Genetic Algorithm

Breast Cancer Detection

Introduction

Breast cancer is one of the leading causes of mortality globally, and, last year, it victimized around 685,000 women [1]. Projections indicate that around 12.9 million people will die due to cancer by 2030 and around 857,000 caused by breast cancer, according to the Global Cancer Observatory (GLOBOCAN) [2]. These high mortality rates are associated with the late diagnosis of the disease. To improve that perspective, affirmative actions to make the population aware have been employed encouraging prevention and early diagnosis. Early detection and treatment improve such prognosis [3]. Breast cancer diagnosis is carried out mainly by imaging examinations. Mammography highlights the formation of denser tissues (calcifications) and has been widely used worldwide to detect breast cancer. However, this examination is not so efficient for dense breasts and presents substantial false-positive rates [4]. Moreover, this acquisition could cause discomfort to the patient and uses ionizing radiation (X-Rays), increasing by up to 2% the patient's chance of developing cancer [5]. Thermography appears as a low-cost alternative, which does not use ionizing radiation, venous access, or any invasive procedure and also can act as a complement to mammography in several scenarios, contributing to the patient's diagnostic process. This examination considers the temperature variation on the skin surface [6] to identify the presence of anomalies. However, human interpretation of those images is complex and, for this reason, artificial intelligence algorithms must be used to help the medical team to identify the differences between sick and healthy breast patterns.

Malignant tumours generally are more serious than benign tumours. As mentioned, early detection of breast cancer leads to much higher chances of successful treatment. In order to reach this goal, it is necessary to have diagnostic systems with high levels of accuracy and reliability that help doctors to distinguish between benign breast tumours and malignant ones.

In this project, a wrapper feature selection method is proposed based on genetic algorithm-based feature selection. Our proposal builds an ensemble method for selecting models and features by a Genetic Algorithm. Support Vector Machine (SVM), K-nearest neighbor, Logistic regression, AdaBoost, Decision Tree, and Random Forest classifiers are used for ensembling the best classifier for breast cancer prediction. After a few generations of the Genetic Algorithm, this method selects the models that best fit the problem of breast cancer classification. Subsequently, these models are employed in the features selection process, which seeks to find the smallest features that best classify the data. In addition to generating an effective breast cancer diagnosis method, the combination of these strategies optimizes identifying the parameters and features that best adapt and describe the classification problem. We minimized the most demanding task and maximized fine-tuning in choosing the most significant features for breast cancer classification.

Methodology

Breast Cancer dataset overview

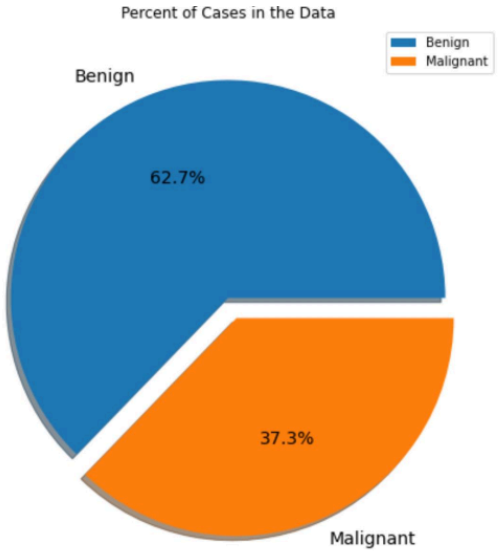
Breast cancer is a malignant tumor arising from breast cells. Even though some of the risk factors (e.g., aging, genetic risk factors, family history, menstrual periods, not having children, obesity) that raise a woman's possibility of developing breast cancer are known, it is not known yet what causes most of the breast cancers and how various factors initiate cells to change into cancerous. Many studies are conducted to learn more, and scientists are having great improvement in understanding how certain alterations in DNA which can affect healthy breast cells to change into cancerous [7, 8].

In this project, we used Wisconsin Breast Cancer Datasets (obtained from UCI Machine Learning Repository) Wisconsin Breast Cancer (Diagnostic). This dataset contains 569 different instances and 32 attributes. Three hundred and fifty-seven cases are benign, and 212 cases are malignant. All attributes are calculated from a digitized image of a fine-needle aspirate (FNA) of patients' breast tissues. All cell nuclei in breast tissues are described by ten real-valued features, and for all these features, the mean, the standard error, and the "worst" (mean of the three largest values) are calculated. As a result, a total of 30 attributes for all images were obtained :

- Radius (mean of distances from the center to points on the perimeter)
- Texture (standard deviation of grayscale values)

- Perimeter
- Area
- Smoothness (local variation in radius lengths)
- Compactness
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions of the contour)
- Symmetry
- Fractal dimension

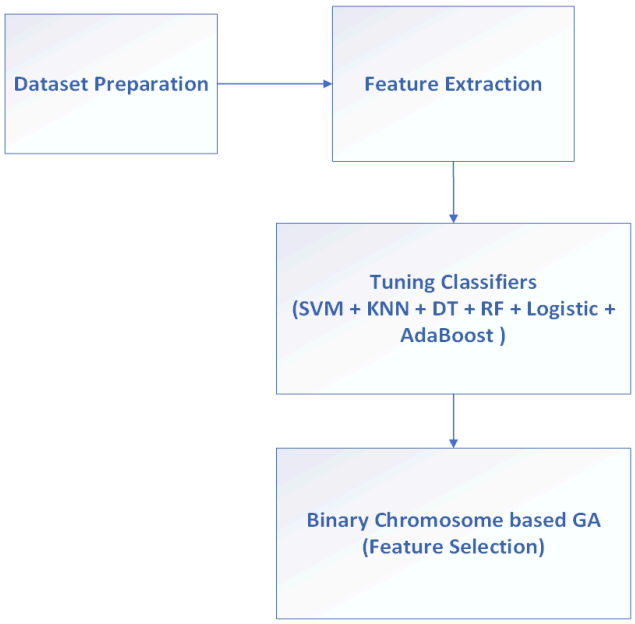
The mean, standard error, and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius.



Genetic Algorithms for Feature Selection

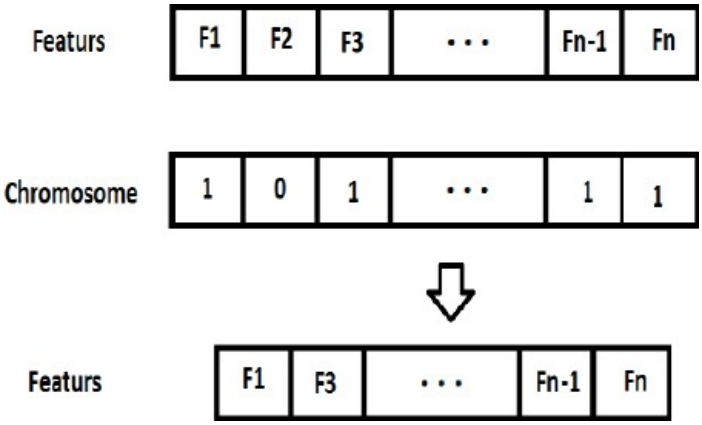
The main purpose of feature selection is to reduce the number of features used in classification while maintaining acceptable classification accuracy. Feature selection reduces the size of the data input into the prediction model; provides an understanding as to which features are required to allow for accurate differentiation between benign and malignant tumors; and can improve the performance of the prediction model. Feature selection methods have therefore become an important step in clinical diagnosis systems

The success of the feature selection process mainly depends on considering two aspects: search strategy and criteria. Different feature selection approaches use different methods to generate subsets and progress the search processes. A number of comparative studies have demonstrated that, for large finite spaces, finding the optimal solution is computationally expensive due to the resulting exponential search space. For this reason, most search strategies attempt to find sub-optimal and near-optimal solutions. Recent research interest has shifted toward global search algorithms (or, metaheuristics). Metaheuristic search strategies have been used to find an optimal solution to a given problem by searching a full space rather than partial feature spaces. Metaheuristic algorithms are especially effective when the information is uncertain and dynamic. Therefore, these advanced approaches are efficient in dealing with biological predictive systems which require handling the uncertainty and imprecision which is apparent in clinical data and medical images. Since the Genetic Algorithm is one of the most widely used optimization methods for finding solutions in complex and nonlinear search spaces, it has been naturally employed to solve feature selection problems. Genetic Algorithms are the main paradigm of evolutionary computing, and a rapidly growing area of Artificial Intelligence (AI). It turns out that there is no accurate definition of “Genetic Algorithms” formally accepted by the evolutionary computation community. However, it can be said that Genetic Algorithms are adaptive heuristic search algorithms which are invented by Holland in the 1960s and inspired by Darwin’s theory of evolution “survival of the fittest”. Five important factors can change how the optimization scheme is performed by Genetic Algorithms (GA): population creation, fitness function, selection schema, genetic operators, and stopping criteria. The algorithm starts with a population of binary strings which are called chromosomes. These are possible candidates for an optimization problem. During each iteration, the populations are evaluated based on their fitness quality, and then the crossover and mutation genetic operators are utilized to select fitter solutions. The following figure shows the proposed method for our project.



Feature Selection

Most commonly, the candidate solutions are encoded in a binary string of 0 and 1. In the binary string, 0 indicates that an associated feature has not been selected, whereas 1 shows that its corresponding feature has been selected.



For efficient feature selection, we must consider the importance of the evaluation criteria for measuring classification performance. All feature selection methods need to use evaluation criteria together with a search strategy to obtain the optimal feature set. Among all of the evaluation criteria, mutual information has attracted the most attention because it is a good indicator of the correlation between features and class labels, and it is

considered to be least sensitive to noise or outlier data than other approaches. The basic idea of the feature selection algorithm is to select one optimal subset from the original dataset by maximizing the joint between the input features and target output. Estimation of high-dimensional is very difficult and has high computational complexity, and this consequently limits the applications of this method. Most of the existing methods adopt low-dimensional. In this section, method-based feature selection is shifted to global search algorithms for breast cancer diagnosis. The main steps of the Genetic Algorithm applied to the breast cancer dataset are described in the following steps:

- **Step 1**: Create an initial population. The generation of the initial population is straightforward and created by setting each bit of the Chromosome to 0 or 1 randomly. The number of strings in the Chromosome equals n, where n is the total number of features of the breast cancer dataset
- **Step 2**: Compute the classification accuracy between features and target classes using the best fine-tuning classifier (Random Forest). In the main loop, each Chromosome in the population is evaluated based on the correlation with the class. The chromosomes with the highest values, i.e. correlations with the target class are used in step3.
- **Step 3**: Fitness proportional selection is used for reproduction. The probability for a Chromosome to be selected is proportional to its fitness score.
- **Step 4**: Crossover and mutation operators reproduce some fitter chromosomes and generate a new chromosome (offspring) to be used in the next generation.
- **Step 5**: Steps 2 to 4 are repeated until the maximum generation (this has been experimentally set to 80) is reached.
- **Step 6**: The individual in the final generation with the best fitness value is selected as the optimal solution.

Experimental results

In order to fine-tuning the best classifier, we did train and test Support Vector Machine (SVM), K-nearest neighbor, Logistic regression, AdaBoost, Decision Tree, and Random Forest classifiers were applied to Wisconsin breast cancer databases. The following table shows the different accuracies for each one.

Classifier	Accuracy
RandomForest	0.972028
KNeighbors	0.965035
LinearSVM	0.958042
Logistic	0.958042
GradientBoosting	0.958042
RadialSVM	0.951049
AdaBoost	0.951049
DecisionTree	0.930070

The RandomForest was the best classifier to continue working on feature selection to measure fitness score, then the proposed feature selection method is presented and The following list shows the best selected relevant features from 30 features.

- radius_mean
- fractal_dimension_mean
- smoothness_worst
- area_mean
- compactness_mean
- perimeter_worst
- texture_mean
- perimeter_mean
- smoothness_mean
- perimeter_se

Conclusion

In this project, a feature selection model with GA-based on feature selection is designed to identify relevant features. GA has more recently developed in comparison to different feature selection algorithms. GA can be useful to feature selection when the problem has

exponential search space. There are many advantages of the GAs for feature selection that have been published in various literatures.

References

[1] Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J. Clin. 2021, 71, 209–249.

[2] International Agency for Research on Cancer (IARC). Cancer Tomorrow; International Agency for Research on Cancer (IARC): Lyon, France, 2021.

[3] Borchartt, T.B.; Conci, A.; Lima, R.C.; Resmini, R.; Sanchez, A. Breast thermography from an image processing viewpoint: A survey. Signal Process. 2013, 93, 2785–2803.

[4] Doi, K. Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential. Comput. Med. Imaging Graph. Off. J. Comput. Med. Imaging Soc. 2007, 31, 198–211.

[5] Arabi, P.M.; Muttan, S.; Suji, R.J. Image enhancement for detection of early breast carcinoma by external irradiation. In Proceedings of the 2010 Second International Conference on Computing, Communication and Networking Technologies, Bangkok, Thailand, m23–25 April 2010; pp. 1–9.

[6] Damiao, C.; Gonzalez, J.R.; Moran, M.B.H.; Fontes, C.P.; Balarine, G.; Cruz Filho, R.; Conci, A. On the possibility of using temperature to aid in thyroid nodule investigation. Sci. Rep. 2020, 2045–2322

[7] Jerez-Aragones J, Gomez-Ruiz JA, Ramos-Jimenez G, MunozPerez J, Alba-Conejo E (2003) A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med 27(1):45–63

[8] Marcano-Ceden˜o A, Quintanilla-Domı´nguez J, Andina D (2011) WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst Appl 38(11):9573–9579

[9] UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set (2012) Retrieved 15 Mar 2012, from UCI Machine Learning Repository: Breast Cancer Wisconsin

(Diagnostic) Dataset:
http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)