INTERNATIONAL JOURNAL OF TECHO-ENGINEERING

ISSN : 0975-4520

DETECTION OF BREAST CANCER DIAGNOSIS USING MACHINE LEARNING

YARRAMSETTY VINAYA[1], YAMINI MANIKONDA[2], MOTUMARRI JAHNAVI[3], PALLA SWATHI[4], Dr. M.BHEMA LINGAIAH [5]

[1, 2, 3, 4]UG Scholars, Department of CSE, ***MALINENI LAKSHMAIAH WOMEN'S ENGINEERING COLLEGE, GUNTUR***, India.

[5] Professor, Department of CSE, ***MALINENI LAKSHMAIAH WOMEN'S ENGINEERING COLLEGE, GUNTUR***, India.

ABSTRACT:

In the field of assisted cancer diagnosis, it is expected that the involvement of machine learning in diseases will give doctors a second opinion and help them to make a faster / better determination. There are a huge number of studies in this area using traditional machine learning methods and in other cases, using deep learning for this purpose. This article aims to evaluate the predictive models of machine learning classification regarding the accuracy, objectivity, and reproducible of the diagnosis of malignant neoplasm with fine needle aspiration. Also, we seek to add one more class for testing in this database as recommended in previous studies. We present six different classification methods: Multilayer Perceptron, Decision Tree, Random Forest, Support Vector Machine and Deep Neural Network for evaluation. Fo this work, we used at University of Wisconsin Hospital database which is composed of thirty values which characterize the properties of the nucleus of the breast mass. As we showed in result sections, DNN classifier has a great performance in accuracy level (92%), indicating better results in relation to traditional models. Random forest 50 and 100 presented the best results for the ROC curve metric, considered an excellent prediction when compared to other previous studies published.

Keywords: ROC, DNN, multi layer, ANN.

1. INTRODUCTION:

Breast cancer is one of the most dangerous and common reproductive cancers that affect mostly women. The oldest documented cases of breast cancer were in Egypt in 3000 BC. Breast tumor is an abnormal growth of tissues in the breast, and it may be felt as a lump or nipple discharge or change of skin texture around the nipple region. Cancers are abnormal cells that divide uncontrollably and are able to invade other tissues. Cancer cells have the ability to spread to other parts of the body through the blood and lymphatic systems. It is the leading cause of death among middle aged and older women. According to cancer statistics, breast cancer is the second most common and the leading cause of cancer deaths among women, second only to lung cancer. Around 1 in 36 (3%) women dies due to breast cancer. It has become a major health issue in the past 50 years, and its incidence has increased in recent years in Malaysia, breast cancer is the most frequent type of cancer among women. It has an incidence rate of about 26% (more than 4400 women) among cancer affecting women. Around 40% of the women who suffered from breast cancer in Malaysia have died (IARC). Hence, determining the right decision from a right diagnosis is crucial.

In today's world with the advent of personalized medicine, it increases the workload and complexity of the doctors in cancer diagnosis. Radiologic and pathology are the key players in making decision for cancer diagnosis. Based on the radiology diagnosis, the results will be submitted to pathology for further diagnosis. Pathology and radiology form the core of cancer diagnosis, yet based on our observation at our studied hospital and under current process of diagnostic medicine, the communication among them remained on papers. That paper contains their respective report of the case on the same patient. This scenario is in parallel with what James et al. Had highlighted in their paper. The working flows of both specialties remain ad hoc and occur in separate "silos," with no direct linkage between their case accessioning and reporting systems, even when both departments belong to the same host institution. Since both radiologists' and pathologists' data are essential to make correct diagnoses and appropriate patient management and treatment decisions, the isolation of radiology and pathology work flows can be detrimental to the quality and outcomes of patient care. These detrimental effects underscore the need for pathology and radiology work flow integration and for systems that facilitate the synthesis of all data produced by both specialties. With the enormous technological advances currently occurring in both fields, the opportunity has emerged to develop an integrated diagnostic reporting system that supports both specialties and, therefore, improves the overall quality of patient care. In this chapter, we are focusing on breast cancer diagnostic for data collected from UKMMC. Hence, breast radio-pathological correlation is essential. The covered topics would include radio-pathological correlation with recent imaging advances such as machine learning with use of technical methods such as mammography and histopathology.

As a standard, the current diagnostic screening consists of a mammography to identify suspicious regions of the breast, followed by a biopsy of potentially cancerous areas. A breast biopsy is a diagnostic procedure that can determine if the suspicious area is malignant or benign. Although criteria for diagnostic categories of radiologic and pathology are well established, manually detection and grading respectively is a tedious and subjective process and thus suffers from inter-observer and intra-observer variations. Early detection via mammography increases breast cancer treatment options and the survival rate. However, mammography is not perfect. Detection of suspicious abnormalities is a repetitive and fatiguing task. For every thousand cases analyzed by a radiologist, only three to four are cancerous, and thus an abnormality may be overlooked. As a result, radiologists fail to detect 10–30% of cancers. Approximately two thirds of these false-negative results are due to missed lesions that are evident retrospectively. Due to the considerable amount of overlap in the appearance of malignant and benign abnormalities, mammography has a positive predictive value (PPV) of less than 35%, where the PPV is defined as the percentage of lesions subjected to biopsy that were found to be cancer. Thus, a high proportion of biopsies are performed on benign lesions. Avoiding benign biopsies would spare women anxiety, discomfort, and expense. As mentioned earlier, with the advent of personalized medicine, the process becomes more complex. Not only that, the emerging of 4th Industrial Revolution (4IR) technology allowed huge amount of data to be captured, and this contributes to the complexity of the radiology and pathology workload. To address these challenges, many researchers are leveraging artificial intelligence to improve medical diagnostics. Machine learning is a sub discipline in the field of artificial intelligence (AI) that explores the study and design of algorithms that can learn from data.

2. PREVIOUS STUDY:

ML comprises a broad class of statistical analysis algorithms that iteratively improve in response to training data to build models for autonomous predictions. In other words, computer program performance improves automatically with experience. ML algorithm’s aim is to develop a mathematical model that fits the data. It comprises of two types of learning which are supervised and unsupervised. Supervised learning algorithm required the data to be labeled for training purposes. For example, in training a set of medical images to identify a specific breast tumor type, the label would be tumor pathologic results or genomic information. These labels, also known as ground truth, can be as specific or general as needed to answer the question. The ML algorithm is exposed to enough of these labeled data to allow them to move into a model designed to answer the question of interest. Because of the large number of well-labeled images required to train models, curating these data sets is often laborious and expensive. Unsupervised ML clusters the data that have similar characteristics, and the unlabeled data are exposed to the algorithm with the goal of generating labels that will meaningfully organize the data. This is typically done by identifying useful clusters of data based on one or more dimensions. Compared with supervised techniques, unsupervised learning sometimes requires much larger training data sets. Unsupervised learning is useful in identifying meaningful clustering labels that can then be used in supervised training to develop a useful ML algorithm. This blend of supervised and unsupervised learning is known as semi-supervised.

3. PROPOSED SYSTEM:

The data set was provided in a CSV file, containing 837751 registers. It was performed from a data set of 569 women, being: The first column of the patient identification code, which is not being used in the training process. The second column is Diagnosis, where 1 indicates Malignant, and 0 indicates benign. The rest of the columns are 30 numeric values that show the measurements of the cell nucleus. The last column was deleted due to contained only NaN values. For the cell nucleus, the inclusion, texture, perimeter, area, softness, compactness, concavity, symmetry and large fractal are measured ten times. The significant error, default, and lower values are the properties calculated, resulting in 10 x 3, 30 columns of input data. In our feature selection/extraction, we opted for the crossvalidation method. Cross-Validation is a technique that aims to understand how your model generalizes, or how it behaves when you predict a data you have never seen. This metric creating different training ‘ s and testing sets, to make sure that the model is performing well. In this case, instead of using only one test set to validate our model, we will use N others from the same data.

RESULTS:

Towards the analysis of our algorithm, we used Jupyter Notebook, python modules (pandas, matplotlib, bumpy) and a scikit-learn framework to process ML algorithms. The following evaluated methods were: Multilayer Perceptron, Decision Tree, Random Forest, Support Vector and Deep Neural network. We divided Random Forest into two sizes: 50 and 100 Trees, aiming to test the different size of trees to verify if their accuracy prediction would be different. We start our experiment splitting our base for training and testing, separating training set in 70 % (398 randomized records) and 30% for test. In this step, we apply one more process for the testing set, splitting into two parts, 50/50. The main idea was to verify in two stages if we obtained a significant difference among the groups. Still, we seek to reduce the chance of over fitting. We need to highlight that DNN model not participated in this split, to verify if without this process the algorithm could present a behavior much different from others.

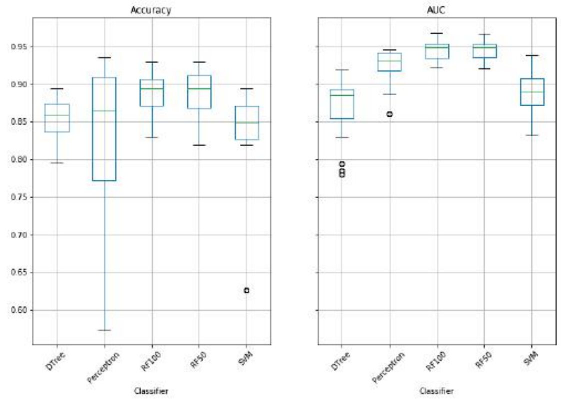

Fig. 4.1. Accuracy and AUC comparative models.

4. CONCLUSION:

Our study presented a set of classification models, trying to find the best model to classify Breast Cancer according to our data set (WDBC). For this proposal, we selected five different techniques of machine learning, which were considered in other studies with similar proposals. Random Forest was divided between two models: 50 and 100 trees collections. Also, we add Deep Neural Network to visualize their performance in comparison to other classifier methods. Furthermore, we use a group of metrics to evaluate all results. In this sense, we gave special attention to accuracy and ROC curve measures, proposing a comparison and discussion between these metrics. The outcomes obtained from experiments have been analyzed across, data tables and charts. Regarding our results, Random forest models and Neural Network models presented the best results for the accuracy and the ROC curve. Other models such as Decision Trees and Support Vector produced lower results. Which model has the highest accuracy, objectivity, and reproducibility? It is not so easy to see if one algorithm is better than another only by looking at the error - rate and accuracy values, since there is no classification algorithm for all the challenges to be overcome.

REFERENCES:

[1] M. Da Saude, " Incidencia de cancer no brasil – estimative 2018," comentarios.asp, p. 130, 2018. [Online]. Available: {http://www1. inca.gov.br/estimativa/2018/sintese-de-resultados-comentarios.asp}

[2] J. Hwang and C. M. Christensen, “Disruptive innovation in health care delivery: a framework for business-model innovation,” *Health Affairs*, vol. 27, no. 5, pp. 1329–1335, 2008.

[3] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, “ Using machine learning algorithms for breast cancer risk prediction and diagnosis, ” *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

[4] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado *et al.*, “ Detecting cancer metastases on gigapixel pathology images, ” *arXiv preprint arXiv:1703.02442*, 2017.

[5] E. Ali ˇ ckovi´c and A. Subasi, “ Breast cancer diagnosis using ga feature selection and rotation forest, ” *Neural Computing and Applications*, vol. 28, no. 4, pp. 753–763, 2017.

[6] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, “ A support vector machine-based ensemble algorithm for breast cancer diagnosis, ” European Journal of Operational Research, vol. 267, no. 2, pp. 687–699, 2018. [Online]. Available: https://doi.org/10.1016/j.ejor.2017.12.001

[7] Y.-Q. Liu, C. Wang, and L. Zhang, “ Decision tree based predictive models for breast cancer survivability on imbalanced data,” pp. 1–4, 2009.

[8] B. Diniz et al., “Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks, ” Computer Methods and Programs in Biomedicine, vol. 156, pp. 191 – 207, mar 2018.

[9] S. Sharma and P. Khanna, “ Computer-aided diagnosis of malignant mammograms using zernike moments and svm,” Journal of Digital Imaging, vol. 28, no. 1, pp. 77–90, 2015.

[10] R. W. D. Pedro, A. Machado-Lima, and F. L. Nunes, “ Is mass classification in mammograms a solved problem? - a critical review

over the last 20 years, ” Expert Systems with Applications, vol. 119, pp. 90 – 103, 2019.