BREAST CANCER DETECTION USING NEURAL NETWORKS

INT 246 CA-1

Submitted by

MANDAPATI VENKATA SAI SUDARSANA VARMA

Registration No : 11805129

Programme Name : Btech. CSE

School of Computer Science & Engineering

Lovely Professional University, Phagwara

ABSTRACT

It is important to detect breast cancer as early as possible. Breast cancer is the most commonly occurring cancer in women and the second most common cancer overall. There were over 2 million new cases in 2018, making it a significant health problem in present days.

Developed a sequential CNN using keras API using the dataset from Kaggle. The accuracy of the new trained model is 96% and accuracy against the validation dataset is 97.3%. So the model performed quite well.

INTRODUCTION

The key challenge in breast cancer detection is to classify tumors as malignant or benign. Malignant refers to cancer cells that can invade and kill nearby tissue and spread to other parts of your body. Unlike cancerous tumor(malignant), Benign does not spread to other parts of the body and is safe somehow. Deep neural network techniques can be used to improve the accuracy of early diagnosis significantly.

A mass can be either benign or malignant. The difference between benign and malignant tumors is that the benign tumors have round or oval shapes, while malignant tumors have a partially rounded shape with an irregular outline. In addition, the malignant mass will appear whiter than any tissue surrounding it

Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called an **artificial neural network.**

A **Convolutional Neural Network (CNN)** is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms.

METHODOLOGY

A sequential model in Keras is used to build the model. It allows us to build the model layer by layer. Each layer has weights that correspond to the layer the follows it. We add the layers at each layer using add() function.

Also used 1D Convolution Layer Conv1D(), this layer is very effective for deriving features from a fixed-length segment of the overall dataset.

In the first Conv1D Layer, learning a total of 36 filters with size of the convolutional window as 3.

Used ReLu activation function. The rectified Linear activation function or ReLU for short is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero.

This function returns 0 if it receives any negative input, but for any positive value x it returns the value back.

Flattening is converting the data into a 1-D array for inputting it to the next layer. flattened the output of the convolutional layers to create a single long feature vector.

The Sigmoid function takes a value as input and outputs another value between 0 and 1. Used the sigmoid function for the final layer before the output.

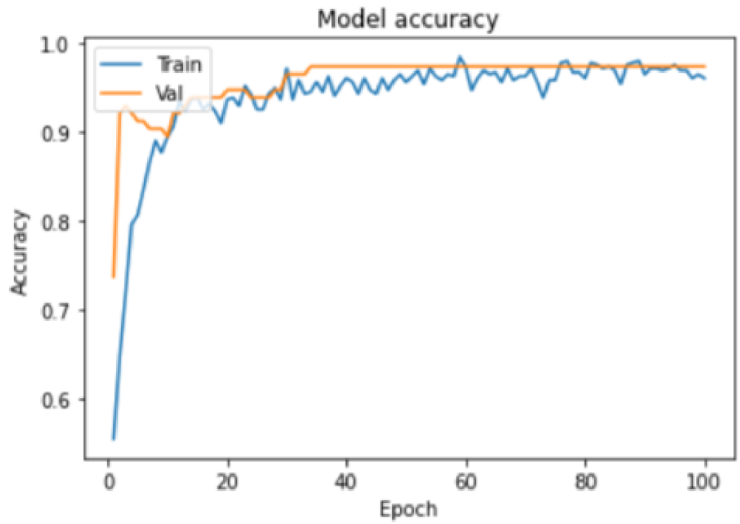
RESULTS

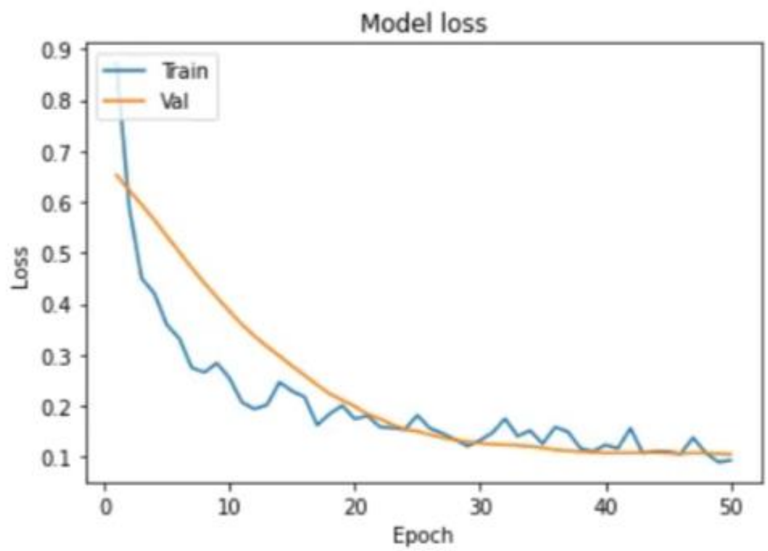
```
Breast cancer wisconsin (diagnostic) dataset
--------------------------------------------

**Data Set Characteristics:**

    :Number of Instances: 569

    :Number of Attributes: 30 numeric, predictive attributes and the class

    :Attribute Information:
        - radius (mean of distances from center to points on the perimeter)
        - texture (standard deviation of gray-scale values)
        - perimeter
        - area
        - smoothness (local variation in radius lengths)
        - compactness (perimeter^2 / area - 1.0)
        - concavity (severity of concave portions of the contour)
        - concave points (number of concave portions of the contour)
        - symmetry
        - fractal dimension ("coastline approximation" - 1)

        The mean, standard error, and "worst" or largest (mean of the three
        largest values) of these features were computed for each image,
        resulting in 30 features.  For instance, field 3 is Mean Radius, field
        13 is Radius SE, field 23 is Worst Radius.

        - class:
                - WDBC-Malignant
                - WDBC-Benign

:Class Distribution: 212 - Malignant, 357 - Benign
```

This is the data from the summary of the dataset loaded.

The data was split into 80-20 ratio for training and testing datasets.

To train the CNN, the maximum number of Epochs was set to 100. As I found that the at 100 the accuracy is more.





In Model accuracy graph, validation accuracy is always greater than train accuracy that means our model is not overfitting.

In Model accuracy graph, validation loss is also very lower than training loss so unless and until validation loss goes above than the training loss than we can keep training our model.