Artificial Intelligent Models for Breast Cancer Early Detection

Erwin Halim
Information Systems Department,
School of Information Systems
Bina Nusantara University
Jakarta, Indonesia 11480
erwinhalim@binus.ac.id

Pauline Phoebe Halim
Faculty of Medicine
University of Indonesia
Jakarta, Indonesia
pauline.phoebe61@ui.ac.id

Marylise Hebrard
Directrice France
Centre sino-français de Formation et d'Échanges notariaux et juridiques à Shanghai
Shanghai, China
marylise.hebrard@cnfr-notaire.org

***Abstract*—Cancer prevalence increases year by year and the medical management develops along with it. As the number increases, breast cancer which contributes most to the number become major attention. Breast cancer detection can be done using many methods but their main purpose are all same. Considering the current situation, Machine Learning (ML) application to support cancer detection have the ability to provide better diagnostic result as in 97.4% of the cases without ML participation, malignancy was to be found on surgery while 30.6% surgeries conducted on benign lesion can be prevented. This paper propose an early detection model for breast cancer which will combine models from previous researches to become functional for many detection methods. Methods that had been proven to show accuracy in their specific field and suggested to be used in parallel: DWT- based multi-resolution MRF (MMRF) segmentation for mammography, MLP for histologic examination, and k-NN – SVMRFE method for gene identification which will be effective by learning from experience of at least 450 datasets of positive breast cancer result for each method and refers to Wisconsin Breast Cancer Diagnosis (WBDC). The research itself will be conducted from October 2018 until March 2019 and the final result will be applied to conduct further research in ML application in breast cancer early detection.**

Keywords—artificial intelligent, breast cancer, machine learning, wisconsin breast cancer diagnosis.

I. Introduction

Cancer is defined as malignant tumor that have unlimited potential growth and expands locally and systemically. [1] It ranks second in causing global death. As many other cancer, breast cancer had become a massive problem in health. It account for 8.15 million people globally in 2016 and the number is growing as the most common cancer. Although breast cancer considered to have lower mortality rate and quite high five years survival rate compare to some other cancer types, the disease burden for breast cancer is the 5th highest in the world. [2]

As many other cancer, breast cancer need to be treated by chemotherapy, radiotherapy, hormonal therapy, and in worse cases, breast surgery or mastectomy. [3] All of these treatment are costly, bears pain and discomfort towards the patient to live their daily life normally. Thus, preventive measures of getting late breast cancer diagnosis remains important to reduce patients suffering, especially early detection which is very helpful in improving breast cancer outcome and survival. [4]

Early detection of breast cancer can be done when early sign and symptoms are present or through screening. Screening methods varies from radiologic imaging, histologic examination to DNA and gene analysis. [4] These methods resulted in many data that need to be processed to determine whether a person is healthy, have high risk of getting breast cancer, or already have breast cancer. [5]

However, in the manual data processing, human error or simply technical incapability may happen which cause information received from data to be less sensitive or specific to the cause. In cancer diagnosing case, mammography data for example, cannot differentiate between benign and malignant. The incompleteness of lesion information effect the next step to be conducted: invasive needle biopsy have to conducted to determine whether the lesion is high risk lesions (HRLs) where lesion incision surgery need to be conducted to prevent malignant spreading. Yet, in 97.4% of the cases, malignancy was to be found on surgery while 30.6% surgeries conducted on benign lesion can be prevented.[6]

Meanwhile, the information system field had develop an artificial intelligence (AI) which help us in processing data to get faster information. AI recreates human perception in processing data thus reduce the need of human resource in information processing. [7] As medical science development is expressed particularly through numerous discoveries in anatomy, pharmacology, and detection, analytic, and surgical technique, AI which masters information processing to produce analysis and defies human limitation when it comes to memorizing, tiredness, and faults; have the advantage in supporting the development of detection technique.

The usage of artificial intelligence in medicine had been introduced way back at the 1940s when neurophysiologist McCulloch and Pitts create model for brain neural interaction with electric circuits. Now, AI had been applied for electronic medical record (EMR), gene chip for cancer gene detection and deep learning analysis of histopathologic skin image for skin

978-1-5386-5821-5/18/$31.00 ©2018 IEEE

carcinoma detection that can reduce human error in detecting cancer by 85%, and many more. [8]

In order to detect breast cancer accurately and reduce the need to do invasive detection method, artificial intelligence assistance is needed to create a system where detecting breast cancer can be done so even with many available detection methods.

In this paper, we are proposing an early detection model for breast cancer which will combine models from previous researches to become functional for many detection methods.

II. Literature review

A. *Modalities to Diagnose Breast Cancer*

Variety of breast cancer screening modalities including mammography, ultrasound, CT scan, MRI, PET, radar-based microwave imaging, RF sensors, biosensors, arrays, biopsy and biomarker detection in DNA and genes. Traditionally, mammography plays an important role on detecting clinically occult disease such as breast cancer and had shown to decrease mortality, however mammography imaging can be unclear and thus usually accompanied by other types of imaging. On the other hand, DNA and gene analysis requires systematic and detailed process to acquire the data itself as cancer is connected to a lot of genes depending on its type. [5]

B. *Machine Learning*

Automated learning or what commonly known as Machine Learning (ML) is a way of computer programing which allows it to learn from available input. As learning can be defined as a process of experience conversion into knowledge, ML input algorithm is in form of training data that represents experience which will be converted into an expertise with help of other program. [9]

ML as an interdisciplinary field is associated with statistics, game theory, information theory, and optimization which joins together as a subfield of computer science. In a sense, ML which general ability is to turn experience into an expertise, can be viewed as a branch of AI. However, between ML and other types of AI, there is a major difference as ML do not build automated intelligent behavior imitation and instead focuses in complementing human intelligence with the strengths and special abilities of computers. [9]

Existence of ML may assist human in doing routines such as speech recognition and image understanding; and also in doing beyond-ordinary task including complex and very large data analysis including weather prediction, astronomical data, genomic analysis, electronic commerce, and more. With sufficient training examples, all of this can be achieved by ML. ML had a flexible nature that enables it to adapt with changes which help in keeping the precision of task even when samples become varied such in detecting spam e-mails, speech recognition programs, and handwritten text decoding with varieties of handwriting. [9]

III. Methods

The methods that we proposed to be used in parallel for this research design including DWT- based multi-resolution MRF (MMRF) segmentation for mammography, MLP for histologic examination, and k-NN – SVMRFE method for gene identification. We recommend to give input from WBDC and at least 450 dataset of breast cancer positive sample for each method. The research itself is planned conducted from October 2018 until March 2019. Every hypothesis in this research will be investigated using Sequence Equation Modeling (SEM) LISREL to determine its significance level to the output. This research will also investigate qualitatively on legal and social issues.

A. *Role of Machine Learning for Cancer Diagnosis*

TABLE I. Comparison Between Traditional Diagnostic Feature and Machine Learning Feature for Breast Cancer Detection based on

***Traditional* Structural Feature**	**Machine Learning Structural Features**
Age	Pathologic Result (atypical ductal hyperplasia)
Age at First Menses	Age
Age at First Pregnancy	Biopsy type (stereotactic core biopsy)
Age at Menopause	Pathologic Result (lobular carcinoma in situ)
Ashkenazi Jewish Ancestry	Pathologic Result (atypical lobular hyperplasia)
Biopsy Type	Prior Biopsy
Breast Density	**Text Features in Pathologic Report**
Breast Imaging Reporting and Data System Category	Atypical ductal
Drinking Habits	Severely
Family History of Cancer	Atypical
Finding Type	Severely Atypical
First Mammogram	
Height	
Hormone Treatments	
Number of Children	
Pathologic Result	
Previous Breast Cancer	
Previous Other Cancer	
Prior Biopsies	
Procedure Code	
Race	
Smoking Habits	
Weight	

Bahl et al. [6]

To reach a complete diagnosis of cancer, accuracy and speed is fundamental. Beth Israel Deaconess Medical Center (BIDMC) and Harvard Medical School research found that using AI in combination with pathologist work may serve 99.5% accurate diagnosis compared to using AI alone (92%) and pathologist work alone (96%). This combination will give advantage as artificial intelligence usage reduces health costs and human resources need, as well as universal access to medical care, while pathologist analysis goes beyond the appearance of the injuries, such as the disease stage, lesion depth, and possible evolution of the cancer. [10] [11]

Machine Learning (ML) is an AI technology that have been developed for the past decades for this purpose. It enables depth analysis of data whether in form of words or picture by searching for patterns according to previous input. Originally, ML techniques for breast cancer diagnosis are tested in open source database but lately the Wisconsin Breast Cancer Diagnosis (WBCD) had surged as a benchmark data sets that proved to support ML technique development into accuracy ranging between 96.36% and 99.90% based on the WBCD's breast cancer attributes. [12] It collects data consist of computed features from fine needle aspirate (FNA) digitized image of breast mass.

Bayesian Network, a graphical representation of uncertain domain merges probability theory and graphical theory to specify a set of conditional independence assumption along with the conditional probabilities in order to describe probability distribution of a set of variables. [13] In Aloraini (2012) research, she uses Bayesian Network to further define WBCD. [14]

In accordance with the literature review conducted by Al-shamasneh and Obaidelah (2017), WBCD application had been done since 2004 for many type of ML techniques and shows a relatively good outcome (88% to 99.27%). [7] On the other hand Cruz and Wishart (2006) research shows that amongst all machine learning technique, artificial neural network (ANN) is the most common to be used up until then. [15]

TABLE II. WBDC ATTRIBUTES [11]

Number	Attribute	Domain
0	Sample Code Number	ID number
1	Clump Thickness	1-10
2	Uniformity of Cell Size	1-10
3	Uniformity of Cell Shape	1-10
4	Marginal Adhesion	1-10
5	Single Epithelial Cell Size	1-10
6	Bare Nuclei	1-10
7	Bland Chromatin	1-10
8	Normal Nucleoli	1-10
9	Mitoses	1-10
10	Class	2 for benign 4 for malignant

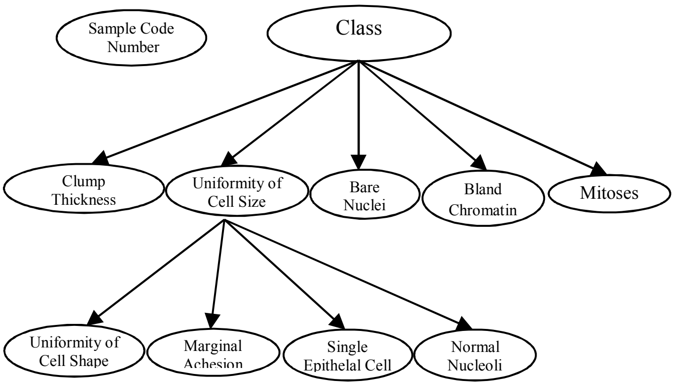

Fig. 1. Bayesian Method Classification of WBCD based on Aloraini (2012) [12]

B. Application of ML Algorithms in WBDC

According to Agarap (2018) experiment on ML Algorithms development with WBDC, out of seven algorithms (GRU-SVM, linear regression, MLP, L1-NN, L2-NN, Softmax Regression, SVM), MLP shows best result with accuracy reaching 99.03%; highest data points at 512896 points; lowest false positive rate at 1.27%; high true positive rate and true negative rate at 99.21% and 98.73% in differentiating benign and malignant tumor. [16]

Multilayered perceptron (MLP) which developed by Rosenblatt in 1958 consist of hidden layers composing numbers of perceptron that enable function approximation through

$$h_{\theta}(x) = \sum_{i=0}^{n} \theta_i x_i + b$$

$$f\left(h_{\theta}(x)\right) = \mathrm{h}_{\theta}(\mathrm{x})^{+} = max(0, \mathrm{h}_{\theta}(\mathrm{x}))$$

activation functions. [16][17]

Fig. 2. MLP Equation modified by Agarap (2018) [16]

C. Application of AI in Mamography Imaging

Zheng and Chan (2001) uses Discrete Wavelet Transform (DWT) that is extended into 2 dimensional function within *x* and *y* axes in creating 4 sub-band images: HH, HL, LH, and LL. These sub-bands contain different information and processed with different algorithm respectively. At the same time, Markov Random Field (MRF) will help in clearing up the image from noise and retrieve original image that benefits in image viewing. [18]

Combination of DWT and MRF, the DWT- based multi-resolution MRF (MMRF) segmentation algorithm, achieved cleared original image and merge it with refined image according to the compactness of each specific pixel. This process resulted in even clearer image compare to the original and support diagnostic decision further. [18]

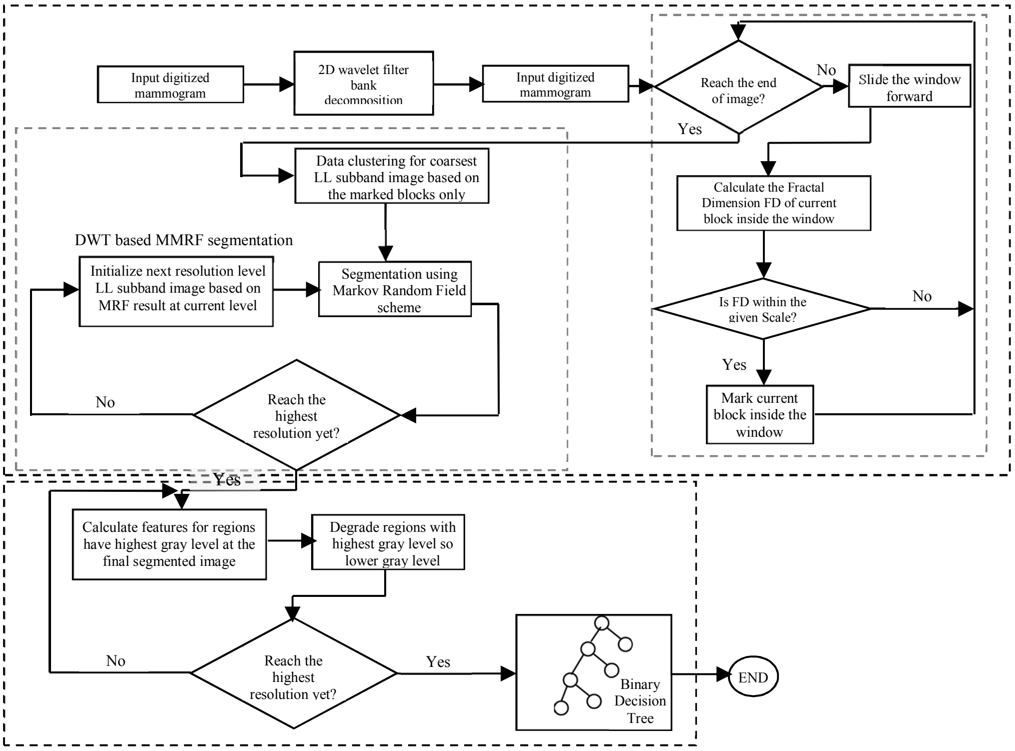
Input digitized mammogram
2D wavelet filter bank decomposition
Input digitized mammogram
Reach the end of image?
No
Slide the window forward
Yes
Data clustering for coarsest LL subband image based on the marked blocks only
Calculate the Fractal Dimension FD of current block inside the window
DWT based MMRF segmentation
Initialize next resolution level LL subband image based on MRF result at current level
Segmentation using Markov Random Field scheme
Is FD within the given Scale?
No
Yes
No
Reach the highest resolution yet?
Mark current block inside the window
Yes
Calculate features for regions have highest gray level at the final segmented image
Degrade regions with highest gray level so lower gray level
No
Reach the highest resolution yet?
Yes
Binary Decision Tree
END

Fig. 3. Tumor detection algorithm based on Zheng and Chan (2001) [18]

D. *Application of ML in Gene Expression Data*

As DNA and gene data develops, it become possible to detect specific genes that shows abnormal mutation in cancer such as P53, BRCA1, and BRCA2. [15] With help of Bayesian classification gene selection approach proposed by Sharma and Paliwal (2012), crucial genes that is related to cancer can be discovered and classified for a specific type of cancer. Sharma proposes that these genes are to be divided into subsets in comparatively smaller sets that can be updated by comparing its relation to other subsets. Related genes subsets will finally merged into an informative subset that is accurate in specifying cancer. [19]

Applying developed gene model by Bhola and Tiwari (2015) with 12625 genes featured, the result accuracy reached 73.33% just by using 24 samples. Using k-Nearest Neighbour (k-NN) as classifier and Support Vector Machine Recursive Feature Elimination (SVMRFE) as feature selection method, Bhola and Tiwari achieved the maximum accuracy of 97.5% compares to with other classifier and methods. [20]

IV. Discussion

According to reviewed literatures, it is possible to combine many algorithms and methods that had been proven to show accuracy in their specific field: MLP for histologic examination, DWT- based multi-resolution MRF (MMRF) segmentation for mammography, and k-NN – SVMRFE method for gene identification. Along with the professionalism of human resources, these methods will provide computerized data and information which support diagnostic purpose for breast cancer.

A. *Hypothesis*

Research hypothesis:

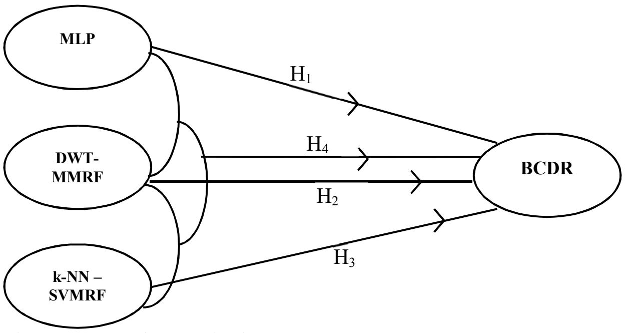

Fig. 4. Research Hypothesis

H1 MLP histopathiologic examination give positive impact to breast cancer detection result (BCDR)

H2 DWT-MMRF segmentation mammography give positive impact to breast cancer detection result (BCDR)

H3 k-NN – SVMRFE gene identification give positive impact to breast cancer detection result (BCDR)

H4 MLP histopathiologic examination, DWT-MMRF segmentation mammography, and k-NN – SVMRFE gene identification give positive impact to breast cancer detection result (BCDR)

B. *Legal and Social Issues*

ML application in breast cancer early detection arises questions: is AI a tool to serve medical staff or to replace them? Who has the final decision to the result of AI analysis? In any case, final decision after analysis is very significant as it affects delivery of treatment and surgical act. The next question is: who is going to provide new discoveries for AI system in order to be included in its calculation? Who and how any shortcoming or misuse of AI analysis result? How will human contact with patient maintained when machine occupies the main role in medical service? Patients may feel more confident going for medical check-ups when they know the diagnostic results comes from the most accurate AI software but after all, patients a human with feeling- not just pure intelligence that is facing the AI, and they will need to be emotionally supported by other humans such as doctors and nurses.

Legally and socially, AI may produce better confidence for the patient with no risk of knowledge deficiency yet at the same time it may frightening as what used to be taken in charge by human now be taken in charge by machine which ignores need of human contact and interconnection, moreover there is a possibility of perceiving the machine as a mask in case of error.

It can be compared with other cases of new technologies that are supposed to assist which in the end produce a source of anxiety for example: self-drive mode for a car or automatic pilot for a plane. In case of accident, it will take time to estimate who is responsible – the provider of the technical system, the car or the plane company? A driver or a pilot can be totally asleep when his car or plane is moving? Similar condition goes for the medical doctor as the human interface between the AI system and the patient- he has to take the final decision, make the final interpretation and more, be the representation of human dimension which includes decision to give or not to give detailed information to the patient, especially in sensitive cases concerning bioethics where there is no hope of survival or possibly inflict disability. A machine will deliver the information without considering limitations of patient but doctor have to appreciate the level of distress that such information will produce on the patient as it can accelerate the fatal issue. Thus, presence of doctor as decision maker and caregiver cannot be obstructed even in the presence of AI.

V. CONCLUSION

MLP histopathiologic examination is expected to give positive impact to breast cancer detection; DWT-MMRF segmentation mammography is expected to give positive impact to breast cancer detection; and k-NN – SVMRFE gene identification is also expected to give positive impact to breast cancer detection.

The interaction of MLP histopathiologic examination, DWT-MMRF segmentation mammography, and k-NN – SVMRFE gene identification is expected to increase the accuracy in early detection of breast cancer.

The result of our research is expected to be applied to conduct further research in ML application in breast cancer early detection.

ACKNOWLEDGMENT

We thank Marylise Hebrard, Directrice France, Centre sino-français de Formation et d'échanges Notariaux et juridiques à Shanghai, Hengshan Rd 692, SHANGHAI 200030, China for her contribution.

REFERENCES

[1] Merriam-Webster Dictionary, Cancer, Springfield: Merriam-Webster Dictionary, 2018, in press

[2] Roser M, Ritchie H, "Cancer", Oxford: Our World in data, 2018, in press

[3] Breast Cancer Organization, “Breast cancer treatment causes severe side effects in many woman”, Ardmore: Breast Cancer Organization, February 2017, in press

[4] Anderson BO et al. “Guideline implementation for breast healthcare in low-income and middle-income countries: overview of the Breast Health Global Initiative Global Summit 2007”. *Cancer*, vol. 113, pp. 2221–2243, 2008.

[5] Wang L, “Early diagnosis of breast cancer”, J. Sensors, vol. 7, pp. 1572, July 2017

[6] Bahl M, Barzila R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. "High-risk breast lesions: amachine learning model to predict pathologic upgrade and reduce unnecessary surgical excision", J. Rad. RSNA, vol 000, pp. 1-9, 2017

[7] Al-shamasneh ARM, Obaidellah UH, "Artificial Intelligence technique for cancer detection and classification: review study", European Scientific Journal, vol. 13, January 2017

[8] Miller DD, Brown EW, "Artificial inteligence in medical practice: the question to answer?", The American Journal of Medicine, vol.131, pp. 129-134, 2018

[9] Shalev-Shwartz S, Ben-David S. Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge Univeristy Press, 2014

[10] Kritz J. Artificial inteligence achieves near-human performance in diagnosing breast cancer: human and computer analyses together identify cancer with 99.5% accuracy, Boston, MA: BIDMC and Harvard University, 2016

[11] Tehrani N, Miller D. "The impact of artificial intelligence on cancer", Global J. Adv. Res., vol. 5, pp. 1-3, January 2018

[12] Yue WB, Wang ZD, Chen HW, Payne A, Liu XH, "Machine learning with applications in breast cancer diagnosis and prognosis", MDPI, vol. 2, May 2018

[13] Wolberg WH, Street WN, Mangasarian OL, “Breast cancer Wisconsin (diagnostic) data set”, UCI Machine Learning Repository, 1992

[14] Aloraini A, "Different machine learning algoriths for breast cancer diagnosis", IJAIA, vol. 3, pp. 21-30, November 2012

[15] Cruz JA, Wishart DS, "Applications of machine learning in cnacer prediction and prognosis", Cancer informatics, vol. 2, pp. 59-78, 2006

[16] Agarap AFM, "On breast cancer detection: an artificial intelligence learing algorithms on the Wisconsin diagnostic dataset", ICMLSC 2018, March 2018.

[17] Frank Rosenblatt. 1958. “The perceptron: A probabilistic model for information storage and organization in the brain.”, Psychological review, vol. 65, pp. 386- 408, April 1958

[18] Zheng L, Chan AK,. "An artificial intelligent algorithm for tumor Detection in screening mammogram", IEEE Trans. Med. Imaging, vol. 20, pp 559-567, July 2001

[19] Sharma A and Paliwal KA. "A gene selection algorithm using Bayesian classification approach." American Journal of Applied Sciences, vol. 9, pp.127-131, 2012

[20] Bhola A, Tiwari AK, "Machine learning based approaches for cancer classificaition using gene expression data", MLAIJ, vol. 2, December 2015