International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

COMPARISON OF MACHINE LEARNING METHODS FOR BREAST CANCER DIAGNOSIS

SUMIT KUMAR[1], HANMANNOLLA AMULYA[2], AZMEERA TEJA[3], SAKE SAI VIKRAM[4], MITTAPALI ABHILASH[5]

[1]Assistent Professor, Department of CSE-IOT, Malla Reddy College of Engineering Hyderabad, TS, India.

[2,3,4,5] UG students, Department of CSE-IOT, Malla Reddy College of Engineering Hyderabad, TS, 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. For 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, perceptron, deep learning, decision Tree.

1. INTRODUCTION:

In Brazil, for the biennium 2018-2019, 59,700 new cases of breast cancer are anticipated. Breast cancer accounts for 25.2% of female malignancies and an incidence rate of 43.3 /100,000 women. An estimated in 522,000 deaths a year, breast cancer is responsible for 14.7% of all deaths. Although it has a higher mortality rate than other malignancies, it has a low fatality because its mortality rate is less than 1/3 of the incidence rate. It is the most surviving

cancer type annually, approximately 8.7 million. In developed countries the numbers have stabilized, followed by a drop in the last decade. In underdeveloped countries, detection occurs in more advanced stages, contributing to the treatment-related morbidity rate .The disruptive technology applications in the health area have been focused on studying the potential impact on human society.

Regarded the assisted cancer diagnosis, it is expected that the involvement of machine learning in diagnosis could provide doctors a second opinion and help them to make a faster/ better diagnosis.Recently, Google reached an accuracy level in identifying skin cancers, suggesting that the cancer accessibility diagnosis could potentially be extended for aside from medical clinics. The application employed Deep Learning to train a neural network classifier with one of the Wisconsin breast cancer data sets (diagnosis), using the classifier to predict the mammary mass prediction with 30 real numerical values that characterize the cell nucleus properties of mammary mass. Although many studies have been studied breast cancer prediction/classification, we propose a study using a specific algorithms group, containing a random forest split for diversified analyzes. The focus in this field is to apply classification techniques and perform classification/prediction directly from the digital image. In our experiment, we showed the classification of breast cancer with numerical data calculated from the digitized image of a fine needle aspirate (FNA) of a mammary mass. This study aims to evaluate the predictive models of machine learning classification regarding accuracy, objectivity, and reproducibility of the malignant neoplasm diagnosis with fine needle aspiration. An experiment was performed with a data set of 569 women diagnosed with breast cancer or not. Throughout the outcomes, it was possible to state that the DNN‘s model has the best results among the other techniques, having a mean accuracy of 92%, while Random Forest collections presenting a ROC curve coefficient of 94%. The primary contribution provided an overview of machine learning models, looking for their outcomes when tested with a breast cancer data set. We selected models previously used in other studies, applying a different workflow in training data phase. Moreover, we add a Deep Neural Network method, which isn’t tested yet for this data set. Some studies have applied this approach in other image datasets, being proved their utility in this field. In our context, we aim to show the network results were evaluated by standard metrics of machine learning and discuss their application when compared to other methods. The comparison of these techniques, adding deep neural networks was expected from other studies in this area. Cancer is the second reason of human death all over the world and accounts for roughly 9.6 million deaths in 2018. Globally, for 1 human death in 6 can be said that is caused by cancer. Almost 70 percent of the deaths from cancer disease happen in countries that have low and middle income. The most common cancer type among women are breast, lung and colorectal, which totally symbolize half of the all cancer cases. Also, breast cancer is responsible for the thirty

percent of all new cancer diagnoses in women. Machine learning (ML) methods ensure analyzing the data and extracting key characteristics of relationships and information from dataset. Also, it creates a computational model for best description of the data. Especially, according to in researches about cancer disease, it can be said that ML techniques can be handled on early detection and prognosis of cancer. Asri et al. have compared some machine learning algorithms for the risk prediction and diagnosis of breast cancer. Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Naive Bayes (NB) and Decision Tree (C4.5) have been applied Wisconsin Breast Cancer (Original) dataset. SVM classification method has been given the highest accuracy value (97.13 %) with least error rate when the experimental results were compared.

2. LITERATURE SURVEY:

1) Detecting Cancer Metastases On Gigapixel Pathology Images

AUTHORS: Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado

Each year, the treatment decisions for more than 230,000 breast cancer patients in the U.S. hinge on whether the cancer has metastasized away from the breast. Metastasis detection is currently performed by pathologists reviewing large expanses of biological tissues. This process is labor intensive and error-prone. We present a framework to automatically detect and localize tumors as small as 100 x 100 pixels in gigapixel microscopy images sized 100,000 x 100,000 pixels. Our method leverages a convolutional neural network (CNN) architecture and obtains state-of-the-art results on the Camelyon16 dataset in the challenging lesion-level tumor detection task. At 8 false positives per image, we detect 92.4% of the tumors, relative to 82.7% by the previous best automated approach. For comparison, a human pathologist attempting exhaustive search achieved 73.2% sensitivity. We achieve image-level AUC scores above 97% on both the Camelyon16 test set and an independent set of 110 slides. In addition, we discover that two slides in the Camelyon16 training set were erroneously labeled normal. Our approach could considerably reduce false negative rates in metastasis detection.

2) Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks

AUTHORS: B. Diniz

The processing of medical image is an important tool to assist in minimizing the degree of uncertainty of the specialist, while providing specialists with an additional source of detect and diagnosis information. Breast cancer is the most common type of cancer that affects the female population around the world. It is also the most deadly type of cancer among women. It is the second most common type of cancer among all others. The most common examination to diagnose breast cancer early is mammography. In the last decades, computational techniques have been

developed with the purpose of automatically detecting structures that maybe associated with tumors in mammography examination. This work presents a computational methodology to automatically detection of mass regions in mammography by using a convolutional neural network. The materials used in this work is the DDSM database. The method proposed consists of two phases: training phase and test phase. The training phase has 2 main steps: (1) create a model to classify breast tissue into dense and non-dense (2) create a model to classify regions of breast into mass and non-mass. The test phase has 7 step: (1) preprocessing; (2) registration; (3) segmentation; (4) first reduction of false positives; (5) preprocessing of regions segmented; (6) density tissue classification (7) second reduction of false positives where regions will be classified into mass and non-mass. The proposed method achieved 95.6% of accuracy in classify non-dense breasts tissue and 97,72% accuracy in classify dense breasts. To detect regions of mass in non-dense breast, the method achieved a sensitivity value of 91.5%, and specificity value of 90.7%, with 91% accuracy. To detect regions in dense breasts, our method achieved 90.4% of sensitivity and 96.4% of specificity, with accuracy of 94.8%.

3) Is mass classification in mammograms a solved problem? - a critical review over the last 20 years

AUTHORS: R. W. D. Pedro, A. Machado-Lima, and F. L. Nunes

Breast cancer is one of the most common and deadliest cancers that affect mainly women worldwide, and mammography examination is one of the main tools to help early detection. Several papers have been published in the last decades reporting on techniques to automatically recognize breast cancer by analyzing mammograms. These techniques were used to create computer systems to help physicians and radiologists obtain a more precise diagnosis. The objective of this paper is to present an overview regarding the use of machine learning and pattern recognition techniques to discriminate masses in digitized mammograms. The main differences we found in the literature between the present paper and the other reviews are: 1) we used a systematic review method to create this survey; 2) we focused on mass classification problems; 3) the broad scope and spectrum used to investigate this theme, as 129 papers were analyzed to find out whether mass classification in mammograms is a problem solved. In order to achieve this objective, we performed a systematic review process to analyze papers found in the most important digital libraries in the area. We noticed that the three most common techniques used to classify mammographic masses are artificial neural network, support vector machine and k-nearest neighbors. Furthermore, we noticed that mass shape and texture are the most used features in classification, although some papers presented the usage of features provided by specialists, such as BI-RADS descriptors. Moreover, several feature selection techniques were used to reduce the complexity of the classifiers or to increase their accuracies. Additionally, the survey conducted points out some still unexplored research opportunities in this area, for example, we identified that some techniques

such as random forest and logistic regression are little explored, while others, such as grammars or syntactic approaches, are not being used to perform this task.

3. EXISTING SYSTEM

In Existing system the mammography mass detection was designed to increase the performance of specialists by serving as double-reading systems and contributing to the reduction of the number of false-positive or false-negative. There is numerous mass segmentation methods in mammograms, a summary of the most relevant methods are selected from dataset, the evaluation metrics presented are the most frequently used in the literature. However, it is considered an unresolved problem, mainly due to the small number of images used in the studies, mass variability and computational limitations.

DISADVANTAGES OF EXISTING SYSTEM:

- To obtaining a consistent dataset and labeled by specialists in the medical field is one of the main challenges in the development of a CAD(Computer-aided detection)
- The amount of images provided by the bases is still insufficient for the generalization of the problems, due to the variability and size of the masses

  Algorithm: Yolo, Full Resolution Convolutional Network (FrCN)

4. PROPOSED SYSTEM:

A deep belief network was used for the detection of breast cancer using a technique of back-propagation supervised path using the Wisconsin Breast Cancer Dataset (WBCD). This approach offers 99% accuracy in the classification task. Compositions using deep learning neural network model and SVDD, a variant of the support vector machine, show experimental results to learn multi-class data without severe over-fitting problems. The random Forest model also presents great results with our implementations. We tested with other models like Decision Tree, Support Vector Machine, Neural Network, and Multi-Layer Perceptron. In this study were used data sets combined and splitting for testing, as well as accuracy indicator as a measure for assessing the results.

ADVANTAGES OF PROPOSED SYSTEM:

- Identifying the use of data-enhancement and transfer learning techniques that indicate an improvement in the performance of deep learning models.
- One of the main advantages of using deep networks techniques when compared to manual resource extraction techniques is the ability to learn a set of high-level attributes and provide high accuracy even in complex problems

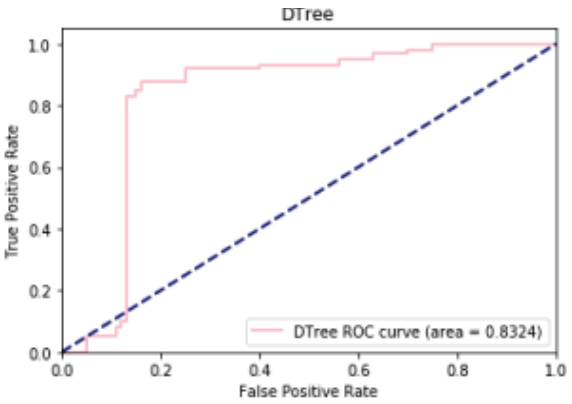
Algorithm: Multi-Layer Perceptron, Decision Tree, Random Forest, Support Vector Machine, Deep Neural Network.

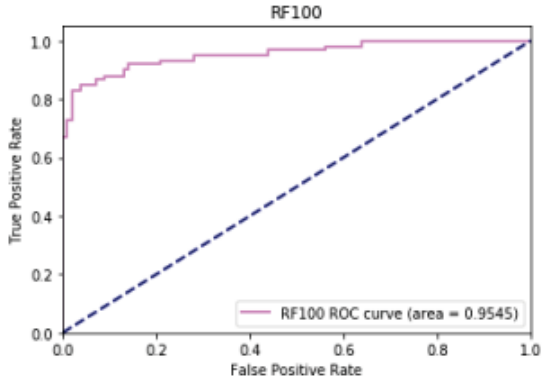
4. RESULTS EXPLANATION

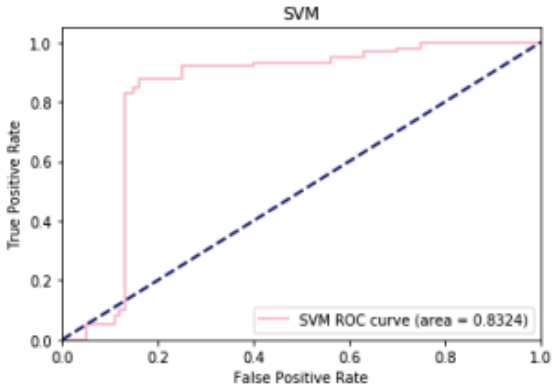
Towards the analysis of our algorithm, we used Jupyter Notebook, python modules (pandas, matplotlib, bumpy) and a scikit-

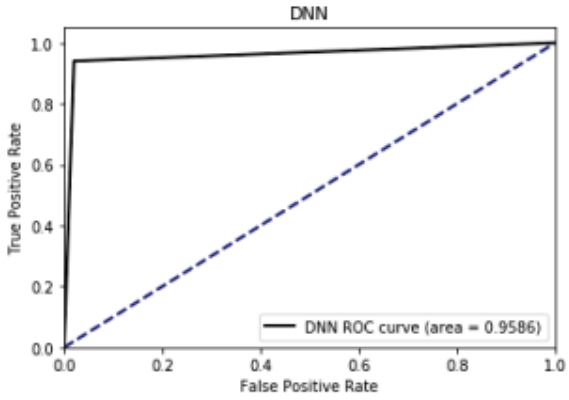
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

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. The mean and log loss coefficient of the two stages test approach was shown in 1 and 2.









5. 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. 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. It is important to understand the power and limitations of different classifiers, and there is a scale for the challenge/community to use it in the best possible way in order to compare the models in question. A good review of algorithm comparison can be found in. Deep Neural Network had a good performance in this study, although their reach better results in studies involving images. Breast Cancer has provided many studies in recent years, through different approaches as computing vision, classification, and prediction. As future work, we considered an improvement in predictions, testing approaches in databases containing images.

FURTHER ENHANCEMENT

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.

REFERENCES:

[1] M. Da Saúde, "Incidência de câncer no brasil - estimativa 2018," http://www1.inca.gov.br/estimativa/2018/sintese-de-resultados 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. Aliˇckovi´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.