**Breast Cancer Diagnosis**

**Introduction**

The aim of study is to construct a system to accurately differentiate between benign and malignant breast tumors. We applied the genetic programming technique to select the best features and perfect parameter values of the machine learning classifiers. The performance of the proposed method was based on sensitivity, specificity, precision, accuracy, and the roc curves. The present study proves that genetic programming can automatically find the best model by combining feature preprocessing methods and classifier algorithms.

Several data mining and machine learning techniques have been developed for breast cancer detection and classification, which can be divided into three main stages:

1. preprocessing
2. feature extraction and
3. Classification

Feature extraction is an important step in breast cancer detection because it helps discriminate between benign and malignant tumors. After extraction, image properties such as smoothness, coarseness, depth, and regularity are extracted by segmentation.

**Challenges**

There were two challenges to automate the breast cancer diagnosis:

1. Determining which model best fits the data and
2. How to automatically design and adjust the parameters of the machine learning model.

**The Proposed Method**

Pipeline is the process of tying together some ordered final modules into one to build an automated machine learning workflow. It provides high-level abstraction of the machine learning process and significantly simplifies the complete workflow. Mostly, it is known as Extract, Transform, and Load (ETL) operations.

The purpose of the work is to optimize the list of data transformations and machine learning algorithms to accomplish the classification transformation. To determine the best combination of machine learning algorithm and data is difficult. As a result of the growth of hyper parameter tuning, genetic programming (GP) is proposed to optimize the data and the control parameters of the proposed model.

The **GP** generates randomly a fixed number of pipelines which constitute the members of the population.

Each individual (pipeline) of the population was evaluated based on its fitness which is chosen in this work as the classification score.

The hyper parameters optimized in this work are the number of kennels function for all the classifiers except linear discriminant analysis. The number of kernels function is chosen randomly. In this work, many applied techniques were tested for the subsequent stages of processing and analysis of the breast cancer dataset.

**Stage 1: Preprocessing.**

Processing was performed on the raw breast cancer data to scale the features using the Standard Scaler module. Standardization of datasets is a common requirement for many machine learning estimators. It transforms the attributes to a standard Gaussian distribution based on (*xi*–mean(*x*))/stdev(*x*) where stdev is the standard deviation. The Robust Scaler depends on the interquartile range to transform the features using (*xi*–*Q*1(*x*))/ (*Q*3(*x*)–*Q*1(*x*)), where *Q*1, *Q*2, and *Q*3 represent quartiles. All the transformations used are

Included in scikit-learn machine learning library.

**Stage 2: Features Selection.**

Usually, feature selection is applied as a preprocessing step before the actual learning. However, no algorithm can make good predictions without informative and discriminative features; therefore, to keep the most significant features and reduce the size of the dataset, we implemented PCA using randomized SVD.

**Stage 3: Machine Learning Algorithm.**

Machine learning algorithms allow better predictive performance compared with a single model. This can be

considered machine learning competition, where the winning solution was used as a model for breast cancer diagnosis.

In this paper, the following machine learning algorithms were used to classify the given data set:

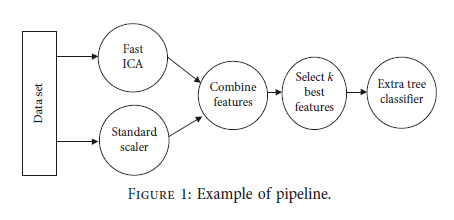
Support vector machine (SVM)

1. K-nearest neighbor, decision tree
2. Random forest (RF)
3. Logistic regression (LR)

**Stage 4: Parameter Optimization.**

Genetic Programming (GP) is a type of evolutionary algorithm (EA) that generalizes the genetic algorithm. GP is a model for testing and selecting the best choice among a set of results. Based on biological evolution and its fundamental mechanism (mutation, crossover, and selection), GP generates a solution. The use of GP is the reason for its flexibility; it can model systems where the structure of the desired models and the key features are not known. In this paper, GP allowed the system to search for models from a range of possible model structures and optimizing the pipelines represented in tree structures for the classification problem. GP first generates a fixed number of pipelines based on the primitives described above, such as features selection decomposition. In other words, the sequence of operators evolves to produce machine learning pipelines that are evaluated to maximize them classification accuracy.

Figure 1 depicts an example of a machine learning pipeline.



After evaluation of the current pipelines machine learning, a new generation is created based on the highest previous pipelines. Each pipeline is considered an individual of GP. The GP is formed by the

Three main operators:

**Mutation operator**:

Changing hyper parameters or adding or removing a primitive preprocessing step such as Standard Scaler or the number of trees in a random forest.

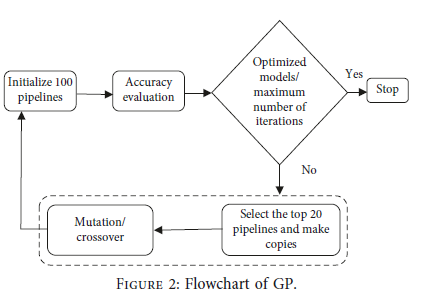
**Crossover operator:**

The crossover operator assumes that 5% of individuals will cross with each other using a 1-point crossover selected at random.

**Selection operator:**

Its main purpose is to select the top 20 individuals and make copies from them. To exchange

Information between the individuals of the population, the crossover or mutation operator can be applied. The subsequent stages of GP are given in Figure 2.



**3. Results**

A single training and test partitions are not effective estimators of a classification error scheme on a limited dataset. Thus, it was decided that a random subsampling scheme should be used in this experiment to minimize any estimation bias.As part of the research, three experiments were set up for training of the input data. In the first case, the point of interest was the feature selection stage. In the second experiment, the focal point was the classification model. Finally,the principal focus of the third experiment was automating the previous experiments into one self-regulating process. In other words, the aim was to automate the process of designing and optimizing machine learning algorithms.

**In the first experiment,** an open-source machine learning software, called WEKA, was employed to extract the features based on the EA, which included

(1) Particle swarm optimization (PSO)

(2) Genetic algorithm (GA)

(3) Evolutionary programming (EP)

(4) Numeric search called best first (BF)

**The second experiment** compared the popular supervised learning algorithms applied for classification of the problem. the metrics used were accuracy, AUC, confusion matrix, and precision-recall

Accuracy (ACC) is the measure of correct prediction of the classifier, and it provides general information about how many samples are misclassified. It is defined as

TP + TN

ACC =

FP + FN + TP + TN

Where TP, FP, TN, and FN are the number of true positives, false positives, true negatives, and false negatives, respectively, when the classifier is predicted. The other metrics derived from a confusion matrix are defined as follows:

TP

Recall =

TP + FN

TP

Precision =

TP + FP

(Precision × recall)

F1 = 2 ×

(Precision + recall)

**Conclusion**

This study attempts to solve the problem of automatic detection of breast cancer using a machine learning algorithm. The present algorithm proceeds in different stages. Three different experiments were conducted using the breast cancer dataset.

In the first test, we proved that the three most popular evolutionary algorithms can achieve the same performance after effective configuration.

The second experiment focused on the fact that combining features selection methods improves the accuracy performance.Finally, in the last experiment, we deduced how to automatically design the machine learning supervised classifier.

Finally, the proposed model looks naturally suited for control parameter setting of the machine learning algorithms in one side and automated breast cancer diagnosis on the other side.