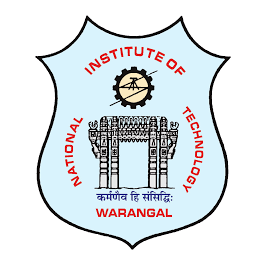
SUMMER INTERNSHIP

DOCUMENTATION ON MACHINE LEARNING

AT

NATIONAL INSTITUTE OF TECHNOLOGY

WARANGAL



**Under the guidance of**  **Submitted By**

**Dr.J. Ravikumar sir**  J.Sujay

**(Professor)**  T.Gopichand

T. Ravali

**Breast Cancer Image Classification: Binary and Multiclass Approaches for Accurate Diagnosis**

**1. Abstract**

Breast cancer, a leading cause of death in women, requires early detection for improved survival rates. However, subjective analysis by pathologists poses challenges in accurately identifying cancerous tissue in biopsy samples. To address this, we propose a breast cancer image classification system that leverages machine learning and computer vision for objective and accurate diagnosis. Our aim is to revolutionize breast cancer detection and contribute to saving lives through early intervention. For this we worked on two datasets namely BreakHis and CBIS-DDSM with the help of deep Convolution Neural networks (CNN).

**2. Introduction**

Breast cancer is a prevalent form of cancer that affects millions of women worldwide. Early and accurate detection is crucial for successful treatment and improved patient outcomes. In this project, we aimed to develop a Multiclass Classification System for Breast Cancer images. The objective was to create a model capable of distinguishing between different types of breast cancer using image analysis techniques and machine learning algorithms.

In this project, our objective was to develop a multiclass classification system for breast cancer images using Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN). We initially worked with a CSV dataset (Wisconsin), achieving high accuracies, and then transitioned to image datasets, specifically CBIS-DDSM and BreakHis, to further evaluate our models.

***3. Dataset description***

***3.1 Wisconsin Dataset (CSV)***

We began our project by working with a CSV dataset containing relevant features extracted from breast cancer images. This dataset provided valuable information for training our ANN and CNN models initially. The CSV dataset likely contained features such as texture, shape, and statistical measures derived from mammograms, ultrasound scans, or histopathology images.

|  |  |  |  |
| --- | --- | --- | --- |
| Optimizer | Epochs | Accuracy (%) | |
| Relu, Sigmoid | Tanh, Sigmoid |
| SGD | 50 | 96.49 | 97.368 |
| Adagrad | 50 | 58.77 | 94.736 |
| RMSProp | 50 | 96.49 | 99.122 |
| Adam | 50 | 95.61 | 96.49 |

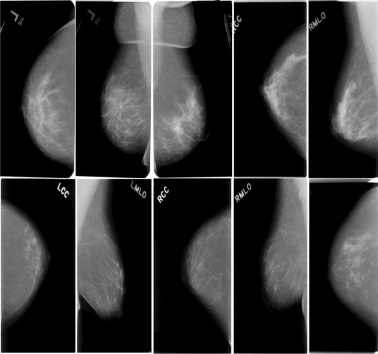
The model was trained with different hyper parameters and the metrics are shown in Table-2.

**Table-2:** Performance Comparison of Different Optimizers, Epochs, and Activation Functions on Breast Cancer Image Classification

*Note:* The table presents a comparison of the performance of different optimizers, varying numbers of epochs, and different activation functions on breast cancer image classification. The accuracy values represent the achieved accuracies on the respective configurations.

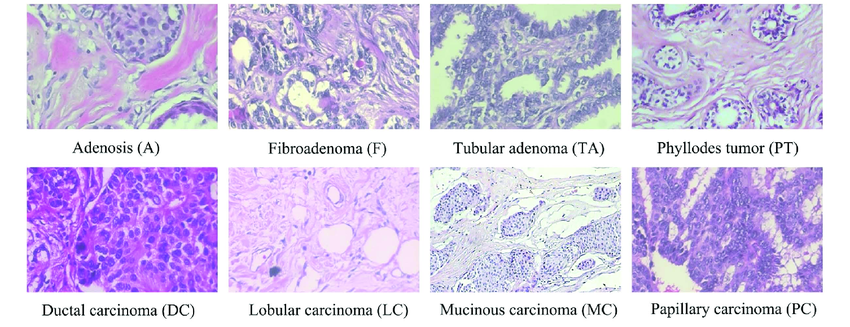
***3.2 CBIS-DDSM Dataset***

To further validate our models, we utilized the CBIS-DDSM image dataset, which is widely used in breast cancer research. This dataset consists of mammograms with annotated labels for different breast cancer subtypes. Our focus was on binary image classification, aiming to distinguish between cancerous and non-cancerous images.



***3.3 BreakHis Dataset***

The BreakHis dataset, which is a publicly available resource of breast cancer Histopathology images. This dataset has been developed in association with the P&D Laboratory – Pathological Anatomy and Cytopathology, Parana, Brazil.



BreakHis dataset contains 7909 samples of microscopic Histopathology images of the breast tissue captured at different magnification levels of 40×, 100×, 200×, and 400×. These samples are collected from 82 patients among which 2480 samples are benign and 5429 are malignant types respectively. Benign type is further divided into four sub-types including Adenosis (A), Fibroadenoma (F), Phyllodes Tumor (PT), and Tubular Adenoma (TA). On the other hand, malignant type is also divided into four sub-types including Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC), and Papillary Carcinoma (PC). Table-1 shows the detailed distribution of images in this dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classes** | **Subtypes** | **Magnification** | | | | **Total** |
| **40x** | **100x** | **200x** | **400x** |
| **Benign(B)** | Adenosis  (A) | 114 | 113 | 111 | 106 | 444 |
| Fibroadenoma  (F) | 253 | 260 | 264 | 237 | 1,014 |
| Phyllodes Tumor  (PT) | 109 | 121 | 108 | 115 | 453 |
| Tubular Adenoma (TA) | 149 | 150 | 140 | 130 | 569 |
| **Malignant(M)** | Ductal Carcinoma (DC) | 864 | 903 | 896 | 788 | 3,451 |
| Lobular Carcinoma (LC) | 156 | 170 | 163 | 137 | 626 |
| Mucinous Carcinoma (MC) | 205 | 222 | 196 | 169 | 792 |
| Papillary Carcinoma (PC) | 145 | 142 | 135 | 138 | 560 |
| Total | | 1,995 | 2,082 | 2,013 | 1,820 | 7,909 |

**Table-1:** Detailed Distribution of BreakHis Dataset

**What is Convolution Neural Network(CNN) ?**

CNN is a class of deep neural networks that extracts features from images, given as input, to perform specific tasks such as image classification, face recognition and semantic image system. A CNN has one or more convolution layers for simple feature extraction, which execute convolution operation (i.e. multiplication of a set of weights with input) while retaining the critical features (spatial and temporal information) without human supervision.

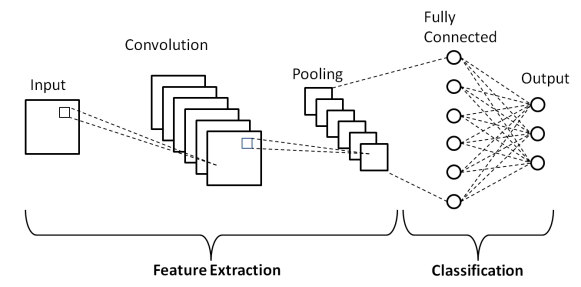
**Why do we need CNN over ANN?**

CNN is needed as it is an important and more accurate way for image classification problems. With Artificial Neural Networks, a 2D image would first be converted into a 1-dimensional vector before training the model.

Also, with an increase in the size of the image, the number of training parameters would increase exponentially, resulting in loss of storage. Moreover, ANN cannot capture the sequential information required for sequence data.

Thus, CNN would always be a preferred way for dealing with 2D image classification problems because of its ability to deal with images as data, thereby providing higher accuracy.

**The Architecture of CNN:**



The three primary layers that define the structure of a convolutional neural network are:

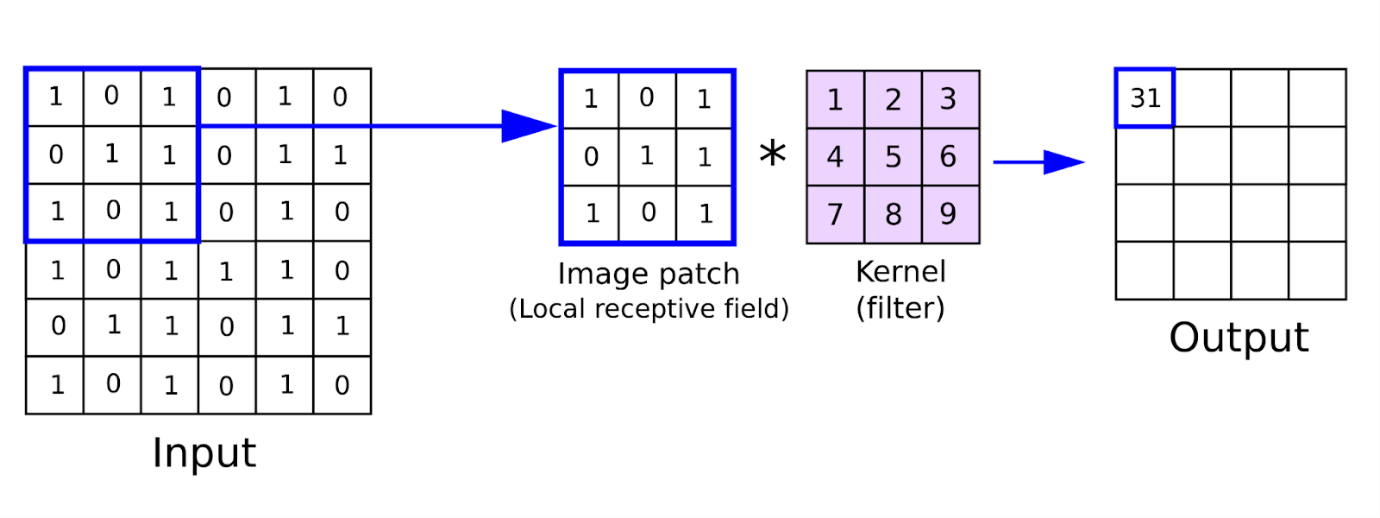
**1)Convolution layer:**

This is the first layer of the convolutional network that performs feature extraction by sliding the filter over the input image. The output or the convolved feature is the element-wise product of filters in the image and their sum for every sliding action.

The output layer, also known as the feature map, corresponds to original images like curves, sharp edges, textures, etc.

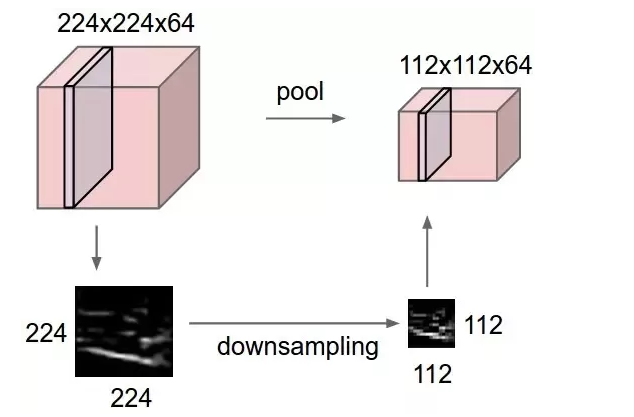
In the case of networks with more convolutional layers, the initial layers are meant for extracting the generic features while the complex parts are removed as the network gets deeper.

The image below shows the convolution operation

**2)Pooling Layer:**

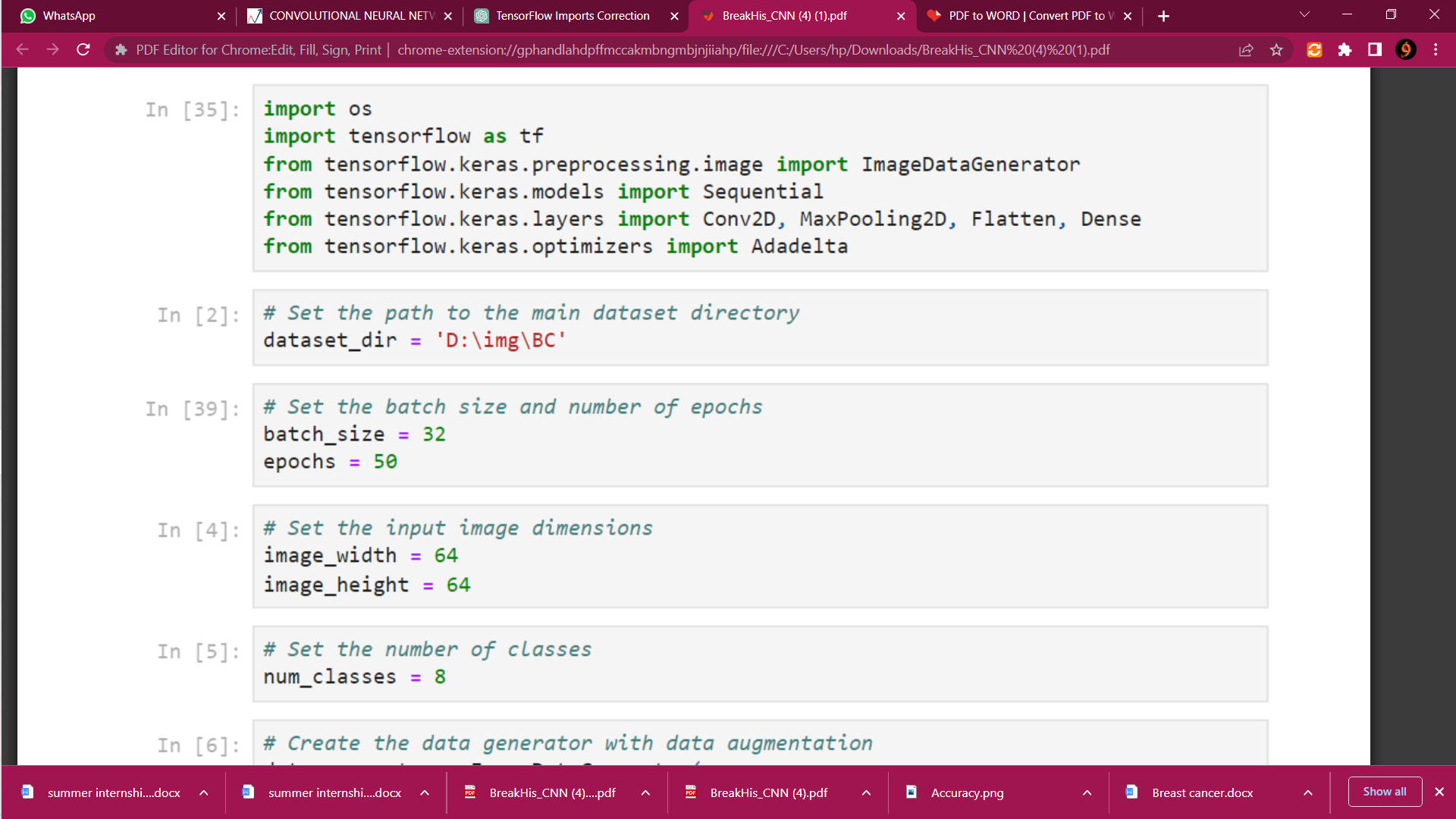
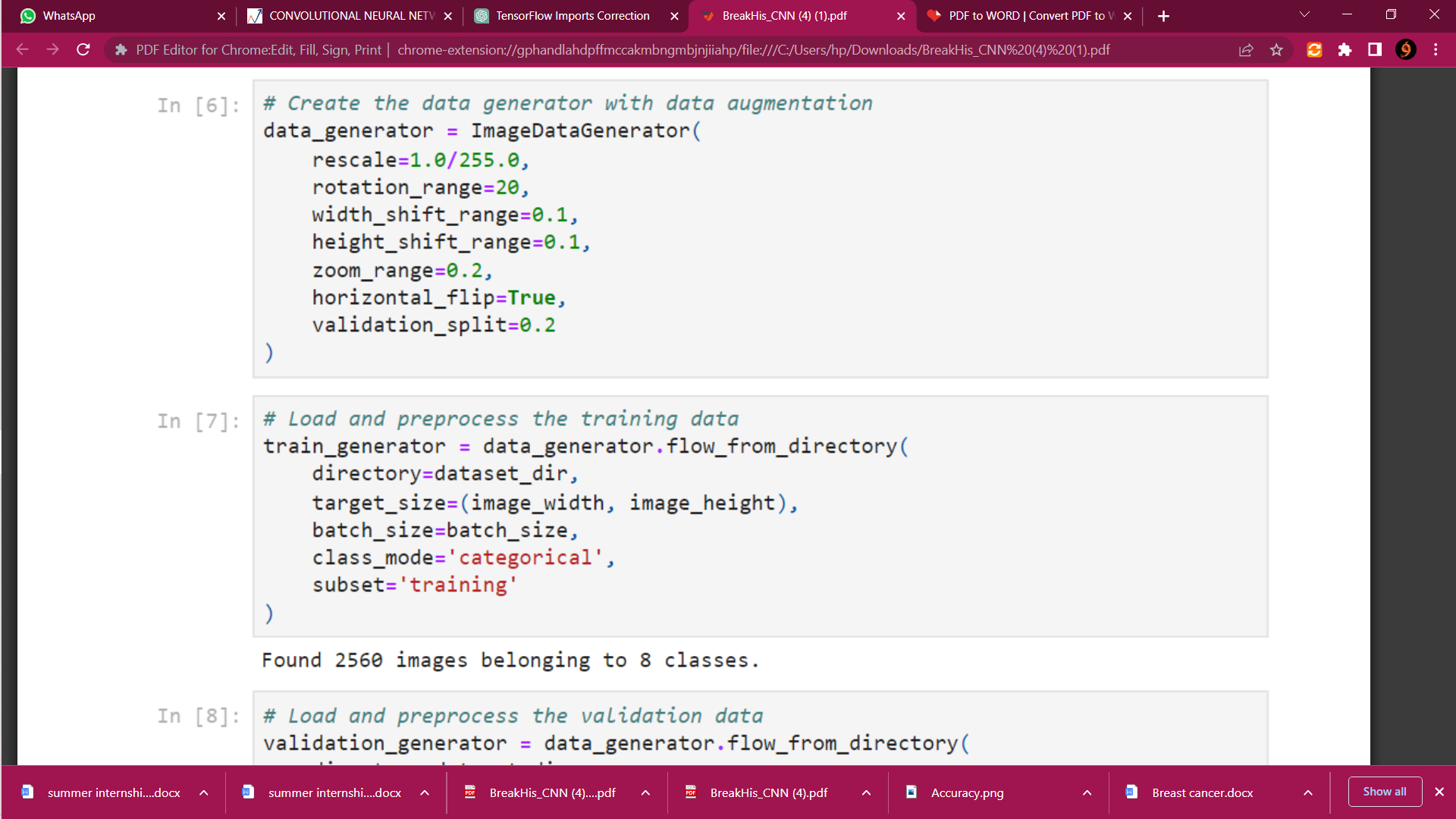
The primary purpose of this layer is to reduce the number of trainable parameters by decreasing the spatial size of the image, thereby reducing the computational cost.

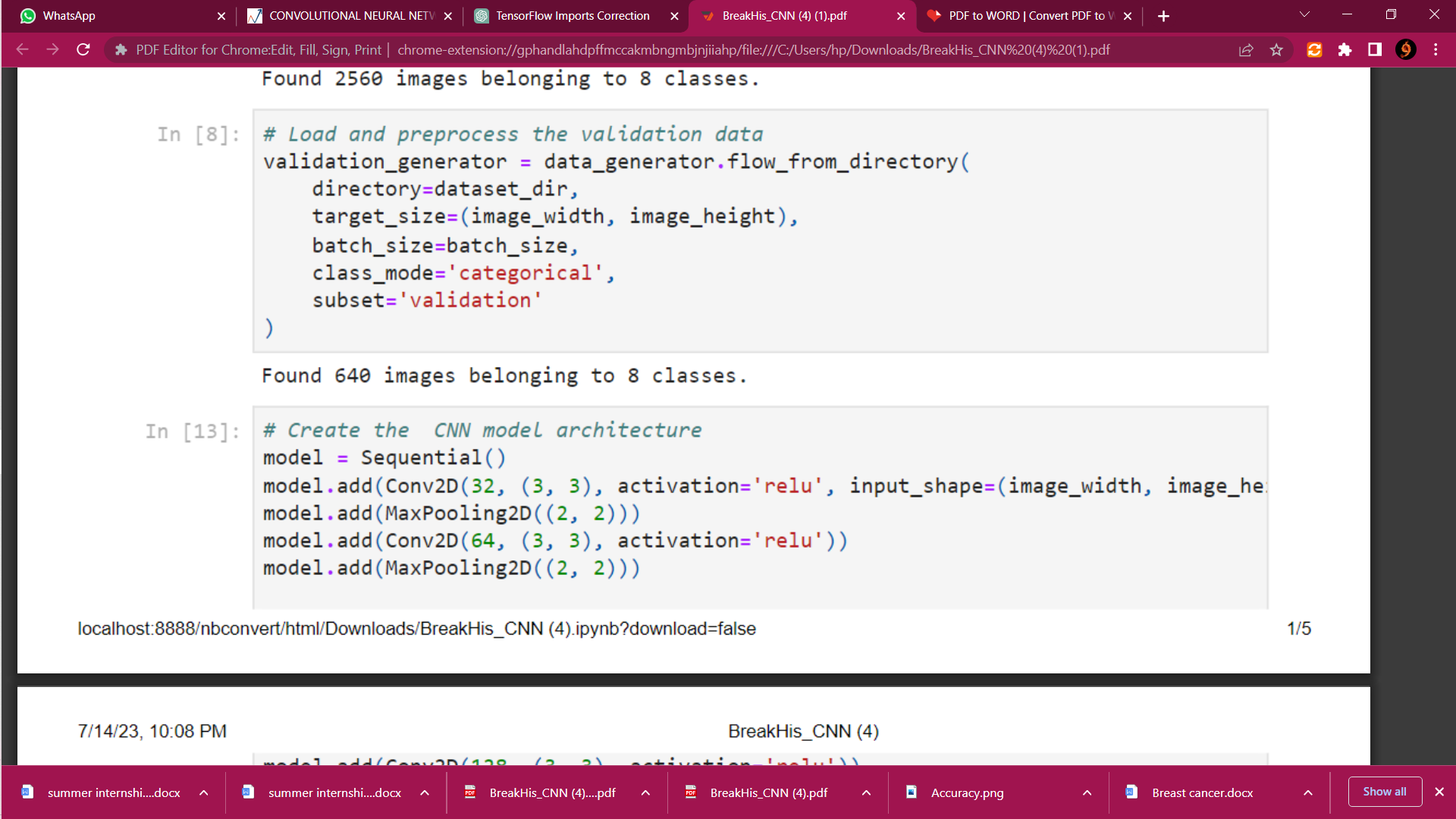
The image depth remains unchanged since pooling is done independently on each depth dimension. Max Pooling is the most common pooling method, where the most significant element is taken as input from the feature map. Max Pooling is then performed to give the output image with dimensions reduced to a great extent while retaining the essential information.

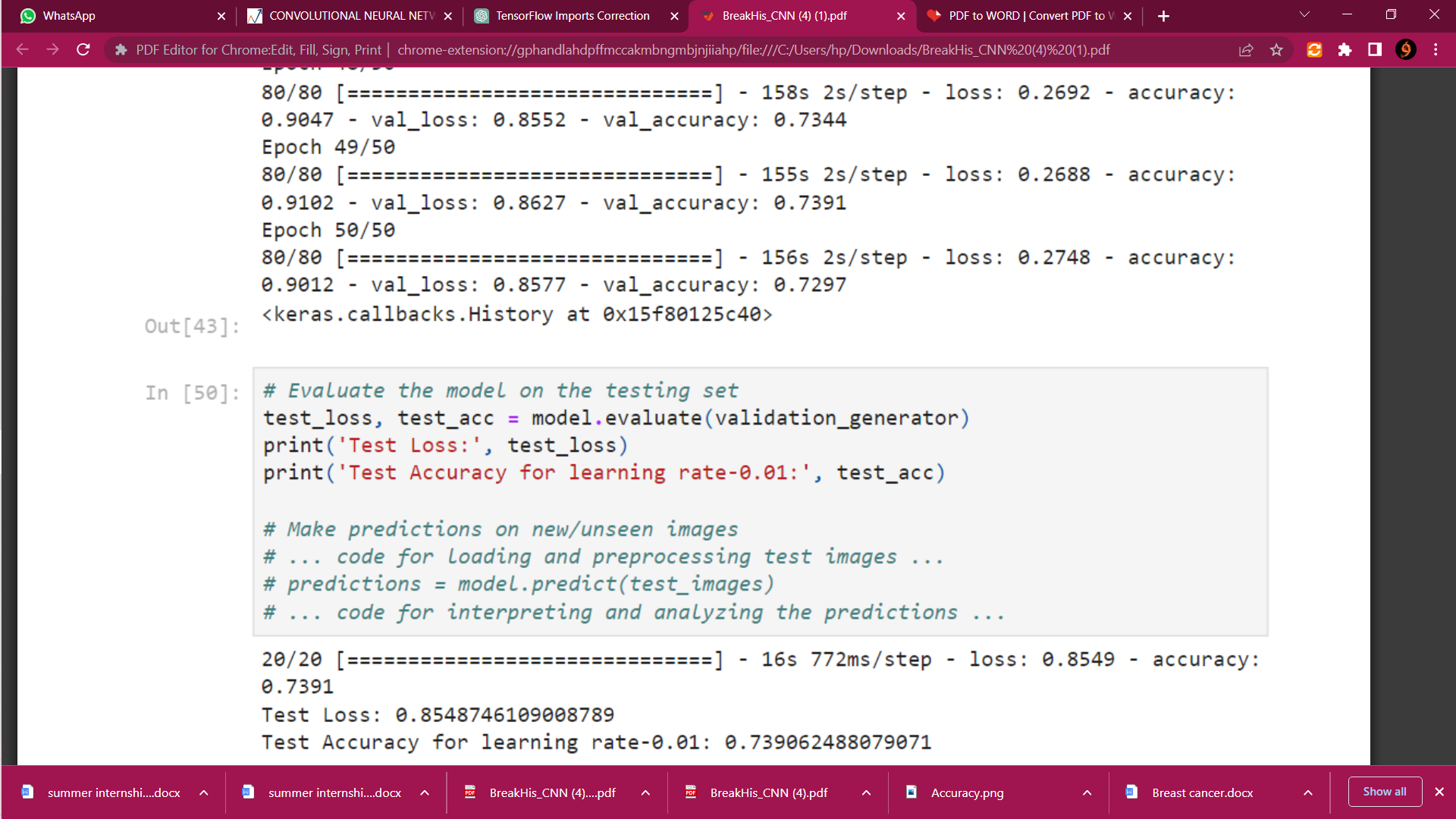


**IMPLEMENTATION:**

Now, let’s implement CNN by **Breast Cancer Multi class Classification** with Break His Dataset .

****

****

****

Here ,the model is trained using the BreakHis dataset. The model is trained with a batch size of 32 and for a total of 50 epochs. The dimensions of the input images are resized to a height and width of 64 pixels.

To improve the model's performance and generalize better, data augmentation techniques are applied. Data augmentation helps increase the diversity of the training data by applying random transformations such as rotations, shifts, flips, and zooms to the images.

The activation function used in the model is ReLU (Rectified Linear Unit), which introduces non-linearity into the network and helps in learning complex patterns.

The model is trained using the Adadelta optimizer with a learning rate of 0.01. Adadelta is an optimization algorithm that dynamically adapts the learning rate based on the gradients' magnitudes. It helps in avoiding the need for manual tuning of the learning rate.

After training the model with the specified settings, the achieved accuracy is reported as 73.90%. This accuracy represents the model's ability to correctly classify the breast tissue samples into the eight different classes based on the histopathology images.

**4. Model Development and Evaluation:**

**a) ANN and CNN Models:**

For the initial stages of our project, we employed ANN and CNN models to train on the CSV dataset. The ANN model achieved an impressive accuracy of **99.122%**, while the CNN model attained 92.98% accuracy, indicating the effectiveness of deep learning techniques for breast cancer classification.

**b) CBIS-DDSM Dataset:**

We then transitioned to the CBIS-DDSM dataset for binary image classification. By fine-tuning our CNN model, we achieved a remarkable accuracy of **98.83%** in distinguishing between cancerous and non-cancerous mammograms. This demonstrated the suitability of CNNs for breast cancer image classification tasks.

**Related Work:**

In our project, we first focused on binary image classification of the BreakHis dataset, specifically differentiating between malignant and benign breast cancer cases. After achieving an accuracy of 82.17% using a deep CNN and the Adam optimizer with a learning rate of 0.001, we explored additional techniques and platforms to further evaluate our models.

One such platform we utilized was Edge Impulse, which allowed us to work with a limited number of images from each class (a total of 8 classes). We experimented with different hyper parameters and varying numbers of epochs to optimize the performance of our models. Through these iterations, we obtained an accuracy of **69.4%** on the binary classification task.

For 84 Images from each class (totally 8 classes)

|  |  |  |
| --- | --- | --- |
| Epochs | Learning Rate () | Accuracy (%) |
| 10 | 0.0001 | 43.5 |
| 50 | 0.0001 | 50 |
| 100 | 0.0001 | 55.6 |
| 50 | 0.001 | 59.3 |

|  |  |  |  |
| --- | --- | --- | --- |
| Architecture | Epochs | Learning Rate () | Accuracy (%) |
| MobileNet-v2 | 20 | 0.0001 | 48.1 |
|  | 20 | 0.0001 | 69.4 |
| EfficientNet | 50 | 0.0001 | 69.4 |
|  | 50 | 0.001 | 26.9 |
|  | 100 | 0.00001 | 22.2 |

|  |  |  |  |
| --- | --- | --- | --- |
| Architecture | Epochs | Learning Rate () | Accuracy (%) |
| MobileNet-v2 | 50 | 0.0005 | 37.5 |
| 10 | 0.0005 | 31.3 |
| 50 | 0.0005 | 42.2 |

Following the binary classification phase, we transitioned to the next stage of our project, which involved multiclass classification of breast cancer images. This task required the classification of images into multiple subtypes or classes. To accomplish this, we leveraged the power of CNNs, as they have demonstrated remarkable capabilities in image classification tasks.

By utilizing a deep CNN architecture and following the methodologies outlined in relevant research papers, we developed a model capable of accurately classifying breast cancer images into their respective subtypes. The accuracy achieved on the multiclass classification task using the CNN model was **73.90%.**

It is important to note that due to the limited availability of images for each class in the Edge Impulse platform, the accuracy obtained for both binary and multiclass classification may be slightly lower compared to working with larger, more diverse datasets. Nonetheless, the results obtained demonstrate the potential of utilizing CNNs and advanced techniques in breast cancer image analysis.

In summary, we initially performed binary image classification on the BreakHis dataset, achieving an accuracy of 82.17% using a deep CNN and the Adam optimizer with a learning rate of 0.001. We then experimented with Edge Impulse, achieving an accuracy of 67% by changing hyper-parameters and epochs. Finally, we progressed to multiclass classification of breast cancer images using a CNN, obtaining an accuracy of 73.90%. These findings contribute to the field of breast cancer image analysis and provide insights into the performance of models under different experimental conditions and platforms.