A Survey on the Recent Feature Selection Methods used for the Detection of Breast Cancer

Ellen Daniel,
Information Technology,
Puducherry Technological University,
Puducherry, India.

Manigandan G.,
Information Technology,
Puducherry Technological University,
Puducherry, India.

Sushma T.,
Information Technology,
Puducherry Technological University,
Puducherry, India.

E. Thamizhselvi,
Research Scholar,
Department of Computer Science and Engineering,
Puducherry Technological University,
Puducherry.

Dr. V. Geetha,
Associate Professor,
Department of Information Technology,
Puducherry Technological University,
Puducherry.

***Abstract* – This paper gives a brief survey about some of the feature selection methods used for the detection of breast cancer. Feature selection is the process of selecting high priority features from a set of features thereby reducing the computational time and improving the performance of the model. The selection of features is based on whether a feature is relevant to the problem being solved by the said model. It reduces redundancy and improves the predictive power of the model. Breast cancer is a high mortality disease with increase in death rate if detected at later stages. The main objective of this paper is to identify the different feature selection methods utilized for the detection of breast cancer in the recent year.**

***Keywords* – Feature selection, breast cancer, detection, hybrid model, performance.**

I. INTRODUCTION

Feature is a property that is measured in a process. There are hundreds of feature domains present in an application. Feature selection helps us to understand the dataset and reduce the computational requirements thereby improving the predictor performance. Feature selection uses supervised and unsupervised learning techniques. In feature selection, a subset of variables is selected from a from the given input that efficiently describes the input data while reducing the effects from noise or any irrelevant variables but can still provide better prediction results. Therefore feature selection is useful for visualization such as image analysis, signal processing and information retrieval. In this paper, some of the recent feature selection methods that are applied for the detection of breast cancer are surveyed.

II. MEDICAL DIAGNOSIS

Feature selection is a preprocessing technique that identifies the key features of a problem. Therefore it is applied for a wide range of problems especially in the field of medical diagnosis. This is because the goal is to reduce the dimensionality of the problem without risking the accuracy, at a lesser cost. Feature selection is applied specifically for medical imaging as the medical datasets have a larger number of features but a fewer samples of a particular disease. In this paper, the focus is on the application of feature selection to breast cancer dataset.

III.BREAST CANCER

Breast cancer is the second main cause of cancer related deaths in women. It is the growth of malignant, cancerous lumps from the breast cells. The death rate increases when it is detected at a later stage. So, detecting at an earlier stage may save a life. Breast cancer is detected by using a biopsy where a tissue is removed and is studied under a microscope for any abnormalities. The histopathologist makes the diagnosis based on his observations from the biopsy. As there is a possible chance for human error, there may happen a misdiagnosis. With recent advancements in the image processing and machine learning techniques, there is the possibility to develop reliable pattern recognition based systems to improve the quality of diagnosis. Feature selection is applied for the image dataset of breast cancer to obtain better prediction of the diagnosis of the disease. This helps in the early diagnosis of the breast cancer which can be treated, thereby improving the survival rate of the breast cancer patients. **Table 1**. shows the inputs of the breast cancer dataset and their description.

Input	Description
Clump thickness	Mono or multi layered cells
Uniformity of cell size	Consistency in size of cells
Uniformity of cell shape	Equality of cell shapes and marginal variances
Marginal adhesion	How much cells on the outside of epithelial stick together
Single epithelial cell size	If epithelial cells are significantly enlarged

Bare nuclei	Proportion of the number of cells that are not surrounded by cytoplasm to those that are
Bland chromatin	Texture of nucleus in the range fine to coarse
Normal nucleoli	Whether the nucleoli are small and barely visible or larger, more visible and more plentiful
Mitoses	Level of mitotic activity
Class	2 for benign 4 for malignant

Table 1. Breast Cancer Dataset

IV. FEATURE SELECTION ALGORITHMS

In this paper, some of the feature selection algorithms that were used for the detection of breast cancer in the recent year with high performance percentage are surveyed. They are Integrated Artificial Immune System and Artificial Bee Colony based breast cancer diagnosis (IAIS-ABC-CDS), Teaching Learning Based Optimization and Salp Swarm Algorithm (TBLO-SSA), Stacking Based Ensemble Framework, Relief Algorithm, Principal Component Analysis and Information Gain, Multi Population based Particle Swarm Optimization (MPPSO), Three-Stage Feature Selection and Twice-Competitional Ensemble (TSFS-TCEM), Correlation Coefficient Function, Grey Wolf Optimizer, Ant Colony Optimization and Tabu search with Fuzzy Rough set for Optimal feature selection (ACTFRO) and Genetic Algorithm and Tabu search with Fuzzy Rough set for Optimal feature selection (GATFRO). Each algorithm gives a better performance based on the dataset and the feature subset selected. The **Figure 1.** shows the graph of the accuracies achieved by the feature selection methods in this survey.

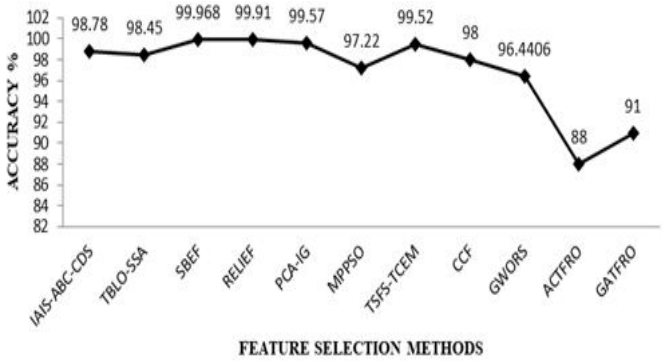

Figure 1. Comparison of the accuracies of the feature selection methods in this survey

A. IAIS-ABC-CDS

Integrated Artificial Immune System and Artificial Bee Colony based breast cancer diagnosis (IAIS-ABC-CDS) uses a wrapper method. This prevents any addition of primary statistical process on a utilized dataset. Here, Artificial Immune System and Artificial bee Colony are integrated to find the cancerous properties in the breast. Feature selection is used to compress the feature set and eliminate the complexity of computation. The number of hidden nodes that is used in the process of optimization of MLP is minimum when feature selection is applied. It is efficient and optimal in determining the highest potential feature subset as the leveraging of artificial immune system and artificial bee colony optimization process merits ensures a rapid and precise diagnosis of breast cancer. But there is no formulation of an integrated ant colony optimization and artificial bee colony approach which results in the onlooker bee of the artificial bee colony to be exploited by the employee. This decreases the explorative capability of the artificial bee colony optimization and de-enhanced intensification process of ant colony optimization. [1]

B. TEACHING-LEARNING-BASED OPTIMIZATION AND SALP SWARM ALGORITHM

Teaching-Learning-Based Optimization algorithm has two phases: teacher and learner. The best solution is taken as the teacher in the teacher phase whereas the knowledge is acquired by the learner from other learners by random interaction among them. Salp Swarm Algorithm solves optimization problems by dividing the population as leader and followers. The front salp is taken as the leader and the remaining as followers. A hybrid model of Teaching-Learning-Based Optimization and Salp Swarm Algorithm (TBLO-SSA) is applied for feature selection. This utilizes very few parameters for the implementation and achieves fast convergence and high accuracy with minimal computation. The selection of a particular value is based on the probability value. The optimal features are selected on the basis of the objective function. The features chosen by TLBO-SSA is used for the characterization of the masses as benign and malignant. The detection of breast cancer diagnosis is efficient and robust. Yet this method has to be tested with datasets of large number of features and samples. [2]

C. STACKING-BASED ENSEMBLE FRAMEWORK

The four machine learning algorithms SVM, k nearest neighbours, Naïve Bayes and perceptron are combined to form a new model called blending or stacking. Stacking needs multiple base learning algorithms to create a sequence of models that gives a new meta dataset. Five feature selection techniques univariate selection, extra tree classifier, correlation matrix with heatmap, recursive feature elimination and random forest are applied to obtain the dataset with reduced features. Univariate selection selects features with strong relationship to the output. Extra tree classifier selects features with higher scores. Correlation sees if a feature has a connection to the target variable and the heatmap spots the feature with major relevance. Recursive feature elimination eliminates the features that are the weakest. Random forest merges the techniques of filter and wrapper. Stacked model's accuracy of breast cancer detection is higher when compared to the accuracy of each ML algorithms separately. But this works the best only when the sub-models are combined together skilfully. Also, the breast cancer datasets are not from different repositories. [3].

D. RELIEF ALGORITHM

Supervised (Relief Algorithm) and unsupervised (Autoencoder, PCA Algorithms) are used to obtain the related features from the dataset. The relief algorithm does feature selection by filter mechanism. It is a distance based filter as the features are ranked on the basis of creation of features that separates the classes. The relief algorithm selects features of high weight and features of low weight are removed from the dataset. The autoender generates output by the application of different transforms to a features set. The PCA linearly transforms correlated features into smaller number for the construction of appropriate features. The results of the Relief Algorithm is more accurate than the unsupervised algorithms. The performance of the SVM classifier is the highest when the features are selected by the relief algorithm. This provides a reliable diagnosis of breast cancer and can easily be incorporated in e-healthcare systems. But the other types of classifiers are not implemented for the features selected by the relief algorithm. [4]

E. PRINCIPAL COMPONENT ANALYSIS – INFORMATION GAIN

Principal Component Analysis generates the principal components by determining the correlation among the features in the dataset. This enables the representation of data with lesser number of variables. Information Gain evaluates a feature set that is selected to find the most relevant features. It selects the final feature set by using a set threshold. [5] discussed that Biomedical and anatomical data are made simple to acquire because of progress accomplished in computerizing picture division. More research and work on it has improved more viability to the extent the subject is concerned. A few tech- niques are utilized for therapeutic picture division, for example, Clustering strategies, Thresholding technique, Classifier, Region Growing, Deformable Model, Markov Random Model and so forth.

F. MULTI POPULATION BASED PARTICLE SWARM OPTIMIZATION

For high dimensional data, Particle Swarm Optimization (PSO) gets stuck in the local optima. So an MPPSO is proposed here. The relief algorithm calculates the distance between each feature and the target to rank the features. In multi population based particle swarm optimization (MPPSO), the multi population starts with the initial solutions that are generated by random and relieff based initializations and searches the solution space simultaneously using both the populations. [6] discussed about the combination of Graph cut liver segmentation and Fuzzy with MPSO tumor segmentation algorithms. The system determines the elapsed time for the segmentation process. The accuracy of the proposed system is higher than the existing system. The algorithm has been successfully tested in multiple images where it has performed very well, resulting in good segmentation.

G. TSFS-TCEM

The three-stage feature selection method is of three stages. First, FC-FDR method and information gain are used for feature filtering. Second, redundancy is eliminated and dimensionality is reduced. Third, the optimal is selected by multiple repeatability tests. In twice-competitional ensemble method, the competition is between the predictive model sets and the optimal model. The final model is selected by the optimal model from the predictive model sets. The three-stage feature selection and twice-competitional ensemble (TSFS-TCEM) learning method is a novel hybrid feature selection and ensemble learning framework. [7] discussed that Automatic liver tumor segmentation would bigly influence liver treatment organizing strategy and follow-up assessment, as a result of organization and joining of full picture information. Right now, develop a totally programmed technique for liver tumor division in CT picture.

H. CORRELATION COEFFICIENT FUNCTION

In [8], breast cancer is detected at its early stage by using machine learning techniques. Correlation-based Feature Selection (CFS) uses ranking method to select the features. PCA removes the extra variables, thereby solving the problem of overfitting. MLP generates a correlation function that is the product of the combination of the two covariance variables. The selection of features is done on the basis of correlation coefficient function in order to obtain an efficient selection of attributes. This deploys multi filtering methods. The significant features are then evaluated by the attribute evaluator using the PCA and the attributes of the highest rank are used for the classification. MLP classifier gives the highest accuracy of classification. This model is efficient in the detection of benign and malignant class. But the research is limited to only a specific dataset here.

I. GREY WOLF OPTIMIZER

The grey wolf optimization algorithm generates the best solution in each iteration till the fulfilment of the stopping criterion (maximum number of iterations). By various combinations, the fitness of the features acquired is calculated. A computer aided diagnosis (CAD) based on the grey wolf optimizer (GWO) is utilized for the detection of breast cancer [9]. GWO with rough set theory derives the appropriate features from an extracted feature set. A hybrid of GWO and Rough Set (GWORS) identifies the significant features from the extracted mammogram images dataset. This method encloses image processing, mass segmentation, feature extraction and classification. The GWORS eliminates the redundancy features in the extracted dataset. The dependency of conditional features is calculated based on the decision feature. The performance is one of the finest for the identification of normal or abnormal data. The accuracy of classification is increased and the high computational cost is decreased. The amount of attribute that is anticipated to a wolf pack of a predefined size improves the ac tness of search with lesser space for global optimal. The pr lem is, a suitable search space has to be defined for an outcome of better prediction with no familiarity about the

dataset. The subsets cannot be chosen without the description of the attribute size.

J. ACTFRO AND GATFRO

In ACTFRO algorithm, the search for the relevant features is done by ant colony optimization. Then the fuzzy rough set evaluates the subset of features. In GATFRO algorithm, the most fitted feature in each generation is chosen by the selector operator in the genetic algorithm. The fuzzy rough set evaluates the fitness function of the subset of features. The two hybrid global and local optimal feature selection algorithms Ant Colony Optimization and Tabu search with Fuzzy Rough set for Optimal feature selection (ACTFRO) and Genetic Algorithm and Tabu search with Fuzzy Rough set for Optimal feature selection (GATFRO) are used to predict breast cancer [10]. The optimal features subset is selected by these algorithms. The ACTFRO selects features that guarantee a consistent convergence and diversification that helps to solve a combinational optimization problem. The GATFRO selects features that are conceptually straightforward and diversified that helps to solve nonlinear optimization problems. The features are then evaluated with fuzzy rough evaluation function. As the selection of global and local optimal features are robust, the level of accuracy of classification is improved with lesser computational time. A credible missing linkage is not found.

V. CONLUSION

In this paper, a survey about the different techniques of the feature selection used for the detection of breast cancer is presented. This survey gives a clear view of the recent techniques available for the feature selection process especially for breast cancer detection. The performance of the feature selection models mentioned in this survey is discussed. From this survey, it is understood that there is scope for the researchers to develop better feature selection algorithms from the existing models to solve complex problems as the models surveyed here are the betterment of the previously available methods for feature selection. A possible feature selection method based on the biological markers could also be developed for the detection of breast cancer in the future.

REFERENCES

[1] Punitha .S, Fadi Al-Turjman and Thompson Stephan, “An Automated Breast Cancer Diagnosis using Feature Selection and Parameter Optimization in ANN”, Elsevier, January 03, 2021.

[2] Shankar Thawkar, “A Hybrid Model using Teaching-Learning-Based Optimization and Salp Swarm Algorithmfor Feature Selection and Classification in Digital Mammography”, Springer, January 03, 2021.

[3] Vikas Chaurasia and Saurabh Pal, “Stacking-Based Ensemble Framework and Feature Selection Technique for the Detection of Breast Cancer”, Springer, February 02, 2021.

[4] Amin Ul Haq, Jian Ping Li, Abdus Saboor, Jalaluddin Khan, Samad Wali, Sultan Ahmad, Amjad Ali, Ghufran Ahmad Khan and Wang Zhou, “Detection of Breast Cancer through Clinical Data using Supervised and Unsupervised Feature Selection Techniques”, IEEE Access, February 09, 2021.

[5] Christo Ananth, S.Aaron James, Anand Nayyar, S.Benjamin Arul, M.Jenish Dev, “Enhancing Segmentation Approaches from GC-OAAM and MTANN to FUZZY K-C-MEANS”, Investigacion Clinica, Volume 59, No. 1, 2018,(129-138).

[6] Christo Ananth, D.R.Denslin Brabin, “ENHANCING SEGMENTATION APPROACHES FROM FUZZY K-C-MEANS TO FUZZY-MPSO BASED LIVER TUMOR SEGMENTATION”, Agrociencia, Volume 54, No. 2, 2020,(72-84).

[7] Christo Ananth, D.R.Denslin Brabin, “ENHANCING SEGMENTATION APPROACHES FROM GAUSSIAN MIXTURE MODEL AND EXPECTED MAXIMIZATION TO SUPER PIXEL DIVISION ALGORITHM”, Sylwan, Volume 164, No. 4, 2020,(15-32).

[8] V. Nanda Gopal, Fadi Al-Turjman, R. Kumar, L. Anand and M. Rajesh, “Feature Selection and Classification in Breast Cancer Prediction using IOT and Machine Learning”, Elsevier, April 18, 2021.

[9] B. Sathiyabhama, S. Udhaya Kumar, J. Jayanthi, T. Sathiya, A. K. Ilavarasi, V. Yuvarajan and Konga Gopikrishna, “A Novel Feature Selection Based on Grey Wolf Optimization for Mammogram Image Analysis”, Springer, May 24, 2021.

[10] L. Meenachi and S. Ramakrishnan, “Metaheuristic Search Based Feature Selection Methods for Classification of Cancer”, Elsevier, June 22, 2021.