The performance of the plain GP classifier without implementation of SMOTE technique is validated on imbalanced satellite images. To discriminate our GP classifier from the developed hybrid GP classifier we use the term "plain GP classifier" instead of "GP classifier" from here on. The results of classification represented in the Figure 4-15 shows that there is a need for a hybrid system that can perform a better job in the minority regions. The results of are evaluated by calculation of mean accuracy, confusion matrix, G-means, and F1 score for all images. Tables 4-(10-13) show the confusion matrix assessed for classification of these images classified by the plain GP classifier (results for the hybrid GP classifier will be reported later). Tables 4.14 itemizes the accuracy, G-mean and F1 scores for classification of each image using the plain GP classifier. Figure 4-16 indicates the average and standard deviation of plain GP classification using 10-fold cross validation in 10 experiments for each image. The classification results propose that the plain GP classifier requires a major improvement to perform successfully in classification of minority classes in imbalanced multiclass datasets; therefore, we developed the hybrid GP classifier to solve this problem.



Figure 4-15. (a), (b), (c), and (d) are the original satellite images. (e), (f), (g) and (h) depicted below them are their corresponding classified images using our plain GP classifier without implementation of SMOTE technique.

Table 4-10.	Confusion matrix for imbalanced satellite image (a) classified
by the plain	GP classifier.

		Actual Class		
		Forest	Village	River
Predicted	Forest	24012	695	458
Class	Village	752	405	358
	River	326	738	12255

		Actual Class		
		Forest	Village	River
Predicted	Forest	4490	647	290
Class	Village	269	113	202
	River	88	368	3533

Table 4-11. Confusion matrix for imbalanced satellite image (b) classified by the plain GP classifier.

Table 4-12. Confusion matrix for imbalanced satellite image (c) classified by the plain GP classifier.

			Actual Class	
		Forest	Village	River
Predicted	Forest	10972	1893	899
class	Village	479	206	216
	River	150	215	24969

		Actual Class		
		Forest	Village	River
Predicted	Forest	105876	4863	1693
Class	Village	532	835	1119
	River	331	597	53897

Table 4-13. Confusion matrix for imbalanced satellite image (b) classified by the plain GP classifier.

Table 4-14. Classification evaluation for the plain GP classifier including accuracy, G-mean, and F1score values calculated for the imbalanced satellite images.

	Image(a)	Image(b)	Image(c)	Image(d)
Accuracy	91.68%	81.35%	90.38%	94.61%
G-mean	0.62	0.52	0.56	0.67
F1 Score	0.71	0.63	0.65	0.71



Figure 4-16. Evaluation of plain GP classifier on imbalanced satellite images in 10 experiments using 10-fold cross validation.

3.3.2 Hybrid GP classifier equipped with SMOTE technique for resampling

The evaluation results of our plain GP classifier based on the calculation of accuracy, G-mean, and F1 score on the imbalanced satellite images revealed that our technique needs to be modified for tackling imbalanced data. While the classification accuracy for the imbalanced data is high because of the correct classification of majority instances, the low G-means and F1 scores confirm poor classification of the minority classes. Therefore, our hybrid system employs SMOTE resampling technique to make the size of the minority class samples balanced with majority class samples. Figures 4-(17- 20) visualize the class distribution in the images before and after resampling using the SMOTE technique.



Figure 4-17. Distribution of the dataset before and after applying SMOTE technique for the satellite image (a).



Figure 4-18. Distribution of the dataset before and after applying SMOTE technique for the satellite image (b).



Figure 4-19. Distribution of the dataset before and after applying SMOTE technique for the satellite image (c).



Figure 4-20. Distribution of the dataset before and after applying SMOTE technique for the satellite image (d).

3.3.3 Results of the hybrid GP classifier

The functionality of our hybrid GP classifier is tested employing SMOTE technique with 4 nearest neighbours. Fig 4-21 shows the results of classification on the satellite images using the hybrid system. The results are evaluated by the confusion matrix, accuracy, G-means, and F1 score. Tables 4-(15-18) illustrate the confusion matrixes for classification of these images. Table 4-19 lists the accuracies, G-means and F1 scores for the classification of the satellite images using the hybrid system which are clearly improved compared with the results of the plain GP classifier described in section 3.3.2. Evaluation of GP classifier on imbalanced satellite images in 10 experiments using 10-fold cross validation is represented in Figure 4-21.



Figure 4-21. (a), (b), (c), and (d) are the original satellite images, and (e), (f), (g) and (h) are their corresponding classified images using our hybrid GP classifier. The results are significantly improved compared with the plain GP classifier.

		Actual Class		
		Forest	Village	River
D	Forest	24600	325	241
Predicted Class	Village	195	1204	116
	River	192	328	12799

Table 4-15. Confusion matrix for imbalanced satellite image (a) classified by the hybrid GP classifier.

Table 4-16. Confusion matrix for imbalanced satellite image (b) classified by the hybrid GP classifier.

		Actual Class		
		Forest	Village	River
Due di sta d	Forest	5020	283	124
Class	Village	83	442	59
	River	53	236	3700

Table 4-17. Confusion matrix for imbalanced satellite image (c) classified by the hybrid GP classifier

		Actual Class		
		Forest	Village	River
Dualistal	Forest	12530	893	342
Class	Village	173	601	127
	River	103	218	25013

		Actual Class		
		Forest	Village	River
Deviltated	Forest	109895	1625	913
Class	Village	331	1503	652
	River	198	297	54350

Table 4-18. Confusion matrix for imbalanced satellite image (d) classified by the hybrid GP classifier.

Table 4-19. Classification evaluation for the hybrid GP classifier including accuracy, G-mean, and F1score values calculated for the imbalanced satellite images.

	Image (a)	Image (b)	Image (c)	Image (d)
Accuracy	96.50%	91.12%	95.36%	97.64%
G-mean	0.813	0.851	0.912	0.931
F1 Score	0.912	0.895	0.932	0.952



Figure 4-22. Evaluation of the hybrid GP classifier on the satellite images in 10 experiments using 10-fold cross validation.

3.3.4 Improvement made by the hybrid GP classifier versus the plain GP classifier

Both hybrid GP classifier and the plain GP classifier are conducted on the classification of imbalanced satellite images. We compared average accuracy, G-mean, and F1 score values associated with their performances and reported the results in Table 4-20. The results confirm the superiority of the hybrid GP system over the plain GP classifier for the classification of imbalanced data.

		Accuracy	G-mean	F1 score
Image (a)	Plain GP classifier	91.68%	0.62	0.72
	Hybrid GP classifier	96.50%	0.82	0.91
Image (b)	Plain GP classifier	81.35%	0.52	0.63
Image (b)	Hybrid GP classifier	91.12%	0.85	0.90
Image (c)	Plain GP classifier	90.38%	0.56	0.65
	Hybrid GP classifier	95.36%	0.91	0.93
Image (d)	Plain GP classifier	94.61%	0.67	0.71
	Hybrid GP classifier	97.64%	0.93	0.95

Table 4-20. Average accuracies, G-means, and F1scores calculated for both hybrid and plain GP classifiers.

3.4 Evaluation of SVM classifier performance on satellite images before and after SMOTE technique

The performance of SVM (Support Vector Machine) algorithm in classification of imbalanced satellite images is shown in the Figure 4-25. The results of classification on the imbalanced satellite images after balancing using SMOTE are represented in Figure 4-26. Also, the results of SVM classifier on the images before and after balancing using SMOTE in 10 experiments including 10-fold cross validation are assessed by average accuracy and standard deviation and listed in Table 4-21.



Figure 4-23. (a), (b), (c), and (d) are the original satellite images, and (e), (f), (g) and (h) are their corresponding classified images using SVM before balancing.



Figure 4-24. (a), (b), (c), and (d) are the original satellite images, and (e), (f), (g) and (h) are their corresponding classified images using SVM after SMOTE resampling technique.

	Accuracy	G-mean	F1 score
Image (a) before balancing	91.42%	0.62	0.73
Image (a) after balancing	94.40%	0.65	0.78
Image (b) before balancing	79.14%	0.50	0.43
Image (b) after balancing	80.26%	0.59	0.52
Image (c) before balancing	85.98%	0.41	0.39
Image (c) after balancing	86.42%	0.57	0.50
Image (d) before balancing	92.78%	0.60	0.66
Image (d) after balancing	95.32%	0.74	0.72

Table 4-21. Evaluation of SVM performance by accuracy, G-mean, and F1score.



Figure 4-25. Accuracy evaluation of SVM classifier on imbalanced satellite images in 10 experiments using 10-fold cross validation.



Figure 4-26. Accuracy evaluation of SVM classifier on balanced satellite images using SMOTE in 10 experiments using 10-fold cross validation.

3.5 Comparison between the hybrid GP system and SVM-SMOTE classifier

In this section, the performance of our hybrid GP classifier on the imbalanced satellite images is compared with the SVM classifier which employs SMOTE (SVM-SMOTE). Table 4-22 itemizes the average accuracy, G-mean, and F1 score values for both hybrid GP and SVM classifiers. Additionally, the classified images by both hybrid GP and SVM-SMOTE classifiers are shown in Figures 4-(27-30). The results approve that our developed hybrid GP classifier provides higher accuracies as well as higher G-means and F1 score values on imbalanced satellite images. Since the GP algorithm performs a global search for a model enabling the algorithm to evolve with respect to satisfying the criteria of the fitness function, it needs a balanced training dataset to produce a strong minority-considering classifier. Our hybrid GP system benefits from the strength of SMOTE

approach to successfully classify both minority and majority classes in an imbalanced dataset.

			Average Accuracy	G- mean	F1 score
	Image (a)	GP classifier	91.68%	0.62	0.71
		SVM classifier	91.42%	0.62	0.73
	Image (b)	GP classifier	81.35%	0.52	0.63
Imbalanced		SVM classifier	79.14%	0.50	0.43
images	Image (c)	GP classifier	90.38%	0.56	0.65
		SVM classifier	85.98%	0.41	0.39
	Image (d)	GP classifier	94.61%	0.67	0.71
		SVM classifier	92.78%	0.69	0.66
	Image (a)	GP classifier	96.50%	0.81	0.91
Balanced images using SMOTE		SVM classifier	94.40%	0.65	0.78
	Image (b)	GP classifier	91.12%	0.85	0.89
		SVM classifier	80.26%	0.59	0.52
	Image (c)	GP classifier	95.36%	0.91	0.93
		SVM classifier	86.42%	0.57	0.50
	Image (d)	GP classifier	97.64%	0.93	0.95
		SVM classifier	95.32%	0.74	0.72

Table 4-22. Comparison between hybrid GP and SVM classifiers for the satellite images using mean accuracy, G-mean, and F1scor values.



Original image (a) Hybrid GP classifier SVM-SMOTE classifier

Figure 4-27. The comparison between classification performance of the hybrid GP and SVM-SMOTE classifiers applied on image (a).

Fig 4-27 shows that SVM can generally classify the satellite image but in some

areas, which are shown by blue arrows perform worse than the hybrid GP.



Original image (b)

Hybrid GP classifier SVM-SMOTE classifier

Fig 4-28 shows that SVM over-classified the minority pixels shown by blue arrows; therefore, the village areas are classified larger than the ground truth image. However, hybrid GP can detect minority pixels with a higher accuracy compared with the SVM-SMOTE.

Figure 4-28. The comparison between classification performance of the hybrid GP and SVM-SMOTE classifiers applied on image (b).



Original image (c) Hybrid GP classifier SVM-SMOTE classifier

Figure 4-29. The comparison between classification performance of the hybrid GP and SVM-SMOTE classifiers applied on image (c)..

Fig 4-29 shows that SVM-SMOTE cannot classify minority pixels in the image and misclassified the majority and minority pixels shown by blue arrows; therefore you cannot see the village areas in some parts of the image like the ground truth and the river is classified as forest in some parts. However, Hybrid GP can detect minority and majority pixels efficiently compared with SVM-SMOTE.



Original image (d)

Hybrid GP classifier

SVM-SMOTE classifier

Figure 4-30. The comparison between classification performance of the hybrid GP and SVM-SMOTE classifiers applied on image (d).

Figure 4-30 shows that SVM-SMOTE misclassified the majority and minority classes. Some parts of the river is classified as the village which is shown by blue arrows on the Figure 4-30. Additionally, SVM-SMOTE ignores the minority class in some parts indicated by blue arrows thus the village inside the forest is not identified correctly. On the other hand, hybrid GP can detect minority and majority pixels with a higher accuracy compared with SVM-SMOTE.

The evaluation results for both hybrid GP and SVM-SMOTE classifiers in terms of accuracy, G-mean, and F1 score on satellite images are represented in Figures 4-(31-33). The results confirm that the hybrid GP system performs with a higher accuracy and can detect minority classes in the imbalanced data more effectively based on the G-mean and F1 score values.



Figure 4-31. The classification comparison for the hybrid GP, and SVM-SMOTE performed in 10 experiments and with 10-fold cross validation.







Figure 4-33. The evaluation of GP-hybrid and SVM-SMOTE classifier in terms of F1 score measurement for each satellite image.

CHAPTER 5

Conclusion and Future Work

To conclude, this thesis explored the classification of multiclass datasets, medical images and imbalanced satellite pictures using a novel genetic programming (GP) system. A GP classifier that autonomously evolves feature equations in the form of trees for the classification of multiclass data has been developed and described. The proposed algorithm uses a pruning mechanism and a new fitness function to solve the classification problem for multiclass datasets. The pruning technique passes the orthogonal least squares (OLS) check in order to maintain the original tree-based structure to the extent that it is possible. This is necessary because the tree structure is an essential element of GP system. Our developed GP classifier was tested on Iris, Wine, Glass, Pima, BUPA Liver and Balance Scale datasets. The results of the six classification problems demonstrate that this method performed very well even when applied on datasets with very small sample sizes. This approach is compared with DT, RF, and RF-SOM for Iris, Wine, Glass, and Pima datasets. In addition, it is compared with DT and SVM for BUPA Liver and Balance Scale datasets with the 10- fold cross-validation. Furthermore, the GP classification method is used for detecting a tumor in MRI brain images. Classification accuracy for the brain tumor data was compared to that of the SVM method to validate the findings, as well as employing 10-fold CV. To extend our algorithm, we combined our modified GP classifier with the SMOTE approach for classification with multiclass unbalanced data and to develop a hybrid GP classifier to address limitations of classifying both minority and majority classes. Our results verified that the hybrid classifiers that evolved using the SMOTE-

balanced training dataset performed with a higher accuracy on the tasks compared to other techniques such as SVM.

Future research will include extending this study to improve the performance of object detection on image datasets, particularly for tumor type classification of MRI brain images. The improved GP classifier was not applied to classify tumor types, and thus future research will aim at developing a predictive model for brain tumor type classification. Furthermore, to reduce the training time of the developed GP algorithm, our method will be modified in order to run on parallel computing systems with a higher speed.

CHAPTER 6

References

- [1] L. Breiman, J. Friedman, R. Olshen, and C. J. L. Stone, "Classification and regression trees (monterey, ca: Wadsworth and brooks/cole)," 1984.
- [2] S. Kumar, Z. Khan, and A. J. I. J. o. A. C. R. Jain, "A review of content based image classification using machine learning approach," vol. 2, no. 3, p. 55, 2012.
- [3] M. Hossin, M. J. I. J. o. D. M. Sulaiman, and K. M. Process, "A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS," vol. 5, no. 2, p. 1, 2015.
- [4] X. Gao, Q. Sun, H. Xu, and J. J. E. S. w. A. Gao, "Sparse and collaborative representation based kernel pairwise linear regression for image set classification," vol. 140, p. 112886, 2020.
- [5] S. Dreiseitl and L. J. J. o. b. i. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," vol. 35, no. 5-6, pp. 352-359, 2002.
- V. Rodriguez-Galiano, M. Chica-Olmo, F. Abarca-Hernandez, P. M. Atkinson, and C. J. R.
 S. o. E. Jeganathan, "Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture," vol. 121, pp. 93-107, 2012.
- B. Tang, H. He, P. M. Baggenstoss, S. J. I. T. o. K. Kay, and D. Engineering, "A Bayesian classification approach using class-specific features for text categorization," vol. 28, no. 6, pp. 1602-1606, 2016.
- [8] M. A. Friedl and C. E. J. R. s. o. e. Brodley, "Decision tree classification of land cover from remotely sensed data," vol. 61, no. 3, pp. 399-409, 1997.
- [9] S.-H. Liao and C.-H. J. E. S. w. A. Wen, "Artificial neural networks classification and clustering of methodologies and applications–literature analysis from 1995 to 2005," vol. 32, no. 1, pp. 1-11, 2007.
- [10] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. J. N. Zhang, "Efficient kNN classification algorithm for big data," vol. 195, pp. 143-148, 2016.
- [11] K. S. Durgesh, B. J. J. o. t. Lekha, and a. i. technology, "Data classification using support vector machine," vol. 12, no. 1, pp. 1-7, 2010.
- [12] S. V. KHALILZAD, A. R. TALEBPOUR, A. KAMALIASL, and S. M. R. AGHAMIRI, "AUTOMATIC ASSESSMENT OF CARDIAC ARTERY DISEASE USING THE DCAD MODULE," 2009.
- [13] V. Khalilzad-Sharghi, A. Talebpour, A. Kamali-Asl, and N. Hendijani, "Automatic Assessment of Cardiac Artery Disease by Using DCAD Module," in 2008 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, 2008, pp. 243-246: IEEE.
- [14] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 2017, pp. 131-135: IEEE.
- [15] L. Mou, P. Ghamisi, X. X. J. I. T. o. G. Zhu, and R. Sensing, "Deep recurrent neural networks for hyperspectral image classification," vol. 55, no. 7, pp. 3639-3655, 2017.
- [16] T. F. Covões and E. R. Hruschka, "Classification with multi-modal classes using evolutionary algorithms and constrained clustering," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, 2018, pp. 1-8: IEEE.

- [17] J. D. Kelly Jr and L. Davis, "A Hybrid Genetic Algorithm for Classification," in *IJCAI*, 1991, vol. 91, pp. 645-650.
- [18] P. G. Espejo, S. Ventura, F. J. I. T. o. S. Herrera, Man, and P. C. Cybernetics, "A survey on the application of genetic programming to classification," vol. 40, no. 2, pp. 121-144, 2009.
- [19] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," 2002.
- [20] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.
- [21] A. M. Cruz, "Evaluating record history of medical devices using association discovery and clustering techniques," *Expert Systems with Applications*, vol. 40, no. 13, pp. 5292-5305, 2013.
- [22] R. Bhattacharyya, "Cohesion: A concept and framework for confident association discovery with potential application in microarray mining," *Applied Soft Computing*, vol. 11, no. 1, pp. 592-604, 2011.
- [23] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., 2011, p. 696.
- [24] C. Massilamany, V. Khalilzad-Sharghi, A. Gangaplara, D. Steffen, S. F. Othman, and J. J. J. Reddy, "Noninvasive assessment of cardiac abnormalities in experimental autoimmune myocarditis by magnetic resonance microscopy imaging in the mouse," no. 88, p. e51654, 2014.
- P. Sharma and M. Kaur, "Classification in pattern recognition: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 4, 2013.
- [26] L. Cai and T. Hofmann, "Hierarchical document categorization with support vector machines," presented at the Proceedings of the thirteenth ACM international conference on Information and knowledge management, Washington, D.C., 2004.
- [27] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3446-3453, 2012.
- [28] P. C. Pendharkar, "A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem," *Computers & Operations Research*, vol. 32, no. 10, pp. 2561-2582, 2005.
- [29] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification," *Machine Learning*, journal article vol. 73, no. 2, p. 185, August 01 2008.
- [30] J. Ham, C. Yangchi, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 492-501, 2005.
- [31] K. A. Nazeer and M. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm," in *Proceedings of the world congress on engineering*, 2009, vol. 1, pp. 1-3: Association of Engineers London.
- [32] C. A. C. Coello, G. B. Lamont, and D. A. Van Veldhuizen, *Evolutionary algorithms for solving multi-objective problems*. Springer, 2007.
- [33] M. S. Varniab, C.-C. Hung, and V. K. J. A. S. A. C. R. Sharghi, "Data mining and image analysis using genetic programming," vol. 19, no. 4, pp. 40-49, 2020.
- [34] J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection*. MIT Press, 1992, p. 680.

- [35] W. Banzhaf, J. R. Koza, C. Ryan, L. Spector, and C. Jacob, "Genetic programming," *IEEE Intelligent Systems and their Applications,* vol. 15, no. 3, pp. 74-84, 2000.
- [36] R. Poli, W. B. Langdon, and N. F. McPhee, *A Field Guide to Genetic Programming*. Lulu Enterprises, UK Ltd, 2008, p. 252.
- [37] J. R. Koza and J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection*. MIT press, 1992.
- [38] K.-H. Liu and C.-G. Xu, "A genetic programming-based approach to the classification of multiclass microarray datasets," *Bioinformatics*, vol. 25, no. 3, pp. 331-337, 2008.
- [39] M. Chabbouh, S. Bechikh, C.-C. Hung, and L. Ben Said, "Multi-objective evolution of oblique decision trees for imbalanced data binary classification," *Swarm and Evolutionary Computation*, vol. 49, pp. 1-22, 2019.
- [40] S. Bechikh, M. Elarbi, C.-C. Hung, S. Hamdi, and L. B. Said, "A Hybrid Evolutionary Algorithm with Heuristic Mutation for Multi-objective Bi-clustering," in *2019 IEEE Congress on Evolutionary Computation (CEC)*, 2019, pp. 2323-2330: IEEE.
- [41] M. Hammami, S. Bechikh, C.-C. Hung, and L. B. Said, "Weighted-Features Construction as a Bi-level Problem," in 2019 IEEE Congress on Evolutionary Computation (CEC), 2019, pp. 1604-1611: IEEE.
- [42] A. Tahmassebi *et al.*, "An evolutionary approach for fMRI big data classification," in 2017 *IEEE Congress on Evolutionary Computation (CEC)*, 2017, pp. 1029-1036.
- [43] H. Al-Sahaf, M. Zhang, M. Johnston, and B. Verma, "Image descriptor: A genetic programming approach to multiclass texture classification," in 2015 IEEE Congress on Evolutionary Computation (CEC), 2015, pp. 2460-2467.
- [44] V. Khalilzad Sharghi, Z. Han, H. Xu, and S. F. J. M. r. i. m. Othman, "MR elastography for evaluating regeneration of tissue - engineered cartilage in an ectopic mouse model," vol. 75, no. 3, pp. 1209-1217, 2016.
- [45] V. Khalilzad-Sharghi and H. J. A. o. b. e. Xu, "Design and fabrication of an MRIcompatible, autonomous incubation system," vol. 43, no. 10, pp. 2406-2415, 2015.
- [46] S. Othman, H. Xu, K. Wartella, V. Khalilzad-Sharghi, and I. Bargar, "MRI-compatible bioreactors and methods of using," ed: Google Patents, 2015.
- [47] V. Khalilzad Sharghi, K. Wartella, and S. J. B. E. S. Othman, "Monitoring bone repair in a mouse calvarial defect model using magnetic resonance imaging," 2014.
- [48] V. Khalilzad-Sharghi, J. Wood, W. Pohlmeier, J. Tart, R. Griess, and S. Othman, "MRE and QPCR indicate changes in liver phenotypes upon consumption of a high fat diet."
- [49] P. Afshar, A. Mohammadi, and K. N. Plataniotis, "Brain Tumor Type Classification via Capsule Networks," in *2018 25th IEEE International Conference on Image Processing* (*ICIP*), 2018, pp. 3129-3133.
- [50] V. Khalilzad Sharghi, K. Wartella, and S. F. Othman, "Monitoring bone repair in a mouse calvarial defect model using magnetic resonance imaging," *Biomedical Engineering Society (BMES)*, 2014.
- [51] V. Khalilzad-Sharghi, "Applications of novel MRI technologies in tissue engineering and disease diagnosis," *ETD collection for University of Nebraska Lincoln,* 2014.
- [52] K. Wartella *et al.*, "Multi modal imaging for assessment of tissue engineered bone in a critical - sized calvarial defect mouse model," *Journal of tissue engineering and regenerative medicine*, vol. 11, no. 6, pp. 1732-1740, 2017.
- [53] W. Abeysinghe, C.-C. Hung, S. Bechikh, X. Wang, and A. Rattani, "Clustering algorithms on imbalanced data using the SMOTE technique for image segmentation," in

Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems, 2018, pp. 17-22.

- [54] N. V. Chawla, Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., "Smote: synthetic minority over-sampling technique.," *Journal of artificial intelligence research,*, vol. 16, no. 1, pp. 321–357, 2002.
- [55] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," in *Kdd*, 1998, vol. 98, pp. 73-79.
- [56] A. S. Solberg and R. Solberg, "A large-scale evaluation of features for automatic detection of oil spills in ERS SAR images," in *IGARSS'96. 1996 International Geoscience and Remote Sensing Symposium*, 1996, vol. 3, pp. 1484-1486: IEEE.
- [57] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1999, pp. 155-164.
- [58] M. Shokri and J. Javidan, "A combination Technique based on Polynomial Method and look-up Table for Linearity Digital Pre-distortion in Power Amplifiers," 2015.
- [59] M. Shokri and J. Javidan, "A novel Technique for Linearizing Digital Pre-distortion in Power Amplifiers for OFDM Applications," *UCT Journal of Research in Science, Engineering Technology*, vol. 4, no. 2, pp. 24-30, 2015.
- [60] O. Boussif *et al.*, "A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine," (in eng), *Proc Natl Acad Sci U S A*, In Vitro

Research Support, Non-U.S. Gov't vol. 92, no. 16, pp. 7297-301, Aug 1 1995.

- [61] Z. Mengjie, G. Xiaoying, and L. Weijun, "Looseness Controlled Crossover in GP for Object Recognition," in 2006 IEEE International Conference on Evolutionary Computation, pp. 1285-1292.
- [62] A. Song, V. Ciesielski, and H. E. Williams, "Texture classifiers generated by genetic programming," presented at the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600), 12-17 May 2002,
- [63] M. Zhang, P. Andreae, and M. Pritchard, "Pixel Statistics and False Alarm Area in Genetic Programming for Object Detection," in *Applications of Evolutionary Computing*, Berlin, Heidelberg, 2003, pp. 455-466: Springer.
- [64] M. Zhang, V. B. Ciesielski, and P. Andreae, "A Domain-Independent Window Approach to Multiclass Object Detection Using Genetic Programming," *EURASIP Journal on Advances in Signal Processing*, journal article vol. 2003, no. 8, p. 206791, July 21 2003.
- [65] M. Zhang and W. Smart, "Using Gaussian distribution to construct fitness functions in genetic programming for multiclass object classification," *Pattern Recognition Letters*, vol. 27, no. 11, pp. 1266-1274, 2006.
- [66] M. Shokri Varniab, C.-C. Hung, and V. Khalilzad Sharghi, "Classification of multiclass datasets using genetic programming," *Proceedings of the Conference on Research in Adaptive and Convergent Systems,* pp. 76-82, 2019.
- [67] S. B. Kotsiantis, I. D. Zaharakis, and P. E. J. A. I. R. Pintelas, "Machine learning: a review of classification and combining techniques," vol. 26, no. 3, pp. 159-190, 2006.
- [68] S. B. Kotsiantis, I. Zaharakis, and P. J. E. a. i. a. i. c. e. Pintelas, "Supervised machine learning: A review of classification techniques," vol. 160, no. 1, pp. 3-24, 2007.
- [69] A. Song, V. Ciesielski, and H. E. Williams, "Texture classifiers generated by genetic programming," in *Proceedings of the 2002 Congress on Evolutionary Computation*. *CEC'02 (Cat. No. 02TH8600)*, 2002, vol. 1, pp. 243-248: IEEE.

- [70] B. Tran, B. Xue, and M. Zhang, "Genetic programming for multiple-feature construction on high-dimensional classification," *Pattern Recognition*, vol. 93, pp. 404-417, 2019.
- [71] S. Picek, E. Hemberg, D. Jakobovic, and U.-M. O'Reilly, "One-Class Classification of Low Volume DoS Attacks with Genetic Programming," in *Genetic Programming Theory and Practice XV*: Springer, 2018, pp. 149-168.
- [72] R. R. Mendes, F. B. de Voznika, A. A. Freitas, and J. C. Nievola, "Discovering fuzzy classification rules with genetic programming and co-evolution," in *European Conference on Principles of Data Mining and Knowledge Discovery*, 2001, pp. 314-325: Springer.
- [73] J. Madár, J. Abonyi, and F. Szeifert, "Genetic Programming for the Identification of Nonlinear Input–Output Models," *Industrial & Engineering Chemistry Research*, vol. 44, no. 9, pp. 3178-3186, 2005/04/01 2005.
- [74] P. Płoński and K. Zaremba, "Visualizing Random Forest with Self-Organising Map," Cham, 2014, pp. 63-71: Springer International Publishing.
- [75] N. Manwani and P. S. Sastry, "Geometric Decision Tree," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 42, no. 1, pp. 181-192, 2012.
- [76] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. J. J. o. a. i. r. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," vol. 16, pp. 321-357, 2002.
- [77] (2020). UCI Machine Learning Archive. Available: [https://archive.ics.uci.edu/ml/datasets/
- [78] (2020). *The Whole Brain Atlas Harvard Medical School*. Available: http://www.med.harvard.edu/AANLIB/
- [79] (2020). Landsat Science NASA. Available: https://landsat.gsfc.nasa.gov/data/
- [80] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
- [81] B. Schölkopf, A. J. Smola, and F. Bach, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.