IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Breast Cancer Detection Using Ensemble Techniques

[1]Amrita Trivedi, [2]Unnati Sheth, [3]Viresh Sawant, [4]Varun Nimje, [5]Dr Ankita Malhotra

[1,2,3,4]Student, [5]Professor

[1,2,3,4,5]Electronics and Telecommunication

[1,2,3,4,5]MCT's Rajiv Gandhi Institute of Technology, Mumbai, India

Abstract: This paper deals with the detection of Breast Cancer. The proposed method reduces the false positive and false negative errors using ensemble techniques that combine several base models to produce one optimal predictive model. The main purpose of this paper is to detect cancerous tissue. It is an extension of image processing and feature extraction. Here, machine learning classifiers have been used, which undergo image processing, image segmentation, and feature extraction with OTSU Thresholding, CLAHE, and GLCM. After extraction other independent algorithms using Logistic Regression, Decision Tree, Naïve Bayes, Support Vector Machine, K-nearest, neighbor, and further on, attempting voting mechanism, ensemble techniques (Bagging and Boosting). Here, three different datasets are used which are compared with accuracies for independent algorithms. Thus to draw an inference, XG Boost is comparatively better than others with an average accuracy of 97.5%

***Index Terms*- GLCM, OTSU, Image Processing, Feature Extraction, Voting**

I. INTRODUCTION

Oncologists have always struggled for detecting breast cancer at an early stage. Although, now with the help of Machine Learning it has become easier to assist the female patient with genetic mutations. Traditional methods used to analyse digital mammograms are quoted to have high false negative rate. With the help of researches already conducted it can be concluded that using algorithm one can get an accuracy level. By providing timely clinical management to patients, early detection of this disease and classification into cases might improve the prognosis and even save lives. A precise diagnosis of the classification of breast cancer into benign, malignant, and normal cases is a difficult task in cancer research. Because of the computer's ability to learn from previous samples in order to recognise and classify patterns, for cancer detection, machine learning and classification algorithms are commonly employed

Than Than Htay and Su Su Maung have presented Early-Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbour (kNN) on Mammography image. They have used the mini MIAS dataset to obtain the mammograms[1]. Muhammad Kashif, Kaleem Razzaq Malik, Sohail Jabbar and Junaid Chaudhary have developed a paper on Application of machine learning and image processing for detection of breast cancer. In this paper they have developed a model to predict breast cancer from mammogram images. They used a hybrid approach having mammogram processing and machine learning (ML) algorithms. The mammogram processing technique is employed to extract features from mammogram images. The images having abnormalities Malignant or Benign are classified by Machine Learning[2]. A. Qayyum presented a simple methodology for breast cancer detection in digital mammograms. In that system, they focused on the purpose of removal of pectoral. After removing the pectoral muscle, GLCM feature extraction was performed and finally, Support Vector Machine (SVM) was trained as a classifier to classify the breast region into normal and abnormal tissues[3]. Dinsha developed a paper on breast tumour segmentation and classification. In this method, pre-processing work is carried out by Contrast Limited Adaptive Histogram Equalization (CLAHE) technique. Using K-means and fuzzy c means, segmentation process would be carried out. Various features are extracted from the segmented images. Finally, classification has been made by using the SVM and Bayesian classifiers[4]. K.N. Nyein Hlaing presented a model for automatic classification of currency notes using K-Nearest Neighbour (k-NN) classifier with second order texture (GLCM) features. The input images utilized in that system are Myanmar paper money Note. That paper showed that the recognition rate of their method is higher than the other method[5]. Biswas designed a Computer Aided Diagnosis (CAD) approach for identifying normal and diseased breast tissues. Artifacts are removed using the ROI extraction technique, and noise is removed using the 2D median filter in this system. CLAHE is used to improve the image's look, and GLCM is used to extract features. Breast tissues are classified as normal or pathological using classifiers such as KNN, SVM, and ANN[6]. Youssef provided a mammography classification based on the extraction of global statistical features. In this study, a new technique is used to capture image samples. Noise removal is done in pre-processing for terahertz imaging, and feature extraction is done using a statistical method[7]. Hybrid Machine Learning

Algorithms for Predicting Academic Performance In this work, a hybrid approach of principal component analysis (pca) as conjunction with four machines learning (ml) Algorithms: random forest (rf), c5.0 of decision tree (dt), and Naïve bayes (nb) of bayes network and support vector machine (svm), to improve the performances of classification by solving the misclassification problem. Three datasets were used to confirm the robustness of the proposed models[8]. A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches presented Wrapper-based feature selection approach along with nature-inspired algorithms such as Particle Swarm Optimization, Genetic Search, and Greedy Stepwise has been used to identify the important features. On these selected features popular machine learning classifiers Support Vector Machine, J48 (C4.5 Decision Tree Algorithm), Multilayer-Perceptron (a feed-forward ANN) were used in the system[9]. Review The Breast Cancer Detection Technique Using Hybrid Machine Learning [10]. Breast Cancer Prediction Using a Hybrid Data Mining Model In this paper, a hybrid model employing three algorithms of Naive Bayes Network, RBF Network, and K-means clustering is presented to predict breast cancer type. In the proposed model, the voting approach is used to combine the results obtained from the above three algorithms[11]. Hybrid Classification Algorithm Approach for Breast Cancer Diagnosis This study suggests a hybrid classification algorithm which is based upon Genetic Algorithm (GA) and k Nearest neighbour algorithm (kNN). GA algorithm has been used for its primary purpose as an optimization technique for kNN by selecting best features as well as optimization of the k value, while the kNN is used for classification purpose[12].

In the proposed work to check the efficiency and accuracy level based on earlier researches Wisconsin datasets using Logistic Regression, Decision Tree, Naïve Bayes, Support Vector Machine, K-nearest neighbour. On procuring it, Digital Mammograms from MIAS and INBREAST Dataset were contracted. Digital Image processing was attempted in order to manipulate the data and interpret the medical data which can be used for early detection of anomalies. Image was then resorted and sharpened using Contrast Limited AHE was adapted as a histogram equalizer for Noise reduction. Further on, OTSU Thresholding was done in order to iterate all the possible threshold values and calculate the pixel levels which either come under foreground or background. For feature extraction GLCM was used for image texture features Contrast, Dissimilarity, Homogeneity, Energy, Correlation, Angular Second Moment.
This model will be used a predictor for diagnosis of breast cancer.

II. METHODOLOGY

The proposed method of detection for masses in digital mammogram is given below in Fig. 1 Fig. 2 and Fig. 3

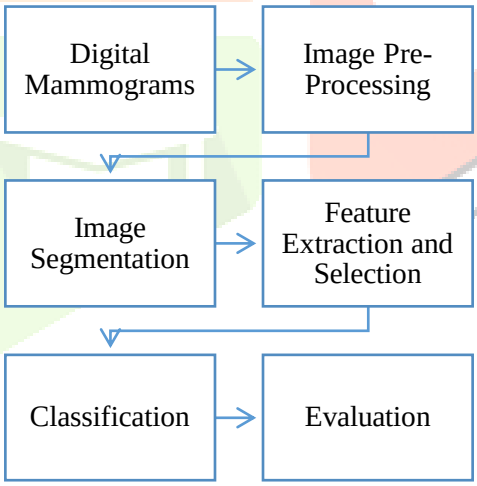

Fig. 1 Method for the proposed work

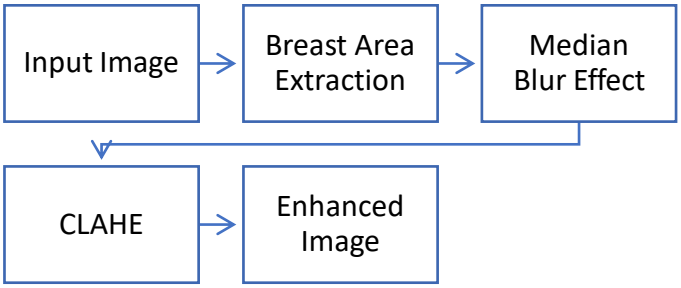

Fig. 2 Image Pre-processing

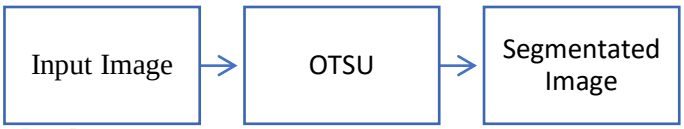

Fig. 3 Image Segmentation

[A] *Mammogram Database*

The dataset used is publicly available and was created by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA. There are 30 features which are used because the dataset to coach and test the algorithms. Further on, MIAS Dataset[13] was used for digital mammograms during this database, the first MIAS database were digitized at 50-micron pixel edge and has been reduced to 200-micron pixel edge and clipped or padded in order that every image is 1024 X 1024 pixels Fig 4. All images were held as 8-bit gray level scale images with 256 different gray levels (0-255) and physically in portable gray map (pgm) format. This study solely concerns the detection of masses in mammograms and, therefore, an entire of 100 mammograms comprising ill-defined, spiculated, circumscribed and normal case were considered. Ground truth of location and size of masses is out there inside the database. Further on, an equivalent process was adapted for Inbreast Dataset.

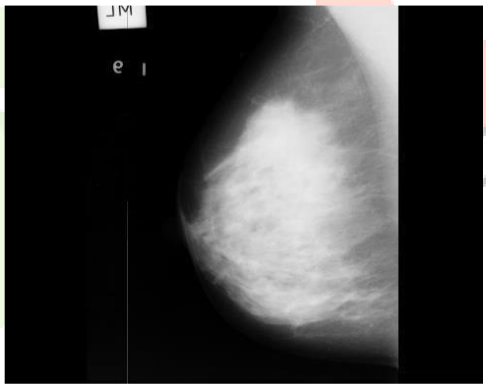

Fig. 4 Mammogram Sample

[B] *Image Processing*

Image Pre-processing is the first phase for this method. This pre-processing phase crucial in removing the noise from the image and improving the quality of the images. There were various image enhancement techniques such as mean filter, Adaptive median filter, Gaussian filter. All of the three techniques used for removal of noise and enhances the edges of the mammogram which further helps in the segmentation phase. In this paper Median Blur Technique was used following with CLAHE as shown below in Fig 5.

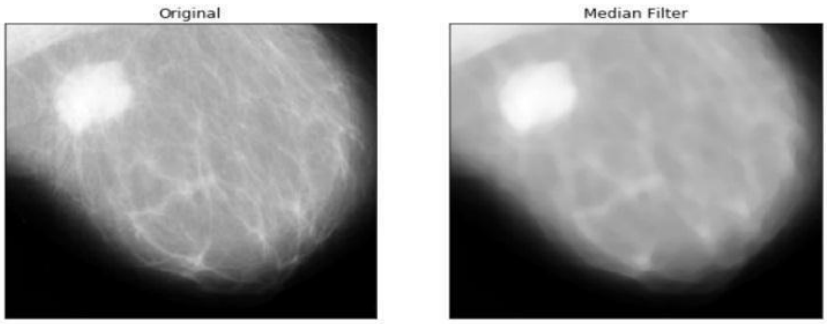

Fig. 5 Median Blur

Contrast Limited Adaptive Histogram Equalization CLAHE
It's a basic variant of adaptive histogram equalization in which the contrast amplification is limited thus reducing the problem of noise amplification. In CLAHE, the contrast amplification within the vicinity of a given pixel value is given by the slope of the transformation function.

[C] *Image Segmentation*
In here, OTSU Thresholding Image was performed segmentation is the division of a mammogram into several integral parts. The main aim in here is to simplify and make it easier to analyse. Also, to procure the location of the suspicious were to assist in diagnosing and further classification of the abnormalities into benign or malignant.

[D] *Feature Extraction*
Textures were always useful in sectoring the normal breast tissues and the masses. It is always capable to differentiate from under the curve (ROC). In here, the texture feature was extracted using GLCM (grey level co-occurrence matrices and they were designed at a distance d=1 and for θ given 0º, 45º, 90º, 135º Shown in Table 1. Four different directions were used as one might not be enough. The texture from here were contrast, energy, homogeneity correlation and ASM of grey level values and image as shown in Fig. 6 and Fig. 7

Contrast- The number of local variations in the image

$$Contrast = \sum_{i,j} |i-j|^2 \, p(i,j)$$

Energy- The sum of squared elements in GLCM

$$Energy = \sum_{i,j} p(i,j)^2$$

Homogeneity- Closeness of distribution of elements in GLCM-to-GLCM diagonal

$$Homegneity = \sum_{i,j} \frac{p(i,j)}{1+|i-j|}$$

Correlation- The way a pixel is with its neighbour pixel in that image

$$\text{Correlation} = \sum_{i,j} \frac{(i-ui)(j-uj)p(i,j)}{\sigma i \sigma j}$$

ASM- Uniformity

$$Entropy = -\sum_{i,j} (p\,(i,j).* \log(p(i,j)))$$

With performing this a CSV file is made

Table 1 Extracted Features (Data)

Features	Arrays			
	0º	45º	90º	135º
Contrast	1.025631	1.025631	1.025631	1.025631
Dissimilarity	0.174315	0.229133	0.135673	0.216145
Homogeneity	0.947509	0.931822	0.956809	0.93663
Energy	0.578407	0.576124	0.581388	0.576226
Correlation	0.967523	0.95605	0.97698	0.957649
ASM	0.334554	0.331918	0.338012	0.332036

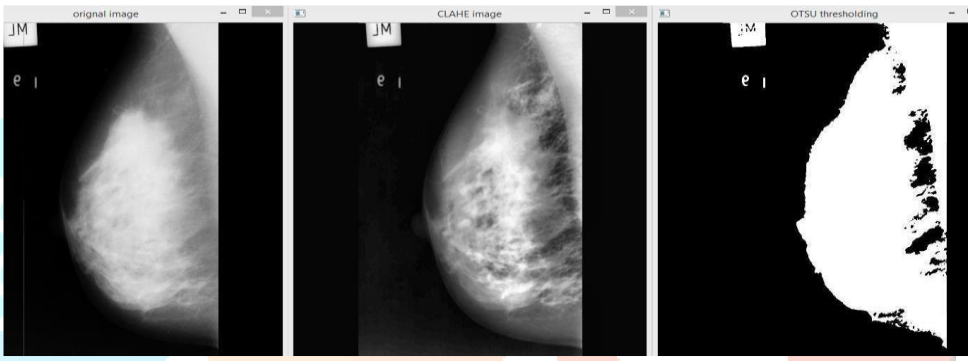

Fig. 7 Feature Extraction (Image)

Further on the 5 techniques namely Logistic Regression, Decision Tree, Naïve Bayes, SVM and KNN were used.

[E] *Voting Classifier*

Here, the model trained the data and predicted the output based on their calculations for majority voting. The main concept is to find an accurate model by creating different dedicated space which trains them and predicts the output class given in Fig. 8

Voting Classifier supports two types of voting

1. Hard Voting- In here, the predicted one is with the class with highest majority of votes given by all of the classifiers i.e., the output class (X, X, C) therefore here X is the output.
2. Soft Voting- In here, the output is predicted using the average of probability given to the class

By using Voting Classifiers, it enabled to train faster and easier to interpret and lesser errors furthermore improving the accuracy given that the correct subset was chosen.

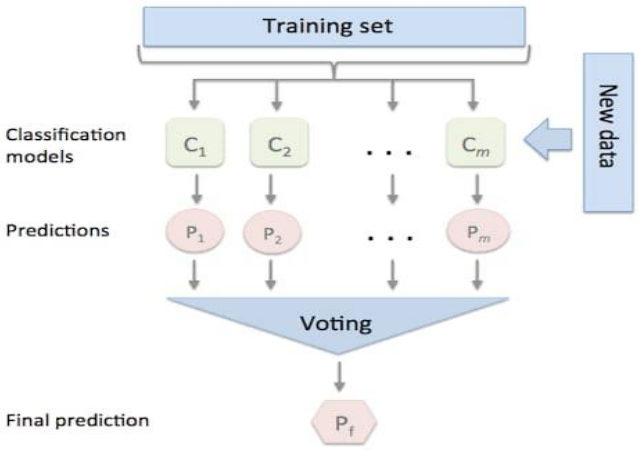

Fig. 8 Voting Classifier

[F] *Ensemble Techniques*

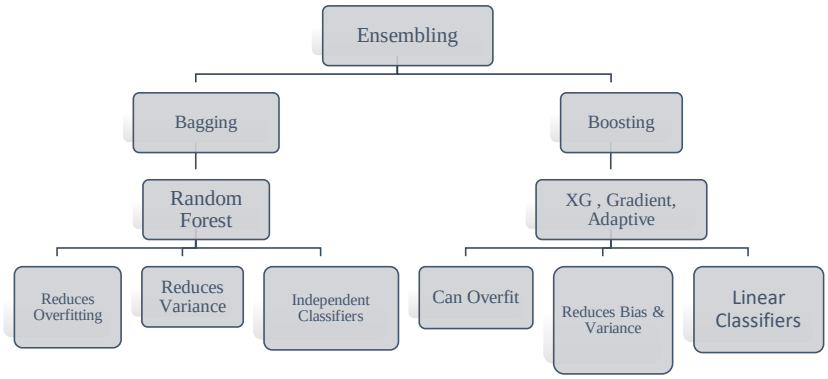

Fig. 9 Ensemble Classifiers

Random Forest

$$RFfi_i = \frac{\sum j \in \text{all trees normalfi}_{ij}}{T}$$

Where;
RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model
normfi sub(ij)= the normalized feature importance for i in tree j
T = total number of trees

XG Boost

$$F2(x) = \sigma(0 + 1^{*}h1(x) + 1^{*}h2(x)))$$

Where;
The resulting value of $F2(x)$ is considered as the prediction from XG Boost model.

Gradient Boosting

$$y_i^p = y_i^p + \alpha * \delta \sum (y_i - y_i^p)^2 / \delta y_i^p$$

Which becomes;

$$y_i^p = y_i^p - \alpha * 2 * \sum (y_i - y_i^p)$$

Where;
α is learning rate and $\sum(y_i - y_i^p)$ is sum of residuals

Adaptive Boosting

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

III. RESULTS AND DISCUSSION

Overall Accuracies of the Algorithms are asserted in the table below. These can be computed and compared. Most of them have given above 80% of precision. It suggests that selecting the correct classifier has a significant impact on the accuracy of mammogram diagnosis. In the proposed method of detecting breast cancer we have considered three datasets. The training and testing sizes taken are 70% and 30% respectively. The basic and ensemble classification algorithms were applied first on Wisconsin dataset having 31 readily extracted features. For the second and third dataset the images from MIAS and Inbreast dataset. These images were given as input to image processing and segmentation methods namely CLAHE and Otsu thresholding. Further the processed images go through texture based feature extraction technique called GLCM. The 24 features extracted from all the images are given as an input to five basic classification algorithms and five ensemble algorithms. The MIAS dataset and XGboost classifier shown the highest accuracy which is 98.26%. The accuracies of classifiers are shown in the table given below. Refer Table 2

Table 2 Comparison of Accuracies of different Algorithms

Data Sets	Training	Testing	Features	Accuracy (%)
Wisconsin Malignant- 212 Benign- 357	455	114	31	Logistic Regression: 98.63 Decision Tree: 100 Naïve Bayes: 94.72 Support Vector Machine: 98.43 K-Nearest Neighbour: 98.04 Voting Classifiers: 96.49 Random Forest Bagging: 95.70 XG Boost:100 Gradient Boosting: 92.98 Adaptive Boosting: 94.73
MIAS Malignant- 271 Benign- 51	257	65	24	Logistic regression: 84.08 Decision tree: 99.7 Naive bayes: 59.86 Support Vector Machine: 84.08 K-Nearest Neighbour: 86.15 Voting Classifier: 84.85 Random Forest Bagging: 84.43 XG Boost: 98.26 Gradient Boosting: 69.69 Adaptive Boosting: 72.72
Inbreast Malignant- 1000 Benign- 2000	2400	600	24	Logistic Regression: 63.63 Decision Tree: 97.83 Naïve Bayes: 63.33 Support Vector Machine: 76.24 K-Nearest Neighbour: 96.96 Voting Classifiers: 81.06 Random Forest Bagging: 92.63 XG Boost: 94.26 Gradient Boosting: 73.42 Adaptive Boosting: 72.75

IV. CONCLUSION

The ultimate goal of this paper is to propose the image preprocessing, feature extraction, classification and ensemble methods. The image preprocessing methods for noise removal and segmentation were Median filtering, Contrast Limited Adaptive Histogram Equalization (CLAHE) and OTSU thresholding. After segmentation the 24 features were extracted using Gray Level Coocurrence Matrix (GLCM) method. Basic classification algorithms namely Logistic regression, Decision tree, Naive bayes, Support Vector Machine, K-nearest neighbor gave 84.03, 99.7, 59.86, 84.08, 84.43, 86.15 accuracy(%) respectively. The proper way of selecting suitable classification ensemble technique can increase the accuracy of detecting the cancer. For that five algorithms were built using the ensemble approach. Accuracy achieved using hard and soft voting classifier is 84.43 for bagging and boosting methods such as Random forest, XGboost, Gradient boosting and Adaptive boosting it is 69.69 and 72.72 accuracy (%) respectively. Thus from the data it can be concluded that XGboost is better in comparison to others for classification as it gives more accurate results than other algorithms. Refer Table 2

V. REFERANCES

[1] T. T. Htay and S. S. Maung, "Early-Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbour (k-NN) on Mammography image," 2018 18th International Symposium on Communications and Information Technologies (ISCIT), 2018, pp. 171-175, doi: 10.1109/ISCIT.2018.8587920.

[2] Kashif, Muhammad & Yaakob, Shahrul. (2020). Deep Learning Applied to Arabic and Latin Scripts: A Review. International Journal of Scientific & Technology Research. 8. 1510-1521.

[3] A. Qayyum and A.Basit, "Automatic breast segmentation and cancer detection via SVM in mammograms'', IEEE,2016.

[4] D. Dinsha, "Breast tumour segmentation and classification using SVM and Bayesian from thermogram images, "Unique Journal of Engineering and Advanced Sciences, vol. 2, 2014, pp: 147-151. J.

[5] K.N. Nyein Hlaing and Anil Kumar K.G, "Myanmar paper currency recognition using GLCM and k-NN'', Second Asian Conference on Defence Technology (ACDT), IEEE,2016

[6] R. Biswas, A. Nath and S. Roy, "Mammogram classification using Gray-Level Cooccurrence Matrix for diagnosis of breast cancer' 'International Conference on Micro-Electronics and Telecommunications Engineering,2016, pp161-166.

[7] Youssef Ben. Y and EI hassane and Jamal. Z and Abdelaziz. B, "Statistical features and classification of normal and abnormal mammograms'', IEEE,2014.

[8] Phauk, Sokkhey & Okazaki, Takeo. (2020). Hybrid Machine Learning Algorithms for Predicting Academic Performance. International Journal of Advanced Computer Science and Applications.11.10.14569/IJACSA.2020.0110104.

[9] Presentation of Novel Hybrid Algorithm for Detection and Classification of Breast Cancer Using Growth Region Method and Probabilistic Neural Network Zeynab Nasr Isfahani,1 Iman Jannat-Dastjerdi,2 Fatemeh Eskandari,1 Saeid Jafarzadeh Ghoushchi ,3 and Yaghoub Pourasad.

[10] Nidhi Mongoriya, Vinod Patel, "Review the Breast Cancer Detection Technique Using Hybrid Machine Learning" SSRG International Journal of Computer Science and Engineering 8.6 (2021): 5-8

[11] Bahmani, Elham & Jamshidi, Mojtaba & Shaltooki, Abdusalam. (2019). Breast Cancer Prediction Using a Hybrid Data Mining Model. JOIV: International Journal on Informatics Visualization. 3. 10.30630/joiv.3.4.240.

[12] Abed, Baraa & Shaker, Khalid & Jalab, Hamid & Shaker, Hothefa & Mansoor, Ali & Alwan, Ahmad & Salman, Ihsan. (2016). A Hybrid Classification Algorithm Approach for Breast Cancer Diagnosis. 10.1109/IEACON.2016.8067390
Pl,

[13] http://peipa.essex.ac.uk/ipa/pix/mias/