

A Novel approach to classifying breast cancer using daggers and ensemble algorithms

¹R.Bhavani²V.Prakash³K.Chitra⁴J.Agalya

¹Research Scholar, SCSVMV University, Enathur, Kanchipuram and Assistant Professor, SASTRA Deemed To Be University, Kumbakonam.

²Assistant Professor, SASTRA Deemed To Be University, Thanjavur.

³Assistant Professor, Govt. Arts and Science College for Women, Melur.

⁴PG Student, SASTRA Deemed To be University, Kumbakonam.

Abstract

In recent days, Breast cancer has become a common disease around the world. Many tools have been successfully implemented to diagnose this disease. In the proposed system, a novel technique, the knowledge-based system is employed. A knowledge-based system (KBS) is used to solve a complex problem in a computer program. Several data mining techniques are used in the medicinal field; classification is one of the important methods for disease diagnosis. Decision tree algorithms are used to classify the breast cancer tumor easily. Ensemble algorithm is used to randomly split the dataset for efficient classification. In Parallel the Dagger's algorithms read the dataset and compare the range with the fuzzy rule. Various datasets are used to improve the prediction accuracy of breast cancer using the fuzzy rule-based technique.

Introduction

Breast cancer is one of the dangerous diseases around the world. In starting stage the normal cell grows into particular size in its life cycle, then a new cell is formed by dividing the existing cell and the old cell will die. The behavior of cancerous cell differs from the normal cell. A normal cell changed into cancerous when a mutation occurs in DNA. Major symptoms of breast cancer are painless breast lump, change in the breast, fluid coming from the nipple, dimpling of the skin and red scaly patch of skin. Sometimes breast cancer is painful and soft. Some causes of breast lump are infection and injuries. Sometimes breast cancer did not reveal any kind of symptoms. A Breast lump is also called as a tumor. There are two classifications of cells. A non-cancerous tumor is known as Benign and the cancerous tumor is known as Malignant. Benign is also the abnormal growth of cells inside the breast,

but it is not life-threatening. It can increase the risk of breast cancer. Early stage detection of cancer is useful for timely treatment and prevention. In this stage, the tumor is small, easy to detect and curable.

Data mining is used to summarize the data in a useful manner, structured order, and understandable. In many fields, data mining is used to make the prediction for future perspective. In the medical field, data mining is used to overcome some issues such as incomplete, dealing with noisy, intensive and heterogeneous data. Data mining has methods to discover the hidden patterns in the large datasets such as machine learning, statistics, artificial intelligence and database systems. Clustering, classification and association mining are common techniques for predictive analysis. The modern medical field has stored a large amount of information in the medical database. It is a necessary need to extract the useful and important details from the medical database. An extensive amount of data and knowledge stored in the medical database. The increase of data volume is difficult to extract the useful information from the database. Data mining has been used to develop specialized tools for accessing, knowledge discovery, data analysis and effective use of stored data and knowledge.

In data mining, classification technique is used to assigns items in a collection to categorize the target. The goal of this technique is to accurately predict the data and identify the level in various sectors. Here the classification is used to classify the breast cancer tumor is a cancerous tumor or a non-cancerous tumor. Because breast cancer is one of the deadly diseases around the world. Decision tree algorithm is part of the supervised learning algorithms. The level of understanding in decision tree algorithm is easy when compared with the classification algorithms. Ensemble learning algorithm is used here to classify the results from the given trained datasets. A fuzzy rule is based on the fuzzy logic is a form of a valuable logic, with variables in constant values 0 to 1. It is employed to find the difference between absolute true and false. By contrast, the absolute value of Boolean logic variables may be 0 or 1. This type of fuzzy rule method is applied to identify a tumor is a cancerous tumor or a non-cancerous tumor from the trained datasets. Dagger algorithm is used for the dataset aggregation process. This algorithm is read the attribute values from the trained dataset and compared with the constructed fuzzy rule to identify the range of the values. Ensemble algorithm is randomly split the dataset into small chunks. Then it processed to classify the dataset into target class, benign and malignant. Benign is known as non-

dangerous and malignant is dangerous, this way the disease diagnosis process is worked in this paper.

Existing system:

In the world, breast cancer becomes a second non-skin cancer disease which was the reason of cancer death. In western countries, 10 percent of women have suffered from breast cancer disease and worldwide millions of women suffered from this life-threatening illness [1]. In 2017, the report of American cancer society was second important cause of death for women is breast cancer and woman get older then the risk of breast cancer increases significantly and statistics released by the breast cancer society of Canada in 2015, which is every week in Canada 481 females will be detected with this cancer[2], [3]. Several kinds of research are going around the world about to detect the reason for breast cancer and for prevention. The mortality rate is reduced by using a preventive measure[3]. Breast cancer affects 1.3 million women worldwide every year. fourteen percentage of cancer-related death, a cause of breast cancer. The number of people affected by breast cancer has increased simultaneously in the coming years, so breast cancer will remain a dangerous disease[4].

Breast cancer is widespread cancer in women everywhere in the world. Since 2011, approximately 1.3 million women diagnosed with breast cancer around the world. Every year 465,000 women had died by the breast cancer. In China, 214,360 females die from breast cancer in 2008. In 2021, it has been valued the number of Chinese women diagnosed with breast cancer will reach 2.5 million. In 2017, a report published by the Chinese national cancer center, breast cancer is common cancer disease in Chinese women. Lung, stomach, liver, oesophageal and colorectal cancer, after this breast cancer is the sixth main killer in small and medium cities, with mortality rate 8.44% and 9.59% respectively[5]. The study of World Health Organization(WHO), breast cancer is topmost cancer amid women equally in developing and developed countries[6]. Mammography is the generally accepted and used screening test for breast cancer in early[7].

In classification there are two schemes are applied image classification and patient classification[8]. Classification And Regression Trees(CART)algorithm with fuzzy rules to the classification of breast cancer[1]. Breast cancer is classification is to identify a tumor is a benign or malignant using training samples. Classification is achieved by a novel GNRBA

method. It is based on sparse representation[3]. Method based proportional sample method (PBS). Identifying ideological data representation. First, the input database a binary relationship is a normal relationship. This relationship binary varieties are modified, then the rate of each form is calculated and replaced to the original data[9]. For diagnosis the breast cancer using case-based system(CBR) system[10]. A deep Convolutional Neural Networks(CNN) to get a discriminative description of breast tissues for breast mass classification[11]. For classification purpose used 'fuzzy quantification subethood-based algorithm'[12]. Bayesian Network(BN) modeling approach and the K2 learning algorithm are used for breast cancer diagnosis[5]. Classification algorithm j48 tree is using, it is a small modification of C4.5 algorithm[6]. Non-Subsampled Contourlet Transform (NSCT) and Super-Resolution (SR) algorithms are used for breast cancer detection and classification[7]. The two techniques are used for classification of breast tissue, radial basis function network(RFBN) and backpropagation learning algorithm(BPNN)[13].

Proposed system:

In this proposed system, mainly ensemble algorithm is used to split the dataset. And ensemble is a decision tree algorithm, it classifies the dataset into cancer and non-cancer. There are ten attributes in the dataset nine attributes are used to identify the disease. Initially, the dataset is pre-processed and empty cell rows and also irrelevant data cell rows are removed in the noise removal step. In the breast cancer dataset attribute value within 1-10, any other value, punctuation marks, and the string will be considered as an irrelevant data. And the pre-processed dataset, which is processed in the proposed method to classify the breast cancer. Breast cancer becomes the second most dangerous among women after the lung cancer. So, classification is important for the treatment of cancer patient to cure the disease in the beginning stage. Because breast cancer is a deadly disease and mostly affects the women, we need more awareness about that disease. And rural area women don't know about breast cancer and they are careless in health problems. In data mining, many techniques are used in medical field. And also, it plays important role in the automatic tool for disease diagnosis for the patient in around the world. In medical camp are related to works they have several patient data are mostly it has converted into the datasets for diagnosis process. In that time the automated tool is used to detect the patient results and easily arranged and listed the results. For that reason, data mining plays a major role in the medical field for disease diagnosis. And many data mining methods playing important role in the medical field like classification, clustering, and association mining. These methods are mostly used for the

predictive analysis in any field like medical, business, and education etc. Here classification technique is used in the medical field for detection of disease.

The fuzzy rule is used for the classification process in this proposed method. The Boolean method true or false is used here for detect cancer stage is benign or malignant. Then the fuzzy logic divides the dataset into positive or negative. And this method has set with the range in the fuzzy set for breast cancer diagnosis and the range for each attribute will be set in the range construction in fuzzy rule phase. It classifies the results as benign and malignant respectively non-cancerous tumor and cancerous tumor. The algorithms are mainly focused on the result classification of this proposed method. The ensemble has bagging and boosting method for easily process the large data set. In working they randomly split the dataset into small chunks after it has to work in the fuzzy rule and classified the data for disease diagnosis. The architecture of the proposed system is given in fig 1.1.

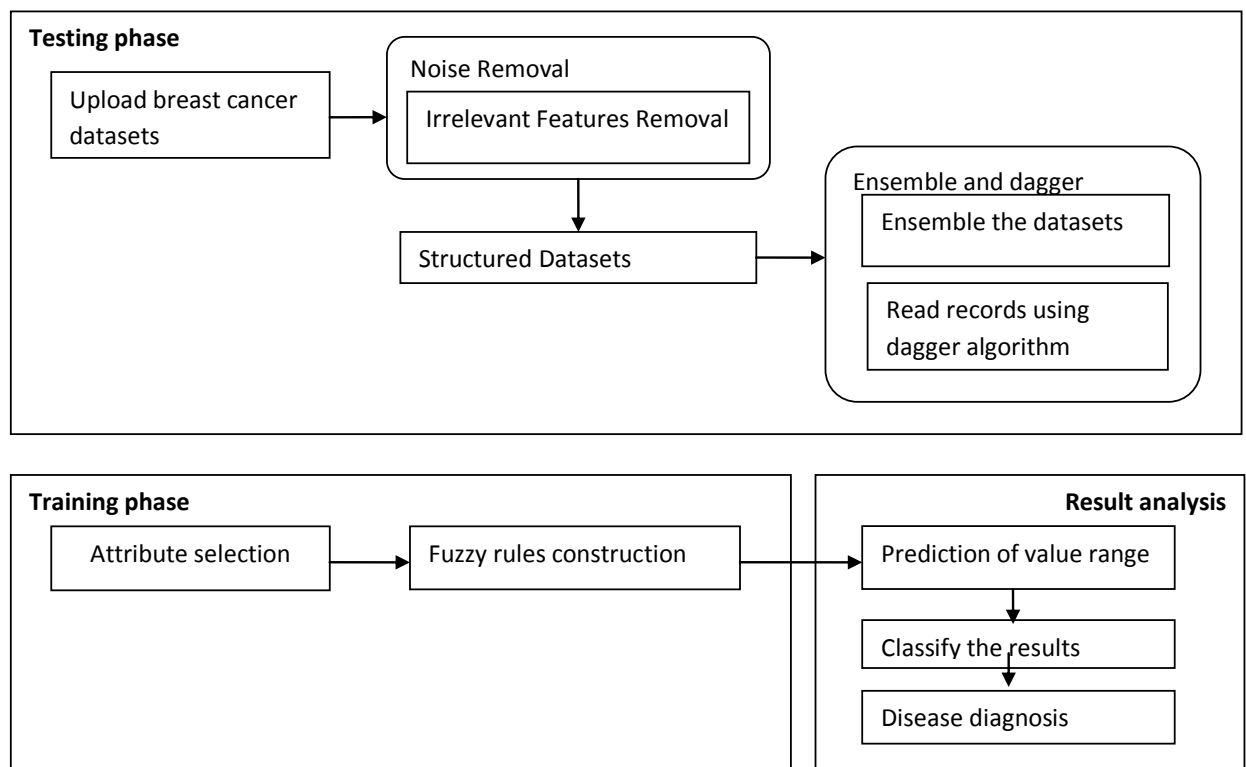


Fig 1.1: Architectural design of proposed system

This method is referred to in this pseudo code:

Pseudocode for Ensemble Algorithm

```

Input: P=Benign

      N=Malignant

      T=Number of dataset

Result: Ensemble classifier = Trained ensemble

Begin

  For i=1 to T do

    Fi = randomly selected features subset

    Pi =P(Fi) //feature sub-spacing

    //for within-class imbalance

    Pi =OVERSAMPLE(Pi)

    //for between-class imbalance

  Repeat

    Randomly sample Ni belongs to N

  Until |Ni| = |Pi|;

  N= N – Ni

  Ni =Ni(Fi) //Feature sub-spacing

  tree-i = train decision tree using Pi and Ni

  return ensemble = 1/T sigma i=1 to T tree-i

```

The fuzzy rule is constructed as a table in the database and then in dagger algorithm is compared the data set values. In this comparison user has found who are all has breast cancer. Each attribute has the unique range of two stages, benign and malignant. In the fuzzy rule, table start range and end range for each attribute will be mentioned and stored in the database. In training data set attributes are used in the disease diagnosis except one is patient identity number. The dagger algorithm is applied inside the ensemble algorithm for automatically classified the disease is a non-cancerous tumor and classified tumor. Dagger read the range of, each attribute in the dataset, the pseudo code follows as:

Pseudo code for Dagger Algorithm

```

Initialize  $D \leftarrow \emptyset$ 

Initialize  $T^1$  to any policy in  $\Pi$ 

for  $i = 1$  to  $N$  do

  Let  $T_i = B_i T^* + (1 - B_i)T^{i-1}$ .

  Sample  $T$ -step trajectories using  $T_i$ .

  Get dataset  $D_i = \{(s, T^*(s))\}$  of visited states by  $T_i$ 
  and actions given by expert.

  Aggregate datasets:  $D \leftarrow D \cup D_i$ 

  Train classifier  $T^{i+1}$  on  $D$ .

end for

Return best  $T^i$  on validation.

```

In the classification method, some important techniques are decision tree method, memory-based reasoning, support vector machine, neural networks, naïve Bayes and Bayesian belief network, and rule-based method. In this proposed method using two techniques, one is rule-based and another one is decision tree method. Using the both the proposed method worked efficiently and easily. Purpose of classification is used here for disease diagnosis and prediction of a tumor. After the algorithms are worked, the result showed with each attribute range is benign or malignant. And the stage of the patient is compared with all nine attributes the higher one will be the cancer stage of the patient. And also, the list of the cancer patient and non-cancer patient displayed individually then graph for the benign and malignant patients. In the results has shown who are all have the breast cancer. The main process of this proposed system classification with decision tree method and rule-based method. Prediction of breast cancer is important to detect and cure the breast cancer disease. In the proposed system, implemented an efficient tool for breast cancer classification.

Result

In data mining many methods are used for prediction, here classification method used for prediction of breast cancer using ensemble and dagger algorithm, the ensemble is a decision tree algorithm and it has worked in an incremental learning method. The pre-processed training dataset has been used in the proposed classification method. In the pre-processing, the empty and irrelevant dataset rows are removed. And it has been efficiently classified the training dataset into cancer patients and non-cancer patients.

Left Window (Cancer Patients):

Result5	Blandchromatin	Result7	NormalNucleoli	Result8	Mitoses	Result9	Result
MALIGNANT	9	MALIGNANT	7	BENIGN	1	BENIGN	CANCER
MALIGNANT	7	MALIGNANT	10	MALIGNANT	1	BENIGN	CANCER
MALIGNANT	7	MALIGNANT	4	BENIGN	3	BENIGN	CANCER
BENIGN	8	MALIGNANT	9	MALIGNANT	1	BENIGN	CANCER
MALIGNANT	7	MALIGNANT	3	BENIGN	3	BENIGN	CANCER
BENIGN	8	MALIGNANT	10	MALIGNANT	1	BENIGN	CANCER
MALIGNANT	7	MALIGNANT	3	BENIGN	7	MALIGNANT	CANCER
MALIGNANT	3	BENIGN	9	MALIGNANT	1	BENIGN	CANCER
MALIGNANT	3	BENIGN	3	BENIGN	1	BENIGN	CANCER
MALIGNANT	8	MALIGNANT	9	MALIGNANT	8	MALIGNANT	CANCER
BENIGN	7	MALIGNANT	8	BENIGN	10	MALIGNANT	CANCER
MALIGNANT	7	MALIGNANT	10	MALIGNANT	7	MALIGNANT	CANCER
BENIGN	4	BENIGN	10	MALIGNANT	10	MALIGNANT	CANCER
MALIGNANT	7	BENIGN	9	MALIGNANT	10	MALIGNANT	CANCER

Right Window (Non-Cancer Patients):

Result5	Blandchromatin	Result7	NormalNucleoli	Result8	Mitoses	Result9	Result
BENIGN	3	BENIGN	1	BENIGN	1	BENIGN	Non-CANCER
MALIGNANT	3	BENIGN	2	BENIGN	1	BENIGN	Non-CANCER
BENIGN	3	BENIGN	1	BENIGN	1	BENIGN	Non-CANCER
BENIGN	3	BENIGN	7	BENIGN	1	BENIGN	Non-CANCER
BENIGN	3	BENIGN	1	BENIGN	1	BENIGN	Non-CANCER
MALIGNANT	3	BENIGN	1	BENIGN	1	BENIGN	Non-CANCER
BENIGN	3	BENIGN	1	BENIGN	1	BENIGN	Non-CANCER
BENIGN	1	BENIGN	1	BENIGN	5	BENIGN	Non-CANCER
BENIGN	2	BENIGN	1	BENIGN	1	BENIGN	Non-CANCER
BENIGN	3	BENIGN	1	BENIGN	1	BENIGN	Non-CANCER
BENIGN	2	BENIGN	1	BENIGN	1	BENIGN	Non-CANCER
BENIGN	4	BENIGN	4	BENIGN	1	BENIGN	Non-CANCER
BENIGN	3	BENIGN	1	BENIGN	1	BENIGN	Non-CANCER
MALIGNANT	5	BENIGN	5	BENIGN	4	BENIGN	Non-CANCER

Fig 1.2: classified list of cancer and non cancer patients

The fuzzy rule has been generated for given more accuracy and decision tree method used for less-time consuming. In the result, the fig 1.2 shows a cancer patient is individually listed and non-cancer patients also listed. Classification is mainly used for separate the dataset by class-based, the two classes are benign and malignant. A benign tumor is not dangerous but malignant is dangerous, malignant leads to death so, breast cancer known as the most common deadly disease.

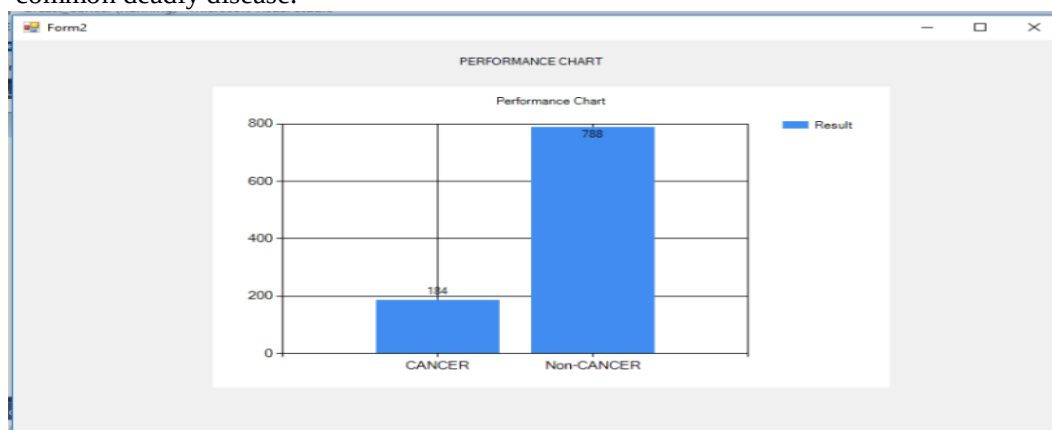


Fig 1.3: Graph shows the comparison of cancerous and non cancerous tumor present in dataset

It is necessary to find the stage of breast cancer and detect the disease. In fig 1.3, the final result was the number of patients has the breast cancer disease in the given dataset and separately view who are all have a non-cancerous tumor.

Conclusion

In the medical field, several data mining algorithms are used for prediction and detect the disease. Various classification algorithms and methods are implemented in this field. In many fields, classification algorithms are mainly used for accuracy. In this proposed system ensemble and dagger, algorithms are worked in the training dataset and it successfully classified. It gives better accuracy for disease diagnosis and efficient manner, reliable to the user. Breast cancer is one of the deadly diseases so, it is important to detect symptoms. Once identified, if any person has a breast tumor immediately proper check-up is needed. So, this method helps to detect, predict the breast cancer. In future, we can extend this approach to implement various algorithms related machine learning approach to improve the accuracy of disease prediction

References

- [1] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," *Telemat. Informatics*, vol. 34, no. 4, pp. 133–144, 2017.
- [2] Ł. Jeleń, A. Krzyżak, T. Fevens, and M. Jeleń, "Influence of feature set reduction on breast cancer malignancy classification of fine needle aspiration biopsies," *Comput. Biol. Med.*, vol. 79, no. October, pp. 80–91, 2016.
- [3] L. Dora, S. Agrawal, R. Panda, and A. Abraham, "Optimal breast cancer classification using Gauss – Newton representation based algorithm," vol. 85, pp. 134–145, 2017.
- [4] S. W. Lam, C. R. Jimenez, and E. Boven, "Breast cancer classification by proteomic technologies: Current state of knowledge," *Cancer Treat. Rev.*, vol. 40, no. 1, pp. 129–138, 2014.
- [5] S. Liu *et al.*, "Quantitative analysis of breast cancer diagnosis using a probabilistic

- modelling approach,” *Comput. Biol. Med.*, vol. 92, no. November 2017, pp. 168–175, 2018.
- [6] N. Dehghan, H. Hassanpour, and M. Reza, “Fully automatic classification of breast cancer microarray images,” *J. Electr. Syst. Inf. Technol.*, vol. 3, no. 2, pp. 348–359, 2016.
- [7] F. Pak, H. Rashidy, and A. Alikhassi, “Breast cancer detection and classification in digital mammography based on Non-Subsampled Contourlet Transform (NSCT) and Super Resolution,” *Comput. Methods Programs Biomed.*, vol. 122, no. 2, pp. 89–107, 2015.
- [8] P. Wang, X. Hu, Y. Li, Q. Liu, and X. Zhu, “Automatic cell nuclei segmentation and classification of breast cancer histopathology images,” *Signal Processing*, vol. 122, pp. 1–13, 2016.
- [9] E. Rezk *et al.*, “Conceptual data sampling for breast cancer histology image classification,” *Comput. Biol. Med.*, vol. 89, no. July, pp. 59–67, 2017.
- [10] D. Gu, C. Liang, and H. Zhao, “A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis,” *Artif. Intell. Med.*, vol. 77, pp. 31–47, 2017.
- [11] Z. Jiao, X. Gao, Y. Wang, and J. Li, “A parasitic metric learning net for breast mass classification based on mammography,” *Pattern Recognit.*, vol. 75, pp. 292–301, 2018.
- [12] D. Soria *et al.*, “Artificial Intelligence in Medicine A quantifier-based fuzzy classification system for breast cancer patients,” *Artif. Intell. Med.*, vol. 58, no. 3, pp. 175–184, 2013.
- [13] A. Helwan, J. Bush, and R. H. Abiyev, “ScienceDirect ScienceDirect,” *Procedia Comput. Sci.*, vol. 120, no. 2017, pp. 402–410, 2018.

