A Deep Learning Technique for Classification of Breast Cancer Disease

Yelepi Usha Rani, Lakshmi Sowmya Kotturi, G.Sudhakar

Abstract: In recent years researchers are intensely using machine learning and employing AI techniques in the medical field particularly in the domain of cancer. Breast cancer is one such example and many studies have proposed CAD systems and algorithms to efficiently detect cancer cells and tumors. Breast cancer is one of the dreadful cancers accounting for a large portion of deaths caused due to cancer worldwide mostly affecting women, needs early detection for proper diagnosis, and subsequent decrease in death rate. Thus, for efficient classification, we implemented different ML techniques on Wisconsin dataset [1] namely SVM, KNN, Decision Tree, Random Forest, Naive Bayes using accuracy as a performance metric, and as per observance, SVM has shown better results when compared to other algorithms. Also, we worked on Breast Histopathology Images [2] scanned at 40x which had images of IDC which is one of the most common types of breast cancers. And to work with the image dataset along with EDA we used high-end techniques like a mobile net where smote a resampling was used to handle imbalanced class distribution, CNN, SVC, InceptionResNetV2 where frameworks like Tensor Flow, Keras were loaded for supporting the environment and smoothly implement the algorithms.

Keywords: Breast cancer, IDC (Invasive ductal cancer), Histopathology Images, Tensor Flow, Keras, CNN, SVC, InceptionResNetV2, SVM, KNN, Decision Tree, Random Forest, Naive Bayes, Wisconsin dataset.

I. INTRODUCTION

Breast cancer is one of the prevalent diseases that most women are affected by worldwide and records the second-highest death rate caused by all types of cancers. The breast is made up of different tissues varying from fatty tissues to dense tissues. Whereas the lobules in the breast are tubes that consist of milk glands that are responsible to produce milk and ducts that carry the generated milk to the outlet(nipple) [3]. The lobules are also addressed as glandular tissues and the tissues that provide fiber connectivity throughout the tissues are also known as stromal tissues. Apart from these, there are also lymphatic tissues that help to remove toxins and waste materials from the breast.

Dr.Yelepi Usha Rani*, VNR Vignana Jyothi Institute of Engineering and Technology,India Department of Information Technology usharani_y@vnrvjiet.in

Lakshmi Sowmya Kotturi, VNR Vignana Jyothi Institute of Engineering and Technology,India, Department of Information Technology kotturisowmya61@gmail.com

Dr.G.Sudhakar, Assistant Professor(c),School of Information Technology,JNTUH ,India Departement of Computer Science Engineering, Sudhakar4321@jntuh.ac.in

Breast cancer develops when the healthy cells start growing uncontrollably and forming lumps due to the abnormal expansion of cells [5]. And this mainly occurs because of mutations in DNA/RNA and changes in the genes. On the other hand, they might occur due to several factors like Obesity, aging, hormonal imbalances (mostly postmenstrual) [4]. And statistics have shown that only 5% of the total ratio is due to inheritance and also occurring of BC in women has 100 times more chances than in men. It might also occur due to conditions like huge exposure to radiation, chemicals (chemists, doctors, engineers, etc.), undernourishment [6], persons with a genetic tendency to mutations [7], and others just with age factor (above 80). Few symptoms that usually show up like terrible pain in the breasts, inward pulling of the nipple, reddish skin, lumps like hard substances which are the stagnated fluids, and toxic materials which can be indications by the body for breast cancer. They can be easily identified by periodically check of their own body and any symptoms of pain can be a strong indication and upon any signs consulting a doctor is always recommended. During advanced stages of breast cancer symptoms like pain in the underarm and lymph bone, inhaling problems, and less consumption of food along with abnormal weight, pain in nerves, and weakness [9]. Unfortunately, diagnosis of any cancer becomes more critical, and painful treatments are to be used and there is no guarantee for life as the death rate is high as they identify at later stages. Thus, doing BSE (breast self-examination) and keeping track of breast mass, texture and keeping a lookout for any abnormal symptoms can help identify cancer in early stages and any negligence can simply complicate the case [8]. Having a healthy lifestyle and getting less exposure to radiation (cell phones, microwaves), and using quality products that are free of chemicals like acetone, arsenic (usually found in makeup products and nail polish removers) can also be found in rice in negligible amounts, benzene, etc. can also reduce the chances of developing cancer. Through advancements in the medical industry have evolved painless treatments like chemotherapy, Oncogenes Inactivation, Quadrantectomy, Lumpectomy, Radiation Therapy, chemotherapy, Nanotechnology, etc but all of these methods are efficient only when cancer is detected in early stages. Thus, there is an immense need to make the process of identifying cancer cells by differentiating them from normal healthy cells. But the system that is used for detection should be reliable and up to standards that the medical field can adapt and use in real-time for breast detection. So, to develop this CAD system Machine learning algorithms come in handy,

and even complex solutions like AI algorithms with high-end frameworks can also be employed to meet the requirements. By developing such a highly reliant and effective system we can detect cancer in the early stages making diagnosis an easy process and also not only detecting but also identifying useful patterns and making analysis by digging in the important information related to patients would help the doctors understand the scenario properly and provide at most relevant treatment seamlessly and save lives. This way it helps doctors to make the right decisions based upon the case and situation of the patient along with early detection which in turn helps reduce the overall death rate. Yet some resources like abundance data (medical data is sensitive and hard to collect from various sources and integrate), considering exceptional cases like macro and microcalcifications formed in the breast, symptoms, and complexities involved a change from person to person so algorithms must be well trained with enough data to handle each data. Apart from all these having proper infrastructure and avoiding imbalances in data, overfitting also developing computational efficient systems are the open challenges that researchers are trying to overcome today.

II. PREVIOUS WORK

In recent times many types of research have been conducted in the domain of breast cancer to make the process of identifying BC in early stages and to assist doctors to make a righteous decision by projecting useful information and links between previous records of patients making the process efficient so that doctors provide exact treatment.Ebru Aydındag Bayrak[10] worked on the Wisconsin dataset and applied classifiers SVM and ANN by comparing them at different stages using accuracy, precision, ROC curve, recall for performance evaluation. After pre-processing they split that data in the ratio 66% for training and 33% for testing. 10-k cross fold was used for cross-validation. Along with which the WEKA tool was used to validate the results. With SVM 97.13% accuracy was obtained. Lal Hussain [11] has used mammogram images for classification into cancerous and non-cancerous using SVM kernel, Navies Bayes method, and decision tree classification methods. An intensive pre-processing was done and features were extracted by doing EDA. To avoid artifacts raised due to change in angle or lighting while collecting mammogram images they used SIFT available in MATLAB. They used metrics like sensitivity, precision, recall, confusion matrix, specificity, and ROC curve and its area for comparison between algorithms applied. Md. Imran Hossain Showrov [12] has conducted his experiment considering WBCD as input and used algorithms like Linear SVM, RBF SVM, Sigmoid SVM, polynomial SVM, Feedforward NN, RBF NN. The performance metric chosen was accuracy and SVM has recorded 96.72% whereas RBF NN has resulted in 95.88%. Vikas Chaurasia[13] have applied Sequential Minimal Optimization (SMO), IBK, BF Tree on WBCD and obtained Accuracy of 96.2%,95.90%,95.46% respectively. They conducted this with the motive to compare the accomplishment of different classification techniques using WEKA software.

R.Chtihrakkannan [14] has done research using X-ray images as input and for pre-processing the images they used Gaussian filter and edge detection to identify the breast tissue and layout of the tumor. To obtain the first-order features they used wavelet transformation and then obtained Grayscale image. Further, they also obtained second-order features using GLCM. Then upon the obtained features, they applied a deep neural network which is a supervised classification algorithm. They also used other algorithms like ANN, SVM, KNN but in comparison with DNN, they recorded average results. Whereas DNN obtained an accuracy ranging from 87%-96%.

III. PROPOSED WORK

a)   Working with Wisconsin breast cancer dataset

As part of this study, we used the WBCD [1] that is available publicly on the UCI repository. WBCD was digitally computed from FNA (fine needle aspirate) where various features about the cell nuclei were extracted and tabulated. Looking at the description of the dataset it has 569 instances with 33 features among which 31 features were floating with 1 integer(id) value and an object value.

An initial step was to analyze data and pre-process so we quickly check for missing values and see that the last feature which is Unnamed is a complete NA, along with the id feature which has no significance in the classification of BC so we drop these features. Furthermore, as one feature (class variable) is in text format representing M-malignant-benign are transformed into numerical using Hot encoding. We use the classical method ". describe "to get insights about the data and its distribution. Moving on we visualize the data corresponding to each feature lets plot a graph showing the changes and distribution is 'Malign' or 'Benign'. Also, Display the corresponding graphs so that they can be used for individual comparison based on features. As we plot a box plot, we observe that features such as texture_se, symmetry_se are not linearly separable as their median is almost the same as these features are not required and would not add many contributions while classifying. Then we apply normalization to the data using a standard scaler. Then we use a swarm plot to visualize data we see some features are highly correlated to validate we then use a correlation heat map. And as predicted some features are highly correlated were as some features are very less correlated to the class variables thus by fixing our threshold as correlation value<0.07   features fractal_dimension_mean,texture_se,smoothness_se/ symmetry_se /,fractal_dimension_se are dropped. As the next step, we apply the feature extraction technique ExtraTreesClassifier and obtain the most important 15 features as seen in figure (1) we drop the remaining features and prepare our final copy of data to further build ML models, test and compare various algorithms.
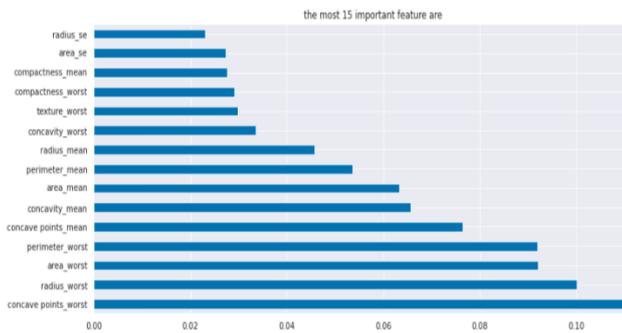
Figure-(i) The 15 important features after applying ExtraTreesClassifier

But before building training models one last step in pre-processing was to remove outliers. To identify outliers, we use a box plot of individual features and see that the feature 'concativity_mean' has outliers thus we remove values in the range (0.3- 0.5). Similarly, the feature 'perimeter worst' has outliers so values x>165 or x<165 are removed.

MACHINE LEARNING APPLICATION

As now our data is clean and ready to develop models, we first split data into train and test in the ratio 70:30 respectively. We applied different classifiers using accuracy as the performance evaluation metric whereas Logistic regression has resulted in 97.43% accuracy.

KNN was applied by fixing the number of neighbors as 3 and metric as Minkowski and obtained 94.8% accuracy. We implemented SVM with RBF kernel and obtained an accuracy of 96.15%. We further applied decision trees and generated rules that can be converted to associate rules and find useful and interpretable inferences. It resulted in 91.6% accuracy but we observed that the decision tree gives higher accuracy if applied before removing the redundant features due to overfitting. Then we employed random forest classification using entropy as the criterion and obtained 96.15% accuracy. Moving further now we employed advanced algorithms like ANN with 4 hidden layers along with adam optimizer by using Keras as back support for implementation and obtained an accuracy of 95%. For ANN we used 100 epochs and batch_size was set to 16. Also, when XGBClassifier was used we obtained an accuracy of 96.7%. On a similar note, we also experimented with CatBoostClassifier and obtained 96.79% accuracy. Below is the table that shows the results obtained for each classifier along with a comparison chart.

Model	Score
Logistic Regression	0.974359
xgboost	0.967949
catboost	0.967949
Support Vector Machines	0.961538
Random Forest	0.961538
ANN	0.955128
KNN	0.948718
Decision Tree	0.916667

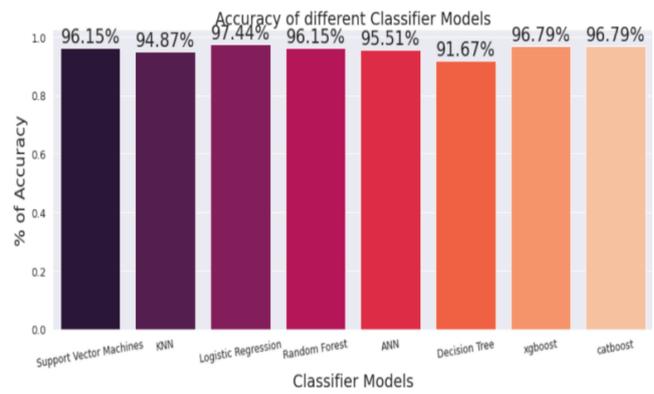**Figure-(2) Accuracies of each classifier**

Figure-(3) comparison chart

b) Working histopathology images to classify IDC (Invasive ductal cancer).

In types of breast cancer, the most commonly occurring type is invasive ductal cancer and on the fortunate side, it has an 80% of diagnosis rate. But early detection plays an important role to improvise the survival rate and also recommend an accurate treatment plan. In recent years with the AI revolution, many models were developed to predict and classify almost all types of cancers. But for efficient model quality data is the primary factor. So here we used one of the reliable datasets which were carefully collected following the procedures with all precautions. The histopathology images for IDC [2] which were scanned at 40X resolution each of size 50 x50 total of 277,524 patches were scanned among which 198,738 IDC negative and 78,786 IDC positive. The dataset is large and is formatted and stored in subfolders in such a way that) each patch's file name is of the format: uxXyYclassC.png→ 10396idx5x1351y1101class0.png. Here if the patient ID is keenly observed (10396idx5) X represents the x-coordinate from where this patch was cropped from, Y represents the y-coordinate from where this patch was cropped from, and C shows the class where 0 is non-IDC and 1 is IDC. The dataset has a total of 280 subfolders i.e., a total of 280 patients were examined while collecting this data and indeed it had two classes [0,1] for each patient.

IMAGE PREPROCESSING:

By performing EDA on the image dataset, the results are as follows and the key observations were:
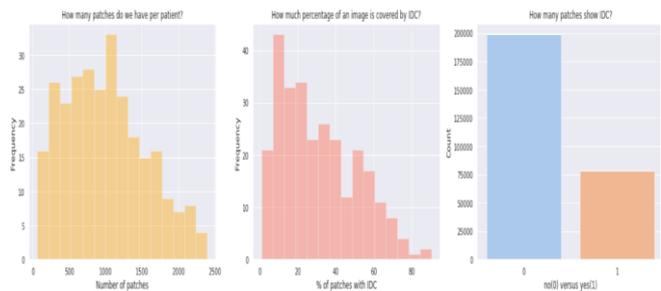

Figure-(3) Exploratory data analysis results

By looking at figure (3) there comes scope for questions like whether all images were scanned at the same resolution or not as there is a lot of difference in no. of patches for each patient.

Some patients have more than 80 % of patches that show IDC. Consequently, the tissue is full of cancer, or only a part of the breast was covered by the tissue slice that is focused on the IDC cancer. Also, we that see the distribution of class variable and notice how the ratio of IDC patches to non-IDC patches is highly uneven. Here we approximately see the split of IDC and non-IDC is 28/70 which may cause biasing towards non-IDC samples relative to IDC samples. Then we try to visually find the differences between cancerous tissue cells, and healthy tissue cells.
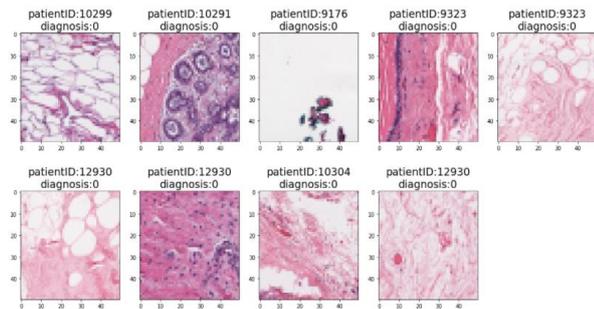

Figure-5 healthy vs cancerous tissue images

From figure-5 we can say that sometimes we can find artifacts or incomplete patches, some images are also less than 50 x 50. Patches with cancer look more violet and crowded than healthy ones. It raises the question that is this typical for cancer or is it more typical for ductal cells and tissue. Exceptionally some of the healthy patches are very extensively violet-coloured too.
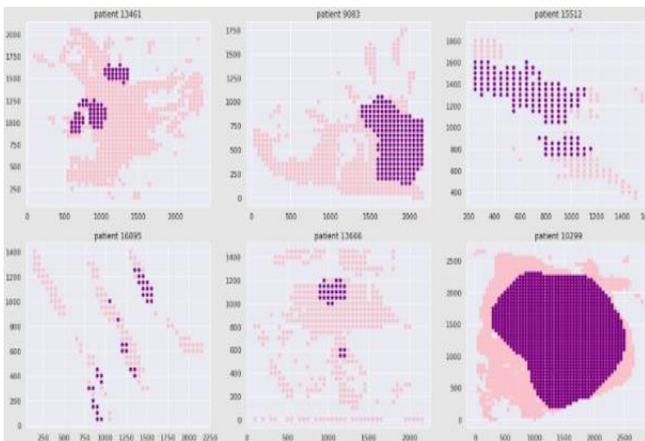

Figure-(6) Binary target visualization for individual tissues

Earlier as discussed, we can extract the x& y coordinates of the cropped patches which can be used to reconstruct the image as shown in figure (6). This way we can explore how the diseased tissue looks when compared to the healthy tissue. We can also identify the most common places that cancer tends to develop. we can plot a heatmap of the most common areas where cancer appears. If the position where the image was cropped has significance then probably, we can use it as an input feature for our model. By looking at figure-(6) we see that sometimes some parts of the tissue are missing which might be lost during the preparation of data or lost. And cancerous tissues tend to appear in dense clusters rather than scattered all over. Further, we move a

step forward and try to visualize the original image by repatching to understand the image from a broad perspective and in the process, we use masking. As observed, there is a high imbalance in the distribution of class variables to avoid overfitting, we can use data augmentation as part of which we can do random rotation, random crop, random flip (horizontal and vertical both), random lighting, random zoom, Gaussian blur. Also, by using test time augmentation (TTA) can be used to improve the result by computing and multiple times for each image. And now we have a better understanding of the dataset it's time to develop a different model. We have dealt with this dataset in various ways using different algorithms each time so the below section walks you through various models built.

1)   Breast cancer classification using SVC

In this model, we first create a data frame of 5 features namely patient id, x-coordinate, y-coordinate, image path, class variable(diagnosis). We then manually shuffle data not disturbing the individual order of patient id. We then divide the data into a 70:30 ratio for train and test. Also, we keep a part of the data from the unprocessed dataset as out of sample data for further validation of the trained model. Here shuffling of data is done to avoid the predominance of non-IDC patches over IDC. But then comes the question about the computational expense for working with around 280,000 images. Thus, we apply any of the methods like Principal Component Analysis (PCA), Single Value Decomposition (SVD), & Latent Dirichlet Allocation (LDA) to reduce the number of dimensions. Or we can do it manually by writing a function that takes an RGB image and converts it to Grayscale by using the weighted sum of R, G, B and returns a single dimension.

We applied PCA which is a dimensionality reduction technique that works by taking covariance between the different variables that we have into account. Here PCA converts RGB to Gray as part of reduction but also takes care that no valuable information is lost and the comparison can be seen in figure-(7). It also reduces no. of dimensions per image from (1,7500) to (1,2500) thus reducing the memory consumed for any further operations. The obtained variance ratio before applying PCA is 93.3% which increases to 99.9% post-PCA also before applying PCA we reshaped our images from (50,50,3) to (2500,3). Now we train the data for which we first set the model parameters where the kernel is set to RBF and the model used is SVC from sklearn.SVM and obtained a score of 0.581.
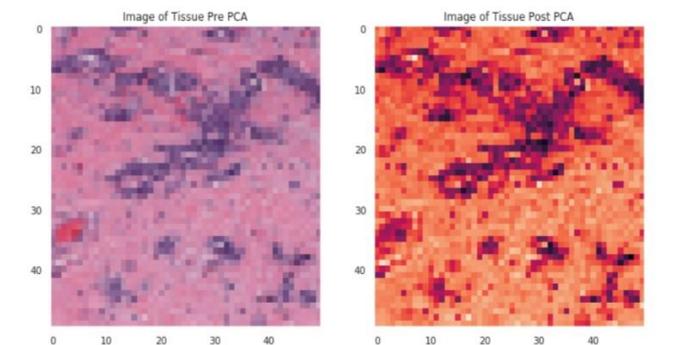

Figure-7   comparing   images pre and post PCA

2) IDC classification using smote and mobile net

This method was mainly to avoid the imbalance in a class variable by doing resampling using Smote technique. In this method, we create a data frame with 4 features namely patient id, image path, diagnosis, pixels (matrix of RGB values in the whole image). By using counterplot, we see the distribution of class variable and there is a huge difference which can bias our training model. So, we first divide our dataset into an 80:20 ratio for train and test. Then we get the pixels to feature ready by reshaping and apply smote which results in the size of the oversampled dataset as 11906. Then we further build the model using MobileNetV2 from Keras. Then we test our data and do performance evaluation using classification_report, confusion_matrix which resulted in 84.8% accuracy as shown in figure-(8).
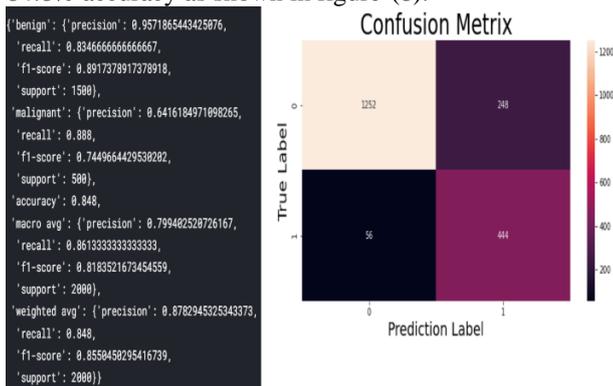

Figure-8 classification report and corresponding confusion matrix of MobileNetV2 model.

3) Using InceptionResNetV2 for classification of histopathology images

After getting a quick sneak peek about file structure and information about the image dataset we then create a data frame with columns diagnosis, file paths. As class distribution is highly unbalanced, we limit each class to around 25000 samples. We then divide the data for training, testing, and validating in the ratio 9:0.5:0.5. We set the test batch size to 50 and run for 50 test steps. We apply batch normalization for each layer and apply our model InceptionResNetV2 from Keras. We also create a subclass for callbacks to control the learning rate of the training model and visualize the training results for each epoch. Then we plot the data along with evaluation on the test set and the results are as shown in figure-(9). Also, we can visualize the results using the classification report and confusion matrix in figure-(10) where we made predictions using the portion of data setback for validating. Using this model resulted in an inaccuracy of 86%.
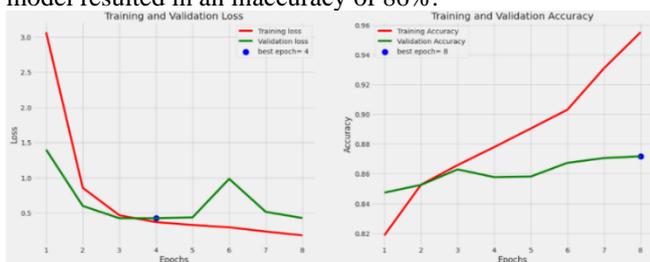
Figure-9 Result of epoch with the lowest validation loss
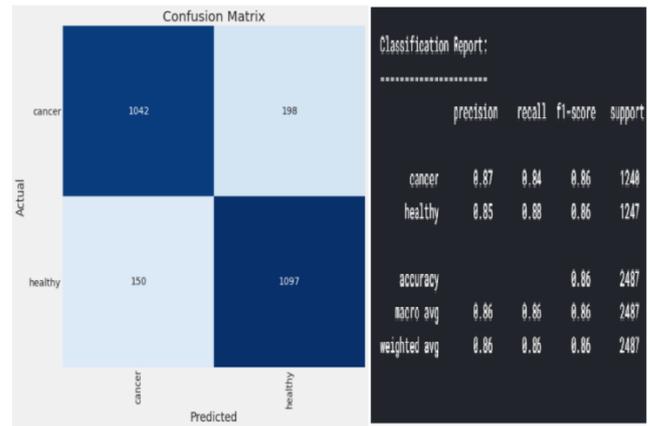

Figure-10 confusion matrix and classification report

4) CNN for classification for Invasive Ductal Carcinoma

CNN is a specialized neural network that is designed for image classification and works well with 2d matrices. In this model, we used transfer learning which simply means that the developed model is restarted for the second task and it is one of the popular approaches in deep learning. We split data in the ratio 0.8:0.2 and use a pre-trained ResNet18 Convolutional Neural Net model. As mentioned, we use and use transfer learning to learn weights of only the last layer of the network because working with advanced techniques like neural networks need a minimum of 3 lakh images for reasonable results, and for higher efficiency, we almost need 1 million images. Then we apply CNN and also use hyperparameter tuning to select the optimized parameters for the algorithm also the learning rate can be seen in figure-11. The accuracy obtained is 89% and we can visualize the test results using the confusion matrix in figure-12.
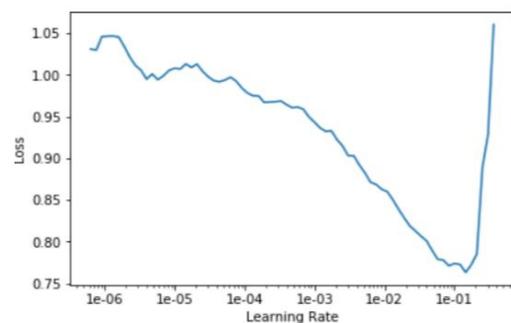
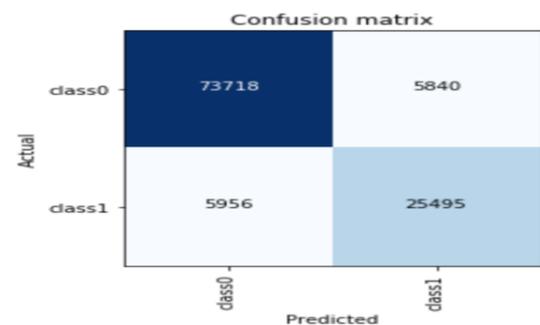

The figure-11 learning rate of the CNN model

Figure-12 confusion matrix of CNN model

IV. CONCLUSION

Breast cancer is one of the most dreadful cancers that occur commonly in the woman and according to statistics in India 1 in 28 women are likely to develop BC and death rate is 2.6% when it is detected early. Thus, early detection plays a key role in improvising the survival rate thus we have implemented various ML techniques, Neural networks, Keras models to investigate and explore the facts and to train and develop efficient models that can accurately classify the input as cancerous or non-cancerous. We also try to add new features to the dataset based on our observations. Intensive pre-processing and EDA are done to get good command and understanding of the data. Also, we worked with both numerical and image datasets and according to our observations generally, cancer cells form clusters and can be easily visualized in the mages. Also collecting data is sensitive in this case and generally finding the balanced number of records for both BC and Non-BC is a near impossible thing thus resampling can help avoid overfitting. And according to the results obtained from SVM, logistic regression algorithms work better with Wisconsin dataset with low error rates. When dealing with image datasets using ML techniques like SVC cannot give us the expected results. Thus, we divide each image into patches to increase the size of the dataset and train advanced models like CNN, InceptionResNetV2 which surely can improvise the results.

REFERENCES

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: the University of California, School of Information and Computer Science, Breast Cancer Wisconsin (Diagnostic) Data Set.
2. Breast Histopathology Images 198,738 IDC(-) image patches;78,786IDC(+)imagepatches https://www.kaggle.com/paultimothymooney/breast-histopathology-images.
3. Breast cancer process India, Breast cancer cost India, Breast cancer, Delhi India. Breast cancer information and resources. 2010. Apr 13, [14 April 2010]. http://www.digforthecure.org/breast-cancer-process-india-breast-cancer-cost-india-breast-cancer-delhi-india. Html
4. Vaka, A. R., Soni, B., & K., S. R. (2020). Breast cancer detection by leveraging Machine Learning. ICT Express. doi:10.1016/j.icte.2020.04.009
5. Mieszkowski M. R. Cancer – A biophysicist's point of view. In: Digital Recordings. 2006. Sep 04, [15 Mar 2010]. http://www.digital-recordings.com/publ/cancer.html.
6. Diet and Physical Activity: What's the Cancer Connection? In: Prevention & Early Detection. 2009. Oct09,[17Mar.2010]. http://www.cancer.org/docroot/PED/content/PED_3_1x_Link_Between_Lifestyle_and_CancerMarch03.asp.
7. Margot New SEER Report Documents High Risk of Second Cancers in Cancer Survivors. Oncology Times. 2007;29(5):8. [Google Scholar]
8. Types of breast cancer. Abviva. 2009. [18 Mar. 2010]. http://abviva.com/1.html.
9. Cancer-Its various types along with causes, symptoms, treatments, and stages, in the cancer info guide. 2009. [15 Mar. 2010]. http://www.cancer-info-guide.com/
10. Bayrak, E. A., Kirci, P., & Ensari, T. (2019). Comparison of Machine Learning Methods for Breast Cancer Diagnosis. 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT). doi:10.1109/ebbt.2019.8741990
11. Hussain, L., Aziz, W., Saeed, S., Rathore, S., & Rafique, M. (2018). Automated Breast Cancer Detection Using Machine Learning Techniques by Extracting Different Feature Extracting Strategies. 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). doi:10.1109/trust com/bigdatase.2018.00057
12. Showrov, M. I. H., Islam, M. T., Hossain, M. D., & Ahmed, M. S. (2019). Performance Comparison of Three Classifiers for the Classification of Breast Cancer Dataset. 2019 4th International Conference on Electrical Information and Communication Technology (EICT). doi:10.1109/eict48899.2019.9068816
13. Chaurasia, Vikas and Pal, Saurabh, A Novel Approach for Breast Cancer Detection Using Data Mining Techniques (June 29, 2017). International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 1, January 2014, Available at SSRN: https://ssrn.com/abstract=2994932
14. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019. R.Chtihrakkannan, Breast Cancer Detection using Machine Learning

AUTHOR PROFILE

Dr.G. Sudhakar, Assistant Professor(c), School Of Information Technology,JNTUH Department of Computer Science and Engineering,Sudhakar4321@jntuh.ac.in
Dr. G. Sudhakar working as Assistant Professor(c) in School Of Information Technology, Jawaharlal Nehru Technological University Hyderabad(JNTUH) since 2011 in the department of Computer Science Engineering. Has 14 years of teaching and research experience in computer science and engineering in various engineering colleges. Pursued B.Tech (Computer Science and Information Technology) from Jaya Prakash Narayan College of Engineering affiliated to JNTUH in 2005, M.Tech (Software Engineering) from School Of Information Technology, JNTUH in 2009, PhD (cloud data Security) from Jawaharlal Nehru Technological University Hyderabad in 2019. Also pursued Post graduate Diploma in telecommunications (PGDTC), Post graduate Diploma in Cyber laws and Intellectual property rights, Post graduate Diploma in Criminal Justice and Forensic Science, Post graduate Diploma in Human Rights from University of Hyderabad in the years 2013, 2015, 2016 and 2017 respectively.
Areas of research interest are cloud computing, cloud data Security, software engineering, artificial intelligence, machine learning and deep learning.

Dr.Yelepi UshaRani, Assistant Professor, VNR Vignana Jyothi Institute of Engineering and Technology Department of Information Technology, ushayelepi@gmail.com Dr.Yelepi UshaRani, Assistant Professor,VNR Vignana Jyothi Institute of Engineering and Technology since 2008 in the department of Information Technology. have 14 years of teaching and research experience in various engineering colleges. Published 4 scopus conferences and 5 sci and web of science Journals and filled one patent. Areas of research interest are Data Mining, cloud computing, artificial intelligence, machine learning and deep learning.

K.L.Sowmya, Research Scholar, VNR Vignana Jyothi Institute of Engineering and Technology Department of Information Technology, Kotturisowmya61@gmail.com
K.L.Sowmya, Research scholar graduated from VNR Vignana Jyothi Institute of Engineering and Technology specialized in information technology 2017-2021 batch. . Published a paper 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) and filled one public patent. Been part of projects under JNTUH and solbolts ltd. and a member of ACM, International student chapter. Areas of research interest are Data Mining, cloud computing, artificial intelligence, machine learning ,IOT and deep learning.