IJIEMR Transactions, online available on 26th Nov 2020. Link :http://www.ijiemr.org/downloads.php?vol=Volume-09&issue=Issue 11

10.48047/IJIEMR/V09/ISSUE 11/54

TITLE: A CRITICAL STUDY ON PREDICTION OF BREAST CANCER WITH THE HELP OF MACHINE LEARNING

Volume 09, ISSUE 11, Pages: 293-296

Paper Authors **K. ANIL KUMAR,DR. ASHWINI KUMAR NAGPAL**

International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

A CRITICAL STUDY ON PREDICTION OF BREAST CANCER WITH THE HELP OF MACHINE LEARNING

NAME- K. ANIL KUMAR

DESIGNATION RESEARCH SCHOLAR, THE GLOCAL UNIVERSITY SAHARANPUR, UTTAR PRADESH

GUIDE NAME- DR. ASHWINI KUMAR NAGPAL

DESIGNATION - PROFESSOR (DEPARTMENT OF MATHEMATICS)

THE GLOCAL UNIVERSITY SAHARANPUR, UTTAR PRADESH

ABSTRACT

Breast cancer remains a significant global health concern, with early detection playing a pivotal role in improving patient outcomes. Machine learning (ML) has emerged as a powerful tool in the field of medical research, offering the potential to enhance breast cancer prediction and diagnosis. This abstract provides an overview of the state-of-the-art techniques and recent advancements in utilizing ML for breast cancer prediction. This study highlights the importance of feature selection and extraction methods in optimizing the performance of ML models. Diverse datasets encompassing clinical, genomic, and radiomic data have been leveraged to train robust ML models. The abstract explores the various ML algorithms commonly employed for breast cancer prediction, such as logistic regression, support vector machines, random forests, and deep learning neural networks.

Keywords: - Breast Cancer, Algorithm, Network, Cancer, Machine.

I. INTRODUCTION

Due to its high fatality rate, cancer necessitates prompt diagnosis and treatment. Breast cancer is a prevalent malignancy affecting women globally, with a significant fatality rate. In this chapter, a novel predictive modelling approach is used to diagnose breast cancer by using a combination of logistic regression, random forest, and deep neural network algorithms.

The performance of the recommended model was evaluated using accuracy, precision, recall, and F1-score. The outcomes achieved via the use of the ensemble approach exhibit remarkable performance in comparison to conventional techniques such as logistic regression and traditional approaches like random forest.

II. DATASET DESCRIPTION

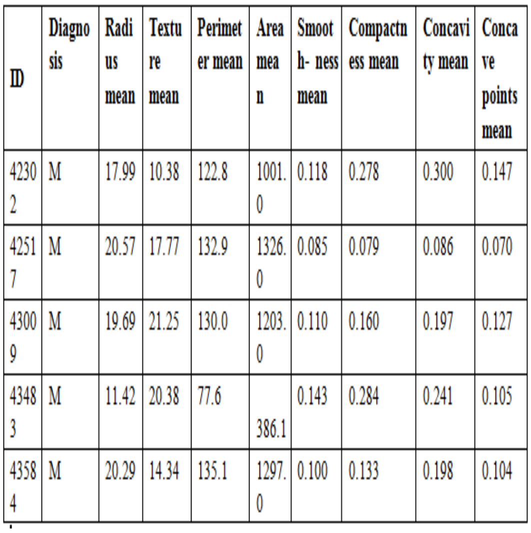
The breast cancer data used in this study was obtained from the Wisconsin Breast Cancer Data Collection (WBCD) (Wolberg, 2019). The data was obtained by the data collection warehouse for machine learning at the University of California, Irvine.

The data collection consists of 569 instances, each representing a data column. These instances are associated with 32 separate attributes, which are represented as rows in the dataset.

The features of a breast lump obtained by fine-needle aspiration (FNA) comprise 10 distinct digital photographs, an image of the nucleus displaying its variation, the average value, and the greatest value (Liu, 2018).

There are a total of 33 kinds of features, which include area, compactness, radius, perimeter, symmetry, texture, smoothness, concave points, fractal dimension, and others.

Table 1. Sample Wisconsin Breast Cancer Dataset



ID	Diagnosis	Radius mean	Texture mean	Perimeter mean	Area mean	Smooth- ness mean	Compactness mean	Concavity mean	Concave points mean
42302	M	17.99	10.38	122.8	1001.0	0.118	0.278	0.300	0.147
42517	M	20.57	17.77	132.9	1326.0	0.085	0.079	0.086	0.070
43009	M	19.69	21.25	130.0	1203.0	0.110	0.160	0.197	0.127
43483	M	11.42	20.38	77.6	386.1	0.143	0.284	0.241	0.105
43584	M	20.29	14.34	135.1	1297.0	0.100	0.133	0.198	0.104

III. DATA PRE-PROCESSING

Data preprocessing is a crucial stage in preparing data for use in a machine learning model, as it serves to cleanse the data and render it appropriate for analysis. This procedure significantly improves the accuracy and efficiency of the model.

- **Checking for missing value**

The first step in the pre-processing phase involves doing a thorough examination of the dataset to identify any instances of missing values. The examination of the Wisconsin breast cancer data collection was conducted. Consequently, the dataset did not include any null entries. According to Fogliatto et al. (2019), all the attributes in the dataset were of a numerical nature.

- **Data scaling (Standardization)**

Data scaling is an essential step in the preprocessing phase. It is necessary to scale the data prior to modeling. The researchers used the standardization method to normalize the data in the study. The data points are shown using the standardization approach, as described by Showrov et al. (2019).

- **Feature Selection**

The third step of the pre-processing procedure is the feature selection phase. The feature selection methodology is a filtering method that identifies and picks the most relevant feature from a given dataset. The use of the feature selection procedure is employed in order to enhance accuracy, mitigate overfitting, and reduce training time (Peker et al., 2015). The present study employs the random forest technique to automatically discern pertinent attributes from the dataset via the use of the random forest algorithm.

The identified trait will have the most impact on the predictive parameter. Instead of using the whole of features, the approach only employs a limited subset of them (Omondiagbe et al., 2019). The random forest methodology leverages the principles of information theory within the field of mathematical communication science to identify the most influential feature by examining a predictive variable.

In the study conducted by Osman and Aljahdali (2020), a random forest classifier approach was used to extract the most significant characteristics from the WBCD dataset. This resulted in the identification of about 15 features that were deemed crucial for analysis.

- **Feature Extraction**

The feature extraction technique is the fourth stage within the preprocessing procedure. Feature extraction is often known as a technique for reducing the dimensionality of data. Feature extraction is a technique used to reduce the dimensionality of data by partitioning a substantial volume of raw data into smaller units for further processing.

The process of dimensionality reduction is used to emphasize significant features within data, hence reducing its overall dimensionality. The present work employs a feature extraction strategy to exclude undesired variables from the WBCD dataset, while acknowledging the accuracy and uniqueness of the original data set (Sethi, 2018).

Principal component analysis (PCA) is used in this particular case with the objective of reducing the number of dimensions. Principal Component Analysis (PCA) is a statistical technique that involves the modification of a collection of observable variables, which are assumed to be connected, using an orthogonal transformation. According to the study conducted by Hajiabadi et al. (2020), The principal component analysis (PCA) technique is used to decrease the dimensionality of the WBCD dataset, resulting in the extraction of two primary features.

IV. CLASSIFICATION (TRAINING AND TESTING THE MODEL)

In order to accurately predict outcomes, it is necessary to train the pre-processed data. The use of machine learning classification methods is necessary for the training of the data. In this study, the researchers used three classification algorithms, namely Deep Neural Network, Random Forest, and Logistic Regression, to train the data. Once the data has through the training process, it proceeds to the testing step, during which it is evaluated using the same set of three machine learning classification techniques. The dataset used in this study, known as the Wisconsin Breast Cancer dataset (WBCD), was divided into two subsets: a training set including 70% of the data and a testing set comprising the remaining 30% of the data. The classifier is responsible for determining the presence or absence of breast cancer in an individual. In cases when an individual is diagnosed with breast cancer, the tumor is anticipated to exhibit malignancy, however in instances where no such diagnosis is made, the tumor is expected to have benign characteristics. Ensemble approach was used to enhance accuracy by using three distinct machine learning algorithms, including Neural Deep Network methods, Random Forest, and Logistic Regression. All three categorizing procedures will ascertain whether the patient's tumor is benign or malignant.

V. CONCLUSION

The utilization of machine learning in the prediction of breast cancer marks a significant advancement in the field of healthcare. This abstract has highlighted the crucial role played by machine learning algorithms, diverse datasets, and imaging modalities in enhancing our ability to detect breast cancer at its earliest stages. With the potential to reduce misdiagnoses and improve patient outcomes, machine learning models have demonstrated their capacity to be valuable tools in clinical practice.

However, it is important to acknowledge the challenges that accompany the integration of machine learning into healthcare. Data quality, bias mitigation, and model explainability remain ongoing concerns that must be addressed to ensure the reliability and ethical use of these predictive tools. Moreover, ongoing research and collaboration between data scientists, clinicians, and healthcare providers are essential to further refine and validate machine learning models for breast cancer prediction.

REFERENCES

1. Zhou, L.-Q., Wu, X.-L., Huang, S.-Y., Wu, G.-G., Ye, H.-R., Wei, Q., Bao, L.-Y., Deng, Y.-B., Li, X.-R., Cui, X.-W. & C.F. Dietrich, 2020. Lymph Node Metastasis Prediction from Primary Breast Cancer US Images Using Deep Learning. Radiology. (294)1,. pp. 19–28.
2. Zheng, X., Yao, Z., Huang, Y., Yu, Y., Wang, Y., Liu, Y., Mao, R., Li, F., Xiao, Y., Wang, Y., Hu, Y., Yu, J. & J. Zhou, 2020. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nature Communications. (11)1,. pp. 1236.
3. Yu, L., Zhao, J. & L. Gao, 2018. Predicting potential drugs for breast cancer based on miRNA and tissue specificity. International journal of biological sciences. (14)8,. pp. 971.
4. Yarabarla, M.S., Ravi, L.K. & A. Sivasangari, 2019. Breast Cancer Prediction via Machine Learning. In: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). April 2019, IEEE, pp. 121–124.
5. Yala, A., Lehman, C., Schuster, T., Portnoi, T. & R. Barzilay, 2019. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. Radiology. (292)1,. pp. 60–66.
6. Xie, J., Liu, R., Luttrell, J. & C. Zhang, 2019. Deep Learning Based Analysis of Histopathological Images of Breast Cancer. Frontiers in Genetics. (10).