International Journal of Precious Engineering Research and Applications (IJPERA)
www.ijpera.com ISSN: 2456-2734,
Volume 7, Issue 1 (May) 2022), PP 83-88

Cancer Treatment Recommendation using Machine LearningTechnologies

Sri Chandana K,Sriram Balakrishna
[1]*Student, R.V College of Engineering, Bengaluru, Karnataka*
[2]*Student,R.V College of Engineering, Bengaluru, Karnataka.*

ABSTRACT*: Breast cancer in India has been ranked number one among the females in India. As per the statistics and studies, the data reports say that 41 per 100000 women for Delhi, Chennai 38 per 100000, Bengaluru 34.4 per 100000 and Thiruvananthapuram 33.7 per 100000 reports as breast cancer patients. In rural areas the ratio is sometimes as high as 66 in the registries and in cities it is as low as 8 in the urban registries. As seen the statistics are very high, prediction of each course of treatment and a follow up treatment for the patient (if required) should be tailored to each patient when it comes to cancer treatment. For the purpose of working for these kinds of patients, we developed a model using both clinical and genomic data which can provide the results for the patients for a given input of clinical and genomic data. After thorough literature survey and consulting with different medical practitioners we found that both clinical and genomic data are important in the treatment of breast cancer. Machine learning (ML) allows us to analyze and visualize the data and discover the relations between the data present and predict the right treatment for a particular patient's clinical and genomic data. We analyzed 1904 patients who underwent breast cancer surgery and obtained their genomic and clinical data. After data pre-processing and visualization, we removed the unnecessary columns which were not helpful for the model by using the relationship matrix(heatmap) and eliminated the values which violated the conditions of the dataset. We developed a single machine learning model (Artificial Neural Network) where this algorithm was tested and obtained an accuracy of 89.83 percent and in few of the evaluation metrics, we obtained Precision = 83.6 percent, Recall = 88.1 percent and F1 Measure = 82.5 percent. According to the literature and the base paper chosen ANN gives the best accuracy for this model building.*

KEYWORDS: *Artificial neural Network (ANN), Cancer, Breast cancer, Genomic data, Clinical data, Cancer prognosis, Treatment Prediction.*

I. INTRODUCTION

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. Breast cancer is a type of cancer that begins in the breast. It can either start in one of the breasts or both breasts. Breast cancer tumors can be of two types: (i) Benign and (ii) Malignant; most of the breast lumps are due to the benign and not the malignant tissues. Benign tumor is a growth which is not cancer and which also does not invade nearby tissues. Malignant tumors are the cells which grow uncontrollably and spread locally or to distant sites. Due to the lack of basic awareness about breast cancers, that is the reason why most of the breast cancers diagnosed are at an advanced stage, due to the lack of health education and awareness in breast cancer inappropriate treatment is not given at the initial stage.

Breast cancer is the most recurrent malignancy among women globally. Now it is outdone the most recurred cancer (lung cancer) in 2020 and has reported 2.3 million new cases and represents 11.7 percent of all types of cancer cases. Few studies and researchers show that by 2030 the rate would have been doubled. As seen above breast cancer can be treated in two ways depending upon the type of cancer caused to the patient. Age, laterality, NPI, oestrogen receptor, menopausal state, cellularity, threegene, histological subtype, Genes- EFGR, ERBB2, BRCA1, CDH1, PTEN, FOXO3, BRCA2 and few other factors are described in the literature as well as used in the work. Some statistical methods allowed us to know the importance between each variable and its importancein the data.

This information was helpful for us to know the behavior of breast cancer and then in turn it was helpful for recommending the type of treatment given for the patient. In the present technological resources, allow us to collect clinical and genomic data for each patient, getting in touch with the necessary doctors for upgrading the importance of the model built and building a robust model. Machine Learning enables us to learn the mutual importance between the data and recommend the treatment as it can learn from previous data and apply that knowledge to recommend in a better manner.

There has been a massive development in the use of machine learning in the various areas in the past few years as the models built using it are robust, reliable and the techniques followed cost less and the output given is based on the data that are already available. In the work it has allowed us to integrate both clinical and genomic data to provide better results. That being said, this study aims at recommending the treatments for the undergone breast cancer surgery which gives a primary evaluation on it by using machine learning in our center to recommend breast cancer treatment.

II. Literature Survey

Carlo Boeri, Corrado Chiappa, Federica Galli, Valentina De Berardinis, Laura Bardelli, Guilio Carcano and Francesca Rovera worked on an approach to give a therapeutic plan and a follow-up tailored to each of the prediction obtained[1]. This was a very complicated approach and in order to address that issue they used a multidisciplinary approach which became widely accepted. They developed two machine learning models - Artificial neural network and Support Vector Machine to obtain relationships between the given data and give an appropriate prediction. They had developed three outcomes - cancer recurrence(both loco-regional and systemic) and death from disease within 32 months. Their study had a limitation where they could not use it for clinical purposes as it lacked sensitivity.

Rehan Rafique,' S.M. RiazulIslam and Julhash U.Kaz have inferred that the current work is carried out only by using the cancer subtypes and the presence of genetic mutations[2]. In the presence of genetic mutation they can't predict the therapeutic response. In order for accurate models to match they've considered pharmacogenomic data. This collectively can be called an adaptive resistance mechanism. Obtaining the results using both genetic and pharmacology data was highly challenging in their work. The pharmacological data was taken from the PDXs(Patient Derived Xenografts) which contains the data.For the adaptive resistance approach, the resistive mechanisms need to be defined for each cancer subtype and also how the genes of the patients will react for the drug response in an individual. Determining the response of the genes of the patients even before the treatment starts is complicated at the initial stages as the technology is lacking advancement.

Yuka Ezaki, K. Nagayama and Ichiro Miura have mentioned that Cancer cells generated by genetic mutations are eliminated by immune cells[3]. Immune cells are the cells that recognize and attack foreign substances, pathogens, and cancers in the body. Immunotherapy is a treatment given to the patients which activates the body's immunity to function to eliminate cancer cells.

So this study assesses the immunotherapy treatment for the patients.In their study, parameters such as the conditions of cancer cell death with immune cells, cell division, and angiogenesis are introduced into the model. They did 30 day research with the cancer patients in which they categorized the patients into 3- 1. No immunotherapy, 2. Innate immunotherapy, 3.Immunotherapy and their study reveals that the person given immunotherapy had their cancer cells reduced compared to the other two.

Muhammet Fatih Ak has researched the data obtained from the University of Wisconsin Hospital which were used for the predictions on breast cancer[4]. This work was done based on two techniques- 1. Data Visualization and 2. ML techniques(logistic regression, k-nearest neighbors, support vector machine, naïve Bayes, decision tree, random forest, and rotation forest were applied to this dataset). Their work aimed to make an analysis using data visualization and machine learning applications for breast cancer detection and diagnosis. They had the highest accuracy for the logistic regression model. Classification and regression were the two categories which they used for their problem approach. Under classification they used the ML techniques and applied them onto the dataset. They divided the work into 3 datasets which had independent features: 1. All independent features,
2. Highly correlated features, 3. Least correlated features. These 3 datasets were tested on the algorithms and results were obtained.Logistic regression is one of the most common algorithms to solve classification problems.

Olow, A.K, Veer, L.v. & Wolf, D.M worked on building a drug recommender system using apriori recommendation system algorithm[5]. This model was built to recommend drugs for breast cancer using previous health data. This work Considers effect of previous drugs or treatment on the patient, to recommend a drug. Hence helping clinicians decide on the next possible treatments.

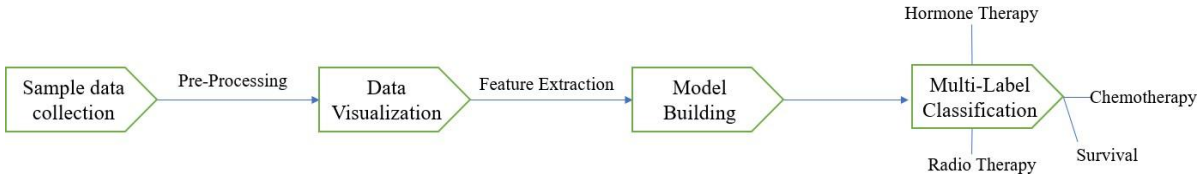
In a review[6] by Zhu, Wan, Longxiang Xie, Jianye Han, and Xiangqian Guo that compares different deep learning methods regarding cancer prognosis prediction. The publications that they reviewed was divided into three different categories: (1) NN models without feature extraction, (2) Fully connected NN models with feature extraction and (3) CNN based models.Their work helps in identifying techniques to predict (i) Tumor

size, (ii) recurrence and (iii) survival using the necessary input. They concluded that Fully connected NN and CNN showed good and effective results.

In a work[7] by Yue, W, Wang, Z, Chen, H, Payne, A, Liu, X. which presents a review of studies that make use of different machine learning approaches, in Breast Cancer Diagnosis and Prognosis. Authors have mainly focused on ANN, SVM, DT and kNN models. By applying different ML techniques on the same data set, ANN achieved 99.59% accuracy; while SVM, DT,k-NN achieved 98.42%, 97.13%, 98.48% accuracy respectively.

III. METHODOLOGY

In this work, ANN model was adopted to train the dataset and evaluate their performance on Cancer treatment recommendation.The implementation is based on the Keras/TensorFlow framework.
This section describes the process of Data Collection, Data Preparation and Model Training in Detail. The overall architecture of the proposed system is shown in Fig. 1.



1. Dataset: The dataset used is obtained from the "AI and Health Informatics Project" repository. This dataset was initially used by them in their study of clinical and genomic data and how it affects the study of breast cancer. This data is real data where their genomic and clinical data was collected by using patient ID. It has 28 features and 1904 instances.The dataset contains all the necessary genomic data which contains(BRCA1,CDH1,PTEN,FOXO3,BRCA2) and also clinical data like pre/post menopausal state, laterality,hormones like ER(Estrogen Receptor), HER(Human Epidermal growth factor Receptor) and type of breast surgery(Breast Conserving or Masectomy) the patient have undergone. Other GEneric Features such as Age, Vital status,IntClust,COHORT are also considered.

2. Data Preparation: The dataset was pre-processed to transform the raw data in a useful and efficient format. Firstly important features were extracted and other columns such as patient id, Vital status, Inferred menstrual state, laterality were deleted. That was followed by deleting rows with null values. Later data in the dataset are normalized by dividing each value with the maximum value of the respective columns. After which, categorical data have been encoded using LabelEncoder() and OneHotEncoder classes of sklearn's preprocessing library. Lastly, data was 85:15 for the training and testing split respectively.

3. Model Training: An Artificial Neural Network with two hidden Layers was used to train the dataset. Model used Adam optimizer at 100 epochs which results in accuracy, loss at each epoch. Overall accuracy of the trained dataset is the accuracy obtained at final epoch. This model gave a training accuracy of 91.4% and loss of0.1789.

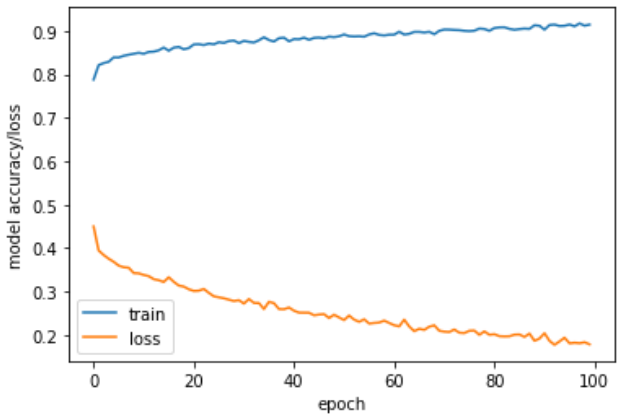


Fig 2: Accuracy/Loss Graph

A graph for accuracy and loss was plotted for different epochs as shown in Fig 2. It was observed that the accuracy was increasing and loss was decreasing for each epoch.

IV. RESULTS

Confusion matrix for each label, roc_auc graph was used to evaluate the models testing accuracy. The final results obtained are as follows:
testing accuracy = 77.78%precision 0.816
Recall = 0.882
F1 Measure = 0.816

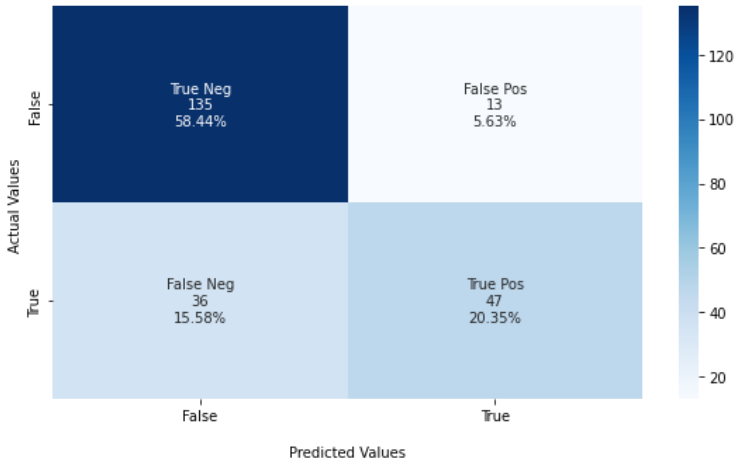


Fig 3: Confusion matrix for chemotherapy

Fig 4: Confusion matrix for Hormone Therapy

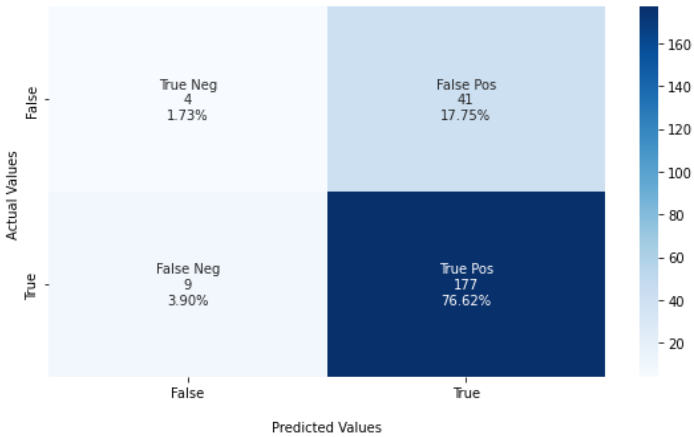


Fig 5: Confusion matrix for RadioTherapy

The confusion matrix has four different results: True Negative, True Positive, False Negative, False Positive. From Fig 3,Fig 4 and Fig 5 it was observed in above figures, majority of the predictions are either True Positive or True Negative.

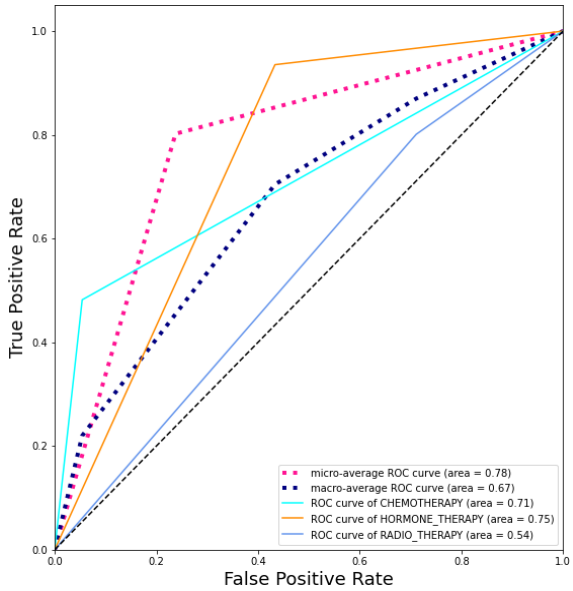


Fig 6: ROC-AUC graph

The ROC graph of the model is shown in Fig 6. From the figure it can be observed that AUC under Chemotherapy, Hormone therapy, Radiotherapy are 0.71,0.75,0.54 respectively.

Table 1: Classification Report

	precision	recall	f1-score	support
CHEMO_THERAPY	0.78	0.57	0.66	83
HORMONE_THERAPY	0.92	0.98	0.95	201
RADIO_THERAPY	0.81	0.95	0.88	186
micro avg	0.86	0.90	0.88	470
macro avg	0.84	0.83	0.83	470
weighted avg	0.85	0.90	0.87	470
samples avg	0.86	0.91	0.86	470

Classification report shown in Table 1 is based on the values obtained in the confusion matrix. The true positive, true negative, false positive and false negative values are used to calculate the precision, recall and F1 scores for each of the 3 labels in the multiclass classification model.

V. CONCLUSION

This study explored the use of various fields in breast cancer, how and why breast cancer occurs and recommended the best suited treatment for a particular patient's clinical and genomic data. It also explored various Machine Learning techniques which can be performed on the dataset. In ANN the results were accurate and specific for a particular input and the work also incorporated on finding the survival percentage of the patient for a particular input of the patient's data. To take this work to a step further, inculcating with the hospitals in real time and take in real time inputs will make the models more precise and robust and the accuracy of the model will increase as the number of patients in dataset will increase by a large value comparatively to the work carried out at present. Nonetheless these predictive models cannot replace the physicians, pathologist, oncologist and surgeons' recommendations as they play a major role in proper adjuvant therapy and follow-up in terms of frequency of the patients and knowledge obtained by them. Therefore, these techniques can be an additional support for them rather than being a primary solution which can be used by them.

REFERENCES

[1]. Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, Rovera F. Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer Med. 2020 May;9(9):3234-3243. doi: 10.1002/cam4.2811. Epub 2020 Mar 10. PMID: 32154669; PMCID: PMC7196042.

[2]. Rafique R, Islam SMR, Kazi JU. Machine learning in the prediction of cancer therapy. Comput Struct Biotechnol J. 2021 Jul 8;19:4003-4017. doi: 10.1016/j.csbj.2021.07.003. PMID: 34377366; PMCID: PMC8321893.

[3]. Yuka Ezaki[1], Katsuya Nagayama[1] and Ichiro Miura[2].Numerical simulations of cancer treatment prediction: modeling the influence of immunity. Published under licence by IOP Publishing Ltd IOP Conference Series: Materials Science and Engineering, Volume 1137, The 11th TSME-International Conference on Mechanical Engineering (TSME-ICoME 2020) 1st-4th December 2020, Ubon Ratchathani, Thailand , **Citation** Yuka Ezaki et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. **1137** 012052

[4]. Ak, M.F. A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare **2020**, 8, 111. https://doi.org/10.3390/healthcare8020111.

[5]. Olow, A.K., Veer, L.v. & Wolf, D.M. Toward developing a metastatic breast cancer treatment strategy that incorporates history of response to previous treatments. BMC Cancer **21,** 212 (2021). https://doi.org/10.1186/s12885-021-07912-7

[6]. Zhu W, Xie L, Han J, Guo X. The Application of Deep Learning in Cancer Prognosis Prediction. Cancers (Basel). 2020 Mar 5;12(3):603. doi: 10.3390/cancers12030603. PMID: 32150991; PMCID: PMC7139576.

[7]. Yue, W.; Wang, Z.; Chen, H.; Payne, A.; Liu, X. Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. Designs **2018**, 2, 13. https://doi.org/10.3390/designs2020013

[8]. https://pubmed.ncbi.nlm.nih.gov/28181405/

[9]. https://bmccancer.biomedcentral.com

[10]. www.sciencegate.app

[11]. https://doctorpenguin.com