International Journal on Emerging Technologies **10**(2): 467-473(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Histopathological Image Classification Using Deep Learning Techniques

K. Sai Prasad and Rajender Miryala
Department of Computer Science and Engineering,
MLR Institute of Technology, Dundigal, Hyderabad, India.

(Corresponding author: Rajender Miryala)
(Received 01 June 2019, Revised 02 August 2019 Accepted 19 August 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Histopathology deals with study and diagnosis of the trimmed or cut tissues collected from the human bodies or plants. Histopathologist collects the tissue samples on glass slides and then manually observe them under the microscope to find the existence of disease. The manual process of identifying a disease exists or not in a tissue sample may be error prone and time consuming. Also, it would become difficult to extract the features from the glass slide after few days. Due to the advent of technology many machine learning libraries are available which can be used to overcome the difficulties of manual process.
We are proposing automated way of identifying the presence of a disease in the tissue samples by creating an image classifier developed by using TensorFlow python library. This image classifier is trained on the histopathology image dataset that contains invasive ductal carcinoma (IDC) and non-IDC categorized images. The trained image classifier model can be used to predict IDC image and non-IDC image. Approaches like this can be much useful in many health care domains and also the efficiency of identifying a tumor can be improved.

Keywords: Convolutional Neural Network, Deep Learning, Image Classification, TensorFlow Concepts

I. INTRODUCTION

TensorFlow. TensorFlow is developed by Google and its distributed as open source platform library. It can be used as a machine learning platform and used in developing deep learning network.
Python is the best language to create a TensorFlow models. Most of the actual TensorFlow operations are implemented in C++ and not in python.
TensorFlow models or applications can be executed on any windows, iOS, android or any google cloud platform.
TensorFlow provides easy ways to debug the applications. TensorFlow provides graphical visualization of models. TensorFlow has got a greater number of contributors and developers compared to other deep learning frameworks.
With TensorFlow, work on the purpose not on accuracy. It all depends on the problem we are trying to solve.
In a simple way TensorFlow program flow goes as follows:

a) Create Data
b) Create placeholder
c) Define Dataset
d) Create pipeline
e) Execute Operation using Session

Each Tensor requires a flattened array of data as input. TensorFlow has over 1800 contributors worldwide and the largest in the deep learning technology community. TensorFlow2.0 has improved in such a way that it's easy to understand and learn compared to older version.

Tensors in the TensorFlow Graph. Tensor contains input data. Required operations are represented as nodes. Several tensors are connected as a flow and given as input to the nodes. An operation is performed on these tensors. An output tensor is produced after the operation is executed. Tensor data flows throws the nodes and it creates a graph of nodes and tensors. Tensors are used to represent and hold the data in the graph. Tensors are represented in n dimensional array as shown in Fig.1

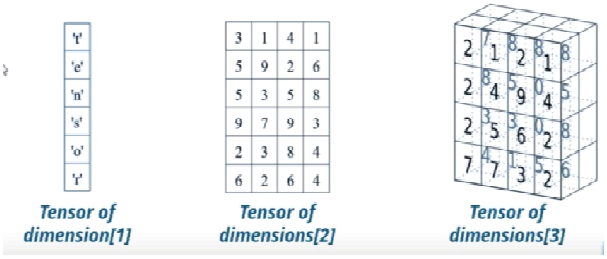

Fig. 1. Tensors of different dimensions.

Tensor Rank. The below diagram shows the rank and math entity of tensor representation and a python example is given to each rank.

Rank	Math Entity	Python Example
0	Scalar (magnitude only)	S = 234
1	Vector (Magnitude and direction)	V = [1.1, 2.2, 3.3]
2	Matrix (table of numbers)	M = [[1,2,3], [4,5,6], [7,8,9]]
3	3-Tensor (cube of numbers)	T = [[[2], [4], [6]], [[8], [12], 11]], [[12], [33], [22]]]
n	n-tensor	---

Fig. 2. Tensor Rank calculation based on the tensor array dimension.

Concept or Transfer Learning. Generally, a new deep leaning model is created by training on a set of images. During the training, the model gains knowledge about the data. If the gained knowledge is used in solving a different problem which is of similar kind, then it's called transfer learning.

Concept of Pre-Training. Consider a knowledge model created on a large dataset. Currently if we have a small dataset then, knowledge model created on large dataset can be used to continue the train on small dataset. So here the pre-trained model is used in training the smaller dataset.

Tensor Data Types. Tensor Data is represented in multi-dimensional array. Tensors are also defined to hold specific data types. For ex: tf.float32, tf.flaot64, tf.int8 etc.

Session in TensorFlow. Session provides an environment where the abstract graph operations are executed by taking the input tensor data.

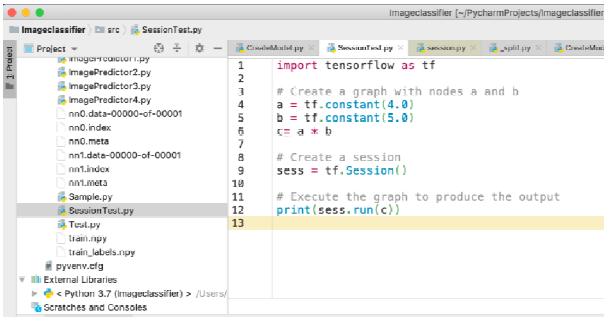

Fig. 3. Simple session creation in TensorFlow.

II. LITERATURE SURVEY

Referenced several papers published on how deep learning can be used in medical domain and how deep learning techniques can help to classify different types of images captured from human or plant tissue samples. Many papers talk about the concepts involved in convolution neural network as well as different technologies that can be used as part of this medical research. Here we discuss about the concepts that helped in our research.

Kumar & Rao, have explained the importance of Convolutional Neural Network (CNN) in digital image world in their work titled 'Breast Cancer Classification of Image using CNN'. The below points were identified and helpful in conducting the further work [4].

a. CNN is a deep learning model that can be used in Image processing, Image classification, video processing and recognition, natural language processing etc.

b. CNN Deep learning models can be used to extract features from image and those features can be used in classifying images.

c. CNN models are proven to be best models in terms of performance. The tests were performed on MNIST dataset.

d. They have used BreakHis dataset (9109 images) for training their model. BreakHis dataset contains images of malignant tumors as well as benign breast cancer tumors.

Kieffer *et al.,* have elaborated about the use of trained models in 'CNN for Histopathology Image Classification: Using Trainingversus Pretrained networks' [1].

a. Pre-trained knowledge models have proven to be very much competitive to train on the new dataset compared to creating a new model from the beginning.

b. Feature vectors are extracted from Pre-trained networks.

c. Discussed about some problems in image classification especially about lack of datasets.

d. Data augmentation concept can be used to generate new set of images from the existing images and these new images can used for training the model. This was with limited set of images; models can be created.

e. They have done their work in Kimia-Path24 dataset which has 27055 histopathology images.

Chang *et al.,* discussed about the importance of deep learning in building an image classification model. They explained about transfer learning methods to detect breast cancer with the help of Googles InceptionV3 model. This model was already trained on different type of images which were not histopathology images [8].

Vu *et al.,* explained about the concept of discriminative feature-oriented dictionary learning (DFDL) in their work titled 'HistopathologicalImageClassification using Discriminative Feature-orientedDictionary Learning' and how DFDL methods can be useful to learn different class or type of histopathology images. These concepts are very helpful in conducting further work [7].

a. They explained about use of morphological features of images, wavelet concepts and many more methods to understand a pathology image.

b. Explained about different machine learning methods in analyzing a pathology image. Also explained about the importance of methods like grey level co-occurrences-Matrix (GLCM) and local-binary-pattern(LBP) are useful in extracting the features of an images.

c. They also spoke on various kinds of problems in understanding histopathological images like large sized images, not availability of the categorized image data etc.

Samah *et al.,* discussed about the following points which are useful in conducting further research [9]:

a. They have used BreakHis dataset for their analysis.

b. They have developed a system to differentiate the histopathology images into two: Benign and Malignant.

c. k-nearest neighbors concept is used bytheir system to measure the performance of the system while extracting the features.

Kumar *et al.* discussed about following concepts that are useful in further research [5]:

a. They have extracted and created a dataset called KIMIA-Path960 containing histopathology-images. It contains a total of 960 images.

b. They have done a comparative research study on local-binary-patterns(LBP), deep-features and bag-of-visual words methodologies that can be used in histopathological image classification.

c. Even though the dataset size is small, the images were high in quality and easily differentiable, which them to achieve models with high accuracy percentage.

Spanhol *et al.,* explained about the following concepts that are useful in further analysis [3]:

a. They have created a histopathological images data set for breast cancer. Number of images are 7909 and they have published over the internet and available publicly.
b. This dataset contains benign and malignant images.
Cruz-Roa *et al.,* used the approach of visually represe-nted information (called by term bag of features) of a histopathological image in their work. BOF(bag of features) term is used in medical image analysis and applications and also in radiology image interpretation. We have followed the similar approach of analyzing the image using visual features [11].
Patel and Gamit discussed about various texture and color feature extraction methods such as color correlation, tamur a texture feature and other ways [12].
Zhang proposed a method to classify microscopic biopsy images by using the concepts of curvelet transform and LBP features with rejection option in 'Breast cancer diagnosis from biopsy images by serial-fusion of random-subspace ensembles [13].
Babaie *et al.* [2] introduced to a new dataset Kimia Path24 which is used for classification of digital pathology images. This dataset was created by to overcome the challenges of storing and retrieving and gigapixel sized pathology image scans.
Komura and Ishikawa, discussed about various types of problems in histopathology image analysis such as non-availability of labeled images, huge image size, color variations, feature extractions. Also discussed about the machine learning methods for histopathological image analysis [6].
Belsare and Mushrif [14] discussed about the histopathology and difficulties in identifying the disease in biopsy image sample by a pathologist. Tissue structures and distribution of tissue cells are done by microscopic examination by a pathologist. To overcome this time-consuming process, image processing techniques are introduced which are discussed in their work. With the increasing availability of digital images, the approaches of automatic identification of disease in the digital pathology images increased. Cruz-Roa A *et, al* [10] discussed about the computerized techniques of detecting disease extent and manual detection of disease extent by a pathologist in their work. Cireşan, *et al.* proposed a method with the help of deep convolutional neural network. This method is used to identify mitosis in breast cancer histology images. This research helped us to proceed further on improving and extending deep learning techniques in image analysis and classification [15].

III. PROPOSED WORK

i) Architecture Diagram

The following diagram in Fig.4 explains about the basic building components or layers required for creating any convolutional neural network. Following are the layers
a. Convolution Layer
b. Pooling Layer
c. Flattening Layer
d. Fully Connected Layer.

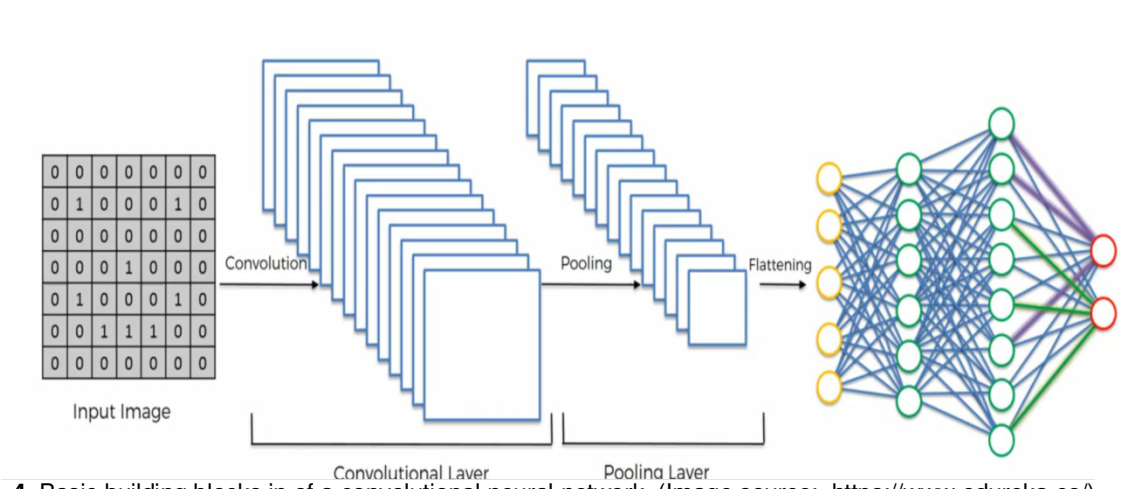

Fig. 4. Basic building blocks in of a convolutional neural network. (Image source: https://www.edureka.co/)

ii) Architecture and Training Process flow Diagram

The following diagram as shown in Fig. 5 explains about the steps involved in creating image classifier model using TensorFlow and Python libraries. Details about each step is explained in the next section.

iii) Proposed Methodology

Algorithm:

Step 1: Read the input data
a. Create two npy arrays with names training.npy and trainining_label.npy.
b. Read the images from the input dataset which has two separate folders containing IDC images and non-IDC images.
c. Append each image from the dataset using cv2.Imread function and add to the train array.
d. If the image read in the previous step is IDC, then mark as IDC image at the corresponding index in the training_label.npy by putting 1 in the same index otherwise put zero to indicate the non-IDC image.
e. Repeat above three steps until all of the images in the dataset are read. Output would be trainingdata.npy and trainingdata_labels.npy files.

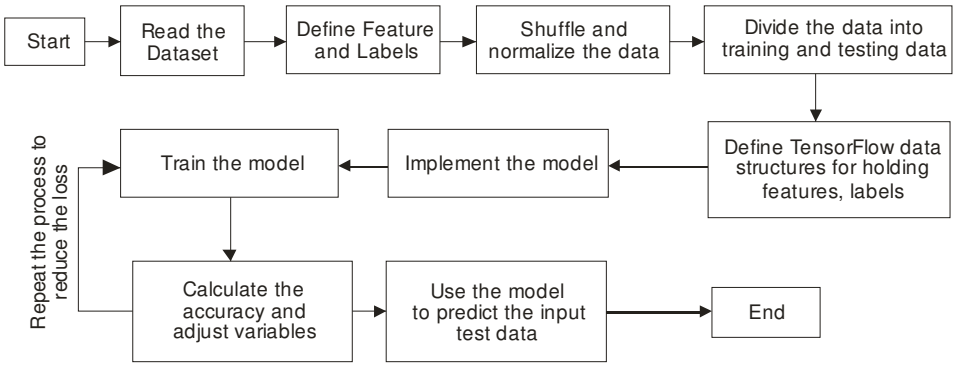

Fig. 5. Dataflow process diagram using TensorFlow implementation.

f. Keep trainingdata.npy and trainingdata_labels.npy files in a separate folder and use them whenever required.

Step 2: Load the input data in arrays for further processing.
a. Create two arrays trainingdata andtrainingdata_labels
b. Load trainingdata.npy and trainingdata_labels.npy created in previous steps into arrays trainingdata and trainingdata_labels

Step 3. Shuffle the input data
a. Shuffle the input arrays trainingdata and trainingdata_labels.
b. Use Np.random.shuffle method to shuffle the data.

Step 4. Plot or display some of the images from the arrays created in the above arrays.

Step 5: Normalize the input data
Input arrays contains integer values from 0 to 255 to hold RGB colors. It's difficult to understand these numbers for TensorFlow. So, divide each number in the array by 255 so that all the numbers would be in the range from 0 to 1.

Step 6: Use one-hot encoding for labels 0, 1
Tensors accepts only arrays as inputs. So, convert 0 and 1 numbers into array representations like [0,1] and [1,0]. This can be done by one-hot encoding method.
Step 7: Build neural network graph with TensorFlow.
Create the following layers in the network:
a. Data input layer
b. Layer1> Convolution1 > ApplyReLu > ApplyMaxPool
c. Layer2> Convolution2 > ApplyReLu > ApplyMaxPool
d. Layer3> Convolution3> ApplyReLu > ApplyMaxPool
e. Layer4> Convolution4 > ApplyReLu > ApplyMaxPool
f. Layer5> Convolution5 > ApplyReLu > ApplyMaxPool
g. Layer6> FullyConnected > ApplyReLu
h. Output Layer> FullyConnected > ApplyReLu

Step 8: Train the model and then validate this model against validation data. Follow the below steps:
a. Divide the input data into training data and validation data.
b. From the training data, generate new set of images by using data augmentation with the help of keras preprocessing library.
c. Rorate, zoom and shift the width and height of the image to create new image using keras library. This new image can be used for training the neural network graph.
d. Read the input data in batches and pass to the network model.
e. Run and train the model using session.run in TensorFlow.

Step 9: Repeat the above process for a required number of epochs. A trained model will be created as an output.

Step 10: Save the classifier model for each run
a. Save the trained model for each epoch. It will create the following files.
For ex, for first epoch with name model0, the following files will be created at the end of the run
model0.data-00000-of-0001
model0.index
model0.meta
Step 11: Use the model to classify the input images
b. Create a new TensorFlow session.
c. Import the saved model into the session.
d. Read input image which we want to categorize.
e. Classify the image by using the saved model.

IV. DATASET USED

Cancer is one of the diseases where number of cases are increased year on year throughout the world. Cancer Statistics in United States of America says that in 2018, over 17 lakh new cases of cancer will be diagnosed (it's a year-old data).
Among all the cancer disease types, in women, breast cancer is diagnosed most compared to other types of cancers like lung or cervix cancer. The situation is almost same in the India and in United States of America. (Gathered this information from google)
Invasive Ductal Carcinoma (IDC) is a type of breast cancer and it's the most commonly observed type.

Pathologists manually identifying this subtype of cancer is very difficult clinically. So instead of manual process, if there an automated process then that can help to identify the IDC cancer and also can reduce the human error. To create an automated process and to train this process, we have taken a sample dataset that consists of digital histopathological images.

This dataset consists of two sets images labelled IDC(cancer image) and non-IDC(non-cancer images). Total images are 5547. Each image is of size 50X50X3. This dataset is taken from the below link http://www.andrewjanowczyk.com/use-case-6-invasive-ductal-carcinoma-idc-segmentation/.

V.RESULTS

We have developed a CNN model using TensorFlow, keras and other python libraries. During the training of the model we have captured and plotted some of the images for reference.

Following are the images from the data set as shown in the Fig. 6.

IDC= 0 indicates non-IDC Image(non-cancerous)

IDC=1 indicates IDC image(cancerous)

Following images in Fig.7. explains about the histogram of RGB pixel intensity of the images [10]. The pixel intensity is ranged from 0 to 255 on the x axis and count of pixels is present on the y axis. Below figure clearly indicated the high intensity of the pixels in IDC image compared to non-IDC image.

Following images in Fig.8. are created as a result of rotating and zooming the images. First row contains images that are generated after zooming a single image. Similarly, for the second-row images. These images are used for training the model.

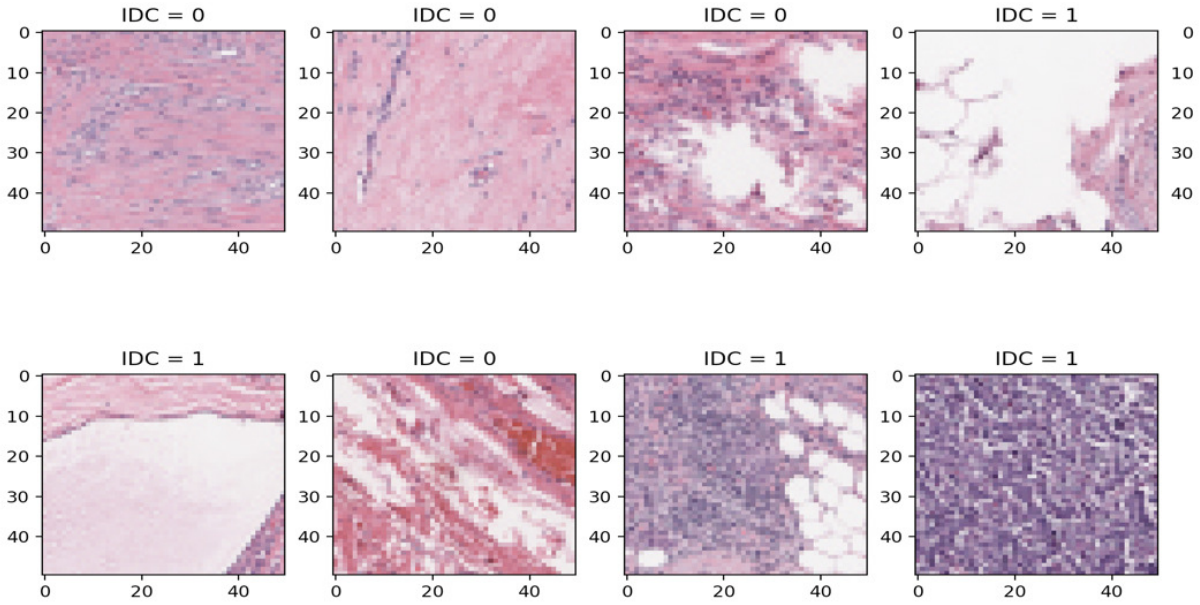

Fig. 6. Images describing the IDC and non-IDC images.

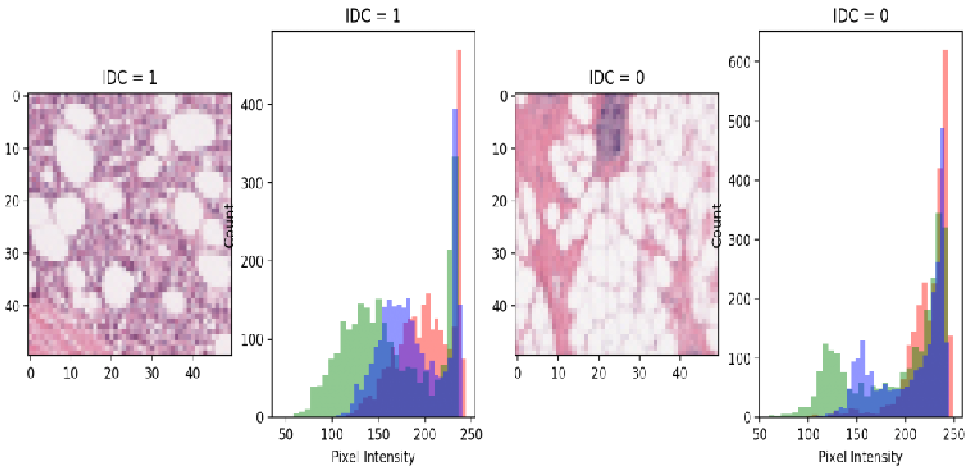

Fig. 7. Images describing the RGB pixel intensity.

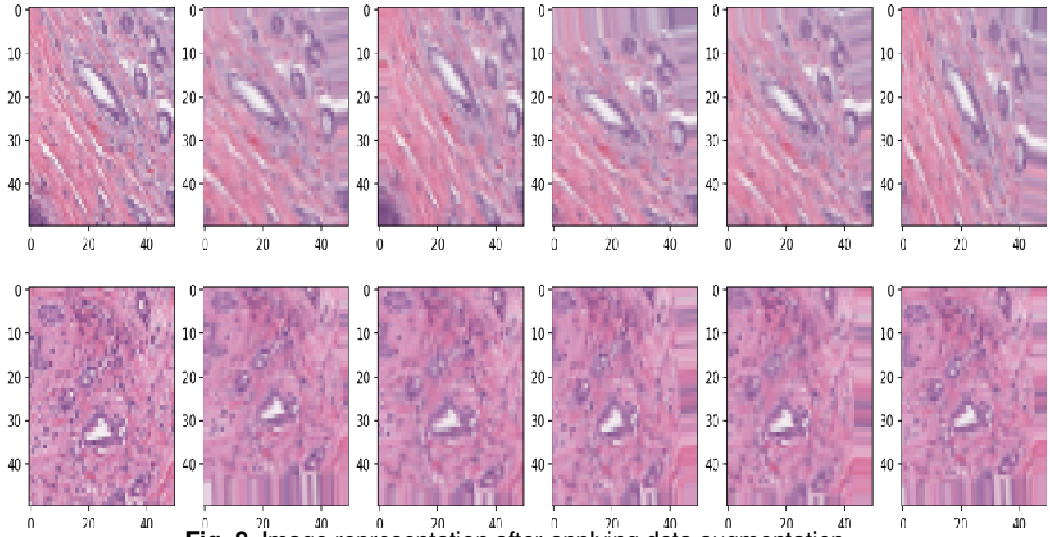

Fig. 8. Image representation after applying data augmentation.

The model is generated after successful training on the input dataset. The generated model is validated against the validation data set. The prediction accuracy is 83%.

VI. CONCLUSION

Here we've discussed about different basic layers present in building a Convolutional Neural Network model as well as how network model can be created with the help of a leading deep learning library TensorFlow. We have used histopathological image dataset. This dataset is divided into two datasets such that one part is used for training and other part of the dataset is used to validate the model created with TensorFlow.

We have created new set of images using the concept of data augmentation. So, with limited data, we can generate a greater number of images and train the model. This model is trained on the system once the training is completed. Afterwards, this model can be used to classify IDC vs non-IDC images.

VII. FUTURE SCOPE

Here we have worked upon on one kind of pathology image dataset and created a model. Similarly, models can be created from different kinds of pathology dataset images and build a framework with these models. Once the framework is ready, it should be able to accept any kind of pathology image and should be able to identify presence of disease.

Conflict of Interest: No

REFERENCES

[1]. Kieffer, B., Babaie, M., Kalra, S., & Tizhoosh, H.R. (2017). Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1-6). IEEE.

[2]. Babaie, M., Kalra, S., Sriram, A., Mitcheltree, C., Zhu, S., Khatami, A., ... & Tizhoosh, H.R. (2017). Classification and retrieval of digital pathology scans: A new dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 8-16).

[3]. Spanhol, F.A., Oliveira, L.S., Petitjean, C., &Heutte, L. A dataset for breast cancer histopathological imageclassification. In *IEEE Transaction on Biomedical Engineering*.(pp. 1-8)

[4]. Kumar, K., & Rao, A.C.S. (2018). Breast cancer classification of image using convolutional neural network. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)* (pp. 1-6). IEEE.

[5]. Kumar, M.D., Babaie, M., Zhu, S., Kalra, S., & Tizhoosh, H.R. (2017). A comparative study of cnn, bovw and lbp for classification of histopathological images. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1-7). IEEE.

[6]. Komura, D., & Ishikawa, S. (2018). Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, **16**, 34-42.

[7]. Vu, T.H., Mousavi, H.S., Monga, V., Rao, G., & Rao, U.A. (2015). Histopathological image classification using discriminative feature-oriented dictionary learning. *IEEE transactions on medical imaging*, *35*(3), 738-751.

[8]. Chang, J., Yu, J., Han, T., Chang, H.J., & Park, E. (2017). A method for classifying medical images using transfer learning: a pilot study on histopathology of breast cancer. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)* (pp. 1-4). IEEE.

[9]. Samah, A.A., Fauzi, M.F.A., Mansor, S., (2017). Classification of benign and malignant tumors in histopathology Images. In *IEEE International Conference on Signal and Image Processing Applications (IEEE ICSIPA 2017).*

[10]. Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N.N., ... & Madabhushi, A. (2017). Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Scientific reports*, *7*, 46450.
[11]. Cruz-Roa, A., Díaz, G., & González, F. (2011). A framework for semantic analysis of histopathological images using nonnegative matrix factorization. In *2011 6th Colombian Computing Congress (CCC)* (pp. 1-7). IEEE.
[12]. Patel, J.M., & Gamit, N.C. (2016). A review on feature extraction techniques in content based image retrieval. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 2259-2263). IEEE.
[13]. Zhang, B. (2011). Breast cancer diagnosis from biopsy images by serial fusion of Random Subspace ensembles. In *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)* (Vol. **1**, pp. 180-186). IEEE.
[14]. Belsare, A.D., & Mushrif, M.M. (2012). Histopathological image analysis using image processing techniques: An overview. *Signal & Image Processing*, **3**(4), 23.
[15]. Cireşan, D.C., Giusti, A., Gambardella, L.M., &Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention* (pp. 411-418). Springer, Berlin, Heidelberg.

How to cite this article: Sai Prasad, K. and Miryala, Rajender (2019). Histopathological Image Classification Using Deep Learning Techniques. *International Journal on Emerging Technologies*, **10**(2): 467–473.