NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY
[IJIERT] ISSN: 2394-3696 Website: ijiert.org
VOLUME 8, ISSUE 9, Sep. -2021

A DATA DRIVEN BREAST CANCER DETECTION METHOD

Ezefosie Nkiru
Department of Computer science,
African University of Science and Technology, Abuja, Nigeria
*nkiru.happiness@gmail.com

Ohemu Monday Fredrick
Department of Electrical and Electronics Engineering,
Air Force Institute of Technology, Kaduna, Nigeria.
*monfred@afit.edu.ng.

Igbax Teryima
Department of Telecommunication Engineering,
Air Force Institute of Technology, Kaduna, Nigeria.
iteryima@yahoo.CA

Zubair Zuleihat Ohunene
Department of Physics, Nigerian Defense Academy, kaduna, Nigeria
*zubairelizabeth@gmail.com

ABSTRACT
One of the most common chronic diseases is breast cancer. Breast cancer is not only widespread, but it is also quite complicated. Treating patients with Breast cancer demands doctors to examine enormous amount of data – often too much for human clinicians to analyze on their own. To accelerate the development of better Breast cancer treatments, a big data analytics model can quickly draw meaningful insights and identify cancer-related patterns in the data which radiologists cannot. Big data analytics can help enhance the accuracy of breast cancer screenings, reduce the number of biopsies needed and increase physicians' confidence in the accuracy of assessments made for screening exams. This article proposes a big data analytics model in breast cancer detection which is capable of processing an enormous amount of data quickly to assist in the early detection of breast cancer with better accuracy. This will augment care delivery, diagnostics and reduce the death rate. In this work, we build a model based on Machine Learning Algorithm to predict breast cancer. The result shows that the machine learning model based on big data analytics can predict better than traditional models, which currently incorporate only a small fraction of patient data.

Keyword: Big data analytics, Breast cancer, Machine Learning, Radiologists, Big data

INTRODUCTION
Breast cancer is estimated to claim the lives of 44,130 people this year. [2]. According to the estimates, breast cancer is one of the most common chronic diseases, necessitating immediate treatment and early detection. The tests listed below can be used to diagnose breast cancer or for follow-up testing after a diagnosis of breast cancer. [3][4];

Imaging tests display images of the interior of the human body. There are different kinds of imaging tests;

- **Breast ultrasound**. A machine that uses sound waves to make detailed pictures, called sonograms, of areas inside the breast.
- **Diagnostic mammogram**. An X-ray showing the breast in greater detail.
- **Magnetic resonance imaging (MRI).** A type of body scan that employs the use of a computer and a magnet. The MRI scan will produce detailed images of the breast's interior.

Biopsy. This is a test that requires removing tissue or fluid from the breast to be examined under a microscope and subjected to additional testing. Biopsies are classified as follows;

- **Fine needle aspiration biopsy.** A thin needle is used to remove a small sample of cells in this type of biopsy.

- **Core needle biopsy.** A wider needle is used in this type of biopsy to remove a larger sample of tissue**.**
- **Surgical biopsy.** This type of biopsy removes the biggest amount of tissue.
- **Image-guided biopsy.** Using an imaging technique such as mammography, ultrasound, or MRI, a needle is guided to the location of the mass or calcifications during this procedure.
- **Sentinel lymph node biopsy.** The sentinel lymph node biopsy procedure is used to determine whether cancer has spread to the lymph nodes near the breast.

Genomic tests. Doctors employ genomic testing to check for certain genes or proteins, which are the molecules produced by genes, in or on cancer cells. There are several types of genomic testing.

- **Oncotype Dx™**. This test is suitable for the people who have ER-positive and/or PR-positive, HER2-negative breast cancer that hasn't advanced to the lymph nodes, as well as some situations where the cancer has spread to the lymph nodes.
- **MammaPrint™**. This test is for patients who have ER-positive and/or PR-positive breast cancer that has not advanced to the lymph nodes or has only spread to 1 to 3 lymph nodes, HER2-negative or HER2-positive breast cancer.
- **Molecular testing of the tumor**. If a patient has locally advanced or metastatic breast cancer, the doctor may recommend testing for the following molecular features:
  - **PD-L1.** It's found on the surface of cancer cells as well as some immune cells in the body. This protein, especially in triple-negative breast cancer, prevents the body's immune cells from destroying the cancer.
  - **Microsatellite instability-high (MSI-H) or DNA mismatch repair deficiency (dMMR).** MSI-H or dMMR-positive tumors have a hard time repairing DNA damage. This means they go through a lot of alterations and mutations. These changes cause aberrant proteins to form on tumor cells, making it easier for immune cells to detect and fight the tumor.
  - **NTRK gene fusions.** This is a type of genetic genetic defect that can be discovered in a variety of cancers, including breast cancer.
  - **PI3KCA gene mutation.** This is a type of genetic mutation that is frequent in breast cancer.
- **Additional tests.** For persons with ER-positive and/or PR-positive, HER2-negative breast cancer that has not progressed to the lymph nodes, further testing may be available. PAM50 (ProsignaTM), EndoPredict, Breast Cancer Index, and uPA/PAI are some of the tests available. They can also be used to predict the likelihood of cancer spreading to other places of the body.

Blood tests. There are several sorts of blood tests**;**

- **Complete blood count** is used counts the amount of different types of cells in a sample of a person's blood, such as red blood cells and white blood cells.
- **Blood chemistry.** This test assesses the health of your liver and kidneys.
- **Hepatitis tests.** These tests are sometimes performed to screen for signs of prior hepatitis B and/or C infection.

Following the completion of diagnostic testing, the doctor will analyze all of the results and describe the cancer. The entire process can take a long time since there is so much data – often too much for human doctors to review on their own – however it is necessary to speed up the development of improved cancer treatments. Big data storage and analysis tools can be utilized to swiftly extract useful information from cancer data. Furthermore, big data analytics techniques can detect pixel-level changes in tissue that are imperceptible to the naked eye, as well as hidden cancer-related patterns in the data. Over the last three decades, medical data storage capacity has increased dramatically, resulting in larger amounts and a broader diversity of recorded medical data (mammography scans, 3D ultrasound, MRI, genomic data, pathological data, etc.). Until recently, these data were frequently employed at the individual level to determine a diagnosis, develop a treatment plan, follow sickness development, and determine a patient's prognosis.

Furthermore, on a statistical scale, only structured data was employed, which represented a small part of the accessible and relevant data sources. The rest was maintained in data graveyards hidden from view of medical personnel. The major promise of Big Data is that it will allow clinicians to use any data source, including unstructured data such as textual medical reports or pictures, to identify unambiguous cancer markers. Big

data analytics will help health care practitioners improve their abilities, resulting in more accurate diagnoses, tailored treatments, and better outcomes.

As a result, this paper provides a big data analytics model for detecting and predicting breast cancer. The proposed system leverages big data Machine Learning model to detect and prediction breast cancer using Logistic Regression algorithm.

The structure of this article is divided into four sections. Section 2 presents the literature and existing works. The proposed methodology is detailed in Section 3. Section 4 presents the results and discussions.

LITERATURE REVIEW

This section presents some of the work done in this area and their outcomes:

Table 1. The table summarizes the work done in this area and their outcomes.

Sr	Paper Title	Author	Objective	Methodology	Conclusion
1.	Prediction of Breast Cancer Using Big Data Analytics[7]	K. Shailaja, B. Seetharamulu, M.A. Jabbar	The goal of this article is to create a classifier model for predicting breast cancer risks using the KNN algorithm.	The KNN algorithm was used to predict the risk of breast cancer.	When compared to other models, the experimental results revealed that the suggested technique improves accuracy, precision, recall, and F-measure.
2.	Prognosis and Diagnosis of Breast Cancer Using Interactive Dashboard Through Big Data Analytics[8]	Gomathi N, and Sandhya P.	The goal of this research is to develop an interactive dashboard approach for determining the presence or absence of a distinct lump known as a tumor in a mammographic picture.	Backpropagation and the Support vector machine system (SVM) were used to analyze benign and malignant cancers, as well as predict cancer stage using mammographic pictures.	The results of nntraintool are used to calculate the test image's performance rate, training state, regression, and error histogram. Furthermore, the cloud and big data analytic ideas offer a next level of cancer stage prediction at the home level.
3.	Big Data Analytics for Early Detection of Breast Cancer Based on Machine Learning[9]	Desislava Ivanova	The paper's major goal is to provide a conceptual model for early detection of breast cancer based on machine learning for processing and analyzing medical large data, as well as subsequent knowledge discovery for personalized treatment.	The Naive Bayes classifier is used to realize the proposed conceptual model.	The Naive Bayes classifier has a better overall performance in predicting diabetic illness.
4.	Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms[10]	Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi	The purpose of the project is to develop a system that can accurately distinguish between benign and malignant breast cancers using genetic programming and machine learning techniques.	Genetic programming (GP) has been used to optimize the data and the proposed model's control parameters	The results show that combining feature processing and modeling improves performance significantly without requiring user intervention.
5.	Breast Cancer Detection using Machine Learning Algorithms[11]	] S. Sharma, A. Aggarwal and T. Choudhury	The purpose of this paper is to compare widely used machine learning algorithms and techniques for breast cancer prediction: Random Forest, kNN (k-Nearest-Neighbor), and Nave Bayes.	The Wisconsin Diagnosis Breast Cancer data set was used as a training set to examine the performance of several algorithms in terms of accuracy and precision.	The outcomes are very competitive and can be used for both detection and treatment.

METHODOLOGY

We used a big data analytics machine learning algorithm to design a machine learning model for breast cancer detection and prediction for better quality and outcomes, and this system runs on Spark.

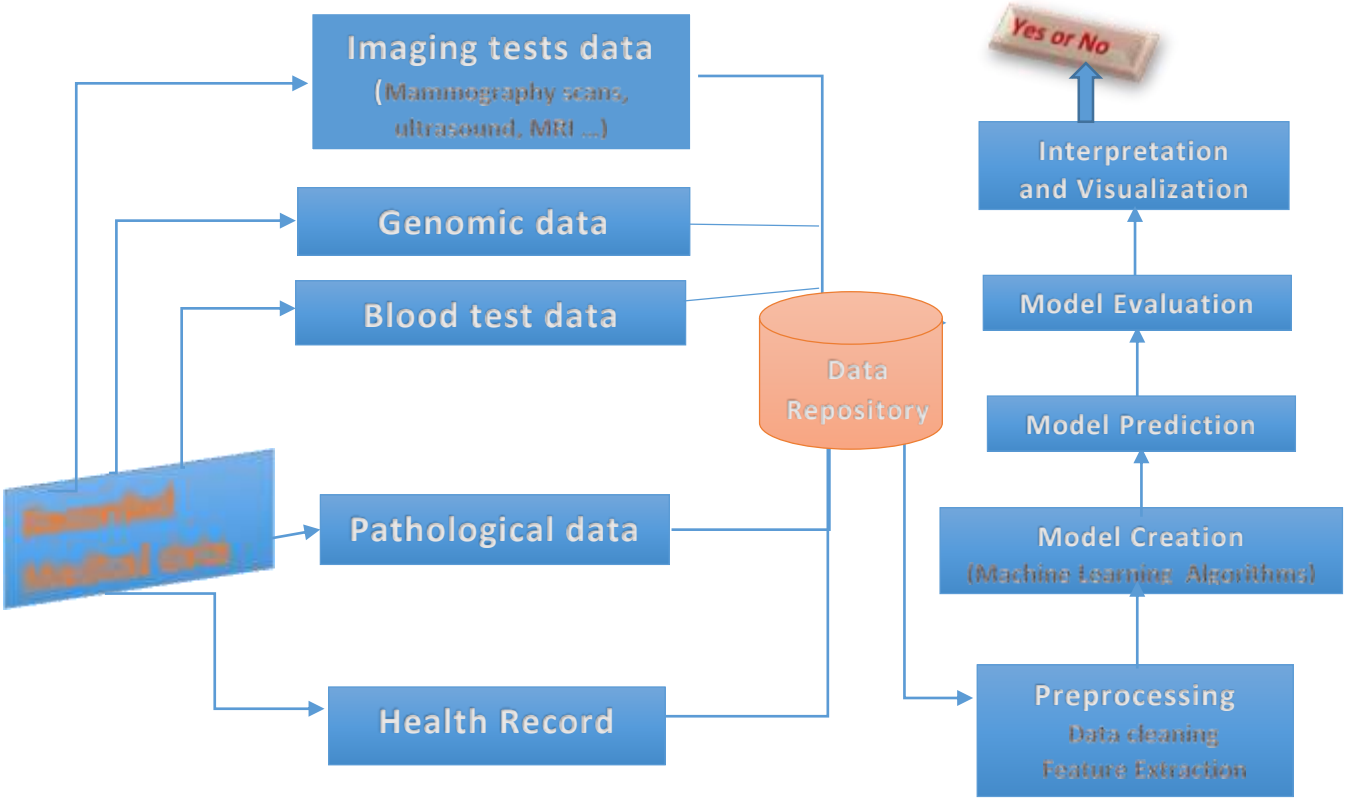

Figure 1: Proposed System Overview

The proposed approach uses big data techniques to process and analyze large amounts of medical data (mammography scans, 3D ultrasound, MRI, genomic data, pathological data, blood tests) in order to extract knowledge for more personalized breast cancer treatment. To comb through all of the available cancer data and find unambiguous combinatorial signatures of cancer, we need big data. The figure 1 and figure 1[12] below show the overview of the system

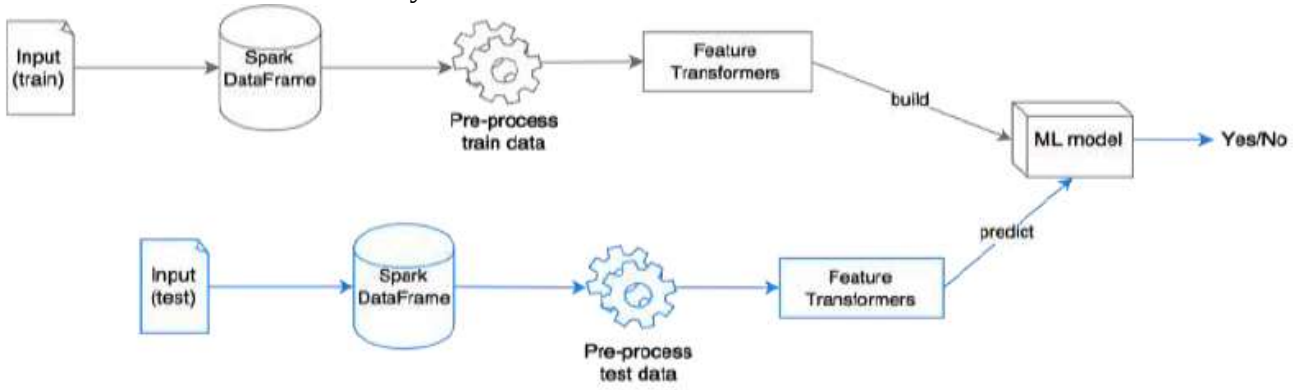

Figure 2: PySpark System flow

Thus the first stage is all the medical recorded are collected, cleaned and gathered in the data repository. The machine learning technique is divided into learning phase and testing phase. In the learning phase, the training datasets are analyzed and trained. Then a model is created in the learning phase. However, in the testing phase, new test data is given as input to the system. The system performs the task and applies the learned knowledge to classify unknown data.

The first stage consists of collecting, cleaning, and storing all medical records in a data repository. Machine learning technique is divided into learning phase and testing phase. In the learning phase, the training datasets

are analyzed and trained, while new test data is provided as input to the created model during the testing phase. The model uses the information it has gained to classify unfamiliar data.

A. Data Collection

The data to be trained in our scenario include collections of imaging test data (Mammography, MRI, Ultrasound...), genomic data, blood test data, pathological data, and health records gathered from multiple individuals previously. They should also be appropriately labeled because we are employing a supervised Learning technique.

The Wisconsin Breast Cancer dataset was collected from the UCI Machine Learning Repository [13] and used in this study.

The features were obtained from digital images of a fine needle aspirate of a breast tumor [1], which describe the nucleus of the current image. The WDBC database has been effected on 569 patients at Wisconsin hospitals, resulting in the discovery of 212 malignant and 357 benign cases.

B. Data Preprocessing

Data Preprocessing is the process of transforming or encoding data so that it may be easily parsed by the machine. In other words, the algorithm can now easily interpret the data's features. As a result, preprocessing our data before feeding it into our model is critical.

At this point, we filtered and cleaned the data that will be used to develop the model. The column's mean value was used to fill in the missing values.

C. Feature Extraction

The dataset features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image [1][13].

The features extracted from the dataset are:

Table 2: Features extracted from the dataset

Attributes	Description
id number	ID number
diagnosis	M= malignant, B= benign
radius_mean	Mean of distances from center to points on the perimeter
texture_mean	Mean of gray-scale values
perimeter_mean	Mean of the perimeter
area_mean	Mean of the area
smoothness_mean	Mean of local variation in radius lengths.
compactness_mean	Perimeter*/area-1.0(mean)
concavity_mean	Severity of concave portions of the contour(mean)
concave points_mean	Number of concave portions of the contour(mean)
symmetry_mean	symmetry(mean)
fractal_dimension_mean	“coastline approximation”-1(mean)
radius_se	Standard error of distances from center to points on the perimeter
texture_se	Standard error of gray-scale values
perimeter_se	Standard error of the perimeter
area_se	Standard error of the area
smoothness_se	Standard error of local variation in radius lengths.
compactness_se	Perimeter*/area-1.0 (standard error)
concavity_se	Severity of concave portions of the contour (standard error)
concave points_se	Number of concave portions of the contour (standard error)
symmetry_se	Symmetry (standard error)
fractal_dimension_se	“coastline approximation”-1 (standard error)
radius_worst	Worst of distances from center to points on the perimeter
texture_worst	Worst of gray-scale values
perimeter_worst	Worst of the perimeter
area_worst	Worst of the area
smoothness_worst	Worst of local variation in radius lengths.
compactness_worst	Perimeter*/area-1.0 (worst)
concavity_worst	Severity of concave portions of the contour (worst)
concave points_worst	Number of concave portions of the contour (worst)
symmetry_worst	Symmetry (worst)
fractal_dimension_worst	“coastline approximation”-1 (worst)

The mean, standard error and worst of these attributes were calculated for each image.
All feature values are recorded with four significant digits.
Missing attribute values: none

D. Model Creation Via Classification

The next step is classification, which classifies the data into two categories based on the selected features: benign and malignant. There are a variety of classification methods based on machine learning that can be used to implement classifiers such as Support Vector Machines (SVM), K-Neighbors (KNN), Naive Bayes, Decision Tree, and so on.
We used the Logistic Regression and Random Forest Machine Learning algorithms to create two classification models, at the end, Logistic Regression model with the best accuracy was taken.

EXPERIMENTAL RESULTS AND DISSCUSSION

We have to build machine learning classification model based on Wisconsin Breast Cancer dataset. It has 32 Columns as shown in table 2. We have to build the model to classify breast cancer into two category. This two categories are (M = malignant, B = benign). There are 32 Columns or features, however, the diagnosis feature is our target variable while others are dependent variables. The pie chart shows the distribution of benign and malignant in the target variable.

Data Preprocessing

We imported the libraries and dataset, then performed data exploration, dealt with missing values, dealt with the categorical data, and then plotted the categorical column's Countplot, Correlation histogram showing how other features correlated with target feature (diagnosis_M), and Heatmap. Correlation histogram shows that many of the features are positively correlated to our target variable.

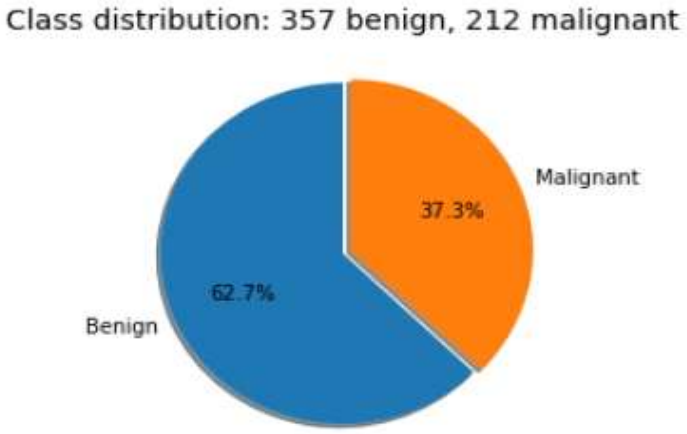

Figure 3: Showing Class distribution malignant and benign

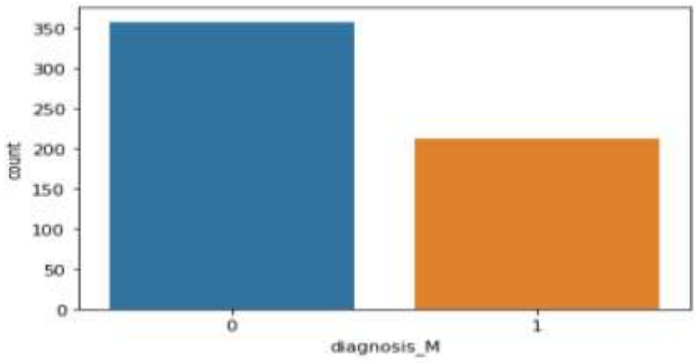

Figure 4: The count plot after the Categorical data has been converted to 0 and 1.

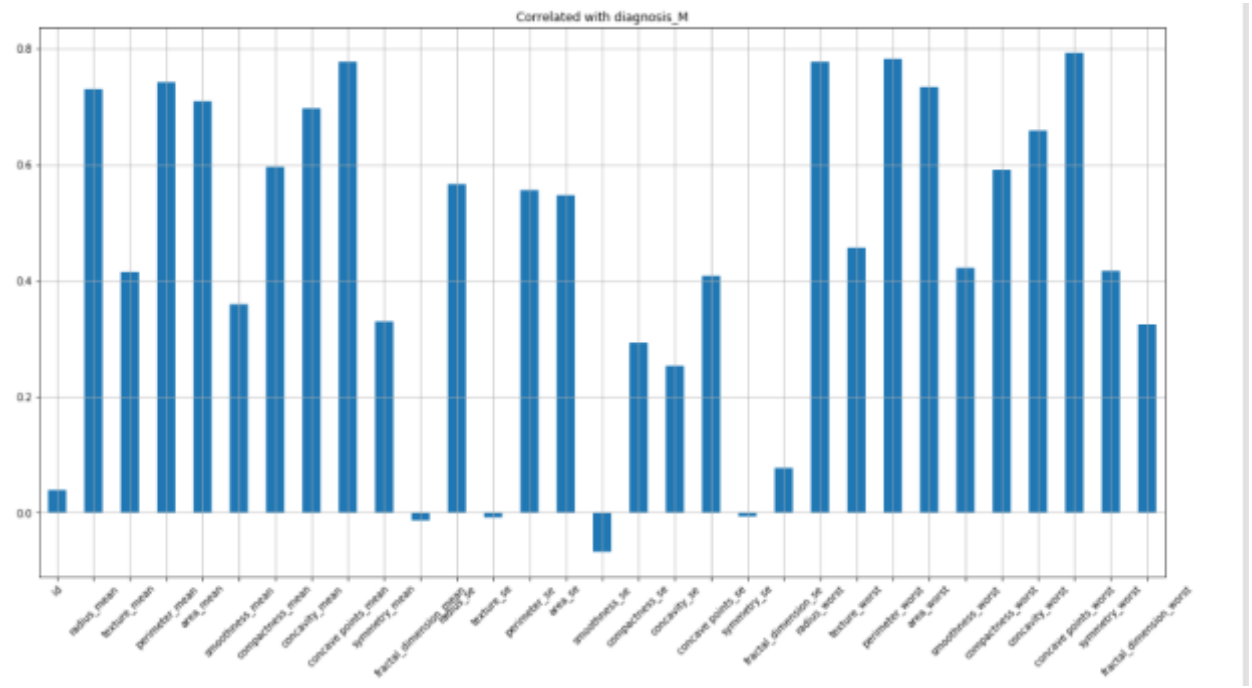

Figure 5: Showing how other features correlated with target feature (diagnosis_M)

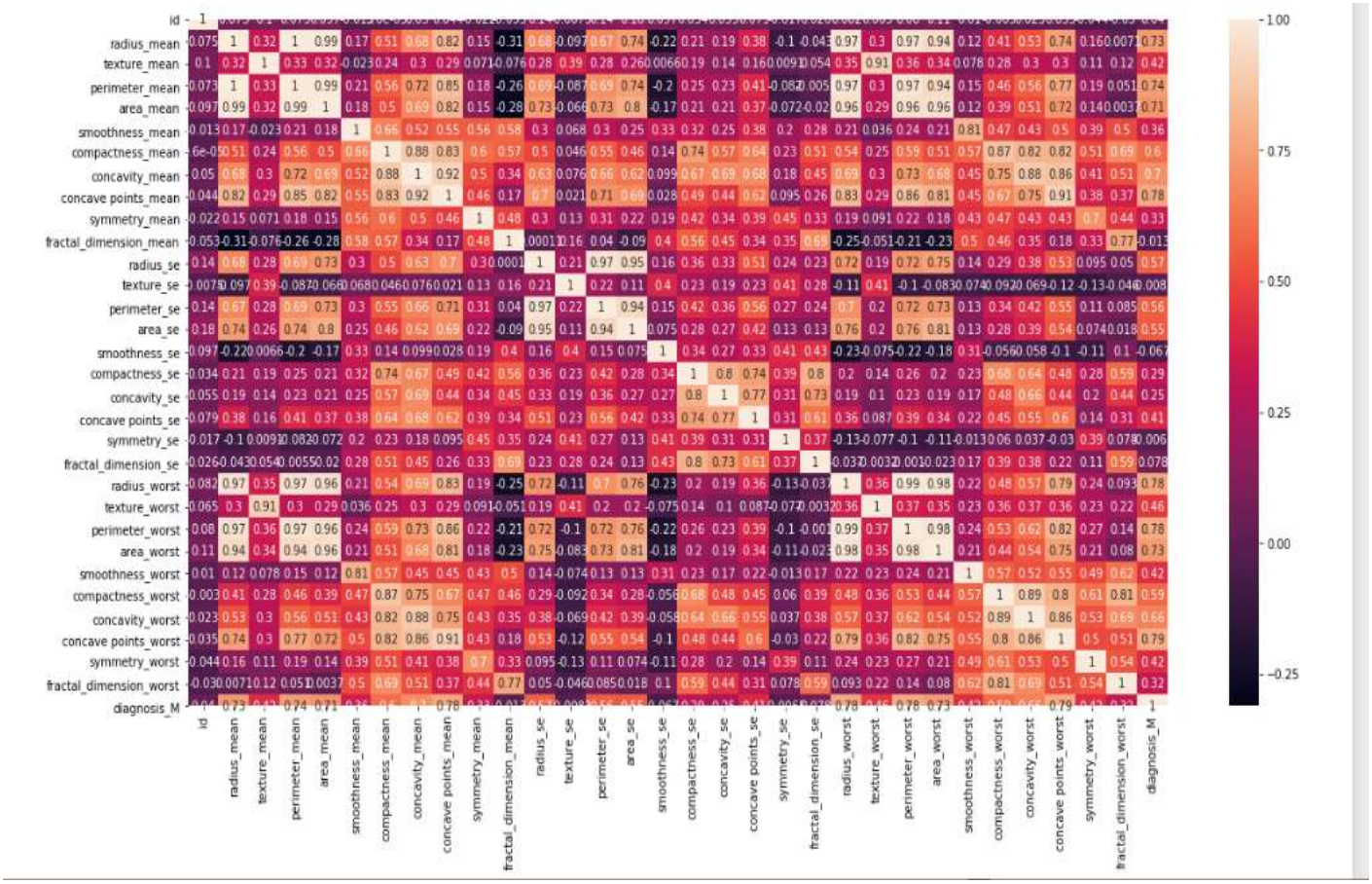

Figure 6: Heap map: Correlation plot showing the attributes in X and Y axis.

We then proceeded on to the next stage, which involved splitting the dataset into training and test sets, as well as feature scaling in both training and test data.

Building the Model

The model was trained or built using two machine learning techniques. We built the model first with the Logistic Regression algorithm and training data, then tested it with the test data to see how well it predicted. The score was then evaluated using cross-validation.

Furthermore, using the training data, we trained the second model with the Random Forest algorithm, and then we validated the model by evaluating its prediction performance using the test data. The score was then evaluated using cross-validation. Finally, the performance of both models was compared to decide which one was best. The table below shows the results:

Table 3: Before Cross Validation

	Model	Accuracy	f1 score	Precision	Recall
0	Logistic Regression	0.95614	0.946237	0.956522	0.936170
1	RandomForest	0.95614	0.943820	1.000000	0.893617

Table 4: After Cross Validation

	Model	Accuracy	Standard Deviation
0	Logistic Regression	97.806763	1.977044
1	Random Forest	95.159420	2.768398

Hyperparameter Tuning using Randomized Search.

We chose Logistic Regression based on performance, and then performed Hyperparameter Tuning for Logistic Regression using Randomized search to identify the optimal parameters. Then, using the Randomized search method, we were able to find the optimal parameters, increasing the accuracy to 98.24%.

```
Accuracy is 98.24 %
Standard Deviation is 2.16 %
```

The results reveal that the model is more capable of identifying and predicting breast cancer than standard models, which only include a limited portion of patient data.

CONCLUSION

The findings show that big data analytics techniques can assist doctors in developing more tailored breast cancer treatment treatments. This demonstrates how big data can help deliver on the promise of precision medicine by increasing access to more accurate, less expensive risk assessment.

REFERENCES

1) https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
2) https://www.cancer.net/cancer-types/breast-cancer/statistics
3) https://www.cdc.gov/cancer/breast/basic_info/diagnosis.htm
4) https://www.cancer.net/cancer-types/breast-cancer/diagnosis
5) https://healthitanalytics.com/news/radiologist-ai-combination-improves-breast-cancer-detection
6) https://healthitanalytics.com/news/big-data-analytics-may-lead-to-more-precise-cancer-treatments#:~:text=The%20findings%20reveal%20the%20ability,treatment%20plans%20for%20breast%20cancer.&text=% 22Our%20deep%20learning%20model%20is,breast%20cancer%2C%22%20Lamb%20said
7) SHAILAJA, K.; SEETHARAMULU, B.; A. JABBAR, M.. Prediction of Breast Cancer Using Big Data Analytics. International Journal of Engineering & Technology, [S.l.], v. 7, n. 4.6, p. 223-226, sep. 2018. ISSN 2227-524X. Available at: <https://www.sciencepubco.com/index.php/ijet/article/view/20480>. Date accessed: 01 sep. 2021. doi:http://dx.doi.org/10.14419/ijet.v7i4.6.20480.

8) Gomathi N, Sandhya P. Prognosis and Diagnosis of Breast Cancer Using Interactive Dashboard Through Big Data Analytics. Biotechnol Ind J. 2017; 13(1):128.
9) Desislava Ivanova , “Big data analytics for early detection of breast cancer based on machine learning”, in Proceedings of the 43rd International Conference Applications of Mathematics in Engineering and Economics: (AMEE'17), 2017, vol. 1910, no. 1. doi:10.1063/1.5014010.
10) Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, Mohammed Faisal Nagi, Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms Journal of Healthcare Engineering SP - 4253641 VL - 2019AB
11) S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.
12) Ezefosie Nkiru, Ohemu Monday Fredrick, "A Data Driven Anomaly Based Behavior Detection Method for Advanced Persistent Threats (APT)", International Journal of Science and Research (IJSR), https://www.ijsr.net/search_index_results_paperid.php?id=SR21726172522, Volume 10 Issue 8, August 2021, 663 – 667
13) https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29