Genetic Algorithm for feature selection in classification of Wisconsin breast cancer dataset.

Venkata Satya Sai Ajay Daliparthi
Department of computer science
Blekinge Tekniska Högskola
Karlskrona, Sweden
veda18@student.bth.se

Sai Ajith Teki
Department of computer science
Blekinge Tekniska Högskola
Karlskrona, Sweden
sate18@student.bth.se

***Abstract*—The main problem in the medical field involves the diagnosis of disease, based on various tests results of the patients. Due to the welter data, the diagnosis becomes difficult for medical experts. Improvements in the medical field resulted in the collection of large databases, which needs to analyze and find out the patterns in them. The automated diagnosis system is most acute in case of deadly diseases like cancer where early detection can improve the survival of patients. This kind of problem is solved by building basic classification machine learning model which generates the cancer detection results. Feature selection is proved to improve the accuracy of the classification models. In this paper we performed a experiment where we used genetic algorithm to select a subset of features that effect the accuracy of the machine learning model. By using the subset of features we build a classification model and compared the accuracy with model build by using all the features. If the model build using the subset of features can improve the accuracy of the machine learning model this can make huge advancements in automation of medical diagnosis and improves lifetime of many people.**

***Index Terms*—Genetic Algorithm, Machine learning, Feature selection, Classification, Medical Diagnosis, and Data preprocessing.**

I. Introduction

The main problem in the medical field involves the diagnosis of disease, based on various tests results of the patients. Due to the welter data, the diagnosis becomes difficult for medical experts. Improvements in the medical field resulted in the collection of large databases, which needs to analyze and find out the patterns in them. One of the applications of Data mining approaches in the medical domain is automated diagnostic systems. The automated diagnosis system is most acute in case of deadly diseases like cancer where early detection can improve the survival of patients. Breast cancer is considered as the most commonly occurring disease in women. It is the second most treated cancer after skin cancer in the United States of America. There are three methods available for the detection of cancer. They are physical examination, mammography, and biopsy. The accuracy of Fine needle aspiration biopsy (FNAC) varies from 30 to 90 percent depending on the expertise of the doctor. So, it is important to develop an accurate identification system that helps doctors to identify breast cancer. To solve this kind of problem, we can build any of the available classification algorithms to build a machine learning model. There is a misconception that more features will result in better accuracy. But in reality, some features will decrease the efficiency and cause overfitting of the model. Feature selection is one of the methods used to solve this problem in classification models.

Feature selection is one of the pre-processing techniques in data mining and used in the fields of statistics and pattern recognition. Feature selection is a process of reducing the number of input variables while building a machine learning model. Selecting the less number of features will reduce the computational cost and increses accuracy. There are three types of feature selection techniques. They are Univariate Selection, feature importance, and correlation matrix. If there are more features we need methods to select the features. One of the most advanced methods for feature selection is the genetic algorithm.

Genetic algorithm is a stochastic method for function optimization based on the concept of biological evolution. In nature, the genes of biological organisims tend to evolve over successive generations to better adapt to the environment. In feature selection, the function is to optimize the performance of the classification model. The design variables are (0) absence and (1) the presence of the feature. The genetic algorithm operates on the population to produce better approximations at each generation. A new generation is created by selecting individuals according to their fitness levels. At each generation, some of the population will learn a function close to solving the problem. The offspring will under mutation. This process results in individuals that are better suited for the environment. A state diagram for the training process of the genetic algorithm is shown below (Figure 1).

(a) Initialization: In the first step, a group of the population who can solve the problem is taken. Each individual is characterized by a set of features called genes. When genes are combined together, they are called chromosomes. the

Identify applicable funding agency here. If none, delete this.

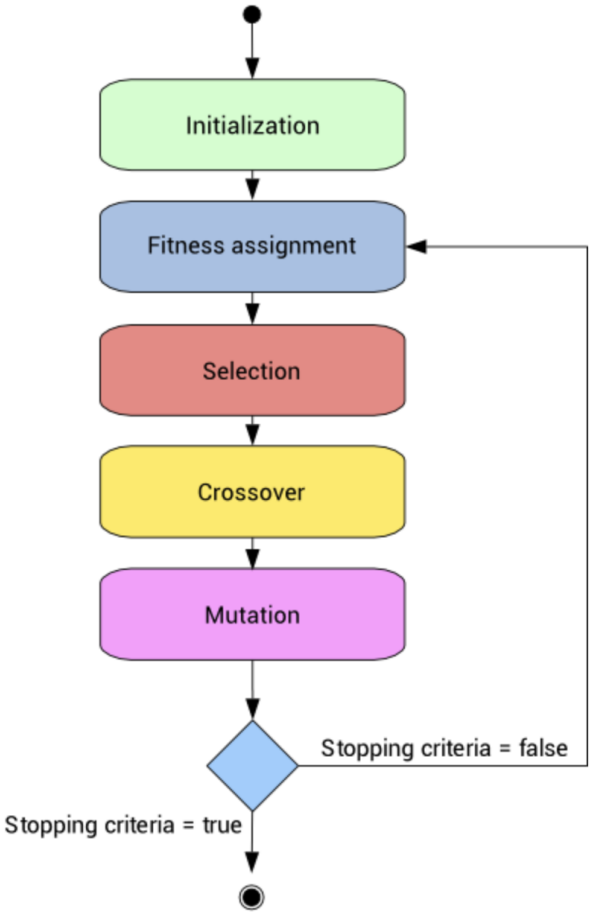

Fig. 1. State Digram of Genetic Algorithm

population is nothing but a group of chromosomes(shown in Figure 2).

(b) Fitness function: Fitness function is an evaluation metric

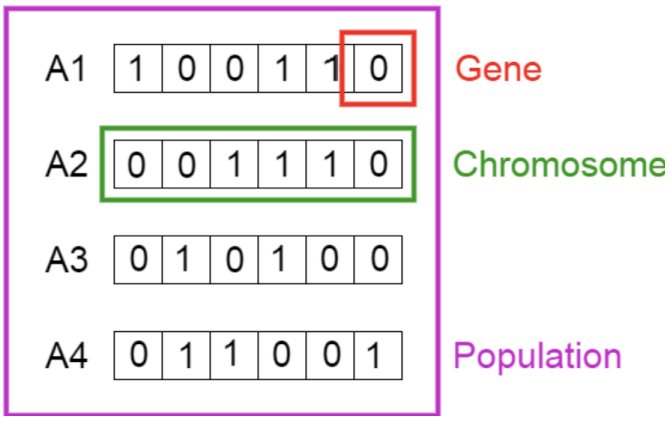

Fig. 2. Initialization of population

used to measure the fitness of each individual. Individuals with high fitness values are selected and reproduced. Lowest fitness value individuals are not selected. Rank based method is the most used method for fitness measurement, where errors of each individuals are sorted.

(c) Selection: Selection is done based on the fitness scores of the individuals. The fit individuals will pass their genes to the next generation and produces offspring.

(d) Crossover: Once the fit individuals are selected then the crossover operator recombines the selected individuals to create a new generation. This operator picks two individuals at random and to create offsprings for the new generation(shown in Figure 3).

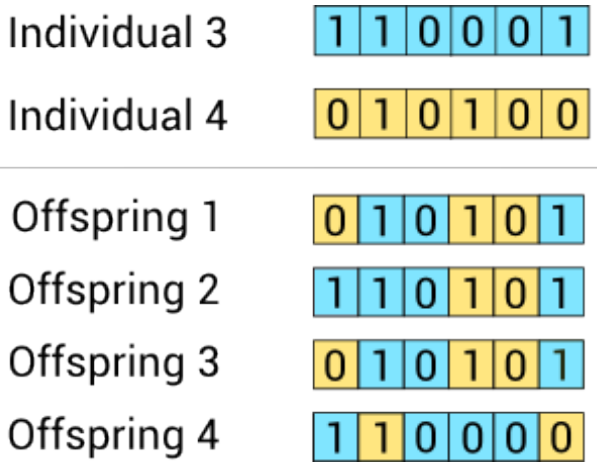

Fig. 3. Crossover

(e) Mutation: Mutation is the process where genes of randomly chosen offsprings are changed (shown in Figure 4). This allows the algorithm to not stuck in the endless loop of crossovers and selection.

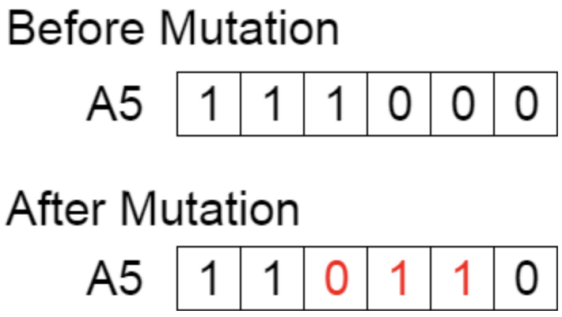

Fig. 4. Crossover

Above steps will continue until the stopping condition is reached. This is the working of genetic algorithm.We are using a Wisconsin breast cancer dataset that consists of 32 attributes. This research involves the usage of a genetic algorithm to find a subset of features that affect the accuracy of the classification model. If we can improve the accuracy of the classification

model that will help the doctors to identify Breast cancer at early stages. This can save or increase the life of many people.

II. RELATED WORK

A wide range of study was done on feature selection for machine learning models. Feature selection has many applications in many fileds such as image recognition [1], image retrieval [2], text mining [3], and bioinformatic data analysis [4].

MA HALL et al. [5] proposed a correlation based feature selection model that can be applied to continuous and discrete problems. The algorithm often out-performs the well-known ReliefF attribute estimator when used as a preprocessing step for many machine learning models.

Michalak and Kwasnicka et al. [6] proposed a relationship-based dual-strategy wrapper feature selection method. A rank criteria system based on class densities for binary data is presented in this method. Xue et al. [7] proposed a PSO feature selection model with new benefit mechanisms that take into consideration both the classification accuracy and the number of selected features.

Vandenbroucke et al. [8] proposed an unsupervised filter feature selection method that uses a competitive learning algorithm to classify the samples and ascertain the number of clusters, and then divides the original feature set into several feature subset. Wang et al. [9] presented a semi-supervised filter feature selection method called SRFS based on information theory, where the unlabeled data are utilized in the Markov blanket as the labeled data through the relevance gain. M Rostami et al. [10] proposed a clustering based genetic algorithm for feature selection (CGAFS) using k-means clustering algorithm.

Genetic algorithm was applied to many machine learning models for feature selection every time the accuracy is improved. S Ahlawat et al. [11] proposed hybrid feature set of statistical and geometrical features is developed in order to get the effective feature set consist of local and global characteristics of sample digits. The method utilizes a genetic algorithm based feature selection for selecting best distinguishable features and k-nearest neighbour for evaluating the fitness of features of handwritten digit dataset.

H Galwani et al. [12] proposed a brain tumour image classification technique based on Genetic Algorithm (GA) for feature selection.The searching capability of genetic algorithms is explored for appropriate selection of features from input data and to obtain an optimal classification.

An experiment conducted by sabah sayed et al. [13] introduced a Nested Genetic Algorithm for feature selection by correlating micro array datasets. An experiment conducted by Y.V.Srinivasa Murthy et al. [14] concluded an approach of genetic algorithms with the support of neural networks has been used to select features rather than considering all the unnecessary dimensions among them. It is called as Genetic Algorithm for Feature Selection(GAFS).It helps in segmentation of vocal and non vocal segments in an audio.

An experiment conducted by Omid Gholami et al. [15] developed a genetic algorithm for train rout- ing and timetabling which obtain effective and strong time table and routes. Wegele and schneider et al. [16]proposed a branch and bound algorithm to achieve an initial solution and genetic algorithm to improve accuracy of solution.

An experiment conducted by Cheun Horng lin et al. [17] developed a feature selection technique in order to reduce the cost and running time of image retrieval to achieve higher rate of recognition. M.loakman et al. [18] proposed a feature selection algorithm based on genetic algorithm (GA) to find the best features that describe EEG signal. The best features are searched among ten statistical features calculated from the cross-correlation of effective channel.

N. Bidi et al. [19] provided an empirical study of a feature selection method based on genetic algorithms for different text representation methods. This feature selection algorithm can accomplished two goals in one hand is the search of a feature subset such that the performance of classifier is best and in other hands is find a feature subset with the smallest dimensionality which achieves higher accuracy in classification.

N. ambarasi et al. [20] proposed a model that predicted more accurately the presence of heart disease with reduced number of attributes. Originally, thirteen attributes were involved in predicting the heart disease. In their work, Genetic algorithm is used to determine the attributes which contribute more towards the diagnosis of heart ailments which indirectly reduces the number of tests which are needed to be taken by a patient. Thirteen attributes are reduced to 6 attributes using genetic search.

H. Handles et al. [21] proposed a model to recognize malignant melanomas and nevocytic nevi (moles), automatically. In the first step, several types of features are extracted by 2D image analysis methods characterizing the structure of skin surface profiles: texture features based on cooccurrence matrices, Fourier features and fractal features. Then, feature selection algorithms are applied to determine suitable feature subsets for the recognition process. A classification performance of 97.7 percent is achieved.

A. Kharrt et al. [22] compared classical sequential methods with the genetic approach in terms of the number of features, classification accuracy and reduction rate. Genetic Algorithm (GA) achieves an acceptable classification accuracy with only five of the available 44 features of MR brain images using wavelet co-occurence.

Genetic algorithm was used to extract important features from the all the features in many fields including the medical filed. But, there are still many datasets and fields in which we have to study the effect of genetic algorithm for feature selection. One such database in medical filed is Wisconsin breast cancer dataset. In this research we are dealing with Wisconsin breast cancer dataset to extract import features by applying genetic algorithm and study the affect in calssification performance by using different metrics like Accuracy, Precision, Recall and F-measure.

III. AIM AND OBJECTIVES

Aim: This research aims to apply genetic algorithm on Wisconsin breast cancer dataset that consists of 32 attributes to select a subset of features that affect the accuracy of the classification model.

Objectives:

1) To Check weather the dataset has any empty values. If there are any empty values fix them with data manipulation methods.
2) Data pre-possessing and Data cleaning has to be done.
3) To build a classification model using the dataset and calculate the accuracy.
4) To apply Genetic algorithm to find a subset of features that affect the accuracy.
5) To initialize a population that can solve the problem by hand and then define a fitness function.
6) To build a classification model depending upon the subset of features selected by the model and caluculate the accuracy.
7) To compare the accuracy results of both models.

IV. RESEARCH QUESTIONS

RQ.1: *"Can genetic algorithm find the subset of features that affect the accuracy of classification machine learning model"*

RQ.2: *"What is the difference in performance of the classification machine learning model with and without feature selection? "*

V. RESEARCH METHODOLOGY

RQ.1: *"Can genetic algorithm find the subset of features that affect the accuracy of classification machine learning model"*

Methodology: Experiment

Justification: To answer this research question we need to perform an experiment where, we apply genetic algorithm on the dataset to find the subset of features and to check weather those subset of features affects the accuracy. So experiment is chosen as research methodology to answer this research question.

RQ.2: *"What is the difference in performance of the classification machine learning model with and without feature selection? "*

Methodology: Experiment

Justification: To answer this research question we need to perform an experiment where, we have to build two classification model with and without using subset of features. After building the models we will generate the results and compare them using the selected metrics.

Metrics: These metrics are applicable to evaluate the results of both the research questions.

1) **Accuracy:** Accuracy is the proportion of true results divided by the total number of cases evaluated.

$$Accuracy = (TP + TN)/(TP + FP + FN + TN)$$

2) **Precision:** Precision is proportion of predicted Positives is truly Positive.

$$Precision = (TP)/(TP + FP)$$

3) **Recall** Recall is proportion of actual Positives is correctly classified.

$$Recall = (TP)/(TP + FN)$$

4) **F-measure** F-measure is harmonic mean of precision and recall.

$$F{-}measure = (2{*}recall{*}precision)/(recall{+}precision)$$

where TP, TN, FP, and FN represents True Positives, True Negatives, False Positives, and False Negatives respectively.

VI. EXPECTED OUTCOMES

After applying genetic algorithm to the dataset, we will get the sub set of features that effect the accuracy of the classification machine learning model. Also after building a classification model using the subset of features, the model is expected to show better accuracy with compared to model build will all the features in the dataset.

VII. WORK DIVISON

The total work of this research is divided between two people in the following way presented below.

Venkata Satya Sai Ajay Daliparthi - 50

Defining research problem, Review previous research findings, Research Methodology, formulating Research Questions, Experiment design and Report writing.

Sai Ajith Teki - 50

Literature study, Research Methodology, Research questions reviewing, Research design, Background study and report writing.

VIII. LIMITATIONS AND RISK MANAGEMENT

Limitations:

This work is limited to deal with Wisconsin breast cancer dataset only and this can be extended to many datasets. This work is limited to implement genetic algorithm to extract the features but doesn't combines other feature selection methods to gain better accuracy. These are the limitations of our research.

Risk Management:

Risk	Impact	Reason and Management
Knowledge risk	Moderate	This risk occurs by insufficient information from the literature review, we will consider Inclusive and exclusive criteria to manage this risk.
Time risk	High	This can delay in experiment, we will make changes in the time and activity plan and try different approaches or change configura-tions to manage this risk.
Data risk	Moderate	This risk occurs if the data in the dataset causes overfitting of the model due to many features. we will try different data cleaning and manipulation methods or find similar dataset to manage this risk.

IX. TIME AND ACTIVITY PLAN

Week	Project Goal	Activity
week 1	Selecting a topic	Literature review and finding supervisor
week 2	Finalizing the topic	Finalize the topic and discuss with supervisor
week 3	Formulating research questions	Formulate research questions related to topic
week 4	Thesis proposal	Prepare thesis proposal and discuss with supervisor.
week 5-8	Data collection	Literature review for the related topic
week 9	Feedback session	Discuss with the supervisor and finalize the data required for the topic
week 10 -13	Experiment	Conduct Experiments that will produce results.
week 14	Analysis of Results	Evaluate the results based on selected metrics.
week 15	Feedback session	Discuss the results with the supervisor.
week 16 -17	Thesis Draft	Prepare draft and get approval from supervisor
week 18 -19	Presentation and opposition	Present the thesis and defend.
week 20	Submission of final thesis	Make final changes and submit the final Thesis.

X. CONCLUSION

Due to the advancements in medical fields the automation of Medical diagnosis plays huge role. while there are many Data mining techniques available to extract useful information form the medical databeses, there is a need for efficient model that will help the doctors to identify diseases. Although there are many feature selection methods avilabe, genetic algorithm is proved to improve the accuracy in models by selecting subset of features. It can be applied to many models where there are more features.

REFERENCES

[1] A. Khotanzad and Y. H. Hong, "Rotation invariant image recognition using features selected via a systematic method," *Pattern recognition*, vol. 23, no. 10, pp. 1089–1101, 1990.

[2] D. L. Swets and J. J. Weng, "Efficient content-based image retrieval using automatic feature selection," in *Proceedings of International Symposium on Computer Vision-ISCV*, pp. 85–90, IEEE, 1995.

[3] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi, "A simultaneous feature adaptation and feature selection method for content-based image retrieval systems," *Knowledge-Based Systems*, vol. 39, pp. 85–94, 2013.

[4] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome informatics*, vol. 13, pp. 51–60, 2002.

[5] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," 2000.

[6] M. Dash and H. Liu, "Handling large unsupervised data via dimensionality reduction.," in *1999 ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, 1999.

[7] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms," *Applied soft computing*, vol. 18, pp. 261–276, 2014.

[8] N. Vandenbroucke, L. Macaire, and J.-G. Postaire, "Unsupervised color texture feature extraction and selection for soccer image segmentation," in *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, vol. 2, pp. 800–803, IEEE, 2000.

[9] Y. Wang, J. Wang, H. Liao, and H. Chen, "An efficient semi-supervised representatives feature selection algorithm based on information theory," *Pattern Recognition*, vol. 61, pp. 511–523, 2017.

[10] M. Rostami and P. Moradi, "A clustering based genetic algorithm for feature selection," in *2014 6th Conference on Information and Knowledge Technology (IKT)*, pp. 112–116, IEEE, 2014.

[11] S. Ahlawat and R. Rishi, "A genetic algorithm based feature selection for handwritten digit recognition," *Recent Patents on Computer Science*, vol. 12, no. 4, pp. 304–316, 2019.

[12] H. Gwalani, N. Mittal, and A. Vidyarthi, "Classification of brain tumours using genetic algorithms as a feature selection method (gafs)," in *Proceedings of the International Conference on Informatics and Analytics*, pp. 1–5, 2016.

[13] S. Sayed, M. Nassef, A. Badr, and I. Farag, "A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets," *Expert Systems with Applications*, vol. 121, pp. 233–243, 2019.

[14] Y. S. Murthy and S. G. Koolagudi, "Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (gafs)," *Expert Systems with Applications*, vol. 106, pp. 77–91, 2018.

[15] O. Gholami and Y. N. Sotskov, "Train routing and timetabling via a genetic algorithm," *IFAC Proceedings Volumes*, vol. 45, no. 6, pp. 158–163, 2012.

[16] S. Wegele and E. Schnieder, "Automatic dispatching of train operations using genetic algorithms," *Publication of: WIT Press*, 2004.

[17] C.-H. Lin, H.-Y. Chen, and Y.-S. Wu, "Study of image retrieval and classification based on adaptive features using genetic algorithm feature selection," *Expert Systems with Applications*, vol. 41, no. 15, pp. 6611–6621, 2014.

[18] M. Lokman, A. Dabag, N. Ozkurt, S. Miqdad, and M. Najeeb, "Feature selection and classification of eeg finger movement based on genetic algorithm," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–5, IEEE, 2018.

[19] N. Bidi and Z. Elberrichi, "Feature selection for text classification using genetic algorithms," in *2016 8th International Conference on Modelling, Identification and Control (ICMIC)*, pp. 806–810, IEEE, 2016.

[20] M. Anbarasi, E. Anupriya, and N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5370–5376, 2010.

[21] H. Handels, T. Roß, J. Kreusch, H. H. Wolff, and S. J. Poeppl, "Feature selection for optimized skin tumor recognition using genetic algorithms," *Artificial Intelligence in Medicine*, vol. 16, no. 3, pp. 283–297, 1999.

[22] A. Kharrat, N. Benamrane, M. B. Messaoud, and M. Abid, "Genetic algorithm for feature selection of mr brain images using wavelet co-occurence," in *International Conference on Graphic and Image Processing (ICGIP 2011)*, vol. 8285, p. 828557, International Society for Optics and Photonics, 2011.