A Survey on Data Mining Feature Selection Approaches towards Breast Cancer

B.Meenapreethi[1], S.G.Nandhini[2], S.Aishwarya[3]
[1] Assistant professor, Dept of BCA and M.Sc. Software Systems
[2, 3] Dept of BCA and M.Sc. Software Systems
[1, 2, 3] Sri Krishna Arts and Science College, Coimbatore-08, India

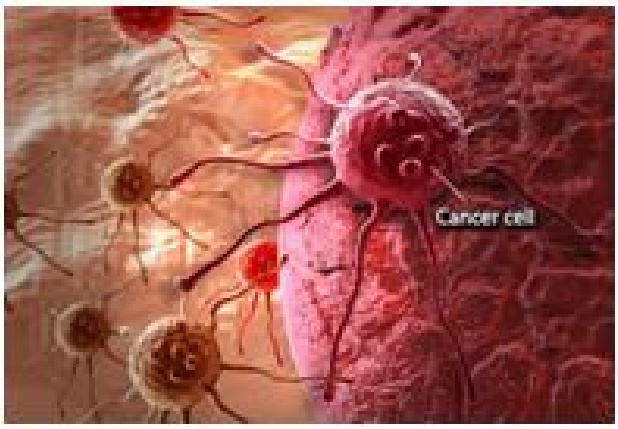
Abstract- *This paper shows a review on medicinal picture selection data mining techniques. In therapeutic field there are various types of issue in medicinal imaging like characterization, segmentation, extraction and selection. Medical datasets are regularly sorted by huge amount of disease measurements and nearly little measure of patient records. These measurements (Feature Selection) are not significant, where these insignificant and excess components are hard to evaluate. Various types of data mining procedures (or) algorithms can be helpful with imprecision and vulnerability in information analysis and can adequately remove the repetitive data.*

Keywords- cancer, data mining, mining techniques, feature selection.

I. INTRODUCTION

1.1 Cancer – The Disease

Cancer is a conceivably lethal sickness caused for the most part by natural factors that transform qualities encoding basic cell-administrative proteins. The resultant unusual cell conduct prompts far reaching masses of irregular cells that crush encompassing ordinary tissue and can spread to imperative organs bringing about dispersed ailment, generally a harbinger of fast approaching patient demise.



All the more altogether, globalization of unfortunate ways of life, especially cigarette smoking and the reception of many components of the cutting edge Western eating regimen (high fat, low fiber) will expand disease frequency.

1.2 Data Mining

Data Mining is the way toward finding intriguing learning from a lot of information put away in database. It is a fundamental procedure where insightful techniques are connected with a specific end goal to remove information designs. Just expressed, information mining alludes to "extricating" or "mining" learning from substantial number of data. [1] The significant purpose behind utilizing information mining has pulled in a lot of focus in the data business. Development of the data innovation has a colossal measure of information and the up and coming requirement for spinning such information into valuable data and learning. These sorts of data accumulated from business administration, building plan, science investigation and therapeutic field.

1.3 Image Mining

Picture Mining assumes crucial part among the analysts in the field of information mining. The picture mining principally centers the extraction of example from expansive gathering of pictures.

The below picture Fig 1.2 demonstrates that the means required in picture handling. The pictures from the database are preprocessed to enhance the picture quality [5]. At that point picture goes under different change, determination and extraction to gauge critical elements. With these elements, mining can utilize information mining procedures to find the critical example.

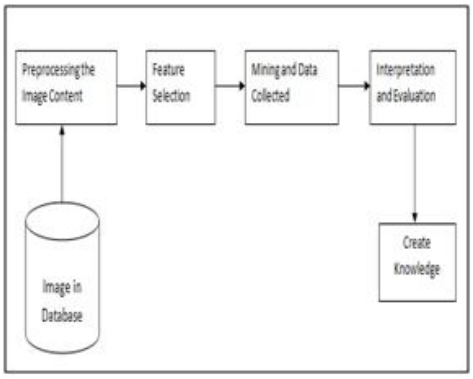


Fig: 1.2 Processes of Image contents

1. 4 Breast Cancer- An Overview

Breast Cancer is the most well-known growth infections among ladies barring no melanoma skin. Cancer are partitioned into two sorts favorable and Malignant. If the cancer is benign under the conditions of early diagnosis. Malignancy status includes the three basic measurement 1.Age 2.Longer tumor length 3.ADC or Apparent Diffusion coefficient (biopsy confirmed).

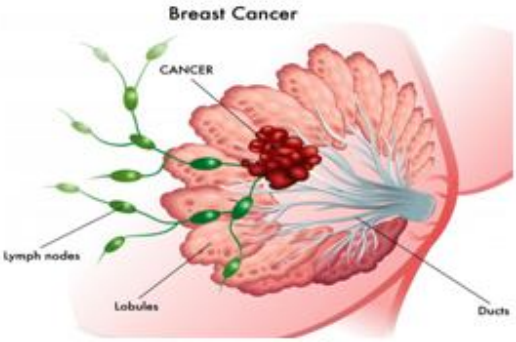


Fig: 1.3 Image of a Cancer Breast

1.4.1 Number phases of Breast tumor

There are 4 number phases of breast cancer, arranging takes into different variables, including

1. The size of the (tumor implies either breast lymph's or Area of disease cells found on a Scan or mammogram)
2. Cancer cells have spread into close-by lymph organs (lymph Node)
3. The tumor cells have spread to some other piece of the body (Metastasized-TNM).

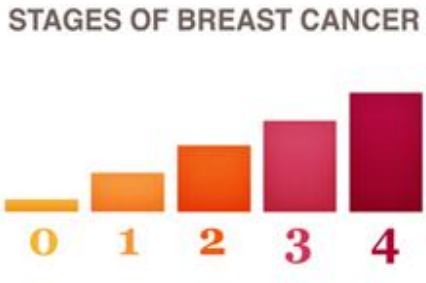


Fig: 1.4 Stages of Breast Cancer

Stage1: Breast growth is part into 2 Stages.
Stage 1A tumor is 2 cm or littler and has spread outside the breast.
Stage 1B: Small regions of bosom Cancer cells are found in the lymph hub shut to the bosom and either the tumor is 2 cm or littler.

Stage 2: breast malignancy: organize
2A:Tumor 2cm or littler in the breast and tumor cells are found in 1 to 3 lymph Node in the armpit or in the lymph Node close to the bosom bone.
Stage 2B: The tumor is bigger than 2 cm however not bigger than 5 cm and little zones of malignancy cells are in the lymph Node.2 to 5 cm spread 1 to 3 lymph hubs in the armpit or close to the breastbones or the tumor is bigger than 5cm and has not spread to the lymph hub.

Stage 3: A breast malignancy No tumor is found in the breast or the tumor might be any size and disease is found in 4 to 9 lymph organs under the arm or in the lymph organs close to the bosom bone. The tumor is more than 5 cm and has spread into up to 3 lymph hubs close to the breast bone.

1.4.2 Execution and Evaluation of Data Mining Techniques in Cancer Diagnosis

Stage 3B: The tumor has spread to the skin of the breast or to the chest divider and made the kinfolk separate or growth swelling. The malignancy may have spread to the up 9 lymph Node in the armpit or to the lymph organs close to the bosom bone.

Stage 4: cancer growth: The tumor can be any size. The lymph Node could possibly contain disease cells. The malignancy has spread to alternate parts of the body, for example, bone, lungs, liver, and cerebrum. The TNM (Tumor, Node, and Metastasis) framework is indicates for each kind of malignancy. Once a patient's T, N, and M classes have been resolved, this data is joined in a procedure called organize gathering to decide a

ladies illness stage.Stage0 (the minimum preferred standpoint arrange) to Stage4 (the most favorable position arrange).

1.5 Prognosis of Breast Cancer

The chances of survival may vary by the stages of breast cancer. The Chances of survival is more in Non-invasive and the early stages than that for the metastatic breast cancer (stage 4) which is the stage where the cancer has spread beyond the neighboring tissues. The Table - Fig 1.5 below shows the 5-year survivability rate of a cancer patient.

Stage	Description	5 – year survival (%)	10 – year survival (%)
Stage – 0	No evidence of Primary Tumor	95	90
Stage – 1	Tumor <= 2cm	85	70
Stage – 2	Tumor > 2cm & <= 5cm	70	50
Stage – 3	Tumor > 5cm	55	30
Stage – 4 (Metastasis)	Any size with extending to - chest wall or skin	5	2

Fig: 1.5 Table of a Prognosis

II. FEATURES OF DATA MINING

The iterative procedure comprises of the following steps:

- **Data cleaning:** otherwise called information purifying it is a stage in which commotion information and insignificant information are expelled from the gathering.
- **Data integration:** at this stage, different information sources, frequently heterogeneous, might be consolidated in a typical source.
- **Data selection:** at this progression, the information significant to the investigation is chosen and recovered from the information gathering.
- **Data transformation:** otherwise called information solidification, it is a stage in which they chose information is changed into shapes suitable for the mining system
- **Data mining:** it is the significant stride in which cunning strategies are connected to remove designs conceivably valuable.
- **Pattern assessment:** this progression, entirely fascinating examples speaking to learning are distinguished in view of given measures.
- **Knowledge representation** is the last stage in which they found learning is outwardly spoken to the client. In this progression perception strategies are utilized to enable clients to comprehend and gives the data mining that comes about.

III. FEATURE SELECTION

3.1 Feature Selection –The Pre-processing

Feature Selection is a pre-preparing step, used to enhance the mining execution by decreasing data dimensionality. Despite the fact that there exist various elements Feature Selection, still it is a dynamic research area in data mining, machine learning. Many element determination calculations stand up to serious difficulties as far as adequacy and effectiveness, due to late increment in data dimensionality (information with a great many elements or properties or factors).

Feature Selection (FS) assumes an imperative part in arrangement. This is one of the Preprocessing strategies in information mining. It is broadly utilized as a part of the fields of measurements, design acknowledgment and therapeutic area. Feature Selection implies diminishing the number of attributes. The attributes are reduced by removing irrelevant and redundant attributes, which do not have significance in classification task. The component determination enhances the execution of the grouping procedures. The procedure of feature selection is

- Generation of applicant subsets of properties from unique list of capabilities utilizing looking methods.
- Evaluation of every applicant subset to decide the pertinence towards the order errand utilizing measures, for example, remove, reliance, data, consistency, classifier mistake rate.
- Termination condition to decide the pertinent subset or ideal element subset.
- Validations to check the chose include subset.

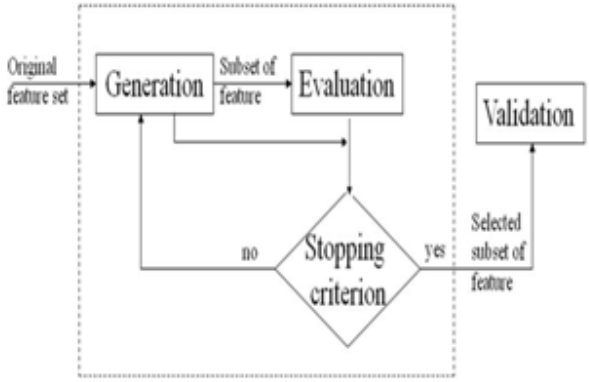


Fig: 2.1 Feature Selection Process

Feature selection methods are classified into different types as filter, wrapper and hybrid approaches. The filter approach applies to the data before the classification. In this method features are evaluated by using methods based on general qualities of the data. In the wrapper approach, the features are evaluated using the classification algorithms. In Hybrid approach features are evaluated using filter and wrapper approaches. Further, the diminished dataset is considered for the order.

3.2 Advantages

- FS is helpful to various computational assignments, for example in procedure of machine learning
- Improve the information control speed
- To lessen the dimensionality issue
- Improve the grouping rate by lessen insignificant and boisterous information
- These are all the a few points of interest of the component segment.

3.3 Disadvantages

- Feature determination technique seeks through the whole subset of elements and locates the best subset highlights among the 2N-1 where N is the aggregate number of highlight.

IV. BASIC DATA MINING TECHNIQUES

Data pre-preparing is applied before the data mining to enhance the nature of the data. Data pre-preparing incorporates data cleaning, data integration, data transformation and data reduction techniques. The components utilized for order purposes matched with the Breast Imaging Reporting and Data System (BI-RADS) as this is the means by which radiologists characterize cancer disease. The BI-RADS elements of thickness, mass shape, mass edge and variation from the norm appraisal rank are utilized as they have been demonstrated to give great grouping exactness. A Classification strategy, Decision tree calculations are broadly utilized as a part of medical field to group the restorative information for determination. Feature Selection builds the accuracy since it disposes the irrelevant attributes.[2]Feature Selection with Decision tree arrangement incredibly improves the nature of the information in restorative analysis. Cart Algorithm with different element choice strategies to see if a similar element choice strategy may prompt best exactness on different datasets of same area. Artificial neural Networks (ANNs) and Support vector machines have been as of late proposed as an extremely viable strategy for design acknowledgment, machine learning and information mining.

The Basic Techniques used in Data Mining are:

1. CART Algorithm
2. Decision tree algorithms
3. ID3
4. C4.5 decision tree algorithms
5. Hunt’s algorithm
6. SVM
7. Naïve Bayes classifier.
8. The back-propagated neural network.

Early Diagnosis needs a precise and dependable determination technique that can be utilized by doctors to recognize benign breast tumors from malignant tumors from threatening ones without going for surgical biopsy. The target of these forecasts is to allocate patients to one of the two gathering either a "generous" that is noncancerous or a "threatening" that is malignant. The visualization issue is the long haul tend to the infection for patients whose malignancy has been surgically evacuated. In this paper, we propose a model-based data mining technique Decision Tree with Feature Selection.

4.1 Decision Tree

Decision tree is a strategy for classifier that is communicated as a recursive division of the instance space. [7] It makes a prescient model, which maps perceptions about anode to decisions about the nodes' objective value. In a tree structure leaves speak to the class marks and branches speak to conjunctions of highlight prompting the class names. Sub section B talks about utilizing decision tree data mining strategy to build a prescient model to analyze whether a tumor

is kind or harmful relying on different qualities related with a specific medicinal record.

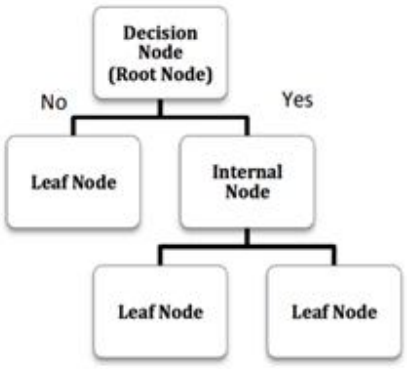


Fig: 2.3 Diagram of Decision Tree

4.2.1 Algorithm

The Decision tree calculation is extremely vigorous and learning effectiveness with its learning time multifaceted nature of O (nlog2n). The result of this calculation is a decision tree that can be effectively spoken to as an arrangement of representative principles (IF….THEN….). This rule can be specifically translated and contrasted and accessible natural learning and giving valuable data to the scientist and clinicians. As indicated by Quinlan, the learning algorithm applies a divide and conquers methodology to develop the tree.[8] The set of examples are related by an arrangement of properties (attributes). A decision tree contains node and leaves where node speak to a test on the values of an attribute and leaves speak to the class of an instance that fulfills the tests. The result is "yes‟ or "no‟ choice. Principles can be gotten from the way from the root to a leaf and using the hubs en route as preconditions for the administrator, to anticipate the class at the leaf. Pruning is important to expel superfluous preconditions and duplications.

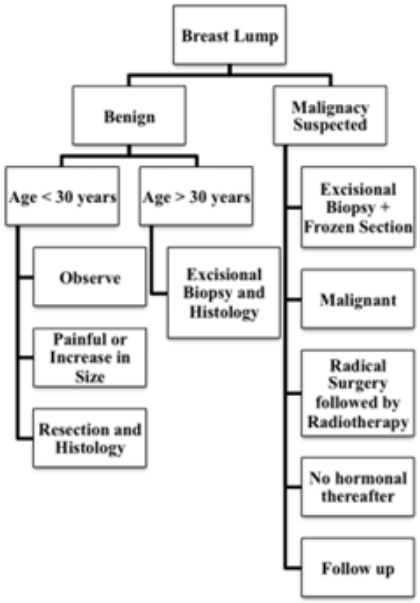


Fig: 2.3 Flow Diagram of Breast Diagnosis

4.2.2 Advantages

- Decision tree utilizes "tree pruning" way to deal with distinguish and expel loud information from the branch and to enhance the arrangement precision.
- The characteristic with the most elevated standardized information is settled on a choice.
- Algorithm utilized consistent and discrete esteems.

4.2.3 Disadvantages

- Efficiency and versatility are low when connected to mining of substantial information bases.
- Decision tree development wasteful while swapping test information from primary memory to reserve memory.

V. CONCLUSION

The automatic diagnosis of Breast malignancy is a vital certifiable medicinal issue. Discovery of cancer growth in its beginning times is the key for treatment. This paper indicates how decision trees are utilized to display real conclusion of Breast malignancy for neighborhood and orderly treatment, alongside exhibiting different methods. Accuracy is most imperative in the field of medicinal finding to analyze

the patient's infection. It is made simpler utilizing the idea of Decision trees.

In future it is intended to gather the information from different areas over the world and make a more exact and general prescient model for cancer malignancy determination. Future examination will likewise focus on gathering information from a later day and age and observe new potential prognostic elements to be incorporated into a Decision tree. The work can be extended and upgraded for the mechanization of Breast tumor analysis.

REFERENCES

[1] Jiewai Han and Micheline Kamber ,"Data Mining Concepts and Techniques", second edition.
[2] Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu, NagarajuOrsu, Suresh B. Mudunuri, " Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification" , (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5,2013.
[3] A.Priyanga and Dr,S.Prakasam " The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness" International Journal of Computer Science and Engineering Communications- IJCSEC. Vol.1 Issue.1, December 2013.
[4] V.Krishnaiah, Dr.G.Narsimha and N.Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013.
[5] M. Vasantha, Dr. V. Subbiah bharathi, R. Dhamodharan," Medical Image Feature, Extraction, Selection and Classification, International Journal of Engineering Science and Technology, Vol.2(6),2010,2071-2076
[6] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2000.
[7] Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria, Decision Tree Analysis on J48 Algorithm for Data Mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.
[8] National Cancer Institute: http://www.cancer.gov/cancertopics/types/breast.
[9] Wikipedia,http://en.wikipedia.org/wiki/File:Mammo_breast_cancer. jpg
[10] American Cancer Society, http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-diagnosis
[11] NHS Choices, http://www.nhs.uk/Conditions/Cancer-ofthe-breast-female/Pages/Treatment.aspx