IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AN ANALYSIS OF MACHINE LEARNING CLASSIFIERS IN BREAST CANCER DIAGNOSIS

P. MANISH[1], M.MONIKA[2], V.TEJA SRUTHI[3], P.NAGA VANDANA[4] AND G.HARIKA[5]

[1]Assistant Professor, [2,3,4,5]Research Scholar

[1]Department of Information Technology,

[1]Pragati Engineering College (A), Surampalem, 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.

Index Terms - ***Diagnosis of malignant neoplasm, Multilayer Perceptron, Decision Tree, Random Forest, Support Vector Machine and Deep Neural Network.***

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.

II. PROBLEM DEFINITION

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%.

III. OBJECTIVE OF PROJECT

- 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 research about cancer disease, it can be said that ML techniques can be handled on early detection and prognosis of cancer.

IV. 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.

V. 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.

VI. REQUIREMENT SPECIFICATION

HARDWARE REQUIREMENTS:

System: Intel i3
Hard Disk : 1 TB.
Monitor : 14' Colour Monitor
Mouse: Optical Mouse.
Ram: 4GB.

SOFTWARE REQUIREMENTS:

Operating system : Windows 10.
Coding Language : Python.
Front-End: Html. CSS
Designing: Html, CSS, javascript.
Database: SQLite.

VII. FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

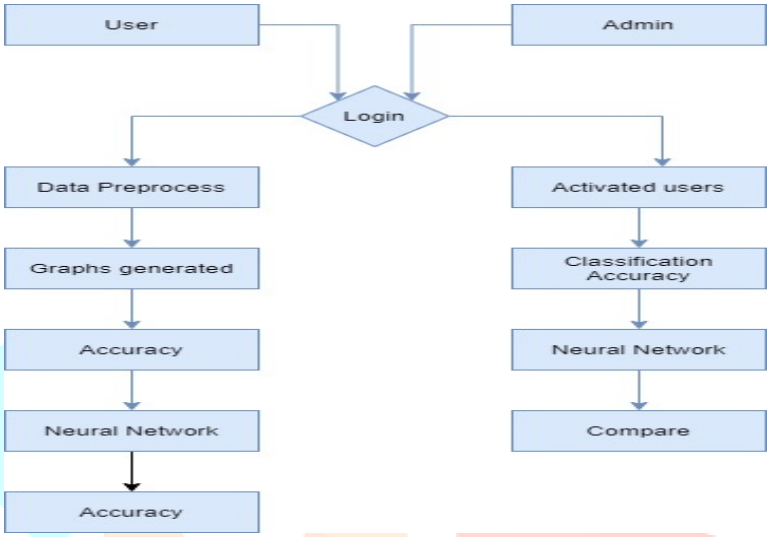
SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

VIII. SYSTEM DESIGN

Food, clothing, and shelter are the essential needs of life. Availability of these needs increases the physical effectiveness and productivity of the people. So housing is a factor of prime importance in human resource development of any economy. At one point in life, everybody has to deal with the housing dilemma. For many people housing is one of the major investments of their life, people pay a fortune to buy their Dream House. Data is at the heart of technical innovations, achieving any result is now possible using predictive models. Machine learning is extensively used in this approach. Machine learning means providing valid dataset and further on predictions are based on that, the machine itself learns how much importance a particular event may have on the entire system based on its preloaded data and accordingly predicts the result. Various modern applications of this technique include predicting stock prices, predicting the possibility of an earthquake, predicting company sales and the list has endless possibilities.

FLOWCHART OF PROJECT



Flowchart of project

IX. MODULES

User:

The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user registers 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 classifiers and 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: Support Vector Machines (SVMs) is a supervised machine learning technique, having great theoretical foundations and excellent empirical successes. The SVM has the constraint which makes the total weight for the positive class equal to that of the negative class. This kind of technique has been applied to different classification tasks such as text classification, object recognition, as well as prediction tasks.

Neural Network

Multi-Layer Perceptron: 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.

X. 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.

XI. 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. Ali˘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.