INTERNATIONAL JOURNAL OF RESEARCH AND DEVELOPMENT IN ENGINEERING SCIENCES
e-ISSN: 2582 - 4201 Volume: 2 | Issue: 2 | Year: 2020

BREAST CANCER DIAGNOSIS AND PROGNOSIS ANALYSIS BY USING MACHINE LEARNING WITH DATA MINING CLASSIFICATION TECHNIQUES

T. CHALAPATHI RAO

Sr. Assistant Professor,
Department of Computer Science and Engineering,
Aditya Institute of Technology and Management, Tekkali.
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India.
chalapathit520@gmail.com

Abstract: *Breast Cancer Diagnosis and Prognosis are two medical applications pose a great challenge to the researchers. The use of machine learning and data mining techniques has revolutionized the whole process of breast cancer Diagnosis and Prognosis. Breast Cancer Diagnosis distinguishes benign from malignant breast lumps and Breast Cancer Prognosis predicts when Breast Cancer is likely to recur in patients that have had their cancers excised. Thus, these two problems are mainly in the scope of the classification problems. This study paper summarizes various review and technical articles on breast cancer diagnosis and prognosis. In this paper we present an overview of the current research being carried out using the data mining techniques to enhance the breast cancer diagnosis and prognosis.*

Keywords: *Breast cancer, Image processing, Diagnosis, Prognosis, Data Mining, Classification.*

1. Introduction

Breast cancer has become the leading cause of death in women in developed countries. The most effective way to reduce breast cancer deaths is detect it earlier. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. The objective of these predictions is to assign patients to either a"*benign*" group that is non-cancerous or a"*malignant*" group that is cancerous. The prognosis problem is the long-term outlook for the disease for patients whose cancer has been surgically removed. In this problem a patient is classified as a 'recur' if the disease is observed at some subsequent time to tumor excision and a patient for whom cancer has not recurred and may never recur [1].

The objective of these predictions is to handle cases for which cancer has not recurred (censored data) as well as case for which cancer has recurred at a specific time [3]. Thus, breast cancer diagnostic and prognostic problems are mainly in the scope of the widely discussed classification problems. These problems have attracted many researchers in computational intelligence, data mining, and statistics fields. Cancer research is generally clinical and/or biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. As the use of computers powered with automated tools, large volumes of medical data are being collected and made available to the medical research groups. As a result, Knowledge Discovery in Databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical cases stored within datasets [2].

The objective of this study is to summaries various review and technical articles on diagnosis and prognosis of breast cancer. It gives an overview of the current research being carried out on various breast cancer datasets using the data mining techniques to enhance the breast cancer diagnosis and prognosis**.**

Computer-aided detection in mammographic screening

Computer-aided detection (CAD) solutions were developed to help radiologists in reading mammograms. These programs usually analyze a mammogram and mark the suspicious regions, which should be reviewed by the radiologist15. The technology was approved by FDA and had spread quickly. By 2008, in the US, 74% of all screening mammograms in the Medicare population were interpreted with CAD, the cost of CAD usage is over $400 million a year. The benefits of using CAD are controversial. Initially several studies have shown promising results with CAD6, 16–20. A large clinical trial in the United Kingdom has shown that single reading with CAD assistance has similar performance to double reading [5]. These controversial results indicate that CAD systems need to be improved before radiologists can ultimately benefit from using the technology in everyday practice [4].

Currently used CAD approaches are based on describing the X-ray image with meticulously designed hand crafted features, and machine learning for classification on top of these features. In the field of computer vision, since 2012, deep convolution neural networks (CNN) have significantly outperformed these traditional methods. Deep CNN-s have reached or even surpassed human performance in image classification and object detection [6]. These models have tremendous potential in medical image analysis. Several studies have attempted to apply Deep Learning to analyze mammograms but the Problem is still far from being solved.

2. Back Ground Work

Breast Cancer: An Overview

Breast cancer is the most common cancer disease among women, excluding non-melanoma skin cancers. The information about the tumor from certain examinations and diagnostic tests are gathered using staging to determine how widespread the cancer is. The stage of a cancer is one of the most important factors in selecting treatment options, and it uses the Tumor, Nodes and Metastasis (TNM) system. When a patient's T, N, and M categories have been determined then this information is combined in a process known as stage grouping to determine a woman's disease stage [7].

Breast cancer is a malignant tumor, grew from cells of the breast. Hence, cancer of breast tissue is called breast cancer. Worldwide, it is the most common form of cancer in females that is affecting approximately 10% of all women at some stage of their life in the Western world. Although significant efforts are made to achieve early detection and effective treatment but scientists do not know the exact causes of most breast cancer, they do know some of the risk factors (i.e. ageing, genetic risk factors, family history, menstrual periods, not having children, obesity)that increase the likelihood of developing breast cancer in females [8].

The survival analysis that is the study of time to an event of interest, such as disease occurrence or death also gives the physicians and the patient better information with which to plan treatment and may eliminate the need for a prognostic surgical procedure.

Knowledge Discovery and Data Mining

This section provides an introduction to knowledge discovery and data mining. We list the various analysis tasks that can be goals of a discovery process and lists methods and research areas that are promising in solving these analysis tasks.

The Knowledge Discovery Process

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process. The following fig. 1 shows data mining as a step in an iterative knowledge discovery process [7].

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning:** also known as data cleansing it is a phase in which noise data and irrelevant data are removed from the collection.

- **Data integration:** at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection:** at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data Transformation:** also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining:** it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation:** this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation:** is the final phase in which the discovered knowledge is visually represented to the user. In this step visualization techniques are used to help users understand and interpret the data mining results.

Data Mining Process

In the KDD process, the data mining methods are for extracting patterns from data. The patterns that can be discovered depend upon the data mining tasks applied. Generally, there are two types of data mining tasks: *descriptive data mining tasks* that describe the general properties of the existing data, and *predictive data mining tasks* that attempt to do predictions based on available data. Data mining can be done on data which are in quantitative, textual, or multimedia forms [8].

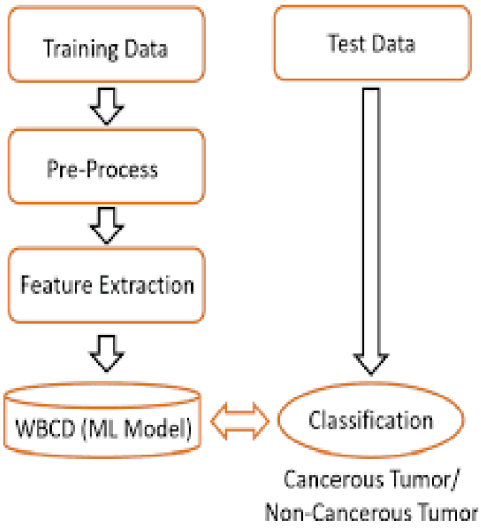

Figure.1: Progressive operations

Data mining applications can use different kind of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects).Data mining involves some of the following key steps.

- ***Problem definition:*** The first step is to identify goals. Based on the defined goal, the correct series of tools can be applied to the data to build the corresponding behavioural model.
- ***Data exploration:*** If the quality of data is not suitable for an accurate model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.
- ***Data preparation:*** The purpose of this step is to clean and transform the data so that missing and invalid values are treated and all known valid values are made consistent for more robust analysis.
- ***Modeling:*** Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next generation techniques such as decision trees, networks and rule based algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.
- ***Evaluation and Deployment:*** Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

3. Related Work

Data Mining Classification Methods

The data mining consists of various methods. Different methods serve different purposes, each method offering its own advantages and disadvantages. However, most data mining methods commonly used for this review are of classification category as the applied prediction techniques assign patients to either a "*benign*" group that is non-cancerous or a "*malignant*" group that is cancerous and generate rules for the same. Hence, the breast cancer diagnostic problems are basically in the scope of the widely discussed classification problems.

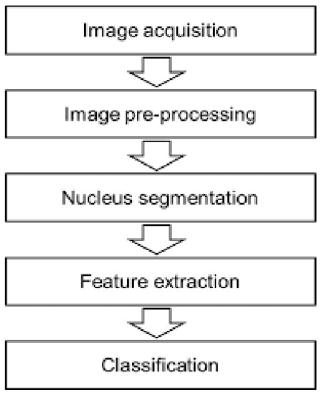

Figure.2: Data mining Classifications

In data mining, classification is one of the most important tasks. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The commonly used methods for data mining classification tasks can be classified into the following groups [9].

Decision Trees (DT's): A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Choice of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached. Decision trees can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.

Support Vector Machine (SVM): Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing

with interactions among features and redundant features.

Genetic Algorithms (GAs) / Evolutionary Programming (EP): Genetic algorithms and evolutionary programming are algorithmic optimization strategies that are inspired by the principles observed in natural evolution. Of a collection of potential problem solutions that compete with each other, the best solutions are selected and combined with each other. In doing so, one expects that the overall goodness of the solution set will become better and better, similar to the process of evolution of a population of organisms. Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about dependencies between variables, in the form of association rules or some other internal formalism [10].

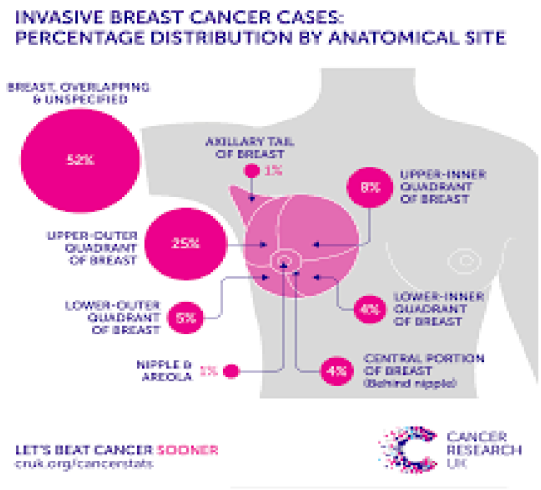

Figure.3: Internal Cancer projection

Fuzzy Sets: Fuzzy sets form a key methodology for representing and processing uncertainty. Uncertainty arises in many forms in today's databases: imprecision, non-specificity, inconsistency, vagueness, etc. Fuzzy sets exploit uncertainty in an attempt to make system complexity manageable. As such, fuzzy sets constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems.

Neural Networks: Neural networks (NN) are those systems modeled based on the human brain working. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input.

Rough Sets: A rough set is determined by a lower and upper bound of a set. Every member of the lower bound is a certain member of the set. Every non-member of the upper bound is a certain non-member of the set. The upper bound of a rough set is the union between the lower bound and the so-called boundary region. A member of the boundary region is possibly (but not certainly) a member of the set. Therefore, rough sets may be viewed as with a three-valued membership function (yes, no, perhaps). Rough sets are a mathematical concept dealing with uncertainty in data. They are usually combined with other methods such as rule induction, classification, or clustering methods.

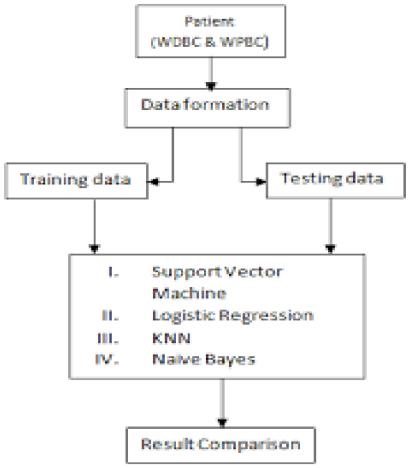

Figure.4: Data mining Classification operations

Data Mining Classification Techniques for Breast Cancer Diagnosis

Clinical diagnosis of breast cancer helps in predicting the malignant cases. A lump felt during the examination roughly give clues as to the size of tumor and its texture. The various common methods used for breast cancer diagnosis are Mammography, Biopsy, Positron Emission Tomography and Magnetic Resonance Imaging. The results obtained from these methods are used to recognise the patterns which are aiming to help the doctors for classifying the malignant and benign cases. There are various data mining techniques, statistical methods and machine learning algorithms that are applied for this purpose. This section consists of the review of various technical and review articles on data mining techniques applied in breast cancer diagnosis.

Orlando Anunciacao, Bruno C. Gomes, Susana Vinga, Jorge Gaspar, Arlindo L.Oliveira and Jose Rueff explored the applicability of decision trees for detection of high-risk breast cancer groups over the dataset produced by Department of Genetics of faculty of Medical Sciences of Universidad Nova de Lisboa with 164 controls and 94 cases in WEKA machine learning tool. To statistically validate the association found, permutation tests were used. They found a high-risk breast cancer group composed of 13 cases and only 1 control, with a Fisher Exact Test (for

validation) value of $9.7*10^{-6}$ and a p-value of 0.017. These results showed that it is possible to find statistically significant associations with breast cancer by deriving a decision tree and selecting the best leaf.

Dr. Medhat Mohamed Ahmed Abdelaal and Muhamed Farouq investgated the capability of the classification SVM with Tree Boost and Tree Forest in analyzing the DDSM dataset for the extraction of the mammographic mass features along with age that discriminates true and false cases. Here, SVM techniques show promising results for increasing diagnostic accuracy of classifying the cases witnessed by the largest area under the ROC curve comparable to values for tree boost and tree forest.

Wei-pin Chang, Der-Ming and Liou explored that the genetic algorithm model yielded better results than other data mining models for the analysis of the data of breast cancer patients in terms of the overall accuracy of the patient classification, the expression and complexity of the classification rule [12].

The artificial neural network, decision tree, logistic regression, and genetic algorithm were used for the comparative studies and the accuracy and positive predictive value of each algorithm were used as the evaluation indicators. WBC database was incorporated for the data analysis followed by the 10-fold cross-validation. The results showed that the genetic algorithm described in the study was able to produce accurate results in the classification of breast cancer data and the classification rule identified was more acceptable and comprehensible.

J. Padmavati performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. Logistic regression was performed using logistic regression in SPSS package and MLP and RBF were constructed using MATLAB. It was observed that neural networks took slightly higher time than logistic regression but the sensitivity and specificity of both neural network models ha a better predictive power over logistic regression. When comparing RBF and MLP neural network models , it was found that RBF had good predictive capabilities and also time taken by RBF was less than MLP [13].

Aboul Ella Hassanien, and Jafar M.H.Ali in their paper presented a rough set method for generating classification rules from a set of observed 360 samples of theWBC data. The attributes were selected, normalized and then the rough set dependency rules were generated directly from the real value attribute vector. Then the rough set reduction technique was applied to find all reducts of the data which contains the minimal subset of attributes that are associated with a class label for classification. They showed that the total number of generated rules was reduced from 472 to 30 rules after applying the proposed simplification algorithm. They also made a comparison between the obtained results of rough sets with the well known ID3 decision tree and concluded rough sets showed higher accuracy and generated more compact rules.

Data Mining Classification Techniques for Breast Cancer Prognosis

Once a patient is diagnosed with breast cancer, the malignant lump must be excised. During this procedure physicians must determine the prognosis of the disease. This is the prediction of the expected flow of the disease. Prognosis is important because the type and intensity of the medications are based on it. Prognosis problem is also called as "analysis of *survival* or *lifetime* data". It poses a more difficult problem than that of diagnosis since the data is *censored*. That is, there are only a few cases where we have an observed recurrence of the disease. In this case, we can classify the patient as *recur* and we know the *time to recur* (TTR).

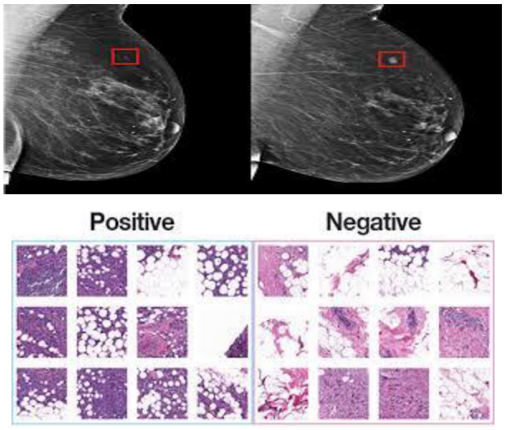

Figure.5: Predictive analysis and Classifications

On the other hand, we do not observe recurrence in most patients. For these, there is no real point at which we can consider the patient a non recurrent case. So, the data is considered censored since we do not know the time of recurrence. For such patients, all known is only the time of their last check-up. We call this the disease-free survival time (DFS). Prognosis helps in establishing a treatment plan by predicting the outcome of a disease.

There are three predictive foci of cancer prognosis:

- prediction of cancer susceptibility (risk assessment),
- prediction of cancer recurrence and
- Prediction of cancer survivability.

The most widely accepted prognostic factor for breast cancer is the American Joint Commission on Cancer (AJCC) staging system based on the TNM system (T,

tumor; N, node; M, metastasis)[16] and survival is considered as any incidence of breast cancer where the person is still living from the date of diagnosis. The objective of prognostic predictions is to handle cases for which cancer has not recurred (censored data) as well as case for which cancer has recurred at a specific time. Thus, breast cancer prognostic problems are mainly in the scope of the widely discussed classification problems. This section consists of the review of various technical and review articles on data mining techniques applied in breast cancer prognosis.

A well known decision tree induction learning technique which has been used by Abdelghani Bellaachia and Erhan Gauven along with two other techniques i.e. Naïve Bayes and Back-Propagated Neural Network. They presented an analysis of the prediction of survivability rate of breast cancer patients using above data mining techniques and used the new version of the SEER Breast Cancer Data. The preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. They have adopted a different approach in the pre-classification process by including three fields: STR(Survival Time Recode), VSR(Vital Status Recode), and COD(Cause Of Death) and used the Weka toolkit to experiment with these three data mining algorithms. Several experiments were conducted using these algorithms [14].

The achieved prediction performances are comparable to existing techniques. However, they found out that model generated by C4.5 algorithm for the given data has a much better performance than the other two techniques.

M. Lundin et al has applied ANN on 951 instances dataset of Turku University Central Hospital and City Hospital of Turku to evaluate the accuracy of neural networks in predicting 5, 10 and 15 years breast cancer specific survival. The values of ROC curve for 5 years was evaluated as 0.909,for 10 years 0.086 and for 15 years 0.883, these values were used as a measure of accuracy of the prediction model. They compared 82/300 false prediction of logistic regression with 49/300 of ANN for survival estimation and found ANN predicted survival with higher accuracy.

W. Nick Street applied ANN classification to Wisconsin Prognostic Breast Cancer and SEER datasets for the analysis of survival. He developed a novel encoding as good and poor prognosis of censored data in an ANN architecture to provide a framework for prognostic prediction.

In [20] Chih-Lin Chi et al used the Street's ANN model for Breast Cancer Prognosis on WPBC data and Love data. In their research they used recurrence at five years as a cut point to define the level of risk. The applied models successfully predicted recurrence probability and separated patients with good(>5 yrs) and bad(<5 yrs) prognoses.

In [22] Jong Pill Choi et al compared the performance of an Artifical Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis. The hybrid Network combined both ANN and Bayesian Network. The Nine variables of SEER data which were clinically accepted were used as inputs for the networks. The accuracy of ANN(88.8%) and Hybrid Network(87.2%) were very similar and they both outperformed the Bayesian Network. They found the proposed Hybrid model can also be useful to take decisions.

In [23] Muhammad Umer Khan et al investigated a hybrid scheme based on fuzzy decision trees on SEER data, they performed experiments using different combinations of number of decision tree rules, types of fuzzy membership functions and inference techniques. They compared the performance of each for cancer prognosis and found hybrid fuzzy decision tree classification is more robust and balanced than the independently applied crisp classification.

4. Conclusion

This paper provides a study of various technical and review papers on breast cancer diagnosis and prognosis problems and explores that data mining techniques offer great promise to uncover patterns hidden in the data that can help the clinicians in decision making. From the above study it is observed that the accuracy for the diagnosis analysis of various applied data mining classification techniques is highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy.

The prognostic problem is mainly analyzed under ANNs and its accuracy came higher in comparison to other classification techniques applied for the same. But more efficient models can also be provided for prognosis problem like by inheriting the best features of defined models. In both cases we can say that the best model can be obtained after building several different types of models, or by trying different technologies and algorithms.

References

[1]. El-Sebakhy A. Emad, Faisal Abed Kanaan, Helmy T., Azzedin F. and Al-Suhaim F., "Evaluation of breast cancer tumor classification with unconstrained functional networks classifier," Computer Systems and Applications, IEEE International Conference, 2006, pp. 281 – 287.

[2]. Osmar R. Zaïane, Principles of Knowledge Discovery in Databases. [Online].

[3]. webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf

[4]. Morton, M. J., Whaley, D. H., Brandt, K. R. & Amrami, K. K. Screening mammograms: interpretation with computer-aided detection—prospective evaluation. Radiol. 239, 375–383 (2006).
[5]. Warren Burhenne, L. J. et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography 1. Radiol. 215, 554–562 (2000).
[6]. Gilbert, F. J. et al. Single reading with computer-aided detection for screening mammography. New Engl. J. Medicine 359, 1675–1684 (2008).
[7]. Fenton, J. J. et al. Influence of computer-aided detection on performance of screening mammography. New Engl. J. Medicine 356, 1399–1409 (2007).
[8]. Fenton, J. J. et al. Effectiveness of computer-aided detection in community mammography practice. J. Natl. Cancer
[9]. institute 103, 1152–1161 (2011).
[10]. Hologic. Understanding ImageCheckerR CAD 10.0 User Guide – MAN-03682 Rev 002 (2017).
[11]. Hupse, R. & Karssemeijer, N. Use of normal tissue context in computer-aided detection of masses in mammograms. IEEE Transactions on Med. Imaging 28, 2033–2041 (2009).
[12]. Hupse, R. et al. Standalone computer-aided detection compared to radiologists' performance for the detection ofmammographic masses. Eur. radiology 23, 93–100 (2013).
[13]. Kooi, T. et al. Large scale deep learning for computer aided detection of mammographic lesions. Med. image analysis 35, 303–312 (2017).
[14]. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Advancesin neural information processing systems, 1097–1105 (2012).

Author's Profile

Mr. **T. Chalapathi Rao** has been working as Assistant Professor in the Department of Computer Science and Engineering, at Aditya Institute of Technology and Management, Tekkali, affiliated with JNTU Kakinada, Andhra Pradesh. He Completed Master in Technology (Computer Science & Engineering) in October 2010 from JNTUA, Ananthapur, India.

He worked as a Sr. Asst. Professor in AITAM, Tekkali since 11 years. He participated in 4 international and 2 national conferences. He guided 12 UG academic projects and participated 7 international and 4 National workshops and 10 FDPs. He published 9 research papers in various reputed international and national Journals, magazines and conferences. He is a member of CSI and ISTE.