Diagnosis of breast cancer using machine learning algorithms based on features selected by Genetic Algorithm: Assessed on five datasets

Saranya N, Kavi Priya S

Full Time Research Scholar, Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi

Associate Professor, Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi

Abstract

Breast Cancer is one of the chronic diseases occurred to human beings throughout the world. Early detection of this disease is the most promising way to improve patients' chances of survival. The strategy employed in this paper is to select the best features from various breast cancer datasets using a genetic algorithm and machine learning algorithm is applied to predict the outcomes. Two machine learning algorithms such as Support Vector Machines and Decision Tree are used along with Genetic Algorithm. The proposed work is experimented on five datasets such as Wisconsin Breast Cancer-Diagnosis Dataset, Wisconsin Breast Cancer-Original Dataset, Wisconsin Breast Cancer-Prognosis Dataset, ISPY1 Clinical trial Dataset, and Breast Cancer Dataset. The results exploit that SVM-GA achieves higher accuracy of 98.16% than DT-GA of 97.44%

Keywords Breast cancer, Machine learning algorithms, Feature Selection, Genetic Algorithm, SVM, Decision tree

1. Introduction

Diagnosis of a disease is a difficult task that necessitates numerous tests on patients in order to reach a precise conclusion. This may lead to the use of analytic devices designed to assist doctors in making decisions. People in cities consume a diet high in calories but low in nutrients [20]. Breast cancer is one of the diseases that destroy a large number of people worldwide each year. Early detection of such diseases is a difficult task in order to reduce the number of deaths. In the fields of chronic diseases such as cancer, various machine learning techniques are used for medical diagnosis, which is supported by prediction. Most of the medical datasets should be considered for prediction because of the unique features each dataset possesses to predict cancer with high accuracy. This is only possible by selecting crucial features from complex breast cancer datasets.

Machine learning employs scientific algorithms to recognise patterns in large datasets and to iteratively improve in performing this recognisable proof with additional data. These algorithms are commonly used in a variety of spaces and applications, such as commercial, security, fund, internet-based life, and misrepresentation detection. In medical applications it plays a crucial task in predicting disease and risks at a right time.

Genetic Algorithm and Machine Learning algorithm used in this research work for feature selection and prediction of breast cancer. The primary goals of this research work are:

➢ To create a feature selection framework based on genetic algorithm in order to select the best features from various benchmark datasets in order to predict breast cancer.
➢ Using machine learning algorithms such as Support Vector Machine, Logistic Regression, k-Nearest Neighbour, and Decision Tree for breast cancer prediction based on GA- selected features and evaluation of their performance.

One of the provocations in prediction algorithms is the multiplicity of features in a dataset, which leads to inaccuracy [10,11]. Feature selection is one method for dealing with this problem [12, 15]. Table 1
show the datasets used in this study, which include Wisconsin Breast Cancer diagnosis, original, and Prognosis dataset, ISPY 1 clinical trial dataset, and Breast cancer dataset, and Breast cancer Coimbra dataset,

| S. No | Datasets                              | Number of attributes | Number of Instances | Number of classes |
|-------|---------------------------------------|----------------------|---------------------|-------------------|
| 1     | Wisconsin Breast Cancer-Diagnosis Dataset | 32                   | 569                 | 2                 |
| 2     | Wisconsin Breast Cancer-Original Dataset | 11                   | 699                 | 2                 |
| 3     | Wisconsin Breast Cancer-Prognosis Dataset | 34                   | 198                 | 2                 |
| 4     | ISPY1 Clinical trial Dataset           | 17                   | 169                 | 2                 |
| 5     | Breast Cancer Dataset                  | 5                    | 569                 | 2                 |

Table 1 Breast Cancer Datasets

The remaining part of this paper is organised as follows. Section 2 summarises what has been done in the literature. Section 3 elaborates the dataset and the techniques used in proposed work. Section 4 details the proposed work. In section 5 the performance of the proposed work is assessed. In the end section 6 details the conclusion.

2. Literature Review

Several algorithms such as k-Nearest Neighbour (kNN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree(DT) are used in Machine Learning. While each of these algorithms processes data differently, this section reviews a few recently proposed machine learning candidates in the field of malignant growth detection in chronological order.

The author of [1] presented a fuzzy system for Parkinson's disease (PD) diagnosis based on kNN(FkNN). Furthermore, they used principal component analysis to identify the most discriminative features upon which the best FkNN model was built. They compared their system to the SVM algorithm and discovered that their proposed method outperformed the SVM algorithm. Their FkNN achieved the highest classification accuracy of 96.07 percent.

[2] asserted that the precision of the kNN method varies with the number of neighbours and the level of information used for classification. Meanwhile, they demonstrated the variation of the maximum and minimum accuracy values with classification set sizes and the number of neighbours. Lynch et al. [3] used the SEER database to the survival rate of lung cancer patients using supervised learning classification techniques such as linear regression, decision trees, GBM, SVM, and a custom ensemble. The results showed that GBM was the most precise of the five individual models used, with a root mean square error value of 15.32.

[4] compared the performance of C4.5, NBs, and kNN classification algorithms on 670 data points, each with 9 attributes, to detect breast cancer diagnosis. They discovered that, while NBs and kNN have the same accuracy of 98.51 percent, C4.5 has the lowest accuracy of 91.79 percent. [5] used decision tree and kNN algorithms to predict diabetes using the Prima Indian Diabetes Dataset, which included 768 data points, each with 8 attributes, and achieved 90.43 percent and 76.96 percent accuracy, respectively.

Pradeep et al. [7] used SVM, NB, and C4.5 techniques on The North Central Cancer Treatment Group (NCCTG) lung cancer data set to assist specialists in drawing more accurate conclusions about cancer survivability rates. The results show that C4.5 performs better in predicting lung cancer as the training data set is increased. In [8] a combined genetic fuzzy algorithm is proposed to detect lung cancer. He
used that algorithm on 32 patients with 56 attributes and achieved 97.5 percent accuracy with a confidence level of 93 percent.

The author of [9] developed a new solution to accelerate the kNN algorithm on the breast cancer database that is dependent on clustering and attribute separation. He compared this proposed algorithm to others in the field, including SVM, ANN, NB, and kNN. The results showed that while ANN performed the best, its execution time was 2.2 times that of the proposed algorithm.

In [13] the authors examined the performance of a well-known machine learning algorithm, kNN, on the Wisconsin Breast Cancer dataset. And this dataset consists of 569 instances and 32 attributes. They used two important dimensionality reduction strategies, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) and found that kNN with LDA performed better than kNN and kNN with PCA, with accuracies of 97.06 percent, 95.29 percent, and 95.88 percent, respectively.

Akben used kNN, SVM, and NB to preprocess data in order to detect chronic kidney disease in [14]. They started with raw data and discovered that the classification was not accurate enough to encourage doctors. As a result, they used the methods after the data had been preprocessed using the k-means clustering approach. The results showed that accuracy was significantly improved, particularly for the kNN classifier, which achieved 96 percent accuracy.

3. Materials and Methods

3.1. Dataset Description

- **Wisconsin Breast Cancer-Diagnostic Dataset (WBC-DD)**
  
  WBC-DD is accessible in UCI Machine Learning Repository and it is a cancer recognition dataset. It is a quantitative dataset containing information about breast masses. It consists of 30 attributes, 569 instances, and two class variables such as Benign or Malignant. The designers of this dataset are Dr. Wolberg, Street and Olvi. In this dataset, the characteristics of the image’s cell nuclei. There are 357 cases of benign breast cancer and 212 cases of malignant breast tumours among the cases diagnosed.

- **Wisconsin Breast Cancer-Original Dataset (WBC-OD)**
  
  WBC-OD is accessible in UCI Machine Learning Repository, and it is also cancer recognition dataset. It consists of 10 attributes, 699 instances. The designer of this dataset is Dr. William H Wolberg. From 1981 to 1991, this dataset represents the chronological categorization of data. It includes a numerical representation of clump thickness, cell size, shape, nuclei, and a diagnosis column for determining if a tumour is benign or malignant.

- **Wisconsin Breast Cancer-Prognostic Dataset (WBC-PD)**
  
  If these predictors are true, prediction models based on them could be utilised as a biomarker for breast cancer. The presence and absence of malignant cells are predicted using this dataset.

- **Breast Cancer Dataset (BPD)**
  
  This is a subset of the Wisconsin Diagnostic Breast Cancer containing 5 attributes for determining if a breast cancer is malignant or benign. The malignancies are also identified in this dataset.

- **ISPY1 Clinical Trial Dataset (ISPY 1)**
  
  Investigation of serial studies to predict your therapeutic response with imaging and molecular analysis i.e. ISPY Trial provides innovative, better, and faster individualised treatment. The cancer imaging archive, the breast imaging research programme, and the University of California at San Francisco (UCSF) provided all data for the 222 patients treated for breast cancer in the ISPY-1 clinical study. To aid in the dissemination and reproducibility of this analysis, the raw data and all code were made available on the websites such as Data World and GitHub. This dataset contains characteristics of breast cancer patients survival data.
3.2. Genetic Algorithm

The Genetic Algorithms (GA) operations reflect the process of natural selection in which the fittest individuals are chosen for reproduction in order to produce offspring for the next generation. Genetic Algorithms are search based because they look for the best gene in the population for the next generation.

GA consists of five parts as follows: Initial population, Fitness function, Selection, Crossover and mutation. The core GA continues as follows: an underlying population of chromosomes is produced in an indiscriminate or heuristic manner. In each iteration i.e., generation, a fitness function is used to depict the streamlining issue in the search space then decodes and evaluates the chromosomes. Based on the fitness values, the chromosomes for the next generation are selected. There are numerous options available here, one of the simplest being the fitness proportionate choice, in which chromosomes are chosen with a likelihood proportional to their relative fitness. This ensures that a selected individual is present at the normal number of times corresponding to its relative performance in the population. As a result, high-fitness chromosomes have a better chance of reproducing and transmitting new individuals to the population, whereas low-fitness chromosomes will not.

Cross over and mutation are hereditary processes that introduce new chromosomes into the population. The cross over operation is carried out with the possibility of two chosen individuals (parents) trading parts of their genomes to form two new chromosomes (offspring). Meanwhile, the mutation operation avoids premature union to nearby optima by randomly examining new focuses in the hunt space; it is carried out by flipping bits at random, with some low likelihood. GA is a stochastic iterative process that does not guarantee the optimal point. Furthermore, the stopping condition could be specified as a maximum number of generations or a desired fitness value.

3.3. Breast Cancer Prediction models

Machine Learning is primarily concerned with the development of computer programmes that can access data and use it to learn on their own. The first step in this process is to examine the data for patterns based on these parameters; this will allow you to make better decisions in the future. This is where computers learn on their own, without any outside intervention, support, or assistance. The computers will learn to take and adjust their own actions based on the circumstances. There are two types of machine learning algorithms: supervised and unsupervised. Supervised machine learning algorithms were majorly used. It can be used when the system needs to learn from the past and apply it to current data with labelled samples in order to predict future events. In this work Support Vector Machine and Decision Tree algorithms are used for juxtaposition.

3.3.1. Support Vector Machine (SVM)

SVM is a type of machine learning model that can be used to solve classification and regression problems. In the case of a classification problem, the hyperplane \( g(x) = \Theta^T x + y + B \) is built using training data and functions as a decision boundary for determining the class of a data point (a multidimensional feature vector) where \( B \) is the bias and \( \Theta \) is the weight vector. In the case of binary classification, SVM finds the closest vectors (data points) of two classes in order to generate a margin and those vectors are known as support vectors. Margin is calculated as the perpendicular distance between the lines passing through the support vectors and is represented by \( \frac{1}{\|\Theta\|^2} \). The main goal is to build an optimised SVM model that will produce an optimal hyperplane with the highest margin.

Support Vector Machines attempts to maximise \( \|\Theta\|^2 \) the misclassification errors by using a set of slack variables denoted by \( \epsilon \), where \( x = 1, 2, \ldots, n \) and a cost parameter \( C \). It is expressed as

\[
\min_{\Theta, B, \epsilon} \frac{1}{2}\|\Theta\|^2 + C \sum_{x=1}^{N} \epsilon_x \\
\text{s.t.} \begin{cases} 
O_x(\Theta x + B) \geq 1 - \epsilon_x \\
\epsilon_x \geq 0, \quad x = 1 \ldots N
\end{cases}
\]
3.3.2. Decision Tree (DT)

The Decision Tree algorithm is a supervised machine learning algorithm that is used to solve classification problems. It is applicable to both specific and continuous input and output variables. Based on the most important splitter in the input variable, it divides the population or sample into two or more homogeneous sets called sub-populations. It predicts and classifies using nodes and internodes. Root nodes categorise instances based on their features. The root node may have two or more branches, whereas the leaf node represents classification. At each stage, the decision tree selects each node by weighing the highest information gain among all attributes.

Information gain, abbreviated as IG, is a statistical property that quantifies how well a given attribute separates training examples based on their target classification. The goal of building a decision tree is to find an attribute with the highest information gain and the lowest entropy. Entropy decreases as a result of information gain. Based on the attribute values provided, it computes the difference between the dataset's entropy before and after splitting, as well as the average entropy after splitting. The information gain decision tree algorithm is used by the ID3 (Iterative Dichotomiser) decision tree algorithm. IG is denoted mathematically as:

\[
IG(N,I) = \text{Entropy}(N) - \text{Entropy}(N,I) \tag{2}
\]

Where Entropy is a measure of how random the information being processed is. The greater the entropy, the more difficult it is to draw any conclusions from the information. Entropy can be calculated as

\[
\text{Entropy}(C) = \sum_{i=1}^{n} -p_i \log_2 p_i \tag{3}
\]

\[
\text{Entropy}(N,I) = \sum_{c \in I} \text{Prob}(c) \times \text{Entropy}(c) \tag{4}
\]

4. Proposed Approach

Figure 1 depicts the general structure of the proposed diagnosing procedure. After obtaining the dataset, the pre-processing method is used to remove missing values or replace them with appropriate data. Then, using the cleaned dataset, a genetic algorithm is used to find the best combination of features that provides the highest correlation between the features and the targets.

![Figure 1 Flow chart of the Proposed Work](image-url)
Following the application of GA, the each prediction models is applied individually to the training dataset in order to learn how to recognise the targets. After each iteration the models are evaluated using test dataset. The trained models are then compared to assess the performance.

5. Results and Discussion

The above-mentioned strategy was applied to five different types of datasets. Table 2 depicts the parameters used by Genetic Algorithm for feature selection. The confusion matrix was used to calculate metrics of accuracy, specificity, and sensitivity for evaluating classification accuracy. They are calculated as

\[
Acc = \frac{True\_Positive + True\_Negative}{True\_Positive + True\_Negative + False\_Positive + False\_Negative} \times 100
\]

(5)

\[
Sen = \frac{True\_Positive}{True\_Positive + False\_Negative} \times 100
\]

(6)

\[
Spec = \frac{True\_Negative}{True\_Negative + False\_Positive} \times 100
\]

(7)

| S. No | Genetic Algorithm Parameters | Values |
|-------|-----------------------------|--------|
| 1     | Population size             | 20     |
| 2     | Number of offsprings generated | 14    |
| 3     | Number of mutants           | 6      |
| 4     | Probability of Crossover operator | 0.7   |
| 5     | Mutation probability        | 0.02   |
| 6     | Maximum number of iteration | 10     |

Table 2. Parameters used in the proposed work

Table 3 shows the relevant features chosen using a genetic algorithm on different datasets. Among them in WBC-PD more number of features i.e. 19 features are selected by GA. In BPD and ISPY 1 few number of features are selected.

| S. No | Datasets   | Number of features |
|-------|------------|--------------------|
| 1     | WBC-DD     | 12                 |
| 2     | WBC-OD     | 7                  |
| 3     | WBC-PD     | 19                 |
| 4     | BPD        | 3                  |
| 5     | ISPY 1     | 7                  |

Table 3. Number of features selected in the dataset

Figure 2 shows the accuracy comparison of SVM and DT in percentage with and without feature selection for different datasets.
Figure 2 Accuracy comparisons of datasets

Figure 3 shows the specificity comparison of SVM and DT in percentage with and without feature selection for different datasets.

Figure 3 Specificity Comparison of datasets

Figure 4 shows the sensitivity comparison of SVM and DT in percentage with and without feature selection for different datasets.
Both SVM and DT show an increase in accuracy. Decision Tree performed better in ISPY 1 dataset than in other four datasets. All other datasets show improved accuracy in the SVM model, which is greater than the DT accuracy rates. SVM with feature selection has nearly higher sensitivity rates than other datasets in terms of sensitivity. In terms of specificity, the outcomes are the same. As a result of this research, we can conclude that the SVM model with GA-selected features is capable of accurate breast cancer prediction. Table 6 compares the accuracy of the proposed work to that of the considered existing papers, demonstrating that feature selection with the proposed prediction models outperforms the others.

| Dataset   | ANN [6] | PS-classifier [6] | GA-Classifier [6] | This work |
|-----------|---------|-------------------|-------------------|-----------|
|           |         |                   |                   | SVM       | DT        |
| WBC-DD    | 95.6    | 95.8              | 95.5              | 98.16     | 97.44     |
| WBC-OD    | 96.2    | 96.1              | 95.5              | 97.13     | 85.1      |
| WBC-PD    | -       | -                 | -                 | 85.1      | 77.22     |
| BCDD      | -       | -                 | -                 | 78.95     | 88.62     |
| ISPY 1    | -       | -                 | -                 | 96.94     | 96.59     |

Table 4 Comparision with Existing Technique

6. Conclusion

In this study, we compared the accuracy of two machine learning algorithms, Support Vector Machines and Decision Tree, in predicting breast cancer using features selected by a genetic algorithm. To compare the performance of the proposed work five different datasets are used. The features for each dataset are chosen using a Genetic Algorithm to improve machine learning prediction accuracy. Both SVM and DT show an increase in accuracy. SVM-GA achieved better accuracy than DT-GA. Future work may involve the use of other machine learning algorithms or the application of other population-based feature selection meta-heuristics and comparing their results to those obtained by the proposed approach.

References

1. Chen H.L, Huang C. C, Yu X. G, Xu X, Sun X, Wang G, Wang S J, “ An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-Nearest Neighbour approach”, Expert Systems with Applications, 40(1), 263-272, 2013.
2. Odajima K, Pawlovsky A. P, “A detailed description of the use of kNN method for breast cancer diagnosis”, Proceedings of the 7th International Conference on Biomedical Engineering and Informatics (BMEI), 2014

3. Lynch C. M, Abdollahi B, Fuqua J, D, de Carlo A. R, Bartholami J. A, Balgemann R. N, …. Frieboes H. B, “Prediction of lung cancer patient survival via supervised machine learning classification techniques”, International Journal of Medical Informatics, 108, 18, 2017.

4. Sapitani N. W. P, Wulan R, Lestari M, “Breast cancer detection using data mining classification methods”, Journal of Mathematics, J(1), 185-191, 2017

5. Hashi E. K, Zaman S. U, Hasan R, “An Expert clinical decision support system to predict disease using classification techniques”, Proceedings of the 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2017.

6. Aalaei S, Shahraki H, Rowhanimanesh AR, Eslami S, “Feature selection using Genetic Algorithm for breast cancer diagnosis: Experiment on three different datasets” Iran J Basic Med Sci. 19: 476-482.

7. Pradeep K. R, Naveen N. G, “Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4.5, and Naïve Bayes algorithm for healthcare analytics”, Procedia. Computer Science, 132, 412-420, 2018.

8. Alharbi A, “An automated computer system based on genetic algorithm and fuzzy systems for lung cancer diagnosis”, International Journal of Nonlinear Sciences and Numerical Simulation, 19(6), 583-594, 2018.

9. Cherif W, “Optimization of k-NN algorithm by clustering and reliability coefficients: Application to breast cancer diagnosis”, Procedia Computer Science, 127, 293-299.

10. Abe N, Kudo M, Toyamma J, Shimboo M, “A divergence criterion for classifier-independent feature selection”, Adv Pattern Recognit: Springer: 668-676, 2000

11. Kohavi R, John GH, “Wrappers for feature subset selection”, Artif Intell 97: 273-324, 1997

12. Aalaei S, Shahraki H, Rowhanimanesh AR, Eslami S, “Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets”, Iran J Basic Med Sci, 19:476-482, 2016

13. Joshi A and Mehta A, “Analysis of k-Nearest Neighbour technique for breast cancer disease classification”, International Journal of Recent Scientific Research, 9(4), 26126-16130, 2018

14. Akben S. B, “Early-stage chronic kidney disease diagnosis by applying data mining methods to urinalysis, blood analysis, and disease history”, IRBM, 39(5), 353-358.

15. Guyon I, Elisseeff A, “An introduction to variable and feature selection”, J Mach Learn Res, 3: 1157-1182.

16. Chic-Lin Chi, Nick Street W and William H. Wolberg,” Application Saof Artificial Neural Network based survival analysis on two breast cancer datasets”, AMIA Annu Symp Proc. 130-134, 2007

17. West D, Mangiamelli P, Rampal R and West V, “Ensemble strategies for a medical diagnosis decision support system: A breast cancer diagnosis application”, European Journal of Operational Research (162), 532-551, 2013.

18. Huang C. L, Liao H. C, “Prediction model building and feature selection with support vector machines in breast cancer diagnosis”, Experts Systems with Applications, 2008.

19. Rong L, “Diagnosis of Breast Tumour using SVM-kNN classifier”, Intelligent Systems, 2010.
20. Saranya N, & Kavi Priya S. (2021). Deep Convolutional Neural Network Feed-Forward and Back-Propogation (DCNN-FBP) algorithm for predicting heart disease using internet of things. International Journal of Engineering and Advanced Technology (IJEAT), 11(1), 83-87, https://doi.org/10.35940/ijeat.A3212.1011121