International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET2310375

A Breast Cancer Diagnosis Framework Based on Machine Learning

Dr. Nikhat Akhtar*[1], Dr. Hemlata Pant[1], Apoorva Dwivedi[2], Vivek Jain[3], Dr. Yusuf Perwej[4]

[1]Associate Professor, [2,3]Assistant Professor, [4]Professor

[1]Department of Computer Science & Engineering, Ambalika Institute of Management & Technology, Lucknow, Uttar Pradesh, India

[2,4]Department of Computer Science & Engineering, Ambalika Institute of Management & Technology, Lucknow, Uttar Pradesh, India

[3]Department of Computer Science & Engineering, IMS Engineering College, Ghaziabad, Uttar Pradesh, India

ARTICLEINFO

Article History:

Accepted: 10 April 2023
Published: 19 May 2023

Publication Issue
Volume 10, Issue 3
May-June-2023

Page Number
118-132

ABSTRACT

Breast cancer is becoming the leading cause of mortality among women. One of the most prevalent diseases in women, breast cancer is brought on by a variety of clinical, lifestyle, social, and economic variables. Predictive approaches based on machine learning offer methods for diagnosing breast cancer sooner. It may be found using a variety of analytical methods, including Breast MRI, X-ray, thermography, mammograms, ultrasound, etc. The most prevalent technique for performance evaluation uses accuracy measures, and the Convolutional Neural Network (CNN) is the most accurate and widely used model for breast cancer diagnosis. The Wisconsin Breast Cancer Datasets (WBCD) were used to evaluate the suggested method. Out of a total of 569 samples, 273 samples were chosen for this experiment as the test data, while the other samples were utilized for training and validation. The review's findings showed that the Convolutional Neural Network (CNN) is the most effective and widely used model for finding breast cancer, and that the most often used technique for judging performance is accuracy metrics. The application of deep learning to such a wide range of real-world issues is astounding.

Keywords: Machine Learning, Breast Cancer, Malignant, Classification, Convolutional Neural Network (CNN), Wisconsin Breast Cancer Datasets (WBCD), UCI Machine Learning Repository.

I. INTRODUCTION

Breast cancer, which has overtaken all other types of cancer as the leading cause of mortality in women, is one of them [1]. According to the fatality rate, breast cancer is the fourth most prevalent cancer worldwide. Many variables that serve many purposes and cause breast cancer are to blame. Diet, genetics, hormones,

obesity, cell proliferation from aberrant cells, and even radiation therapy are some of these variables. Previous research has shown that when breast cancer is discovered at an advanced stage, about half of the patients pass away [2]. As a result of the rapidly expanding population and the exponential daily growth of medical photos, the existing methods for BC categorization are no longer adequate to handle the rising demand for medical images [3]. In order to analyses and find the irregularities within the medical image, pathologists regularly visually inspect and explore the complete pathological image. Furthermore, determining if a medical picture is carcinogenic or not using clinical diagnostic [4] procedures take a considerable amount of time. This procedure is incredibly time-consuming and tiresome. Additionally, each medical practitioner has a different subjective mood for assessing weariness, and the human eye is less good at detecting small changes in the tissue.

This might lead to various medical professionals to various diagnostic findings on the same medical imaging. Medical picture diagnosis is heavily reliant on the human aspect, which is not error-free. Due to the doctor's small error, the patient can experience serious consequences. Studies indicate that if women can identify breast cancer early and undergo therapy at an early stage, the situation may improve [5]. They must accomplish this by accurately forecasting how the illness would develop from a mild condition to breast cancer. Making precise forecasts early on is made easier by machine learning technologies. A key role in the categorization of breast cancer is played by machine learning, a branch of artificial intelligence.

A number of research employ machine learning methods including linear discriminant analysis (LDA), support vector machines (SVM), and artificial neural networks (ANN) [6] to classify data, construct models, or improve the performance of existing models. Convolutional neural networks (CNNs) [7] have gained popularity in recent years for a variety of machine learning applications involving pictures and videos, including image classification, face recognition, object identification, and image segmentation [8]. As a result, the authors of this research have suggested a convolutional neural network (CNN) [9] model for the categorization of breast cancer. By identifying the tumors and categorizing it as benign or malignant [10], Wisconsin Breast Cancer Datasets (WBCD) breast cancer datasets advanced to the next level of quality.

II. BACKGROUND

Due to the recent increase in mortality, the categorization of breast cancer has become a crucial problem [11]. Several state-of-the-art studies that assisted in carrying out our suggested method of diagnosing and categorizing the disease using mammogram images have been conducted in this field, and they are discussed in this section. the previous research on machine learning-based breast cancer diagnosis [12]. For the purpose of identifying and categorizing breast cancer from mammography pictures, a deep learning-focused method [13] has been put forth. In this method, the feature extraction was carried out using the k-means clustering algorithm [14] to speed up robust features, and the mammogram was classified into three classes—normal, benign, and malignant—with 95%, 94%, and 98% accuracy using the multiclass support vector machine classifier.

Utilizing machine learning techniques such as the Support Vector Machines (SVM), Random Forest (RF), Convolutional Neural Network (CNN) method for breast image classification, and conventional Neural Network (NN) [15]. By employing lesion annotations just during the first training phase and omitting them during the latter phases, a deep learning algorithm [16] has been devised that can more reliably detect breast cancer from mammography images. Numerous studies have been published and are based on various methods that could enable early cancer investigation and prediction, including SVM, Decision Tress [18], Artificial Neural Network [19] [20], Minimum Distance Classifier [21], Fuzzy Classifier [22], Fuzzy Rough Neural Network [23][24], Particle Swarm

Optimization [25][26], microRNA and biomarkers [27], and Deep Learning approaches. A survey on the use of deep learning in medical image processing was given by Litejens et al. in 2017 [28]. Modern classification algorithms like CNN, Principal Component Analysis (PCA), and K Nearest Neighborhood algorithms are used to diagnose breast cancer using various datasets of mammogram images [29]. In order to differentiate breast cancer, Meet et al. created a neural network based on the ICA-RBFNN [30], and they were successful in achieving an overall accuracy of 90.49% [31].

CNN [32] is used to diagnose breast cancer. There are two alternative training scenarios: one uses pre-trained weights, and the other uses a random procedure that is evaluated on two separate datasets. Investigational findings show that pre-trained networks accomplish their tasks more effectively. With NB classifier, M. Amare et al.'s [33] diagnosis of breast cancer had an overall accuracy of 97.5%; with K-NN classifier, they reached 96.1%. RBMs are generative deep learning-based techniques that train networks and learn features greedily, layer by layer, using blocks [34]. The same feature selection technique (PCA) was used to obtain 96.4% accuracy using KNN in different research by Yang & Xu et al. [35].

Histopathological pictures were employed to identify breast cancer in another study [36]. Images of the histopathology reveal observations made during the biopsy. Local and hidden patterns can be seen in these pictures. Unsupervised approaches like as Convolutional Neural Networks (CNN), Long-Short-Term Memory (LSTM) [37], and a combination of the CNN and LSTM models are used to find hidden patterns. The photos were then classified using SVM. To distinguish benign tumors from malignant breast cancer, M. Abdar et al. suggested an ensemble technique using vote/voting classifier. For two or three various machine learning algorithms, it created a two-layer voting classifier [38]. The outcomes of different voting methods showed that the straightforward classification algorithm performed adequately [39].

III. AIM

Breast cancer is one of the most prevalent malignancies that has been identified globally, including in India. Despite the high mortality rate, 97% of women with early diagnosis may expect to live for more than 5 years. Statistics show that the number of deaths brought on by this sickness has significantly increased in recent years. The objective of this study is to find general trends that might guide us in selecting the best model and its parameters, as well as to ascertain which qualities are most helpful [40] in predicting whether a tumour is benign or malignant. The goal is to determine if a breast cancer is aggressive or benign. We are used machine learning classification methods to fit a function that can forecast the discrete class of new data in order to do this. To determine if a breast cancer is aggressive or benign, the goal is to do a breast cancer screening.

IV. PREREQUISITE

This section outlines which area employs or adheres to whatever piece of technology, making it crucial when beginning a project. Before defining the hardware and software requirements, the development team should be aware of all the features and applications of the project.

4.1. UCI Machine Learning Repository

In this project, we will use data mining and machine learning techniques to examine the data to discover breast cancer. Breast cancer (BC) is a disease that often affects women everywhere. Early diagnosis of BC can dramatically improve prognosis and survival chances by motivating patients to pursue therapeutic therapy. Using the UCI Machine Learning Repository, we'll analyses the breast cancer dataset. The UCI Machine Learning Repository shown in figure 1 [41] offers free access to a database of machine learning problems. It is hosted and maintained by the Centre for Machine Learning and Intelligent Systems at the University of

California, Irvine. David Aha created it first as a PhD candidate at UC Irvine. For more than 25 years, it has been the go-to source for machine learning practitioners and researchers that want a dataset.

There is a homepage for each dataset that contains all the information that is currently accessible about it, including any relevant research publications. The actual datasets are accessible as ASCII files for download, typically in the useful CSV format. Dr. William H. Wolberg, a physician at the University of Wisconsin Hospital in Madison, Wisconsin, created the dataset that was utilized in this work, and it is freely available online. In order to create the dataset, Dr. Wolberg used fluid samples taken from patients with solid breast masses and Xcyt, an intuitive graphical computer application that can examine cytological traits based on a digital scan.

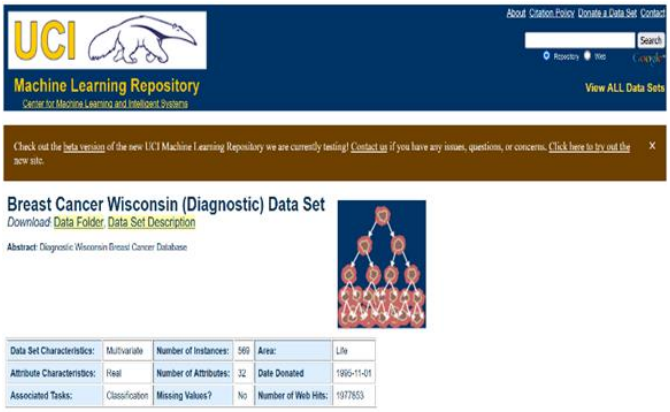

Figure 1 The Wisconsin Breast Cancer Diagnostic (WBCD) Dataset

4.2. Dataset

The investigation was conducted using the Wisconsin Breast Cancer Diagnostic (WBCD) dataset [42]. The dataset's simplified size is 56932 and was downloaded from the well-known machine learning repository UCI-Repository. The number 569 denotes the dataset's sample count, while the

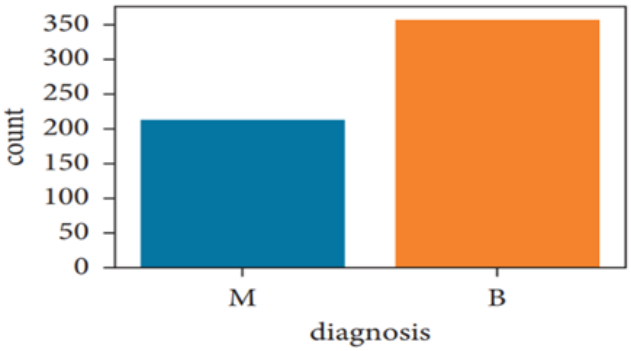

Figure 2: Total number of malignant and benign data

number 32 denotes the number of features displayed in figure 2. The atomic properties of fine needle aspirations (FNAs) taken from patients' breasts and presented make up the example dataset [43]. A small needle is placed into a body fluid or tissue that seems aberrant to get a sample for diagnosis or sickness prediction, such as cancer. Figure 2 displays the total quantity of malignant and benign data in the WBCD dataset. The dataset has no missing characteristics, and the distribution of classes is 357 benign and 212 malignant.

4.2.1. Attribute Information

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from centre to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

4.3. Data Mining and Machine Learning

The term "data mining" is deceptive since the goal is not the actual extraction (mining) [44] of data but rather the extraction of patterns and information from enormous amounts of data. It's also a buzzword that's frequently used to describe any type of extensive data or information processing (collection, extraction, warehousing, analysis, and statistics), as well as any application of computerized decision support systems like artificial intelligence and business intelligence [45] (such as machine learning). The research used the following machine learning algorithms.

4.3.1. Decision Tree Algorithms

Decision tree algorithms are among the most effective machine learning classification techniques. They are methods of supervised learning that use gathered and edited data to improve consequences. Furthermore, a variety of studies, including those in the fields of medicine and health issues, frequently use decision tree algorithms for categorization. There are several different types of decision tree algorithms, including ID3 and C4.5 [46]. Nevertheless, J48 is the most popular decision tree algorithm. J48 is an improved implementation of C4.5 and an extension of ID3.

4.3.2. Random Forest

An incredibly common supervised machine learning technique used for Classification and Regression issues in machine learning is called the Random Forest technique [47]. A random forest is a meta estimator that employs averaging to increase predicted accuracy and reduce overfitting after fitting several decision tree classifiers to diverse dataset subsamples. With the use of bootstrap aggregation and random feature selection, the predictions from the forests are averaged. It has been shown that RF models are reliable predictors of outcomes for both small sample numbers and high dimensional data.

4.3.3. K-nearest-Neighbors (KNN) Algorithm

It is an easy method for pattern recognition supervised learning. It is one of the most popular neighborhood classifiers because of how simple it is to use and how good it is in the field of machine learning. The KNN approach looks in the pattern space for the k training tuples that resemble the unidentified tuples the most [48]. Then, based on similarity measurements, it categorizes fresh instances and saves all previous cases. Performance determines the right number of neighbors (k), which changes from data sample to data sample.

4.3.4. Support Vector Machine (SVM)

It is a statistical learning theory-based supervised learning method for classifying both linear and nonlinear data. SVM [35] separates data into two classes over a hyper plane while preventing over-fitting data by increasing the margin of hyper plane splitting.

4.3.5. Naïve Bayes (NB)

It is a probabilistic classifier that use Bayes' theorem to apply to one of the most efficient classification techniques. The value of the feature is expected to be independent of the values of any other characteristics given the class variable. based on the highest probability. It decides if the supplied tuple belongs to a particular class.

4.3.6. Logistic Regression

In statistics, the logistic model, often referred to as the genuine model, is used to calculate the probability that a certain class or event, such as pass/fail, win/lose, alive/dead, or healthy/ill, would take place. This might be developed to replicate a number of event classes, like determining whether a cat, dog, lion, etc. is present in a photo. The probability assigned to each identified object in the picture would range from 0 to 1, with the sum equaling 1 [49]. Logistic regression was first used in the biological sciences around the turn of the 20th century. Then, it was used for a variety of social scientific purposes. Logistic regression is used when the dependent variable (target) is categorical.

V. PROPOSED SYSTEM

The second most frequent cancer in both men and women globally is breast cancer. Breast cells start to proliferate uncontrollably with the onset of breast

cancer. These cells typically develop into tumors, which are frequently palpable lumps or visible on x-rays. If the tumor's cells have the ability to metastasis to other parts of the body or to neighboring tissues, the tumor is considered malignant (cancer). A feed-forward neural network called a convolutional network analyses visual pictures by processing data in a grid-like architecture. It is sometimes referred to as a ConvNet. In order to handle data with a grid pattern, such as photographs, CNN is a form of deep learning model shown in figure 3. CNN is inspired by the structure of animal visual cortex and was created to automatically and adaptively learn spatial hierarchies of characteristics, from low- to high-level patterns [50]. Convolution, pooling, and fully linked layers are the three types of layers (or "building blocks") that make up a standard CNN. Convolution and pooling layers in order one and two do feature extraction, whereas a fully connected layer in order three translates the retrieved features into the output, such as classification.

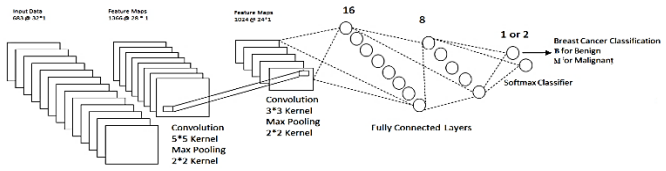

Figure 3. The Suggested CNN Architecture for Classifying Breast Cancer

The dataset was sent to CNN as input for the breast cancer classification application. We trained the deep convolutional kernels in the proposed CNN architecture after feeding them the input. We do employ LeakyRELU [51] nonlinearity for the convolutional layers in this study, and this is what it is:

$$f(\mathrm{p}) = \begin{cases} \mathrm{p}, if\ \mathrm{p} > 0 \\ a\mathrm{p}, \text{else} \end{cases}$$

The convolutional layer can usually be stated as:

$$y^j = f(b^j + \sum_i k^{ij} + p^i)$$

In this section the ith input map is represented by Pi, while the jth output map is represented by Qi. The terms * signify the convolutional operation between the two functions, bij displays the convolutional kernel used between the i and j maps, and bj represents the bias parameter of the jth map. The max-pooling layer was added following the convolutional layer. Each neuron [52] in this layer pools over a s * s non-overlapping region in the input map Pi to form the output map Qi. The max-pooling layer is often defined as:

$$Q_j^i = \max_{0 \le m \le s} \left\{ p_{j,s+m}^i \right\}$$

Max-pooling and convolutional layers are fully linked, and these layers are connected to the Softmax classifier after that. The number of output classes in the Softmax [53] classifier is the same as the number of outputs. The k-dimensional input vector (dataset) is re-normalized, and the Softmax function functions as a function of squashing to give results in the real value range [1,2,3].

$$\delta(R)_j = \frac{e^{R^j}}{\sum_{k=1}^{K} e^{R^k}}$$

Here, $[j = 1, 2, 3, 4, \ldots\ldots k]$

We have two classes in the suggested architecture, first a benign class and second a malignant class. The suggested CNN classifier was trained using the following weighted loss function.

$$\alpha(\xi, P_n, Q_n) = -\frac{1}{N} \sum_{n=1}^{N} \vartheta_n \sum_{k=1}^{K} t_{kn} Q_{kn}$$

Specifically, Pn is the input vector, Qn is the classification output for the nth clinical input data, and tn is the actual clinical sample response. N is the total number of clinical samples, and K is the number of classes.

VI. RESULT AND ANALYSIS

Figure 4 illustrates the difference between the validation folder's 250 photographs for each category and the training folder's 1000 images for each category. [WIDTH, HEIGHT, CHANNELS] is the form of a matrix of pixel values. [32x32x3] is our input. Typically, while detecting low-level features, we start with a small number of filters. We employ additional filters to

find high-level characteristics as we delve further into the CNN. In order to create a feature map, feature detection is dependent on "scanning" the input using a filter of a specific size and executing matrix calculations. When pooling-size= (2, 2), the pooling layer will downscale along the spatial dimensions (width, height), producing an output like [16x16x12].

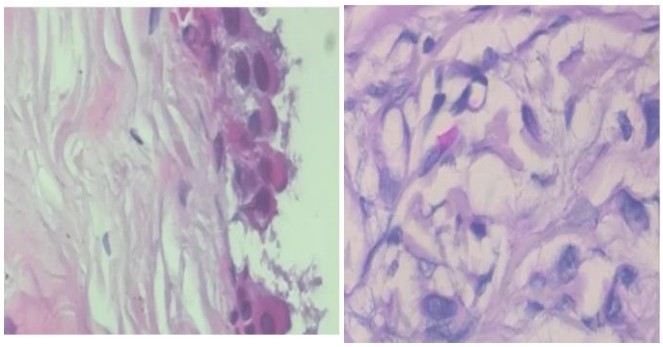

Figure 4. The Benign and Malignant Sample

Our input is a training dataset made up of N images [54][55], each of which has been assigned to one of two classes. The classifier is then trained using this training set to discover the characteristics of each class. Finally, we assess the classifier's performance by asking it to forecast labels for a fresh batch of photos that it has never seen before. Then, we will contrast the actual labels on these pictures with those that the classifier predicted. The photos were loaded into the appropriate directories.

```
def Dataset_loader(DIR, RESIZE, LeakyRELUX=10):
    IMG = []
    read = lambda imname: np.asarray(Image.open(imname).convert("RGB"))
    for IMAGE_NAME in tqdm(os.listdir(DIR)):
        PATH = os.path.join(DIR,IMAGE_NAME)
        _, ftype = os.path.splitext(PATH)
        if ftype == ".png":
            img = read(PATH)
            img = cv2.resize(img, (RESIZE,RESIZE))
            IMG.append(np.array(img))
    return IMG
benign_train = np.array(Dataset_loader('data/train/benign',224))
malign_train = np.array(Dataset_loader('data/train/malignant',224))
benign_test = np.array(Dataset_loader('data/validation/benign',224))
malign_test = np.array(Dataset_loader('data/validation/malignant',224))
```

After that, the data set is divided into two train and test sets, each having 80% and 20% of the images being benign and malignant, respectively.

```
x_train, x_val, y_train, y_val = train_test_split(
    X_train, Y_train,
    test_size=0.2,
    random_state=11
)
w=60
h=40
fig=plt.figure(figsize=(15, 15))
columns = 4
rows = 3

for i in range(1, columns*rows +1):
    ax = fig.add_subplot(rows, columns, i)
    if np.argmax(Y_train[i]) == 0:
        ax.title.set_text('Benign')
    else:
        ax.title.set_text('Malignant')
    plt.imshow(x_train[i], interpolation='nearest')
plt.show()
```

In this section, the (M) stands for dangerous malignant cells, whereas the (B) stands for benign, or healthy, cells. Now, a connection between the various attributes is apparent. How much one column in this heat map influences every other column (for example, radius means have a 32% influence on texture mean). Training and testing. Following that, the datasets were split into independent (P) and dependent (Q) datasets. Where P = df.iloc[:, 2:31]values, Q = df.iloc[:, 1].values. They have a type of array. While the dependent data set (Q) provides the malignancy diagnosis for the patient, the independent dataset (P) contains the qualities that are utilized to forecast the outcome.

We right now split the dataset in half, using 20% for testing and 80% for training. On the training set, we employ a variety of machine learning [56] models, such as K-Nearest Neighbor, decision trees, logistic regression, Naive Bayes Classifier, Random Forest Classifier, and radial basis function neural networks

(RBFNN) [57]. depending on the characteristics provided by the data and the training, whether a tumor is malignant (M) (hazardous) or benign (B) (not harmful). Breast cancer was obtained using the Wisconsin Breast Cancer Datasets (WBCD). The programme will evaluate the datasets based on a number of factors.

6.1. Attributes

- **diagnosis:** The diagnosis of breast tissues (M=malignant,B=benign)
- **mean_radius:** mean of distances from center to points on the peremeter
- **mean_texture:** standard deviation of gray-scale
- **mean_perimenter:** mean size of the core tumor
- **mean_area mean_smoothness:** mean of local variation in radius length.

6.2. Performance Assessment

A classification analysis is used to assess the precision of predictions provided by a classification algorithm. Table 1 displays the performance of the suggested architecture for classifying breast cancer using the Wisconsin Breast Cancer Datasets (WBCD). The metrics listed below are used to assess the CNN.

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

$$F-measure = 2 * \frac{Precision*Recall}{Precision+Recall}$$

True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and Precision were used to measure the effectiveness of the CNN breast cancer classifier [58]. Clinical samples that the constructed classifier accurately identified as benign are referred to as True Positives (TP). True Negative (TN) clinical samples are those in which the proposed classifier accurately identified the malignant clinical data. False-negative and false-positive cases occur when the recommended design incorrectly assigns the data to the benign class or the malignant class, respectively. It displays the classification mistake made. A competent classifier is able to accurately diagnose each sample. Nevertheless, due to the uncertainty of the classifier, a model cannot be employed in clinics if it properly predicts real negative samples but is unable to locate the true positive ones. The constructed classifier must thus have very high accuracy. We also provide precision-recall curves and receiver operating characteristic (ROC) curves to assess the effectiveness of the breast cancer categorization offered by Wisconsin Breast Cancer Datasets (WBCD) [59]. Following the use of numerous classification models, we used a variety of models to achieve the accuracy shown in figure 6.

Table 1. Performance Assessment of proposed CNN and Comparison with previously studies (Wisconsin Breast Cancer Datasets (WBCD))

Method	Accuracy (%)
PSOWNN	93.67
Decision Trees	95.70
K-Nearest Neighbour	97.40
Logistic Regression	95.74
CNN	98.86
Naïve Bayes Classifier	96.10
RBFN	90.49

After building our classification model, it is evident that the CNN Classification algorithm delivers the best results for the data we have. However, it doesn't applicable to all datasets.

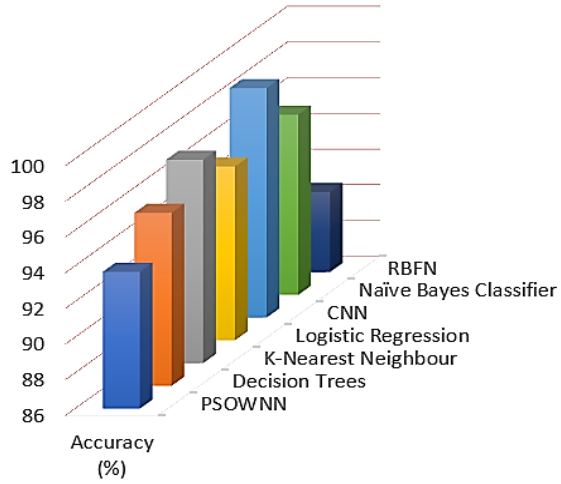

Figure 6 The Accuracies with Different Models

6.3. Receiver Operating Characteristics (ROC) Curves

Recognising the effectiveness of the created classifier requires the use of Receiver Operating Characteristics (ROC) curves [60]. It employs a graphical plot to show how a suggested Convolutional Neural Network (CNN) can discriminate between samples of benign breast cancer and samples of malignant breast cancer. Using this technique, we were able to choose the best Convolutional Neural Network (CNN) model for categorising breast cancer. The ROC curves also show how well a powerful neural network-based model can differentiate across classes. Since there is no misclassification in any class and the entire area under the ROC curve is 1, these models have flawless classification. The Convolutional Neural Network (CNN) performs better with greater Area Under Curve (AUC) [61] values. This further suggests that the amount of training samples affects how well the suggested model performs; as the number of training examples rises, so does the classifier's performance on the test. The classifier's performance degrades as the number of training samples grows smaller.

```
from sklearn.metrics import roc_auc_score, auc
from sklearn.metrics import roc_curve
roc_log = roc_auc_score(np.argmax(Y_test, axis=1),
np.argmax(Y_pred_tta, axis=1))
false_positive_rate, true_positive_rate, threshold =
roc_curve(np.argmax(Y_test, axis=1),
np.argmax(Y_pred_tta, axis=1))
area_under_curve = auc(false_positive_rate,
true_positive_rate)
plt.plot([0, 1], [0, 1], 'r--')
plt.plot(false_positive_rate, true_positive_rate,
label='AUC = {:.3f}'.format(area_under_curve))
plt.xlabel('False positive rate')
plt.ylabel('True positive rate')
plt.title('ROC curve')
plt.legend(loc='best')
plt.show()
#plt.savefig(ROC_PLOT_FILE, bbox_inches='tight')
plt.close()
```

Area Under the Curve, or AUC, is 0.5 on the random line at 45 degrees shown in figure 5. The AUC and model quality increase when the curve deviates more from this line. The curve creates a right-angled triangle at an AUC of 1, which is the maximum a model may achieve. The ROC curve [62] can aid in model debugging. For instance, it suggests that the model is misclassifying at Y=0 if the bottom left corner of the curve is closer to the random line. In contrast, if it is random, it suggests that the mistakes are happening at Y=1.

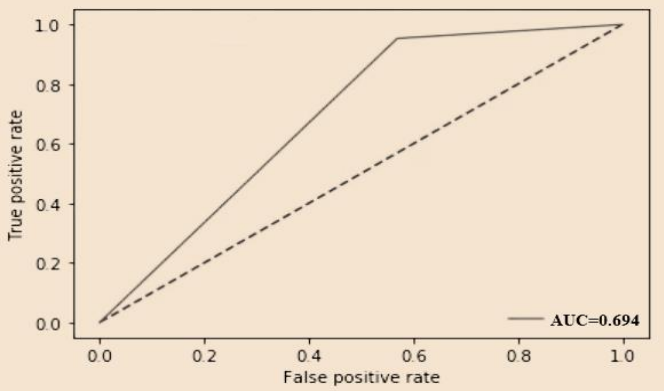

Figure 5. The Receiver Operating Characteristics (ROC) Curves

Out of the 569 samples utilized in this experiment, 273 samples were chosen as the test data while the other samples were used for training and validation. There are 273 samples total, 175 of which are malignant, and 98 of which are benign. The confusion matrices for the test data utilizing the Wisconsin Breast Cancer Datasets (WBCD) are shown in figure 6. The suggested classifier successfully differentiated between all of the benign and cancerous samples. Figure 7 illustrates the results of this model, which had an overall accuracy of 98.86%, precision of 98.78%, recall of 99.26% [63], and F-measure [64] value of likewise 99.07% and Roc-Auc of 0.694.

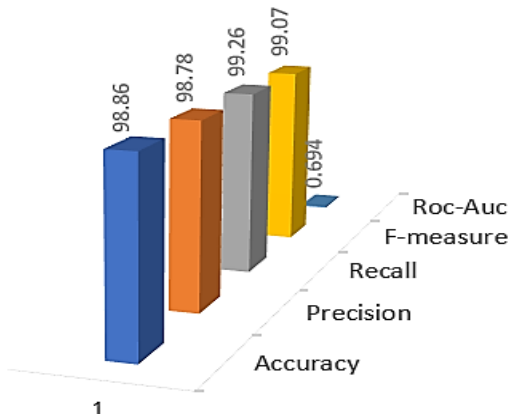

Figure 7 The overall accuracy, precision, recall, F-measure, and Roc-Auc Value

VII. CONCLUSION

The automated diagnosis of breast cancer is a problem that this work attempts to address using a machine learning system. In the current investigation, multiple machine learning techniques for spotting breast cancer were examined. The various research was carried out using the breast cancer dataset. With the help of the Wisconsin Breast Cancer Datasets (WBCD), we learned how to create graphs and results for the same breast cancer tumors predictor in this Python project. Convolutional, max-pooling, and fully linked layers were used in the pre-training phase, and this pre-training phase was. A classification layer is next used to separate the benign from the malignant samples. Accuracy has been demonstrated to increase with a solid dataset. The use of appropriate algorithms with a robust home dataset will lead to the development of prediction systems. These methods can aid in choosing the most appropriate course of treatment when a patient is diagnosed with breast cancer. There are several medicines available depending on the stage of a patient's breast cancer; data mining and machine learning may be of great assistance in choosing the course of therapy to be pursued by extracting knowledge from such applicable databases. The obtained findings for this investigation show how well the classifier performs in comparison to other cutting-edge approaches. This model produced an overall accuracy of 98.86%, with a precision 98.78%, recall 99.26%, and the F-measure value also 99.07% and Roc-Auc 0.694. Automating breast cancer detection to enhance patient care is a challenging task. Lastly, the proposed model seems to be perfectly suitable for controlling parameter settings for machine learning algorithms and automated breast cancer diagnosis.

VIII. WORK IN THE FUTURE

Breast cancer is one of the major factors that lead to death in women. Breast cancer is the most significant issue that women face. Data from the International Agency for Research on Cancer (IARC) in December 2022 show that breast cancer has surpassed lung cancer as the most common cancer in women diagnosed globally. Future studies can concentrate on developing the chosen approach into a potentially practical method for providing doctors with a rapid second opinion when diagnosing breast cancer. In the future, we would like to increase the dataset and assess the efficiency and scalability of the algorithm..

IX. REFERENCES

[1]. El-Nabawy, A., El-Bendary, N., Belal, N.A. “A feature-fusion framework of clinical, genomics, and histopathological data for METABRIC breast cancer subtype classification”, Appl. Soft Comput., 91, 20, 2020

[2]. L.A. Altonen, R. Saalovra, P. Kristo, F. Canzian, A. Hemminki, P Peltomaki, R. Chadwik, A. De La Chapelle, "Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease", N Engl J Med, vol. 337, pp. 1481-1487, 1998.

[3]. B. Liu, K. Yao, M. Huang, J. Zhang, Y. Li and R. Li, "Gastric Pathology Image Recognition Based on Deep Residual Networks," 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, 2018, pp. 408-412. doi: 10.1109/COMPSAC.2018.10267.

[4]. N. Akhtar, S. Rahman, H. Sadia and Y. Perwej, "A Holistic Analysis of Medical Internet of Things (MIoT)", Journal of Information and Computational Science (JOICS), vol. 11, pp. 209-222, 2021

[5]. I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines", Machine Learning, vol. 46, pp. 389-422, 2002

[6]. Yusuf Perwej, "An Evaluation of Deep Learning Miniature Concerning in Soft Computing", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 2, PP. 10 - 16, 2015, DOI: 10.17148/IJARCCE.2015.420

[7]. Shubham Mishra, Versha Verma, Nikhat Akhtar, Shivam Chaturvedi and Yusuf Perwej, "An Intelligent Motion Detection Using OpenCV", Journal of Scientific Research in Science Engineering and Technology, Volume 9, Issue 2, Pages 51-63, 2022, DOI: 10.32628/IJSRSET22925

[8]. Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[9]. Vipin Rawat, Vishal Verma, Jai Pratap Dixit, Nikhat Akhtar, Neeta Rastogi, Susheel Kumar, "Face Mask Identification Using a Machine Learning Approach", Journal of Emerging Technologies and Innovative Research (JETIR), ISSN-2349-5162, Volume 9, Issue 8, Pages 842-847, 2022, DOI: 10.6084/m9.jetir.JETIR2208393

[10]. Liang-Chieh Chen, Jonathan T. Barron, George Papandreou, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[11]. G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Transactions on Medical Imaging, 37(7):1562–1573, 2018

[12]. Khadeeja Naqvi, Divyanshi Gautam, Ashish Kumar Srivastava, Prof. (Dr.) Syed Qamar Abbas, Dr. Nikhat Akhtar, "A Machine Learning-Based Rational Breast Cancer Diagnosis" , Journal of Emerging Technologies and Innovative Research (JETIR), ISSN-2349-5162, Volume 9, Issue 7, Pages 558-567, 2022, DOI: 10.6084/m9.jetir.JETIR2207677

[13]. Liu R., Sun Z., Wang A., Yang K., Wang Y., and Sun Q., "Lightweight Efficient Network for Defect Classification of Polarizers," Concurrency and Computation: Practice and Experience, vol. 32, no. 11, pp. e5663, 2020

[14]. Z. Lv, T. Liu, C. Shi, J. A. Benediktsson and H. Du, "Novel land cover change detection method based on k-Means clustering and adaptive majority voting using bitemporal remote sensing images", IEEE Access, vol. 7, pp. 34425-34437, 2019

[15]. Yusuf Perwej, Nikhat Akhtar, Firoj Parwej, "The Kingdom of Saudi Arabia Vehicle License Plate Recognition using Learning Vector Quantization Artificial Neural Network", International Journal of Computer Applications (IJCA), USA, ISSN 0975 – 8887, Volume 98, No.11, Pages 32 – 38, 2014, DOI: 10.5120/17230-7556

[16]. Nahid A. and Kong Y., "Involvement of Machine Learning for Breast Cancer Image Classification: A Survey," Computational and Mathematical Methods in Medicine, vol. 2017, pp.1-29, 2017

[17]. Dawar Husain, Dr. Yusuf Perwej, Satendra Kumar Vishwakarma, Prof. (Dr.) Shishir Rastogi, Vaishali Singh, Nikhat Akhtar, "Implementation and Statistical Analysis of De-noising Techniques for Standard Image", International Journal of Multidisciplinary Education Research (IJMER), ISSN:2277-7881, Volume 11, Issue10 (4), Pages 69-78, 2022, DOI: 10.IJMER/2022/11.10.72

[18].Salama, G.I., Abdelhalim, M.B., and Abd-elghany Zeid, M., Breast cancer diagnosis on three different datasets using multi-classifiers. Int. J. Comput. Inf. Technol. 1(Issue 01):2277–0764, 2012.

[19].Jafari-Marandi, R., Davarzani, S., Gharibdousti, M.S., and Smith, B.K., An optimum ANNbased breast cancer diagnosis: Bridging gaps between ANN learning and decision-making goals. Appl. Soft Comput. 72:108–120, 2018.

[20].Nikhat Akhtar, "Artificial Intelligence and Machine Learning in Human Resource Management for Sales research Perspective", IEEE International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Electronic ISBN:978-1-6654-7413-9, SCOPUS, ISBN:978-1-6654-7414-6, Chennai, India, 15-16, 2022, DOI: 10.1109/ICSES55317.2022.9914086

[21].Guo, H., and Nandi, A.K.: Breast cancer diagnosis using genetic programming generated feature. 2005 IEEE Workshop on Machine Learning for Signal Processing, Mystic, CT. , pp. 215–220, 2005.

[22].F. A. Mazarbhuiya, Dr. Yusuf Perwej, "The Mining Hourly Fuzzy Patterns from Temporal Datasets", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Volume 4, Issue 10, Pages 555-559, 2015, DOI: 10.17577/IJERTV4IS100576

[23].Zhao, J.Y., and Zhang, Z.L.: Fuzzy rough neural network and its application to feature selection. In: The Fourth International Workshop on Advanced Computational Intelligence, Wuhan. pp 684–687, 2011

[24].Yusuf Perwej, "An Optimal Approach to Edge Detection Using Fuzzy Rule and Sobel Method", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJAREEIE), ISSN (Print) : 2320 – 3765, ISSN (Online): 2278 – 8875, Volume 4, Issue 11, Pages 9161-9179, 2015, DOI: 10.15662/IJAREEIE.2015.0411054

[25].Xue, B., Zhang, M., and Browne, W.N.: New fitness functions in binary particle swarm optimisation for feature selection. In: WCCI 2012 IEEE World Congress on Computational Intelligence June, 10–15, 2012 - Brisbane, Australia, 2012.

[26].Nikhat Akhtar, "Perceptual Evolution for Software Project Cost Estimation using Ant Colony System", International Journal of Computer Applications (IJCA) USA, ISSN 0975 - 8887, Volume 81, No.14, Pages 23 – 30, 2013, DOI: 10.5120/14185-2385

[27].Zen K, Zhang CY. Circulating micro RNAs: a novel class of biomarkers to diagnose and monitor human cancers, Med Res Rev, vol. 32, pp. 326-348, 2012

[28].Litjens, G. et al. A survey on deep learning in medical image analysis. Med. Image Anal. Vol. 42, pp: 60–88, 2017. https://doi.org/10.1016/j.media.2017.07.005

[29].Shamy S. and Dheeba J., "A Research on Detection and Classification of Breast Cancer using k- means GMM and CNN Algorithms," International Journal of Engineering and Advanced Technology, vol. 8, no. e-6S, pp. 501- 505, 2019

[30].Asif Perwej, Yusuf Perwej, Nikhat Akhtar, "A FLANN and RBF with PSO Viewpoint to Identify a Model for Competent Forecasting Bombay Stock Exchange", COMPUSOFT, An International Journal of Advanced Computer Technology, ISSN:2320-0790, 4 (1), Volume-IV, Issue-I, PP 1454-1461, 2015, DOI : 10.6084/ijact.v4i1.60

[31].Mert,A., Kılıç,N.Z.,Bilgili,E.,&Akan, A, Breast cancer detection with reduced feature set. Computational and Mathematical Methods in Medicine, pp. 1–11. 2015

[32].Tsochatzidis L., Costaridou L., and Pratikakis I., "Deep Learning for Breast Cancer Diagnosis from Mammograms-A Comparative Study," Journal of Imaging, vol. 5, no. 3, pp. 37, 2019

[33].M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine

learning," 2018 Electric Electronics, Computer Science, Biomedical Engineering' Meeting (EBBT), pp. 1-4, 2018

[34].Shibata, H.; Takama, Y.; Ieee. Behavior Analysis of RBM for Estimating Latent Factor Vectors from Rating Matrix. In Proceedings of the 6th International Conference on Informatics, Electronics and Vision (ICIEV)/7th International Symposium in Computational Medical and Health Technology (ISCMHT), Univ Hyogo, Himeji Engn, Himeji, Japan, 1–3, 2017

[35].Fu, Y.; Jung, A.W.; Torne, R.V.; Gonzalez, S.; Vöhringer, H.; Shmatko, A.; Gerstung, M. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat. Cancer, 1, 800–810, 2020

[36].Nahid AA, Mehrabi MA, Kong Y. Histopathological breast cancer image classification by deep neural network techniques guided by local clustering. BioMed research international, 2018, PMID: 29707566 17

[37].Y, Perwej, "The Bidirectional Long-Short-Term Memory Neural Network based Word Retrieval for Arabic Documents", Transactions on Machine Learning and Artificial Intelligence (TMLAI), Society for Science and Education, United Kingdom (UK), ISSN 2054-7390, Volume 3, Issue 1, Pages 16 - 27, 2015, DOI: 10.14738/tmlai.31.863

[38].Y. Perwej, Firoj Parwej, "A Neuroplasticity (Brain Plasticity) Approach to Use in Artificial Neural Network", International Journal of Scientific & Engineering Research (IJSER), France, ISSN 2229 – 5518, Volume 3, Issue 6, Pages 1- 9, 2012, DOI: 10.13140/2.1.1693.2808

[39].Abdar, M.; Zomorodi-Moghadam, M.; Zhou, X.; Gururajan, R.; Tao, X.; Barua, P.D.; Gururajan, R. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognit. Lett., 132, 123–131, 2020

[40].O. Golubnitschaja, M. Debald, K. Yeghiazaryan, W. Kuhn, M. Pešta, V. Costigliola, et al., "Breast cancer epidemic in the early twenty-first century: evaluation of risk factors cumulative questionnaires and recommendations for preventive measures", Tumor Biology, vol. 37, no. 10, pp. 12941-12957, 2016

[41].Sun Chang, Yue Shihong," Clustering Characteristics of UCI Dataset", 39th Chinese Control Conference (CCC), IEEE, Accession Number: 19948911 , China,2020

[42].Dataset, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

[43].W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Analytical and Quantitative Cytology and Histology, Vol. 17 No. 2, pages 77-87, 1995

[44].anal Bani Issa, Omar Darwish, Doaa Habeeb Allah, Farah Shatnawi, Dirar Darweesh, Yahya M. Tashtoush, "Analysis of Jordanian University Students Problems Using Data Mining System", 2022 13th International Conference on Information and Communication Systems (ICICS), pp.220-225, 2022

[45].Asif Perwej, Prof. K. P. Yadav, Prof. Vishal Sood, Yusuf Perwej, "An Evolutionary Approach to Bombay Stock Exchange Prediction with Deep Learning Technique", IOSR Journal of Business and Management (IOSR-JBM), USA, Volume 20, Issue 12, Ver. V, Pages 63-79, 2018, DOI: 10.9790/487X-2012056379

[46].Y. Perwej, Firoj, Nikhat Akhtar, "An Intelligent Cardiac Ailment Prediction Using Efficient ROCK Algorithm and K- Means & C4.5 Algorithm" , European Journal of Engineering Research and Science, Belgium, Vol. 3, No. 12, Pages 126 – 134, 2018, DOI: 10.24018/ejers.2018.3.12.989

[47].M. Liu, X. Xu, Y. Tao and X. Wang, "An improved random forest method based on RELIEFF for medical diagnosis", 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International

Conference on Embedded and Ubiquitous Computing (EUC), vol. 1, pp. 44-49, July 2017

[48]. Hao Zhang, A. C. Berg, M. Maire and J. Malik, "SVM-KNN: Discriminative Nearest Neighbour Classification for Visual Category Recognition", 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)

[49]. I. Ruczinski, C. Kooperberg and M. Leblanc, "Logic regression", Journal of Computational and Graphical Statistics, vol. 12, no. 3, pp. 475-511, 2003

[50]. Fukushima K.,"Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position", Biol Cybern 36, pp. 193–202, 1980

[51]. A. L. Maas , A. Y. Hannun and A. Y. Ng , Proc. ICML , 2013

[52]. Y. Perwej, Asif Perwej, "Forecasting of Indian Rupee (INR) / US Dollar (USD) Currency Exchange Rate Using Artificial Neural Network", International Journal of Computer Science, Engineering and Applications (IJCSEA), Academy & Industry Research Collaboration Center (AIRCC), USA, Volume 2, No. 2, Pages 41- 52, 2012, DOI: 10.5121/ijcsea.2012.2204

[53]. R. Hu, B. Tian, S. Yin and S. Wei, "Efficient hardware architecture of softmax layer in deep neural network", Proc. IEEE 23rd International Conference on Digital Signal Processing (DSP), pp. 1-5, 2018

[54]. Y. Perwej, Firoj Parwej, Asif Perwej, "Copyright Protection of Digital Images Using Robust Watermarking Based on Joint DLT and DWT", International Journal of Scientific & Engineering Research (IJSER), France, ISSN 2229-5518, Volume 3, Issue 6, Pages 1- 9, 2012

[55]. Y. Perwej, Asif Perwej, Firoj Parwej, "An Adaptive Watermarking Technique for the copyright of digital images and Digital Image Protection", International journal of Multimedia & Its Applications (IJMA), Academy & Industry Research Collaboration Center (AIRCC), USA, Volume 4, No.2, Pages 21- 38, 2012, DOI: 10.5121/ijma.2012.4202

[56]. Y. Perwej, Ashish Chaturvedi, "Machine Recognition of Hand Written Characters using Neural Networks", International Journal of Computer Applications (IJCA), USA, ISSN 0975 – 8887, Volume 14, No. 2, Pages 6- 9, 2011, DOI: 10.5120/1819-2380

[57]. K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770-778, Jun. 2016

[58]. B. Sahiner et al., "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images", IEEE Trans. Med. Imaging, vol. 15, no. 5, pp. 598-610, 1996

[59]. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set, [online] Available:http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%2528Diagnostic%2529.2011

[60]. C.-I Chang, S.-S. Chiang, Q. Du, H. Ren and A. Ifarragaerri, "An ROC analysis for subpixel detection", Proc. IGARSS. Scanning Present Resolving Future. IEEE Int. Geosci. Remote Sens. Symp., pp. 2355-2357, Jul. 2001

[61]. X. Zhang, X. Li, Y. Feng and Z. Liu, "The use of ROC and AUC in the validation of objective image fusion metrics", Signal Processing, vol. 115, pp. 38-48, 2015

[62]. S. Wang, C.-I Chang and S. Yang, "3D ROC analysis for medical diagnosis evaluation", Proc. 27th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBS), pp. 7545-7548, Sep. 2005

[63]. J. Euzenat, "Semantic precision and recall for ontology alignment evaluation", Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), pp. 348-353, 2007

[64]. Wenduan Xu, Michael Auli and Stephen Clark, "Expected f-measure training for shift-reduce

parsing with recurrent neural networks", HLT-NAACL, pp. 210-220, 2016

Cite this Article
Dr. Nikhat Akhtar, Dr. Hemlata Pant, Apoorva Dwivedi, Vivek Jain, Dr. Yusuf Perwej, "A Breast Cancer Diagnosis Framework Based on Machine Learning", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 3, pp. 118-132, May-June 2023. Available at doi : https://doi.org/10.32628/IJSRSET2310375
Journal URL : https://ijsrset.com/IJSRSET2310375