DOI: https://doi.org/10.46632/daai/2/4/19

Data Analytics and Artificial Intelligence
Vol: 2(4), 2022
REST Publisher; ISBN: 978-81-948459-4-2
Website:http://restpublisher.com/book-series/data-analytics-and-artificial-intelligence/

Machine learning and deep learning for breast cancer risk prediction and diagnosis: a Survey

R.Shankari, J. S. Leena Jasmine, S. MaryJoans
Velammal Engineering College, Chennai, Tamil Nadu, India.
*Corresponding author Email: **leena@velammal.edu.in**

Abstract: Breast cancer is the widest spreading disease among women globally. The prevalence rate of breast cancer continued to rise in the last few decades. The mitotic count is a relevant factor for grading invasive breast cancer. Early analysis is an extremely imperative step in treatment. However, it is not an easy one due to several skepticisms in detection which employ mammograms. Since it is subject to human prone error, requires more time for completion and the nuclei look similar during all stages of mitosis, automatic detection of mitosis is a good solution to overcome these problems. Detailed analysis of breast cancer normally requires medical images of different methods. The sensitivity and specificity of the diagnosis largely depend on the experiences of the radiologists, and uncertain diagnosis is quite frequent because of resolution limitations and the concerns of lawsuits arisen from wrong diagnosis or undetected lesions. In this paper, the top methodologies used for mitosis detection are analyzed. There are many algorithms for classification and prediction of breast cancer: Support Vector Machine (SVM), Decision Tree (CART), k Nearest Neighbors (KNN), Random Forest (RF), and Bayesian Networks (BN). The Wisconsin data set was used to analyze breast cancer as a training set to assess and measure the performance of the three ML classifiers in terms of key frameworks such as accuracy, recall, precision, and ROC. The outcome obtained in this paper provides a critique of the stateofart ML techniques for breast cancer detection. It was also found that the ensemble classifier gives better performance. A preliminary experiment conducted on cascaded RF and Artificial Neural Network (ANN) results in better accuracy than individual classifiers. The paper shows how we can use deep learning technology diagnosis of breast cancer using MIAS Dataset. A deep learning approach is almost used for immense task objective Image processing, Computer Vision, Medical Diagnosis, and Neural Language Processing.

Keywords: Support Vector Machine (SVM), Decision Tree (CART), k nearest Neighbors (KNN), Random Forest (RF), and Bayesian Networks (BN).

1. Introduction

Breast cancer is considered to be the most common cancer around the world and is considered to be one of the major reasons for an increased death rate among women. The etiology of breast cancer is not yet fully understood, but an earlier diagnosis of breast cancer through periodic screening could improve the chance of recovery. The initial stage cancer disclosure is required to provide proper regimen to the patient and curtail the risk of death due to cancer as the detection of these cancer cells at later phases leads to increase chances of death. There is a great need for an automated system that could analyze these images accurately and rapidly. The current screening strategy of breast cancer is based on classic X-ray imaging. Medical imaging is rapidly gaining importance in clinical diagnostics and biomedical research. Computer-based analysis of histopathological images is being used in medical laboratories. Further improvement in image analysis will lead to new algorithms with less human intervention. The result expected would be perfect diagnostic assistance to the pathologist as well as it can predict the patient's outcome without any errors. Breast cancer is the most deadly cancer in women. It is estimated that over 508,000 women died across the globe in 2011 due to breast cancer and in 2018, nearly 266,120 new reports of aggressive breast cancer and 63,960 new reports of non-aggressive (in-situ) breast cancer are identified in women in the United States of America (GHE). Analytical prediction describes a rise of 75% in the coming years. Accurately predicting a cancerous tumor remains a challenging task for many physicians.

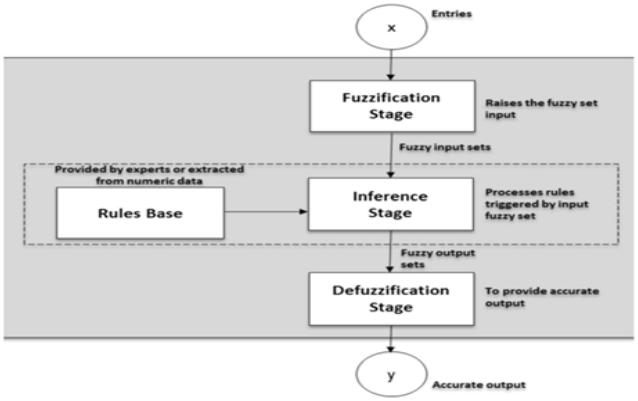

FIGURE 1. Technologies in use to detect cancer cells

Cancerous cells are detected by performing various tests like MRI, mammogram, ultrasound, and biopsy. The development of advanced medical automation and the excessive amount of patient data has excited the path for the development of new approaches in the detection and prediction of cancer. Although data assessment that is collected from the patient and a physician's intake greatly contributes to the diagnostic process, supportive tools could be added to help facilitate accurate diagnoses. These gizmos aim to cancel possible indicative errors and provide a gile way for scrutinizing large chunks of data. Analysts have been employed on and developing assorted deep learning solutions to produce reassuring results. Machine learning is intermittently used in medical applications such as the detection of the type of cancerous cells.

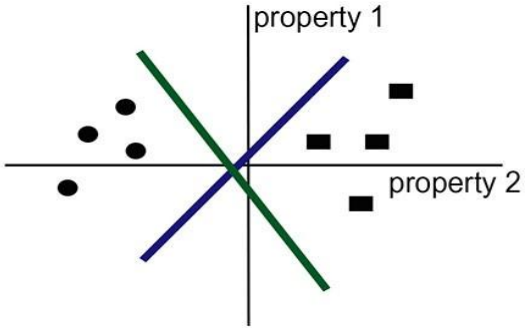

FIGURE 2. SVM generated hyper-planes

Research activities on deep convolution neural networks, image-processing, ultrasound-based cancer detection has been done in the past years. A lot of progress has been made in image processing and recognition, mainly due to the availability of large annotated datasets and deep convolution neural networks (CNN). There is a lack of raw, latest, and updated data for image processing in deep learning which seems to be the major challenge for data scientists for transferring the recent developments in the machine and deep learning to the medical domain. In cancer research, these techniques could be used to identify different patterns in a data set and consequently predict whether a cancer is malignant or benign. A strong prediction model is needed to assess the possible outcomes and thereby formulate procedures that can be carried out to curtail the possibility of cancer inpatients.

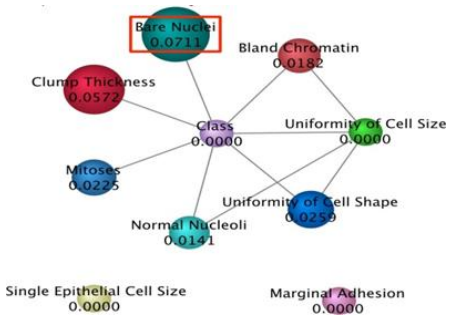

FIGURE 2.

Various machine learning and deep learning models are trained and tested for various different types of diseases and the results have shown that prediction and analysis sensitivity has improved multi-fold. Machine learning has seen a tremendous explosion of interest in the modern medical domain. The goal of this paper is to acquaint the reader with various modern technologies that have been employed in the deep learning field to solve medical issues like cancer cell detection and also to illustrate have these technologies can be used in medical imaging to improvise the analysis and diagnosis of medical conditions. The performance evaluated based on the classification accuracy, recall, precision, and area under the ROC. Machine Learning (ML), is a subfield of Artificial Intelligence (AI) that allows machines to learn without explicit programming by exposing them to sets of data allowing them to learn a specific task through experience. Different algorithms used are Support Vector Machine (SVM), Decision Tree (CART), Naive Bayes (NB) and k Nearest Neigh bours (k-NN).This paper explains the fundamental concept of the ML methods, simulation setup that has been used for carrying out the comparative study and concludes. And the future works that can be performed in the deep learning domain to come up with a more refined and unique solution to predict and diagnose cancer among patients.

2. Machine Learning techniques

The learning process in ML techniques can be divided into two main categories, supervised and unsupervised learning. In supervised learning, a set of data instances are used to train the machine and are labeled to give the correct result. However, in unsupervised learning, there are no pre-determined data sets and no notion of the expected outcome, which means that the goal is harder to achieve. Classification is among the most common methods that go under supervised learning. It

uses historical labeled data to develop a model that is then used for future predictions. In the medical field, clinics and hospitals maintain large databases that contain records of patients with their symptoms and diagnosis. Therefore, researchers make use of this knowledge to develop classification models that can make inferences based on historical cases. The medical assumption has become a much elementary task with machine-based support using the precipitate amount of medical data that is accessible today.

3. K – Nearest Neighbour (Knn)

K Nearest Neighbor makes prophecy using the training dataset directly. Prognosis is made for a new detail input value x by searching through the integrated training set for the K most similar details and compiling the output variable for those K instances. For regression this might be the mean output variable, in distribution, this might be the mode class value. To regulate which of the Kinstancesin the training dataset are most analogous to a new input a distance measure is used. For real input variables, the most popular distance measure is Euclidean distance. Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (xi) across all input attributes j.

$$\text{Euclidean Distance}(n, n_i) = \text{sqrt}(\text{sum}((nj - nij)\text{^}2))$$

In the analysis phase, user-defined constant, and a test point vector is analyzed by assigning the label which is most continual among the training samples and adjacent to that reservation point. Classification and Regression Trees (Cart): A Classification and Regression Tree is an anticipating exemplary, which explains how a result variable's values can be concluded based on other code values. A decision tree is the output of CART where each value is split in an analyzing variable and each node at the end contains an outcome variable with the predicted value. Binary tree representation is used in the regression tree model. Single input variable x denotes every root node and a split point is always a numeric variable. Output variable y is enclosed in the leaf nodes of the regression decision tree to make a prediction. Support Vector Machine (Svm): Support Vector Machine is one of the supervised classification techniques of machine learning that is widely used in the field of cancer prognosis. Critical samples are selected in SVM functions from all classes of support vectors. The linear function is applied for separating the classes that divide them as broadly as support vectors. SVM algorithm, data items are plotted as a point in n-dimensional vector space with the value of each particular coordinate being the value of a feature. Hyper-plane differentiates the two classes toper form classification. Support Vectors are individual observation coordinates. Support Vector Machine is a frontier that separates the line and hyper plane. Support Vector Machine maps the feature space by input vector into a high dimensional hyper plane that separates the two-class data points. The marginal values between the hyper plane and the line that are close to the boundary are maximized. The resulting classifier provides a reliable classification of new samples. The probabilistic outputs can also be obtained for SVMs figure below illustrates how an SVM might work to classify tumors among benign and malignant based on their size and patients' age. The identified hyper plane can be thought of as a decision boundary between the two clusters. The existence of a decision boundary allows for the detection of any misclassification produced by the method.

4. Random Forest (RF)

Random decision forest combines many decision trees to assemblage a forest of trees. Using RF results in increased cohesion compared to decision trees. Random decision forest results are insensitive to noise characteristics in the input data set. It is widely used in cancer prognosis and detection is to handle minorities in the input data set. The recursive approachis used in the RF method based on which everyemphasisinvolves selecting one sample of random size *n* from the data set with reinstatement, and another random sample from the predictors without reinstatement. The data that is selected is partitioned. The unnecessary data is then dropped and the above steps are iterated many times depending on the number of trees. The count is taken over the decision trees that classify the observation data in one category and in the other. Cases are then classified based on a majority vote over the decision trees. Bayesian Net Works (Bn): Bayesian Network is a graphical probabilistic sub-model used in knowledge representation and prediction of ambiguous domains. This network corresponds to an extensively used structure in machine learning known as the DAG directed acyclic graph. This graph consists of several nodes, each corresponding to a random variable and the node, edges represent direct dependence among the corresponding nodes in the graph. An edge between X_1 and X_2 represent a direct dependence between these two nodes. Statistical methods are often used to obtain an estimate for these dependencies. Every variable must have a conditional probability table that shows its probability distribution knowing its direct antecedents. Also, all variables are conditionally independent of their non-descendants given their instant predecessors in the nodal frame. The parents of a node, represents their direct predecessors in the network and the conditional probability is obtained from the probability table associated with each node.

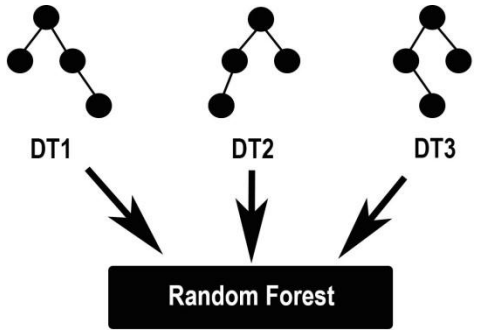

FIGURE 3. Random decision strategy

5. Deep Learning Methods

Multimodal, Brain Tumour Image Segmentation Benchmark (Brats): BRATS stands for Brain Tumour Image Segmentation Benchmark. The researchers applied 20 tumor segmentation algorithms to sixty-five MRI scans of low and high-grade gliomas patients. Gliomasare the most frequent primary brain tumors in adults. The results showed that different algorithms worked best for different parts and regions of the brain, but no single algorithm stood out the best from the rest. Benchmarking means to compare different machine learning algorithms against a particular solution. Deep Convolution Neural Networks (Cnn) For Cad (Computer aided-Detection): Convolution Neural Networks have come up as the latest trend in the deep learning domain for diagnosis and analysis of cancerous tissues and tumors. There are basically three methods to implement the CNN model- training/building the CNN model from scratch, the CNN model learns from the data available in a supervised fashion, and the unsupervised training. The researcher Hoo -Shin, and his team employed CNN to abdominal lymph node detection on the mediastina and abdominal regions and interstitial lung disease ILD detection. The above research under Hoo-Shin found that there is a trade-off between using either good training models or using more of the training data. If less training data is used then it may lead to the poor advancement of the CAD models. Deep CNN architectures with 8, even 22 layers can be useful even for CAD problems where the available training datasets are limited.

TABLE 1. Deep learning approach for Cancer Detection

Technology	Description	Results
Fuzzy method for pre- diagnosis from the Fine Needle Aspirate Analysis.	The method employed is Fuzzy Method. It is based on computational intelligence appropriate description of the Entire process of diagnosis.	This method provides 98.59% sensitivity and hence the diagnosis is more reliable.
K-Nearest Neighbours Algorithm application on Breast Cancer Diagnosis Problem.	The machine Learning classifier employed is the 'K- nearest neighbors' algorithm'. There are various perks of using this method: It is simple to implement, efficient for the small training set and no need to retrain the model if the new training pattern is added to the existing pattern.	This paper analyzes the Wisconsin-breast Cancer diagnosis problem. Though the same Algorithm cannot be applied to all problems based on diagnosis. The KNN method has it set of disadvantages like storage issues etc.
Micro calcification classification assisted by content-based image retrieval breast cancer diagnosis.	The researchers work edupon micro calcification classification method assisted by Mammogram retrieval for breast cancerd election.	This method proved to be very useful in improving the performance of the classifier by 75%-80%
Multichannel Markov Random Field Frame work for Tum our segmentation with an Application to Classification of Gene Expression-Based Breast cancer Recurrence Risk.	The aim is to have reliable and accurate segmentation, efforts are to automate the process since the segmentation is Tedious and expensive. It is a crucial process for image classification System.	Improved segmentation can lead to better extract the tumor characterization with applications on predicting the breast cancer.

The researchers conducted a number of experiments based on targeted and non-targeted CEUS data on more than 20 subjects. Each group targeted and non-targeted contained the training set, a validation set, and the test set. The training set was used to fit the learning model and the validation set to evaluate the model. The test set was used to measure the accuracy and sensitivity of the model. The deep learning experiments conducted achieved 91% specificity and 90% average accuracy.

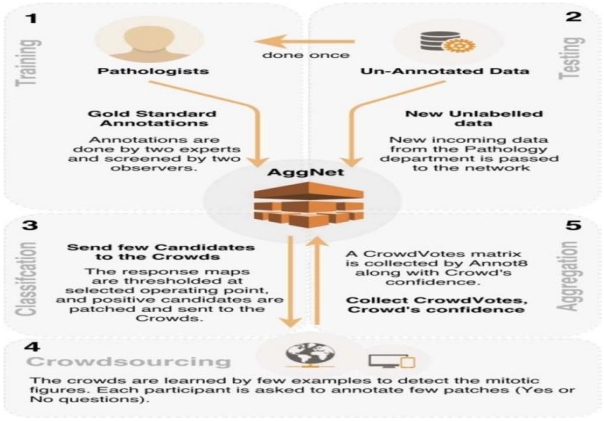

FIGURE 4. Crowdsourcing methodology

Deep learning From Crowd sourcing For Mitosis Detection in Breast cancer: Crowd sourcing is a new way to develop deep learning models. It is a participative activity wherein an individual, a group, or an organization voluntarily undertakes a task, and the data thereby gained from the activity/survey helps to develop machine learning models. It is widely nowadays in the medical domain by researchers to obtain ground-data. The quality of the crowd is very essential in order to restrict "noisy" annotations. The data is collected from both reliable (expert) and unreliable (crowd) sources and thereby the behavior of the machine learning model developed is unpredictable. Unlike typical supervised methods, which learn a model from ground truth labeled data, learning from crowd annotations is different in the sense that there may be possibly noisy multiple labels for the same sample. her2net: a deep learning framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation: Her2Net is a deep learning framework that classifies and segments the cell membranes and nuclei of stained breast cancer images. Breast cancer is the second most common type of cancer only after lung cancer. About 10% of women in the USA and the EU have been diagnosed with breast cancer at some point in their life. Usually, in clinical practice, the doctors analyze and visualize the biopsy tissue slides under a microscope. Such a diagnosis method is error-prone. Also, in many areas where expert pathologists are not available, the HER2 assessment becomes a difficult task. To address these problems, digitization and quantitative image analysis of IHC slides have become essential. Thus, pathologists mostly prefer performing quantitative image analysis using computer-aided digital techniques. The main reasons for adopting digital pathology are its reproducibility and simplicity. Additionally, digital imaging technology is a pixel-based technology, which decreases the inter-observer variability and false positivity by improving detection, segmentation accuracy, and other factors. The researchers concluded that Her2Net is a reliable and feasible deep learning framework. It can also be integrated in to computational frameworks. The Approach: CNN's are the most popular deep learning models for processing multi-dimensional array data such as color images. A typical CNN consists of multiple convolution and pooling layers followed by a few fully-connected layers to simultaneously learn a feature hierarchy and classify images. It uses error back propagation - an efficient form of gradient descent - to update the weights connecting its inputs to the outputs through its multi-layered architecture. Researchers have presented a two-stage approach using two separate CNN. The convolution neural network detects the tissue image in a nucleus while the next CNN takes chunks of data-centered at the detected nucleus as input to predict the possibility that the patch belongs to a case of recurrence. Locality Sensitive Deep Learning for Detection of Nuclei in Colon Cancer Histology Images: Tumors have generally varying degrees of heterogeneity due to their ability to cause inflammatory response, angiogenesis and tumor necrosis. Researchers have developed various locality sensitive deep learning models to detect and classify nuclei in hematoxyl in and eosin (H&E) stained histopathology images of colorectal adenocarcinoma. There are certain existing models like level sets, k-means and graph cuts and the recently proposed method is nucleus detection from H&E histology images. The proposed deep learning approach is based on two premises: distance from the center of the nucleus must be included into the calculation of probability map for detecting that nucleus, and a weighted factor of local predictions for the class label can yield a more accurate result on a nucleus. The experiment was conducted on 20000 nuclei from different histological grades. The CNN proved to be a useful tool for understanding the tumor microenvironment in a better way. The existing computing technologies such as parallel computing and graphical processing unit (GPU) are detrimental factors to improvise the prosed framework for whole-slide histology images. A decision support system to detect changes in the chromatin arrangement in cells. There have been various studies on chromatin arrangement and malignancy-associated changes (MACs) of chromatin arrangement are considered to be an invasive cancer area. MAC analysis and detection is a difficult task as it requires deep knowledge of morphologic features of chromatin network. Researchers are aiming to work on a decision support system (DSS) to automatically and objectively reproduce MAC diagnosis. The researchers took a set of 61 suspected lung cancer patients for clinical diagnosis. Artificial neural network based on DSS is designed to learn the relation between texture and morph metric parameters, computed by image processing techniques on each nucleus, with the MAC prediction each cell.

6. Simulation setup

Data Set: Wisconsin Breast Cancer Data set is accessed from the Machine Learning Repository UCI, an open-source online repository. This data set was collected periodically by Dr.Williamfrom the Wisconsin Hospital University and consists of 669 adduce, where the cases are classified as either benign or malignant. The attribute considered are clump Thickness, bare nuclei, cell Size uniformity, marginal adhesion, cell shape uniformity, epithelial cell size, bland chromatin, normal Nuclei, mitoses and class. All of the above attribute except for the class are of numerical type, and their values range from 1 to 10. The class has avalueof2forbenignand4formalignant. Training Set: The classifier will be tested using the *k fold* cross-validation method. This validation technique will randomly separate the training set into *k* subsets where 1 of the *k* 1 subsets will be used for testing and the rest for training. 10 fold cross-validation is the preferred *k* value used in most validation in ML and will be used in this paper. This means 9 subsets will be used for training of the classifier and the remaining 1 for the testing. This technique is used to avoid over fitting of the training set, which is likely to occur in small data sets and a large number of attributes.

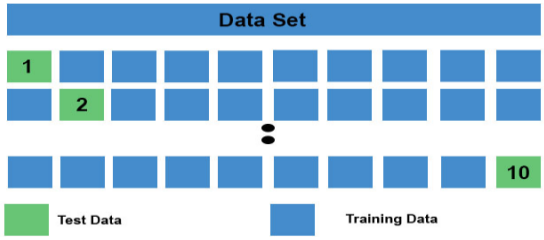

FIGURE 5. Cross validation method.

Simulation Software: Waikato Environment for Knowledge Analysis software WEK Airs used as a machine learning-tool.WEK Aisa Java-based open-source tool that was first released under the GNU General Public License. This tool provides several ML techniques and algorithms including the classification techniques that are being investigated in this paper. Data preprocessing, feature selection evaluation, clustering, and, rule discovery algorithms are processed using WEKA. Data sets are accepted in several formats, such as CSV and ARFF. Besides being an open-source tool, WEKA is also attractive due to its portability and ease of use GUI.

7. Conclusion And Future research

This paper summarizes various technological advancements and deep learning models that are currently under study and also in use to detect and segment the cancer cells and tumors. Image analysis and Convolution Neural Networks (CNN) has resulted in the early and quick diagnosis of the cancer cells so that the treatment is done timely and the patient does not have to suffer at a later stage. The focus is also to create a more unique model that can fit the needs of various diagnosis issues. The survey mainly focused on the common steps used in mitotic detection such as preprocessing, candidate detection and segmentation, feature extraction and classification. This paper presented three of the most popular ML techniques commonly used for breast cancer detection, and diagnosis, namely Support Vector Machine (SVM), Random Forest (RF), and Bayesian Networks (BN). The main features and methodology of each of the three ML techniques were described. Each algorithm executes differently based on the parameter selection and dataset. Overall methodology results that the KNN technique has given the best results. Logistic regression and Naive Bayes have also performed well in the prediction and diagnosis of breast cancer. Predictive analysis in SVM is a strong technique justifies the above finding, we conclude that the Gaussian kernel is the most suited technique for recurrence/non-recurrence prediction of breast cancer. The SVM that is used in the analysis in this paper is only applicable when the number of class variable is binary. To solve this problem scientists have come up with multiclass SVM. Fine-tuning of parameters used in algorithms can result in better accuracy. Furthermore, this can also be implemented on a cloud platform for ease of usage.

References

[1]. UCI depository -http://archive.ics.uci.edu/ml/

[2]. Maglogiannis, I., Zafiropoulos, E., &Anagnostopoulos - An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers, Applied intelligence journal, Volume 30,Issue 1, February 2009

[3]. S.Kharya, D.Dubey, S.Soni -Predictive Machine Learning Techniques for Breast Cancer Detection, IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013,1023-1028

[4]. MandeepRana, PoojaChandorkar, AlishibaDsouza, NikahatKazi – Breast cancer Diagnosis and Recurrence prediction using Machine learning Techniques,IJRET- International Journal of Research in Engineering and Technology, April2015

[5]. National Breast Cancer Foundation Inc.,http://www.nationalbreastcancer.org/about-breast-cancer.

[6]. M. Lazebnik, et al., "A large-scale study of the ultrawideband microwave dielectric properties of normal, benign and malignant breast tissues obtained from cancer surgeries," Phys. Med. Biol., vol. 52, pp. 6093–6115,2007

[7]. T. Sugitani, et al., "Complex permittivities of breast tumor tissues obtained from cancer surgeries," Applied Physics Letters, vol. 104, no. 25, pp. 253702,2014.

[8]. M. Klemm et al., "Clinical trials of a UWB imaging radar for breast cancer," in Proc. Eur. Conf. Antennas Propag. (EuCAP), Barcelona,Spain,2010, pp. 1–4.

[9]. M. Klemm et al., "Development and testing of a 60-element UWB conformal array for breast cancer imaging," in Proc. Eur. Conf. Antennas Propag. (EuCAP), 2011, pp. 1–4.

[10]. H. Song, et al., "A Radar-Based Breast Cancer Detection System Using CMOS Integrated Circuits," IEEE Access, vol.3, pp.2111-2121,2015.

[11]. H. Song, et al., "Detectability of Breast Tumor by a Hand-held Impulse-Radar Detector: Performance Evaluation and Pilot Clinical Study," Scientific Report, vol. 7, pp. 16353,2017.

[12]. S. Sasada, et al., "Portable impulse-radar detector for breast cancer: a pilot study," J. Med. Imag. 5(2), 025502(2018).

