A Breast Cancer Classification Technique Based On Histological Images Using Convolutional Neural Networks

Ms .Neelambary V[1], Ms .Subhashini .G[2]

[1]Assistant Professor, Department of Information Technology, St. Joseph's Institute of Technology, Chennai, India
[2]Assistant Professor, Department of Information Technology, St. Joseph's Institute of Technology, Chennai, India

Abstract:

Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to a disagreement between pathologists. The diagnosis is based on the qualification of histopathologist, who will look for abnormal cells. However, if the histopathologist is not well-trained, this may lead to wrong diagnosis. Computer-aided diagnosis systems showed potential for improving the diagnostic accuracy. In this work, we develop the computational approach based on deep convolution neural networks for breast cancer histology image classification. Hematoxylin and eosin stained breast histology microscopy image dataset is provided as a part of the ICIAR 2018 Grand Challenge on Breast Cancer Histology Images. Our approach utilizes several deep neural network architectures.

Convolutional Neural Networks for Binary class classification and multiclass classification. The Binary class classification is used to classify the cancer cells to malignant and benign. And the Multiclass classification these classes into different subclasses like adenosis, fibroadenoma, phyllodes tumour, tabular adenoma for benign class and ductal carcinoma, lobular carcinoma, mucinous carcinoma, papillary carcinoma for malignant class. The result will show Convolutional Neural Networks outperformed the handcrafted feature based classification with high accuracy in both binary and multiclass classification.

Keywords: histopathologist, Convolutional Neural Networks, adenosis, fibroadenoma, phyllodes tumour.

I INTRODUCTION

Breast cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body. Excluding skin cancer, breast cancer is the most common type of cancer in women in the United States, accounting for one of every three cancer diagnoses. An estimated 211,240 new invasive cases of breast cancer were expected to occur among women in the United States during 2005. About 1,690 new male cases of breast cancer were expected in 2005. The incidence of breast cancer rises after age 40. The highest incidence (approximately 80% of invasive cases) occurs in women over age 50. In addition to invasive breast cancer, 58,590 new cases of in situ breast cancer are expected to occur among women during 2005. Of these, approximately 88% will be classified as ductal carcinoma in situ (DCIS). The detection of DCIS cases is a direct result of the increased use of mammography screening. This screening method is also responsible for detection of invasive cancers at a less advanced stage than might have occurred otherwise. An estimated 40,870 deaths (40,410 women, 460 men) were anticipated from breast cancer in 2005. Breast cancer ranks second among cancer deaths in women. According to the most recent data, mortality rates declined significantly during 1992-1998, with the largest decreases in younger women, both white and black.

In recent years, the classification of breast cancer has been the topic of interest in the field of Healthcare informatics, because it is the second main cause of cancer-related deaths in women. Breast cancer can be identified using a biopsy where tissue is removed and studied under microscope. The diagnosis is based on the qualification of the histopathologist, who will look for abnormal cells. However, if the histopathologist is not well-trained, this may lead to wrong diagnosis. With the recent advances in image processing and machine learning, there is an interest in attempting to develop a reliable pattern recognition-based system to improve the quality of diagnosis.

Types of Breast Cancer

Ninety percent of breast cancer are adenocarcinomas, which arise from glandular tissue. Within this broad category, there is a great degree of variation. For instance, there are about 30 different subtypes of adenocarcinoma. The earliest form of the disease, ductal carcinoma in situ, comprises about 15-20% of all breast cancers and develops solely in the milk ducts. The most common type of breast

cancer, invasive ductal carcinoma, develops from ductal carcinoma in situ, spreads through the duct walls, and invades the breast tissue.

Cancer that begins in the lobes or lobules is called lobular (small cell) carcinoma and is more likely to be found in both breasts. Invasive lobular carcinoma originates in the milk glands and accounts for 10-15% of invasive breast cancers. Both ductal and lobular carcinomas can be either in situ, or self-contained; or infiltrating, meaning penetrating the wall of the duct or lobe and spreading to adjacent tissue.

Less common types of breast cancer include the following:

- Inflammatory breast cancer (diffuse brawny infiltration; breast appears red or inflamed; tends to spread quickly)
- Medullary carcinoma (originates in central breast tissue)
- Mucinous carcinoma (invasive; usually occurs in postmenopausal women)
- Paget disease of the nipple (originates in the milk ducts and spreads to the skin of the nipples or areola)
- Phyllodes tumor (tumor with a leaf-like appearance that extends into the ducts; rarely metastasizes)
- Tubular carcinoma (small tumor that is often undetectable by palpation)

Sarcomas (cancer of the connective tissue) and lymphomas (cancer of the lymph tissue) rarely develop in the breasts.

Scope of The Project

This paper aims to develop a reliable pattern recognition to classify breast cancer based on histological images using Convolutional Neural Networks(CNN) using handcrafted feature encoding models, binary and multiclass classifiers to improve the quality of diagnosis.

II RELATED WORK

The classification of breast cancer has been the topic of interest in the field of Healthcare informatics, because it is the second main cause of cancer-related deaths in women. Breast cancer can be identified using a biopsy where tissue is removed and studied under microscope. The diagnosis is based on the qualification of the histopathologist, who will look for abnormal cells. However, if the histopathologist is not well-trained, this may lead to wrong diagnosis. With the recent advances in image processing and machine learning, there is an interest in attempting to develop a reliable pattern recognition based systems to improve the quality of diagnosis. In this paper, it compare two machine learning approaches for the automatic classification of breast cancer histology images into benign and malignant and into benign and malignant sub-classes. The first approach is based on the extraction of a set of handcrafted features encoded by two coding models (bag of words and locality constrained linear coding) and trained by support vector machines, while the second approach is based on the design of convolutional neural network. It have also experimentally tested dataset augmentation techniques to enhance the accuracy of the convolutional neural network as well as "handcrafted features + convolutional neural network" and "convolutional neural network" and "convolutional neural network features + classifier" configurations[1].

the classification of with this work we explore the possibility to find effective solutions to the above issue by designing and testing two different machine learning approaches. The first one is based on the extraction of a complete set of handcrafted morphometric and radiometric features used within a Support Vector Machines solution. The second one is based on the design and configuration of a Convolutional Neural Networks deep learning architecture. To validate, in a real and challenging clinical scenario, the proposed bacterial load estimation techniques, we built and publicly released a fully labelled large and representative database of both single and aggregated bacterial colonies extracted from routine clinical laboratory culture plates. Dataset enhancement approaches have also been experimentally tested for performance optimization. The adopted deep learning approach outperformed the handcrafted feature based one, and also a conventional reference technique, by a large margin, becoming a preferable solution for the addressed Digital Microbiology Imaging quantification task, especially in the emerging context of Full Laboratory Automation systems[2].

Today, medical image analysis papers require solid experiments to prove the usefulness of proposed methods. However, experiments are often performed on data selected by the researchers, which may come from different institutions, scanners, and populations. Different evaluation measures

may be used, making it difficult to compare the methods. In this paper, it introduce a dataset of 7909 breast cancer histopathology images acquired on 82 patients, which is now publicly available from http://web.inf.ufpr.br/vri/breast-cancer-database. The dataset includes both benign and malignant images. The task associated with this dataset is the automated classification of these images in two classes, which would be a valuable computer-aided diagnosis tool for the clinician. In order to assess the difficulty of this task, we show some preliminary results obtained with state-of-the-art image classification systems. The accuracy ranges from 80% to 85%, showing room for improvement is left. By providing this dataset and a standardized evaluation protocol to the scientific community, we hope to gather researchers in both the medical and the machine learning field to advance toward this clinical application.[3]

Automated breast cancer multi-classification from histopathological images plays a key role in computer-aided breast cancer diagnosis or prognosis. Breast cancer multi-classification is to identify subordinate classes of breast cancer (Ductal carcinoma, Fibroadenoma, Lobular carcinoma, etc.). However, breast cancer multi-classification from histopathological images faces two main challenges from: (1) the great difficulties in breast cancer multi-classification methods contrasting with the classification of binary classes (benign and malignant), and (2) the subtle differences in multiple classes due to the broad variability of high-resolution image appearances, high coherency of cancerous cells, and extensive inhomogeneity of color distribution. Therefore, automated breast cancer multi-classification from histopathological images is of great clinical significance yet has never been explored. Existing works in literature only focus on the binary classification but do not support further breast cancer quantitative assessment. In this study, we propose a breast cancer multi-classification method using a newly proposed deep learning model. The structured deep learning model has achieved remarkable performance (average 93.2% accuracy) on a large-scale dataset, which demonstrates the strength of our method in providing an efficient tool for breast cancer multi-classification in clinical settings.[4]

Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to a disagreement between pathologists. Computer-aided diagnosis systems showed potential for improving the diagnostic accuracy. In this work, we develop the computational approach based on deep convolution neural networks for breast cancer histology image classification. Hematoxylin and eosin stained breast histology microscopy image dataset is provided as a part of the ICIAR 2018 Grand Challenge on Breast Cancer Histology Images. Our approach utilizes several deep neural network architectures and gradient boosted trees classifier. For 4-class classification task, it reports 87.2% accuracy. For 2-class classification task to detect carcinomas it reports 93.8% accuracy, AUC 97.3%, and sensitivity/specificity 96.5/88.0% at the high-sensitivity operating point. To the knowledge, this approach outperforms other common methods in automated histopathological image classification.[5]

Currently, histopathological tissue examination by a pathologist represents the gold standard for breast lesion diagnostics. Automated classification of histopathological whole-slide images (WSIs) is challenging owing to the wide range of appearances of benign lesions and the visual similarity of ductal carcinoma in-situ (DCIS) to invasive lesions at the cellular level. Consequently, analysis of tissue at high resolutions with a large contextual area is necessary. We present context-aware stacked convolutional neural networks (CNN) for classification of breast WSIs into normal/benign, DCIS, and invasive ductal carcinoma (IDC). We first train a CNN using high pixel resolution to capture cellular level information. The feature responses generated by this model are then fed as input to a second CNN, stacked on top of the first. Training of this stacked architecture with large input patches enables learning of fine-grained (cellular) details and global tissue structures. Our system is trained and evaluated on a dataset containing 221 WSIs of hematoxylin and eosin stained breast tissue specimens. [6]

IV PROPOSED SYSTEM

The proposed system is the system which is developed to enhance the shortcomings of existing system by using advanced techniques.Here, it implements Convolutional Neural Networks for Binary class classification and multiclass classification. The Binary class classification is used to classify the cancer cells to malignant and benign. And the Multiclass classification these classes into different subclasses like adenosis, fibroadenoma, phyllodes tumour, tabular adenoma for benign class and ductal carcinoma,

lobular carcinoma, mucinous carcinoma, papillary carcinoma for malignant class. The result will show Convolutional Neural Networks outperformed the handcrafted feature based classification with high accuracy in both binary and multiclass classification. Our Proposing work using CNN topology has to improve the accuracy of the previous ones for the multiclass classification task, where they reached a performance of between 83.31% and 88.23%.

V SYSTEM ARCHITECTURE
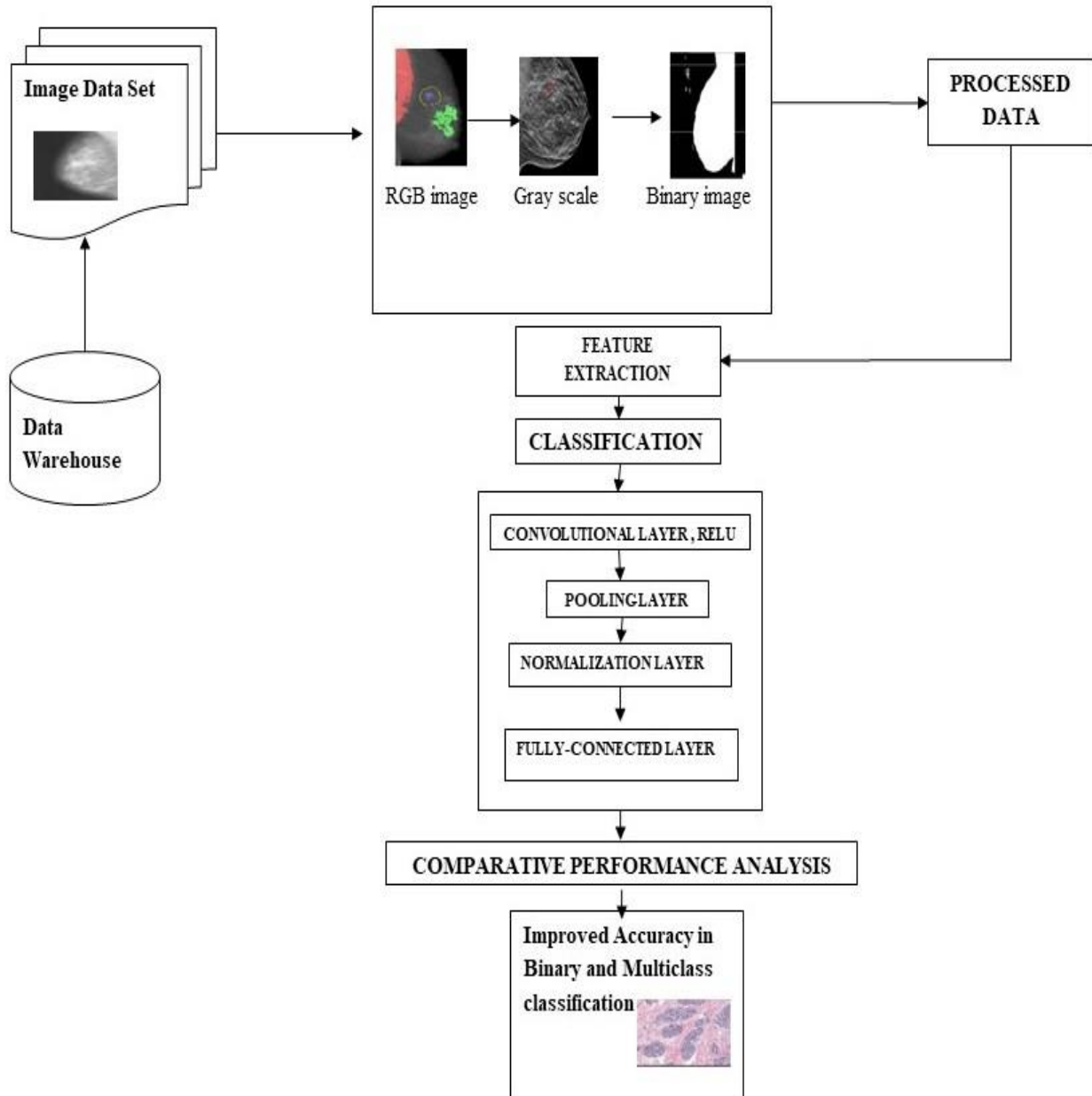
Fig 1 System Architecture A Breast Cancer Classification Technique Based on Histological Images Using Convolution Neural Networks

VI IMPLEMENTATION

This system contains certain modules to execute the proposed system and each module will contain certain algorithms and techniques to be executed. Certainly, this work also contains few

modules to exhibit the breast cancer classification based on histological images and also graph representation.

Dataset is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. This collected data stored in the data warehouse.

DATA PRE-PROCESSING

Pre-processing is defined as the removal of error in the data. It transform raw data and un-structured data into a structured data. Source data is collected and stored in data warehouse and is preprocessed to extract the consistent data. The data is splitted in Trainig and testing Data set.

Fig: 4.1 Data Preprocessing

FEATURE EXTRACTION

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. This is used to reduce the number of resources needed for processing without losing important or relevant information. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set.

Fig:4.2 Feature Extraction

CONVOLUTIONAL NEURAL NETWORKS(CNN)

In machine learning, Convolutional Neural Networks is a class of deep networks, and CNN were inspired by biological processes. The connectivity pattern between neurons and resembles the organization of the animal visual cortex. A CNN consists of an input and an output layer, as well as

multiple hidden layers. The hidden layers of a CNN typically consist of Convolutional layers, Pooling layers, Normalization layers, Fully connected layers .

CONVOLUTIONAL LAYERS

The first layer in a CNN is always a Convolutional Layer.Convolutional layers apply a convolution operation to the input, passing the result to the next layer. The convolution emulates the response of an individual neuron to visual stimuli. Each convolutional neuron processes data only for its receptive field. To learn features as well as classify data fully connected feedforward neural networks can be used.

POOLING LAYERS

Convolutional networks may include local or global pooling layers, which combine the outputs of neuron clusters at one layer into a single neuron in the next layer. For example, max pooling uses the maximum value from each of a cluster of neurons at the prior layer. Another example is average pooling, which uses the average value from each of a cluster of neurons at the prior layer.

FULLY CONNECTED LAYERS

Finally, after several convolutional and max pooling layers, the high-level reasoning in the neural network is done via fully connected layers. Neurons in a fully connected layer have connections to all activations in the previous layer, as seen in regular neural networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset.

COMPARATIVE PERFORMANCE ANALYSIS

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passed , how manyactually passed High precision relates to the low false positive rate.

PRECISION = TP/TP+FP

Precision is used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system. The usage of "precision" in the field of information retrieval differs from the definition of accuracy and precision within other branches of science and technology.

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. **RECALL = TP/TP+FN**

For example, for a text search on a set of documents, recall is the number of   correct results divided by the number of results that should have been returned. It can be viewed as the probability that a relevant document is retrieved by the query.It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by also computing the precision.

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

F1 SCORE = 2*(RECALL * PRECISION) / (RECALL + PRECISION)

This measure is approximately the average of the two when they are close,and is more generally the harmonic mean, which, for the case of two numbers, coincides with the square of the geometric mean divided by the arithmetic mean.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model.

ACCURACY = TP+TN/TP+FP+FN+TN

(i)True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

(ii)True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of   predicted class is also no. E.g. if actual class says this

passenger did not survive and predicted class tells you the same thing. False positives and false negatives, these values occur when your actual class contradicts with the predicted class.

(iii)False Positives (FP) – When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

(iv)False Negatives (FN) – When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

SCATTERED PLOT

A scatter plot also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded (color/shape/size), one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis the figure below shows the scatter plots of two variables of datasets.
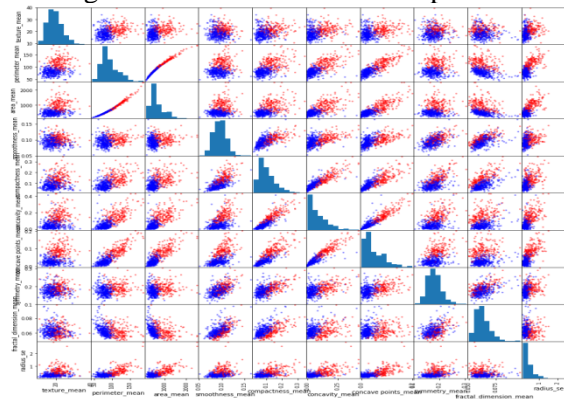
Fig 4.3: Scatter plots of variables

CORRELATION MATRIX

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. Key decisions to be made when creating a correlation matrix include: choice of correlation statistic, coding of the variables, treatment of missing data, and presentation.
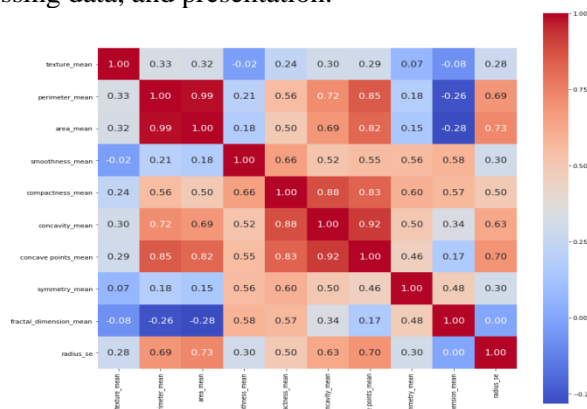

Fig: 4.4 Corelation Matrix

CLASSIFICATION

Classification is a general process related to categorization, the process in which ideas and objects are recognized, differentiated, and understood. In this case, the features are collected and classified by two various methods. Feature based classification ,Convolutional Neural Networks (CNN),**Binary Classification and Multiclass models.** . **Binary classification** refers to those **classification** tasks that have two class labels.**Multiclass** or multinomial **classification** is the problem of classifying instances into one of three or more **classes ie for identifying various types of**

breast cancer. The use of these two types of classification improves the accuracy if the diagnosis. The below figure shows the binary and Multiclass classification result.
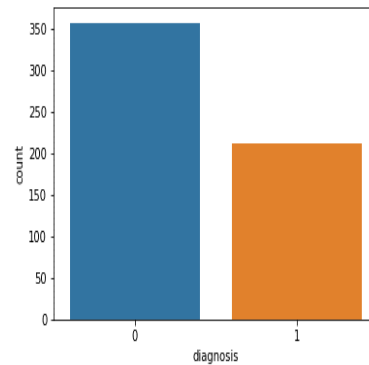
Fig 4.5.1 Binary Classification
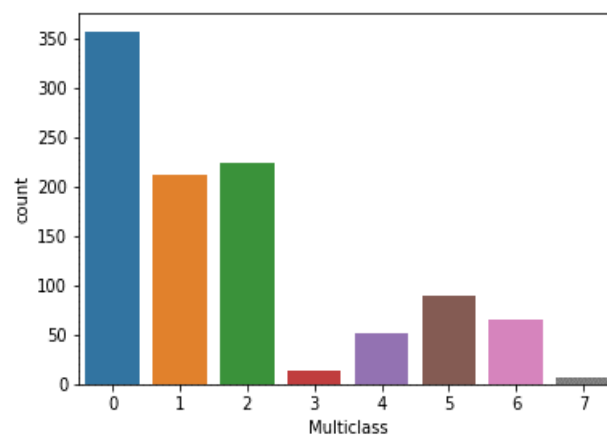
Fig 4.5.2 Multiclass Plots

VII CONCLUSION

Brest Cancer is been widely affecting womens across the world. The early prediction will reduce the death ratio. The proposed work helps to improve the prediction accuracy and thus helps in the quality of the diagnosis. This paper has been developed a reliable pattern recognition to classify breast cancer based on histological images using Convolutional Neural Networks(CNN) using handcrafted feature encoding models, binary and multiclass classifiers to improve the quality of diagnosis. The use of binary and Multiclass classifier has improved the accuracy of the diagnosis. This type of Multi classification model helps to diagnose varios types of Brest Cancers.

FUTURE ENHANCEMENT

A Computer Aided Diagnosis based on the CNN using FPGA or other commercial hardware is one of the challenging task. A deep feature enhancement and Ensemble classification can be implemented so that prediction accuracy of different data sets can be improved efficiently.

REFERENCES

[1] A. Ferrari, S. Lombardi, and A. Signoroni, ''Bacterial colony counting with convolutional neural networks in digital microbiology imaging,'' Pattern Recognit., vol. 61, pp. 629–640, Jan. 2017.

[2] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, ''Breast cancer histopathological image classification using convolutional neural networks,'' in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2016, pp. 2560–2567.\

[3] T. Araújo et al., ''Classification of breast cancer histology images using convolutional neural networks,'' PLoS ONE, vol. 12, no. 6, p. e0177544, 2017.

[4] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, and L. Heutte, ''Deep features for breast cancer histopathological image classification,'' in Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC), Oct. 2017, pp. 1868–1873.

[5] N. Bayramoglu, J. Kannala, and J. Heikkilä, ''Deep learning for magnification independent breast cancer histopathology image classification,'' in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), Dec. 2016, pp. 2440–2445.

[6] A.-A. Nahid and Y. Kong, ''Histopathological breast-image classification using local and frequency domains by convolutional neural network,'' Information, vol. 9, no. 1, p. 19, 2018.

[7] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, ''Breast cancer multiclassification from histopathological images with structured deep learning model,'' Sci. Rep., vol. 7, Jun. 2017, Art. no. 4172.

[8] Y. Zhou, J. Xu, Q. Liu, C. Li, Z. Liu et al., "A radiomics approach with cnn for shear-wave elastography breast tumor classification," IEEE Transactions on Biomedical Engineering, vol. 65, no. 9, pp. 1935–1942, 2018.

[9] H. Qiyuan, H. M. Whitney and M. L. Giger, "A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI," Scientific Reports, vol. 10, no. 1, pp. 1–17, 2020.

[10] E. Chaves, C. B. Gonçalves, M. K. Albertini, S. Lee, G. Jeon et al., "Evaluation of transfer learning of pre-trained CNNs applied to breast cancer detection on infrared images," Applied Optics, vol. 59, no. 17, pp. 1–7, 2020.

[11] K. George, P. Sankaran and K. P. Joseph, "Computer assisted recognition of breast cancer in biopsy images via fusion of nucleus-guided deep convolutional features," Computer Methods and Programs in Biomedicine, vol. 2020, pp. 105531–105532, 2020.

[12] V. J. Kadam, S. M. Jadhav and K. Vijayakumar, "Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression," Journal of Medical Systems, vol. 43, no. 8, pp. 263–264, 2019.

[13] S. Singh and R. Kumar, "Histopathological image analysis for breast cancer detection using cubic svm," in 7th IEEE Int. Conf. on Signal Processing and Integrated Networks, Noida, India, pp. 498–503, 2020.

[14]T. C. Chiang, Y. S. Huang, R. T. Chen, C. S. Huang and R. F. Chang, "Tumor detection in automated breast ultrasound using 3-D CNN and prioritized candidate aggregation," IEEE Transactions on Medical Imaging, vol. 38, no. 1, pp. 240–249, 2018.

[15] S. A. Agnes, J. Anitha, S. I. A. Pandian and J. D. Peter, "Classification of mammogram images using multiscale all convolutional neural network," Journal of Medical Systems, vol. 44, no. 1, pp. 1–13, 2020.