Grenze International Journal of Engineering and Technology, June Issue

An Analysis of Machine Learning and Deep Learning to Predict Breast Cancer

Aiswarya Raghuthaman[1] and Lija Jacob[2]
[1-2]Christ (Deemed to be University), Pune Lavasa Campus, Maharashtra, India
aiswaryaraghuthaman@science.christuniversity.in, lija.jacob@christuniversity.in

***Abstract*—According to the report published by American Cancer Society, breast cancer is currently the most prevalent cancer in women. In addition, it is the second leading cause of death. It needs to be taken into serious consideration. Earlier and faster detection can help in the earlier and easier cure. Normally, medical practitioners take a large amount of time to understand and identify the presence of cancer cells in the human body. This can lead to serious complications even to the death of the individual. Hence there is a need to identify and detect the presence of this disease very accurately and in a shorter span of time. Like every other industry, the medical industry is shifting its paradigm to automation giving excellent results having high accuracy and efficiency, which is achieved using Artificial Intelligence. There are two sets of models developed based on the numerical dataset Wisconsin and image dataset BreakHis. Machine Learning algorithms and Deep Learning algorithms were applied on the Wisconsin dataset. Meanwhile, Deep Learning models were used for analysis of the Breakhis dataset. Machine Learning models- Logistic Regression, K Neighbors, Naive Bayes, Decision tree, Random Forest and Support vector classifiers were used. Deep Learning models- normal deep learning models, Convolutional Neural Network (CNN), VGG16 & VGG19 models. All the models have provided a very good accuracy ranging between 75% and 100%. Since medical research has a requirement for higher accuracy, these models can be considered and embedded into several applications.**

***Index Terms*— Accuracy, Breakhis, Breast Cancer, Classifier, Deep Learning, Machine Learning, WBCD, Wisconsin.**

I. INTRODUCTION

Human breast cancer is a dangerous disease that affects mostly females and is rarely found in males. In females, it has the second largest count of tumors after skin cancer. It is caused by cells in the glandular tissue of the breast called the duct lining cells. General symptoms of breast cancer include: Lump or thickening on the breast, altered appearance of the breast, Alteration in the skin includes dimpling, pitting, redness, or other alterations. Appearance changes in the nipple or alteration in the skin surrounding the nipple, abnormal discharge from nipples.

As per the World Health Organization, earlier detection, timely diagnosis and comprehensive breast cancer management are the three pillars that help in achieving breast cancer. In 2020, approximately 685,000 have died from this tumor, which accounts for 30% of all cancer-related fatalities among women [1]. In India, the low survival rates for this disease is mainly attributed to its very late detection. It occurs due to the lack of awareness and poor early screening and diagnosis rates [2].

Grenze ID: 01.GIJET.8.2.514_1

Earlier detection of breast cancer is highly effective as it can bring significant treatment. The most common treatment suggested is a combination of surgical removal along with radiation therapy, and medication.
Artificial Intelligence is a term that refers to the ability of machines to behave like human beings [4]. It is the core idea of Industrial Revolution 4.0. Artificial Intelligence incorporates several algorithms. These algorithms use a large amount of data to set the parameters which will be set for the model. These algorithms will be further used for the prediction of values. The quality of a model is evaluated mainly based on the reduction of errors. Lower the errors, better the model. Better the accuracy, the better is the model [5].
The medical industry is highly developed due to the introduction of newer and advanced machines into the forefront. The development of these machines helps in the easier and faster detection of diseases in the internal organs. This has led to the faster treatment of various diseases. As the world is moving ahead with industrial revolution 4.0 which incorporates Artificial Intelligence into the system, the medical industry is also shifting its paradigm to Artificial Intelligence [6]. A large number of researches are taking place to detect the presence of diseases automatically using the numerical values computed from various chemical tests, as well as the images of internal organs as well as cells captured using various devices. The major requirement is that models should be the most accurate as the domain is the medical domain.
The cancer cells have dimensions that are numerical. Based on these dimensions, it is understood whether the cancer cell is normal or has a tumor. The images of cancer cells are used to understand whether the cell has a tumor or not. Therefore, the research will attempt to increase the accuracy of determining whether a cell is malignant or not.

II. Related Works

As clinical science has progressed, numerous new strategies for distinguishing breast cancer growth have been created. The following is a brief outline of the research in this field.
Md. Milon Islam et al. [7] conducted a comparative study of machine learning methods on the Wisconsin Breast Cancer (WBCD) data. The algorithms that were compared include, K-nearest neighbours, random forests, artificial neural networks, and logistic regression. The algorithm efficiency was compared using the metrics such as precision, recall, sensitivity, specificity, negative predictive value, false-negative rate, false-positive rate, F1 score, and Matthews Correlation Coefficient. The dataset is partitioned using the ten-fold cross-validation procedure. According to the results, ANN outperformed every other model comparing to its accuracy, precision and F1 scores which are 0.9857, 0.9782, and 0.9890, respectively.
Meriem Amrane et al. [8] performed a comparitive study which used Naive Bayes classifier and K Nearest Neighbor using cross-validation. Both the models performed very well in terms of accuracy, but KNN gives a higher accuracy of 0.9751 when compared to NB Classifier with accuracy 0.9691. The running time for KNN increases with larger dataset. Shubham Sharma et al. [9] tested using the k − fold cross-validation split. It was found that every algorithm had an accuracy of more than 94%, having the ability to determine benign tumor or malignant tumor. When compared to other algorithms, KNN is the most effective algorithm for detecting breast cancer due to its high accuracy, F1 score and precision. Naji et al. [10] that Support Vector Machine outperforms all other algorithms as it achieved a higher efficiency of 0.972, Precision of 0.975, and, AUC of 0.966.
Siham A. et al. [11], for early identification of breast cancer, a Data Mining technique was recommended. Two datasets were used in the study: the breast cancer dataset and the WBC dataset. The training data was discretized, resampled, and missing values were then applied because there was a class imbalance. Then there was a 10-cross validation. Three different classifiers, J48, NB, and SMO, were utilised and their accuracy was compared. The classifier's performance improved after using the resample filter. As performance evaluation criteria, accuracy, standard deviation, ROC Curve, True Positive, and False Positive were used. Experiments reveal that SMO outperforms other models in the WBC dataset with an accuracy of 0.9699, and in the Breast Cancer dataset J48 with an accuracy of 0.7552.
Abdullah-Al Nahid et al. [12] recommended the state-of-the-art Deep Neural Network (DNN) for biological image analysis. Computer-Aided Diagnosis (CAD) approaches are utilised to assist doctors in making more trustworthy decisions. Novel DNN algorithms that are guided by structural and statistical information from the images were employed to classify a set of images from the BreakHis dataset. The algorithms that were used to classify breast cancer images include- CNN, LSTM, and a mixture of CNN and LSTM. While the feature extraction was conducted by proposed novel DNN models, softmax and SVM layers were used for the decision-making. On the other hand, the hidden structure of data was uncovered by K-Means and Mean-Shift clustering techniques. Softmax-Regression techniques and the Support Vector technique were used in the decision layer. The data is analysed using three alternative models. For data analysis, the first model uses CNN techniques, the

second model uses the LSTM structure, and the third model uses the combined CNN and LSTM structures. Upon classifying the images into malignant and benign using DNN models, the best specificity is 0.96, best sensitivity is 0.93, best recall is 0.96, and best F-measure is 0.93.
Steve A. Adeshina et al. [13] focused on the intra-class classification using BreakHis dataset where Breast images are divided into eight classes. Deep Convolutional Neural Networks (DCNN) architecture is more effective in the classification. To offer effective automatic image categorization, the DCNN architecture integrates ensemble learning, backpropagation training, and the ReLU activation function. Furthermore, when compared to splitting the images using train-test split and then combining these images again for final prediction, using whole slide images (WSI) for feature extraction will save time and resources. The accuracy of the inter-class classification is 0.915, with an error rate of 0.845, a precision of 0.6336, and a recall rate of 0.7667.

III. Proposed Model

Breast cancer is a severe disease that affects predominantly women and only a few men. Cancer is formed mainly from cancer cells. The problem is the current tests take a large amount of time to generate a report based on which it is decided whether the patient has breast cancer or not. Thus there exists a need to detect and understand the presence of cancer cells automatically from the images of the cells as well as the dimensions of the cell structures in order to classify them into Malignant or Benign.
Supervised algorithms- Machine Learning models and Deep Learning models are proposed to detect cancer cells in human beings. Fig. 1 and Fig. 2 depicts the flow of the process in both machine learning and deep learning.

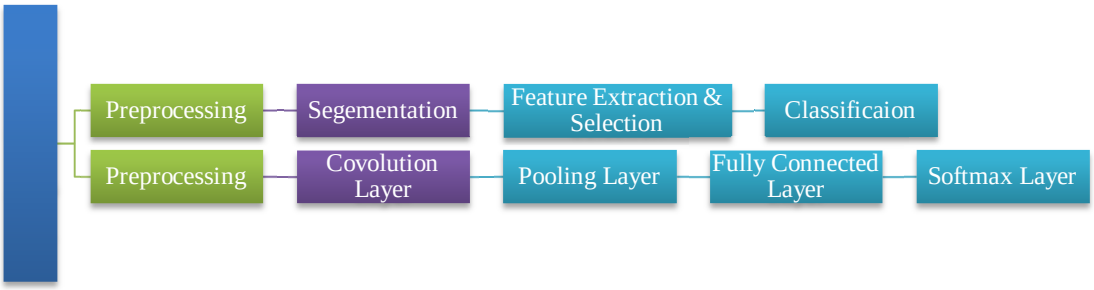

Figure 1. Flowchart of model process

Figure 2. Deep Learning for binary classification

The major purpose is to develop a model that can automatically detect breast cancer from images as well as numerical values. Combining machine learning and deep learning algorithms are mainly used to achieve this. Based on the computation conducted on the data, machine learning algorithms develop models. There are two types of problems: classification and regression. A classification algorithm is used to sort the data, and regression is used for predicting the values. Both techniques use learnings from the training data. Deep learning, which is a subset of machine learning, is similar to machine learning algorithms

A. Naïve Bayes

The algorithm is a Bayes Theorem-based classification technique. Naive Bayes assumes that the predictors are uncorrelated or independent. The Bayes Theorem is defined as in equation (1)

$$P(c \mid X) = P(x_1 \mid c) * P(x_2 \mid c) *\ldots\ldots* P(x_n \mid c) * P(c) \quad (1)$$

Different Naive Bayes classifiers work mainly based on the assumptions that are made on the distribution. Naive Bayes estimates the probability of class based on each feature. Thus the data can be converted in two ways. First, conversion of continuous features to categorical features using a binning process. Second, Gaussian Naive Bayes classification.

B. Logistic Regression

A classification model, logistic regression is. Based on an input variable, it models the likelihood of a class. The output variable has binary outcomes. Logistic regression estimates the parameters of the logistic model. The logistic model calculates the logit function, which is given in equation (2)

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \quad (2)$$

Thus the value falls between -1 and +1. If the logit function value lies below 0, then it is of class 0 and if the value of the logit function lies above 0, then it is of class 1. Major assumptions of logistic regression are, the observations are independent and there exists no multicollinearity.

C. K Nearest Neighbors Classifier

This classifier is a non-parametric supervised classification algorithm. For classification, the principle of distance is applied. The algorithm stores all the data and classifies novel data based on its similarity. During training, the number of neighbours is determined. When a new data point enters, the distance is calculated with respect to each neighbor. Take the k nearest neighbors and then determine the class of each data point. The class is defined as the new data point for the one which has maximum occurrence within all the k neighbors. There are different types of distance which include Euclidean, Manhattan, and Minkowski distances. Also, the number of neighbors selected should be odd to avoid confusion. Also, the numbers such as 1 and 2 will lead to noisy data. It is also suggested to have a higher number of neighbors.

D. Decision Tree Classifier

A decision tree is among the most powerful algorithms, which is a supervised classification technique. The flowchart of the algorithm follows a treelike structure. The internal nodes act as a test on features and each outcome the decision of the test. The leaf node represents the class. During the learning process of the tree, the source dataset is divided based on the attribute value tests followed by recursive partitioning. The recursion is completed in either of the two conditions: subset has same target values or when splitting does not lead to any predictions. When a new data point is tested in an algorithm, it goes through the decision tree and the corresponding class is predicted.

E. Random Forest Classifier

An ensemble learning-based supervised classification algorithm. It consists of a set of decision trees. Ensemble learning is a technique for teaching people how to work together. By combining different classifiers to solve a complex problem. The higher the number of trees, the higher is the accuracy and the lesser is the overfitting. It follows the principle of bagging. The bagging principle creates a different training subset from the training data of sample with replacement and the final output is based on the majority voting.

F. Support Vector Machines

It is one of the most popular classification algorithms. The principal aim of Support vector classifiers is to create a line or boundary that splits the entire space into classes. The line or boundary is called a hyperplane. When a new data point enters based on the space it is present, the corresponding class can be determined. Support vector machines help in choosing those vectors that can creates the hyperplane and these vectors are called support vectors. There are two types of support vector machines based on the separability of the dataset. It is linear SVM if the data is normally distributed and if the data is not separable linearly, then it is non-linear SVM.

G. Convolutional Neural Network

A deep learning algorithm that is used to analyze image datasets. Feature extraction and image classification are the two basic components. The layers under feature extraction include Convolution, Pooling, and Flattening. The convolution layer transforms the image in order to extract features from it. The pooling layer minimizes feature dimensions, reducing the number of parameters to learn and thereby saving computation time. The two types of pooling are maximum pooling and average pooling, with max pooling outperforming average pooling. Flattening layer converts into a one-dimensional vector thereby feeding the image classification. Image classification is done using fully connected layers. It derives nonlinear results as output.

H. Customized Convolutional Neural Network

The kernal of the CNN is 3x3 in size, whereas the input dataset is 64x64x3. The pooling is given by Max pooling and then it undergoes flattening. There are two convolution layers. After that, it's sent to a flattening layer, which is subsequently followed by a completely connected layer. ReLU and softmax are used as activation functions in the convolution layer and fully connected layer respectively.

I. VGGNet

Visual Geometry Group (VGG) is a standard multiple-layer Convolutional Neural Network architecture. It is built on cutting-edge object recognition models and serves as one of the most popular image recognition architectures. The input layer of VGG has a size of 224x224. The convolution layers have minimal receptive filed and are followed by an activation Rectified Linear Unit (ReLU). The number of strides is 1. The hidden layers in VGG have a ReLU activation function. There are three fully connected layers in ReLU. Based on the depth there are two kinds of VGG architectures: VGG16 & VGG19. VGG16 has 16 layers out of which 13 layers are convolutional layers. VGG19 has 19 convolutional layers out of which 16 layers are convolutional layers.

IV. EXPERIMENTS AND RESULTS

A. Data

In order to conduct the analysis, there were two datasets used, Wisconsin and BreakHis dataset.

Wisconsin dataset: Wisconsin is a classification dataset that was fetched from the Irvine Repository. There are mainly two classes of images- malignant and benign. The malignant class is down sampled in this dataset. The major columns of this dataset: ID number, Diagnosis (M = malignant vs B = benign), Radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension are ten real-valued attributes for each cell nucleus. The average, standard deviation, and worst of all 10 attributes are obtained. [14]. Thus there are 32 columns in the dataset. Wisconsin Dataset is shown in Figure 3. Figure 4. Plots the distribution of diagnosis within the dataset.

```
Data columns (total 33 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    object
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
 32  Unnamed: 32              0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

Figure 3. The information regarding the dataset

```
B    357
M    212
Name: diagnosis, dtype: int64
```

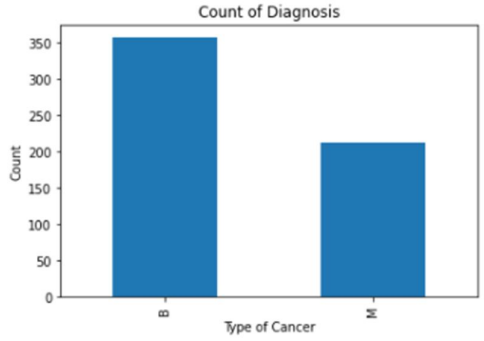

Figure 4. Count Plot of diagnosis for Wisconsin dataset

BreakHis Dataset: The Histopathological Image Classification of Breast Cancer used images of 82 patients. From these patients, a total of 9,109 microscopic were extracted. Four magnification factors- 40X, 100X, 200X, and 400X were used. In the dataset, there are 2480 benign and 5429 malignant samples. There are two main groups for the dataset BreakHis: benign tumors (B) and malignant tumors (M). A benign lesion is one that is neither malignant nor cancerous. The term "malignant tumour" refers to cancerous growth in the human body [15]. Table 15. Elaborates the dataset breakdown structure.

TABLE I. STATISTICAL BREAKHIS STRUCTURE [15]

Magnification	Benign	Malignant	Total
40X	652	1370	1995
100X	644	1437	2081
200X	623	1390	2013
400X	588	1232	1820
Total of Images	2480	5429	7909

B. Data preprocessing

As part of this phase, a consistent format and structure is created, and this is suitable for building a machine learning model. Data preprocessing is done for Wisconsin Dataset, whose columns have numerical and categorical values. Hence the preprocessing steps include:

1) Handling Imbalanced Data: A classification dataset is said to be imbalanced if the count of class values is in unequal class proportions. The majority of classes account for a significant portion of the dataset. Minority classes account for a lesser portion of the dataset. An imbalanced dataset leads to biasedness, where majority classes will be learned more and will lead to overfitting.
2) Data Scaling: It is a technique that is used for normalizing a set of independent values. Data scaling ensures that the set of values belongs to a common range. There are different methods for data scaling which include standardization and normalization. Standardization is used when the data follows the normal distribution. Normalization is used when data does not follow the normal distribution. Data scaling was performed for the Wisconsin dataset only. The data scaling operation performed was Standardization.
3) Cleaning the Data: It is the process of identifying errors from the dataset and removing it, thereby ensuring high quality of data and reduces inappropriate conclusions. Some of the steps in data cleaning involve reducing extra spaces, selecting and treating all NULL cells, conversion of text into numbers, removing duplicates, changing the lower/upper/proper case, and spell-check.
4) Null Value Detection: Null values are values that are unknown in a dataset. It can be a single value, an entire row, or an entire column in a dataset. It needs to be handled separately, as machine learning models cannot handle null values. Some of the methods that can handle null values include removal of null values, imputation using mean, median, and mode. No null values were detected in either of the datasets.
5) Handling Data Redundancy: Data redundancy refers to duplication of observations. The same piece of data is repeated in more than one place. Tracking redundancy can track long-term consistency issues.
6) Random Forest Feature Selection: Random Forest is one of the most extensively used supervised machine learning algorithms. It is based on the concept of interpretability. Embedded approaches include feature selection. Filter and wrapper techniques are combined in these embedded methods. High accuracy, higher generalization, and interpretability are among its benefits. There are 400-1200 decision trees in the random forest. Each tree is constructed using a combination of random observation extraction and random feature extraction. Because each tree does not observe all features or all observations, de-correlated trees have a lower likelihood of overfitting. Each tree is made up of a series of binary questions represented by nodes. The purity of each bucket determines the importance of each attribute. During the training process, a tree calculates how much each feature lowers impurity. The higher the relevance of a trait, the less impurity it reduces. In the case of random forest, impurity decrease from each feature are computed using the average computed across all the trees. Thus, the final importance of the variable is determined from the average value.

C. Performance Metrics

The performance of a classification algorithm is described by a confusion matrix, which is a square matrix. The diagonal elements show the number of correctly predicted observations. The off-diagonal elements show the number of incorrectly predicted observations. It is used to find accuracy, misclassification, precision and sensitivity. Table 2. Shows the Confusion Matrix.

TABLE II. CONFUSION MATRIX

	Predicted True	Predicted False
Actual True	True Positive	False Negative
Actual False	False Positive	True Negative

1) Accuracy: It is a metric in the classification model. It is the percentage of observations that were correctly anticipated. Mathematically shown in equation (3).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

The values lie between 0 and 1. The higher the accuracy, the better is the model.

2) Precision: It is a metric in classification which tells the proportion of count of true observations that are predicted. Mathematically shown in equation (4).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

3) Recall: It is a metric in classification which tells the proportion of count of truly predicted values out of the count of original observations. Mathematically shown in equation (5).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

4) F1 Score: It is the harmonic mean of precision and recall. Mathematically shown in equation (6).

$$\text{F1Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (6)$$

D. Results

The best attributes were chosen using random forest classifiers. The top 7 features were selected based on the importance which includes, mean of radius, mean of compactness, mean of texture, mean of smoothness, mean of fractal dimensions, mean symmetry, texture standard error, and smoothness standard error. Table 3. Shows the important features using Random Forest Classifier.

TABLE III. FEATURE IMPORTANCE USING RANDOM FOREST CLASSIFIER

Feature	Feature Importance
Mean radius	0.379206
Mean Compactness	0.163988
Mean Texture	0.137017
Mean Smoothness	0.095211
Mean Fractal Dimension	0.074691
Mean Symmetry	0.063961
Texture (Standard Error)	0.043059
Smoothness (Standard Error)	0.042868

Upon analysis the Wisconsin dataset using machine learning and deep learning techniques- Random Forest Classifier, Support Vector Classifier, Logistic Regression, Decision Tree Classifier, K Neighbours Classifier, and Naive Bayes Classifier, Random Forest classifier is the most accurate machine learning algorithm with an accuracy of 0.972, followed by Logistic Regression with accuracy 0.965, Support Vector classifiers with accuracy 0.965, K Neighbors Classifier with accuracy 0.9301, Decision Tree Classifier has an accuracy of 0.9161, while the Naive Bayes Classifier has an accuracy 0.9161. Coming to the execution of machine learning algorithms, Decision Tree classifier took the lowest time followed by Naive Bayes Classifier with execution time 0.0067 seconds, Logistic Regression with execution time 0.0073 seconds, Support Vector Classifier with execution time 0.0078 seconds, Random Forest Classifier with execution time 0.1806 seconds, and K Neighbors with execution time 1.0490 seconds. From execution time and accuracy, it can be that Logistic Regression has the highest efficiency in the classification of the Wisconsin dataset. Meanwhile, K Neighbors Classifier showed the lowest efficiency.

Deep learning algorithms are trained with fully connected layers. The hidden layers' activation function is ReLU, while the output layer's activation function is sigmoid. The number of epochs is 100, Adam acts the optimizer,

and binary cross-entropy is used as the loss function. Thus, the accuracy obtained is 0.958. Table 4. Gives the various model accuracy.

TABLE IV. MACHINE LEARNING ALGORITHM WITH EXECUTION TIME AND ACCURACY

Algorithm	Execution Time(seconds)	Accuracy
Logistic Regression	0.00734	0.965
Random Forest Classifier	0.1806	0.972
K Neighbors	1.04899	0.9301
Support Vector Classifier	0.00776	0.965
Decision Tree	0.00353	0.9161
Naive Bayes Classifier	0.00676	0.9161
Deep Learning	5.331	0.958

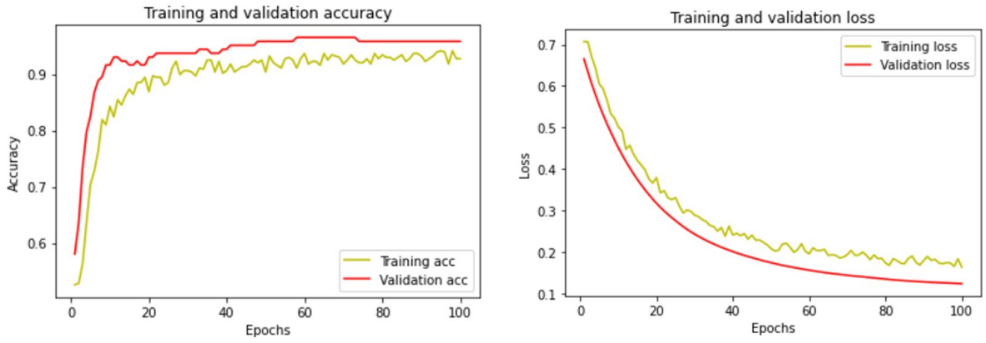

Figure 5. Binary Classification: Training & Validation for Accuracy & Loss

The Breakhis dataset uses Deep Learning models which include regular Customized Convolutional Neural Network and VGG architectures, VGG16 and VGG19. There were 40 epochs for training the algorithm. Adam is the optimizer and loss function is binary cross-entropy. Table 5. Shows the training and testing accuracy of BreakHis Dataset over Customized Convolutional Neural Network, VGG16 and VGG19.

TABLE V. DEEP LEARNING ALGORITHM WITH TRAINING AND TESTING ACCURACY

Algorithm	Training Accuracy	Testing Accuracy
CNN	0.9796	0.7712
VGG16	0.9582	0.7276
VGG19	0.9393	0.7825

V. CONCLUSION

The detection of breast cancer involved two types of datasets, the Wisconsin dataset and the BreakHis dataset which are numerical and image datasets respectively. The Wisconsin dataset was used by both machine learning algorithms and deep learning algorithms. All the models attained an accuracy above 90%. Seven main algorithms- Logistic Regression, Random Forest Classifier, K Neighbours, Decision Tree, Support Vector Classifier, Naive Bayes Classifier and Deep Learning. In the image dataset, CNN architecture exhibits the highest accuracy followed by VGG16 and VGG19. Thus we could see that for the BreakHis dataset, the accuracy obtained was above 85%. In the Collab environment, all of the algorithms were written in Python using the scikit-learn module. In conclusion, the numerical dataset, machine learning algorithms provide better accuracy compared to that of deep learning algorithms. The machine learning algorithms used for numerical dataset provides better accuracy than deep learning algorithms for both numerical and image datasets. In future work, the Usage of multiple modalities for the identification of cancer can be helpful for the development of

applications and the Development of new systems with multiple algorithms. Automated and faster detection of different types of cancer from three dimensional images as well as numerical values helps in the easier and early detection.

REFERENCES

[1] "Breast cancer," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] T. Cytecare, "Statistics of breast cancer in India: Cytecare Hospitals," Cytecare Hospital in Bangalore, 07-Mar-2021. [Online]. Available: https://cytecare.com/blog/statistics-of-breast-cancer/.

[3] Cancer.org. 2022. Types of Breast Cancer | About Breast Cancer. [online] Available at: <https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer.html>

[4] "Alan Turing and the beginning of ai," Encyclopædia Britannica. [Online]. Available: https://www.britannica.com/technology/artificial-intelligence/Alan-Turing-and-the-beginning-of-AI.

[5] "A better measure of relative prediction accuracy for model ..." [Online]. Available: https://www.researchgate.net/profile/Chris-Tofallis/publication/271773969_A_better_measure_of_relative_prediction_accuracy_for_model_selection_and_model_estimation/links/59ae7b22aca272f8a167ae7d/A-better-measure-of-relative-prediction-accuracy-for-model-selection-and-model-estimation.pdf.

[6] T. Davenport and R. Kalakota, "The potential for artificial intelligence in Healthcare," Future healthcare journal, Jun-2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/.

[7] A. Nahid, M. A. Mehrabi, and Y. Kong, "Histopathological breast cancer image classification by deep neural network techniques guided by local clustering," *Macquarie University*, 22-Oct-2018. [Online]. Available: https://researchers.mq.edu.au/en/publications/histopathological-breast-cancer-image-classification-by-deep-neur.

[8] "Analysis of breast cancer detection using different ..." [Online]. Available: https://link.springer.com/content/pdf/10.1007/978-981-15-7205-0_10.

[9] "Breast cancer histopathological database (BreakHis) - Laboratório Visão Robótica e Imagem," *Laboratório Visão Robótica e Imagem - Laboratório de Pesquisa ligado ao Departamento de Informática*, 25-Oct-2019. [Online]. Available: https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/.

[10] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast Cancer Classification Using Machine Learning: Semantic scholar," *undefined*, 01-Jan-1970. [Online]. Available: https://www.semanticscholar.org/paper/Breast-cancer-classification-using-machine-learning-Amrane-Oukid/ea98b9481ec6943b799a89f0647becafa303066f.

[11] M. M. Islam, M. R. Haque, H. Iqbal, M. Hasan, M. Hasan, and M. Kabir, "[PDF] Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques: Semantic scholar," *undefined*, 01-Jan-1970. [Online]. Available: https://www.semanticscholar.org/paper/Breast-Cancer-Prediction%3A-A-Comparative-Study-Using-Islam-Haque/34de24378803dbd4138401800a90f2653d0bdaaa.

[12] "Machine learning algorithms for breast cancer prediction ..." [Online]. Available: https://www.researchgate.net/profile/Olivier-Debauche/publication/352572400_Machine_Learning_Algorithms_For_Breast_Cancer_Prediction_And_Diagnosis/links/61392c04b1dad16ff9f04bbf/Machine-Learning-Algorithms-For-Breast-Cancer-Prediction-And-Diagnosis.pdf.

[13] S. A. Adeshina, A. P. Adedigba, A. A. Adeniyi, and A. Aibinu, "Breast cancer histopathology image classification with deep convolutional neural networks: Semantic scholar," *undefined*, 01-Jan-1970. [Online]. Available: https://www.semanticscholar.org/paper/Breast-Cancer-Histopathology-Image-Classification-Adeshina-Adedigba/d2a64f6b59a2e5e0302e396679606fa21c98321d.

[14] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[15] S. Sharma, A. Aggarwal, and T. Choudhury, "Figure 1 from breast cancer detection using machine learning algorithms: Semantic scholar," *undefined*, 01-Jan-1970. [Online]. Available: https://www.semanticscholar.org/paper/Breast-Cancer-Detection-Using-Machine-Learning-Sharma-Aggarwal/f2caea5a7fd7f48914023f2365b07f119a9793bb/figure/0.