AN ANALYSIS OF MACHINE LEARNING CLASSIFIERS IN BREAST CANCER DIAGNOSIS

Achanta. Harika, Dr.I.R.Krishnam Raju, Sri.V.Bhaskara Murthy,
MCA Student, Professor, Associate Professor
Dept Of MCA
B.V.Raju College, Bhimavaram

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.

I. 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.

Bazazeh and Shubair have investigated the comparative study of machine learning techniques as Support Vector Machine (SVM), Random Forest (RF) and Bayesian Network (BN) for detection and diagnosis of breast cancer. The Original Wisconsin Breast Cancer was used as a dataset and Weka software was used as a Machine Learning tool. The key performance parameters of machine learning classifiers have been compared according to accuracy, recall, precision and ROC area. They have suggested that BN has the best performance according to recall and precision values and RF technique has optimum performance in term of ROC area. Ahmad et al. have exercised machine learning algorithms for predicting the rate of two years recurrence of breast cancer disease. The dataset has been obtained from Iranian Center of BreastCancer (ICBC) program, collected the time period of 1997-2008 years. The dataset is consisted of population characteristics and 22 input variables also the cases have been collected from 1189 women of diagnosed breast cancer. Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (DT) have been applied and SVM has been showed the best performance with highest accuracy and least error rate. Bektas and Babur have studied on diagnosis of breast cancer using machine learning techniques. Kent Ridge Microarray has been used 2 datasets and support vector machine, k-star, random forest algorithm and voted perceptron have been applied. Random forest algorithm has been showed more performance than applied feature selection method. Chen et al. have applied Support Vector Machine classification algorithm on Wisconsin Diagnostic Breast Cancer dataset. In the study, the training and testing sets have been split as 50-50%, 70-30% and 80-20%. According to different training/testing percent, accuracy values have been calculated.

III. 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 are 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)

IV. PROPOSED SYSTEM

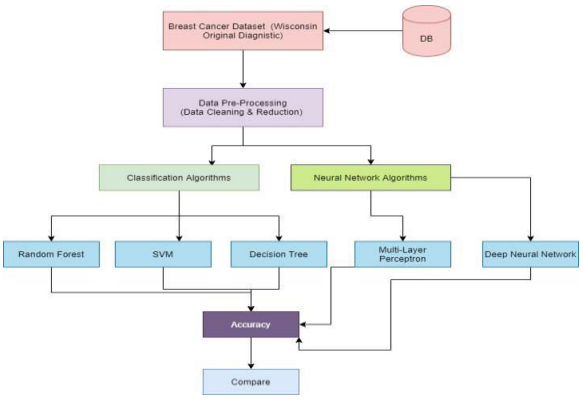
A deep belief network was used for the detection of breast cancer using a technique of back-propagation supervised path using the Wisconsin Breast Cancer Dataset (WBCD). This approach offers a 99% accuracy in the classification task. Compositions using deep learning neural network model and SVDD, a variant 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.

V. SYSTEM ARCHITECTURE



VI. IMPLEMENTATION

MODULES:

- **User**
- **Admin**
- **Classification**
- **Neural Network**

MODULES DESCRIPTION:

User:

The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user register then admin can activate the customer. Once admin activated the customer then user can login into our system. The user can get the data from University of Wisconsin Hospital database.

The data are stored in media folder. Before processing the data we need to preprocess the data. At the time of preprocess we can generate the graphs like hist diagram of the selected attributes. Later we can split the data into training and testing. The 80% data goes to training and 20% we are testing for our results. The sklearn model selection libraries can do the process.

Admin:

Admin can login with his credentials. Once he login he can activate the users. The activated user only login in our applications. The admin can set the training and testing data for the project. In the code the dataset can be found under media folder. The dataset in the format of comma separated values. User can perform the cleaning and fill with its mean values of missing featured from the columns. Admin also check the accuracy scores of proposed algorithms. First admin can test the classification results. To see the graph we need to enable matlotlib.use("TkAgg"). Once it done the he can test the deep neural network algorithm accuracy. It is better to user before running project we need to enable matplotlib.user("Agg") then we can solve the server restarts problems.

Data Classification:

- **Decision Tree**

Decision tree algorithms are considered an alternative for regressions and classifiers tasks. the Decision Tree Algorithms structure can be compared to a set of rules (If-then), classifying new samples and trying to develop an understandable and accurate model. Thereby, the Decision Tree algorithm operates such as others Supervised Learning techniques, working with sets for training and tests.

- **Random Forest**

Defined as an ensemble learner, Random Forest works creating multiple classifiersand regression trees, each one trained based on the subset of training examples and the subset of all given features at random. Each decision tree, the input enters at the root of the tree and traverses down the tree according to the split decision at each node.

Support Vector Machine Multilayer Perceptron (MLP) is a classifier based on the neural network's, very similar to perceptron but with more layers. Each output layer receives the stimulus of the intermediate layer, building a set of appropriate outputs. MLP uses a supervised learning technique knew as backpropagation function, which learns iteratively by processing data set of training examples, comparing the network's prediction for each target value.

Deep Neural Network Lastly, we considered using a deep neural network to verify their performance related to this database. Presenting relevant results in recent studies, this network has been used in many tasks we found value in testing this network due to good results in previous binary classification studies. Also, DNN's algorithms were suggested for application in this database as a way to verify their performance in comparison to traditional methods of machine learning.

VII. 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.

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.