C K PITHAWALA COLLEGE OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER ENGINEERING

BREAST CANCER DETECTOR

Group No : 11

TERM : EVEN 2020-21

TEAM DETAIL

- **Students Detail**

Enrollment No	Name of the Student
170090107 005	Jainesh K. Doshi
170090107 049	Devansh V. Shah
170090107 050	Dhruvil N. Shah
170090107 051	Keneel C. Shah

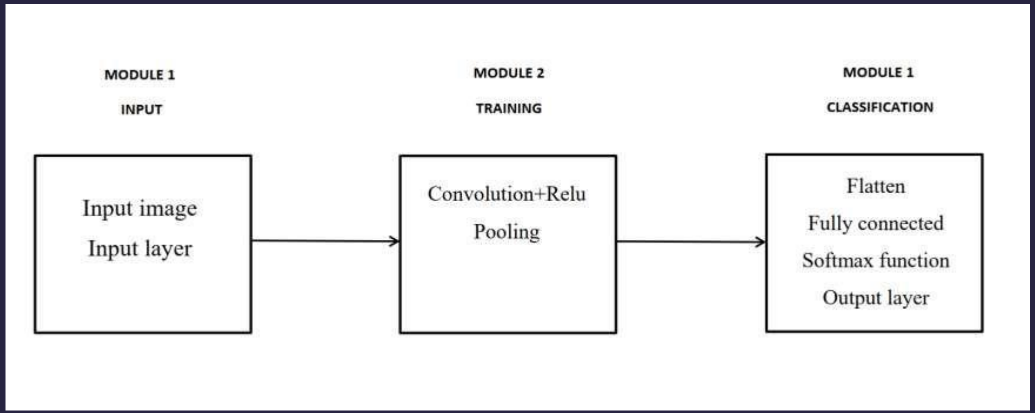
- **Group No** : 11
- ***Project Id*** *: (120711)*
- **Guide** : Prof. Neelam Surti
- **Co-Guide** : Prof. Vishruti Desai

ABSTRACT:

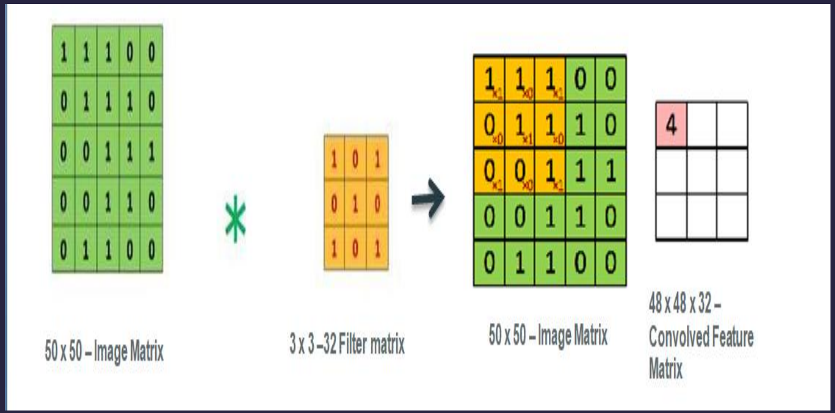
The subject disclosure presents systems and computer-implemented methods for assessing the risk of cancer recurrence in a patient based on the basis of IDC: invasive ductal carcinoma vs non-IDC: invasive ductal carcinoma images, on Mammography images like Benign vs Normal vs Malignant and on Mammography images Fatty vs Fatty-Glandular vs Dense-Glanular together and on Numeric based data for benign vs malignant. These classifications can be done on the basis of the method selection for breast cancer detection. Our aim is to classify cancerous cells from the IDC images, Mammography images, and numeric data so that different types of models can be trained which can provide the different types of facilities in breast cancer detection. Within the opening move, we analyzed the photographs and appearance at the distribution of the pixel intensities as well as numeric data. Then, the pictures were normalized, and we attempted some very well-developed transfer-learning Convolutional Neural Networks like VGG16, VGG19, InceptionV3, ResNet50, and InceptionResNetV2 on IDC and Mammography based images and Linear, Logistics Regression, SVM-Linear, SVM-RBF, Decision-Tree Classifier, Random-Forest Classifier and GaussianNB for Numeric data. We had then validated and compared each of those models for the finalization of the model for respective methods. These all-neural networks were implemented as a python class and therefore the complete TensorFlow session is often saved to or restored from a file. We then also implement tensor summaries, which were used for the visualizations with TensorBoard. The output layer of the IDC model gave the result in form of IDC-Positive vs IDC-Negative, whereas the output layer of the Mammography model gave the result in form of Benign vs Normal vs Malignant and Fatty vs Fatty-Glandular vs Dense-Glanular together and Benign vs Malignant for Numeric model. So as to stop over-fitting the training data we will generate new images by rotations, translations, and zoom or retrain the whole model for both numeric and image-based datasets.

BREAST CANCER DETECTION

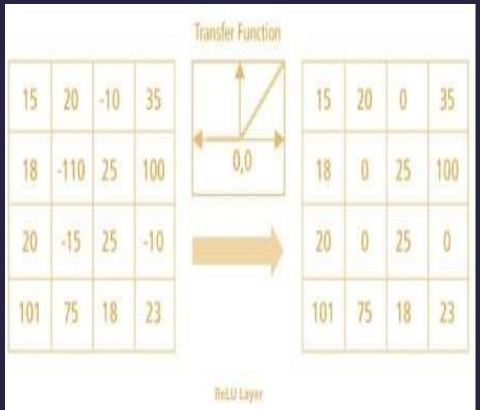
CNN- ARCHITECTURE



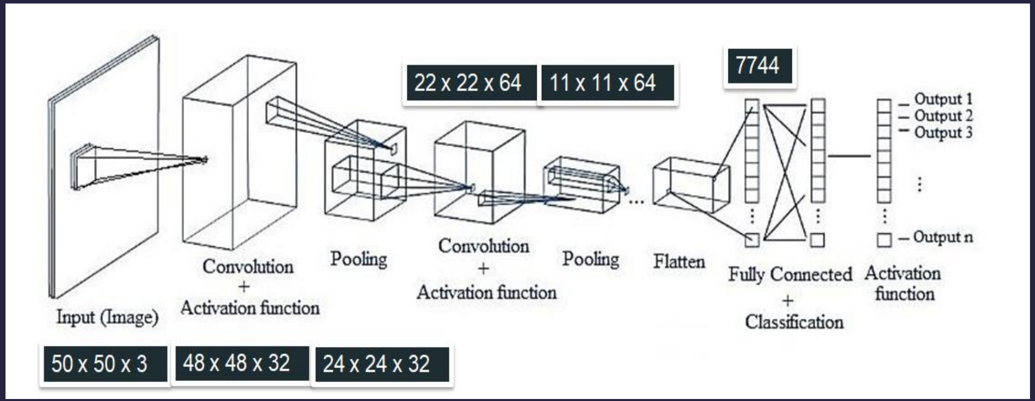
1. **Overview of the Convolutional Neural Network System architecture**



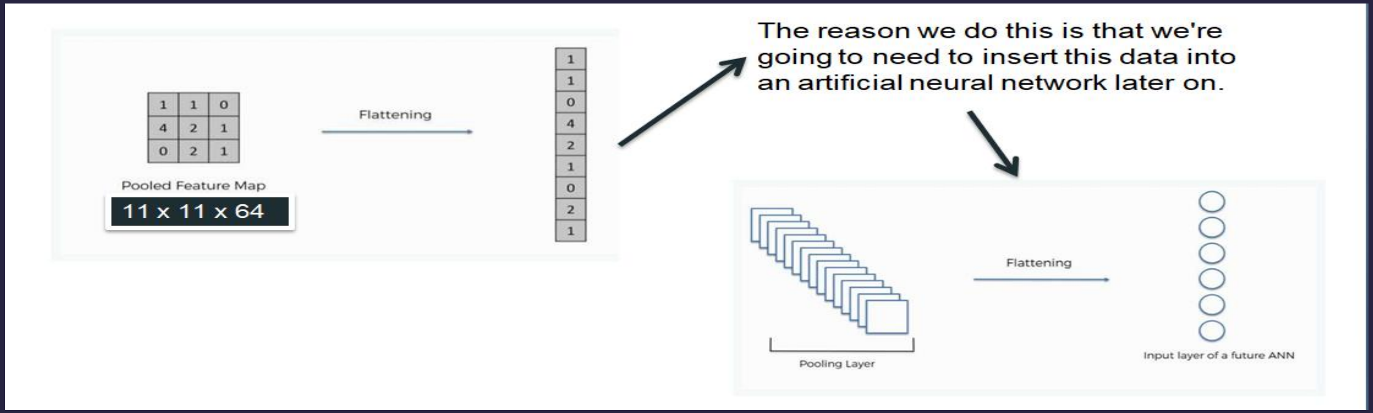
3. **Feature Extraction Process from image using CNN**



RELU Activation Function



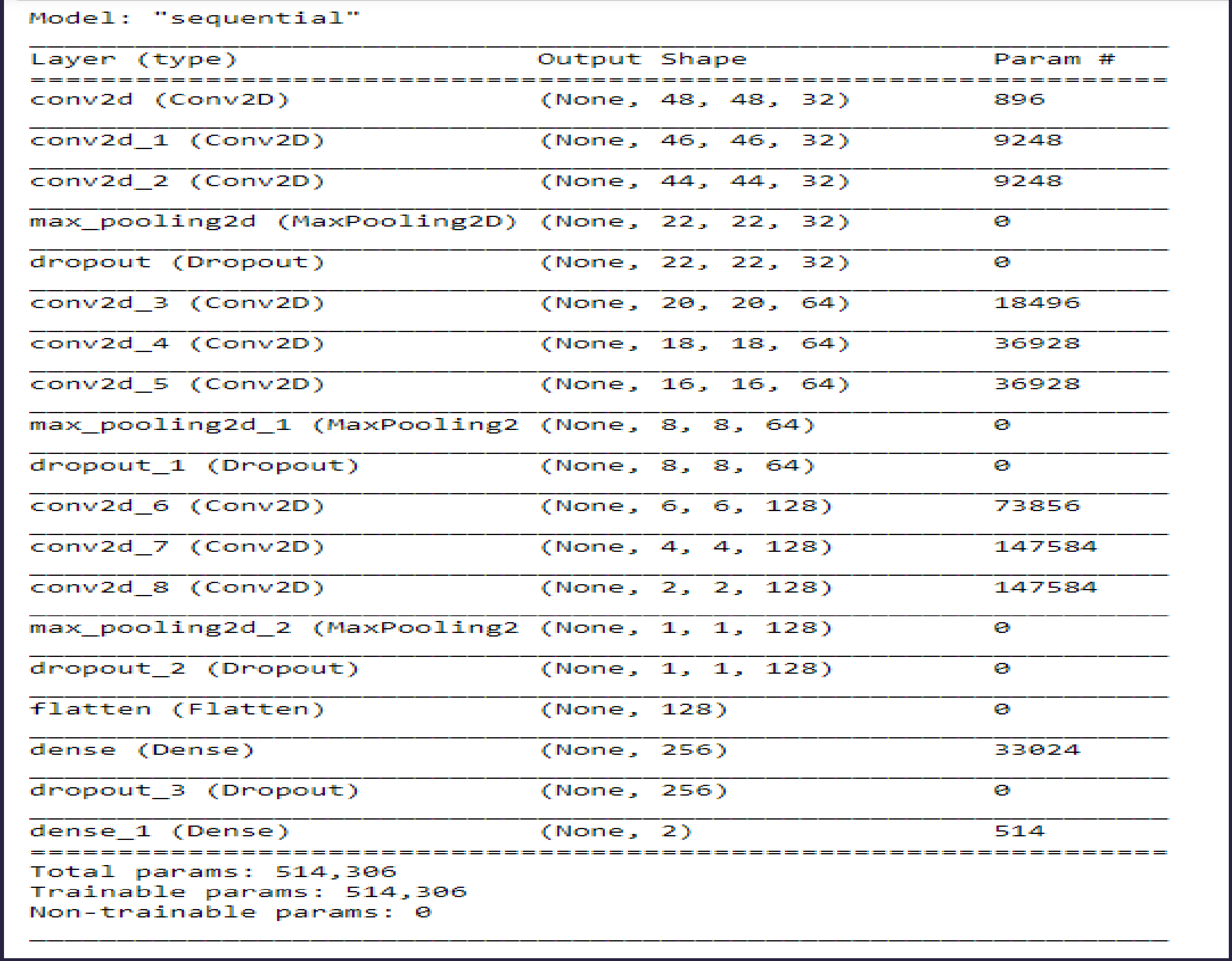
2. **In Depth System Architecture of the CNN with it comprehensive layers**



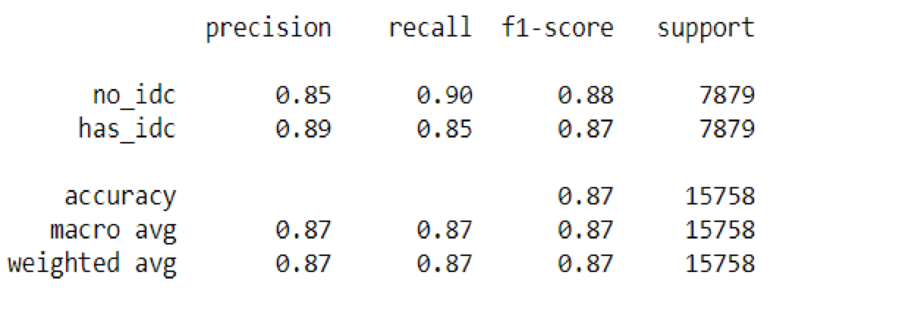
The reason we do this is that we're going to need to insert this data into an artificial neural network later on.

4. **Last Process in CNN – Flattening all Parameters**

9 - LAYER ARCHITECTURE WITH CLASSIFICATION REPORT



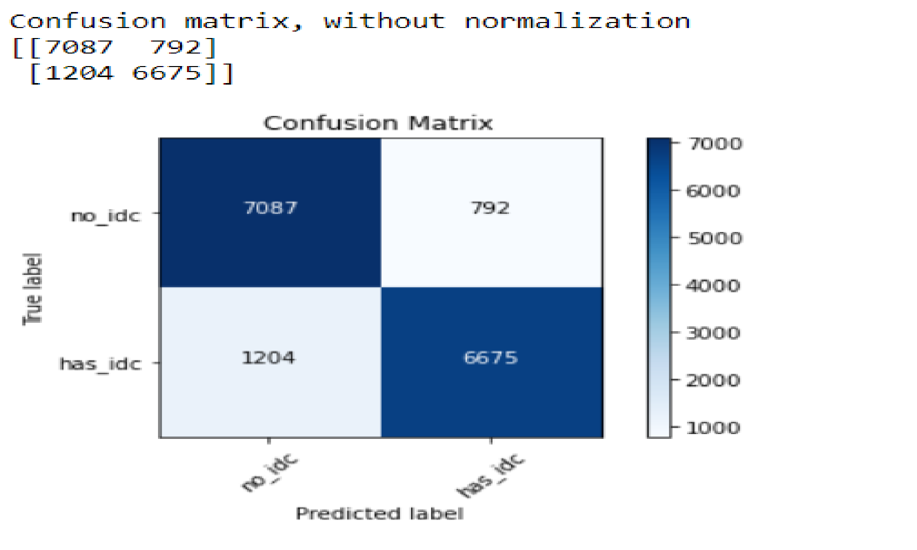
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 48, 48, 32)        896
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 46, 46, 32)        9248
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 44, 44, 32)        9248
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 22, 22, 32)        0
_________________________________________________________________
dropout (Dropout)            (None, 22, 22, 32)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 20, 20, 64)        18496
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 18, 18, 64)        36928
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 16, 16, 64)        36928
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 8, 8, 64)          0
_________________________________________________________________
dropout_1 (Dropout)          (None, 8, 8, 64)          0
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 6, 6, 128)         73856
_________________________________________________________________
conv2d_7 (Conv2D)            (None, 4, 4, 128)         147584
_________________________________________________________________
conv2d_8 (Conv2D)            (None, 2, 2, 128)         147584
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 1, 1, 128)         0
_________________________________________________________________
dropout_2 (Dropout)          (None, 1, 1, 128)         0
_________________________________________________________________
flatten (Flatten)            (None, 128)               0
_________________________________________________________________
dense (Dense)                (None, 256)               33024
_________________________________________________________________
dropout_3 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 514
=================================================================
Total params: 514,306
Trainable params: 514,306
Non-trainable params: 0
_________________________________________________________________
```



```
              precision    recall  f1-score   support

      no_idc       0.85      0.90      0.88      7879
     has_idc       0.89      0.85      0.87      7879

    accuracy                           0.87     15758
   macro avg       0.87      0.87      0.87     15758
weighted avg       0.87      0.87      0.87     15758
```

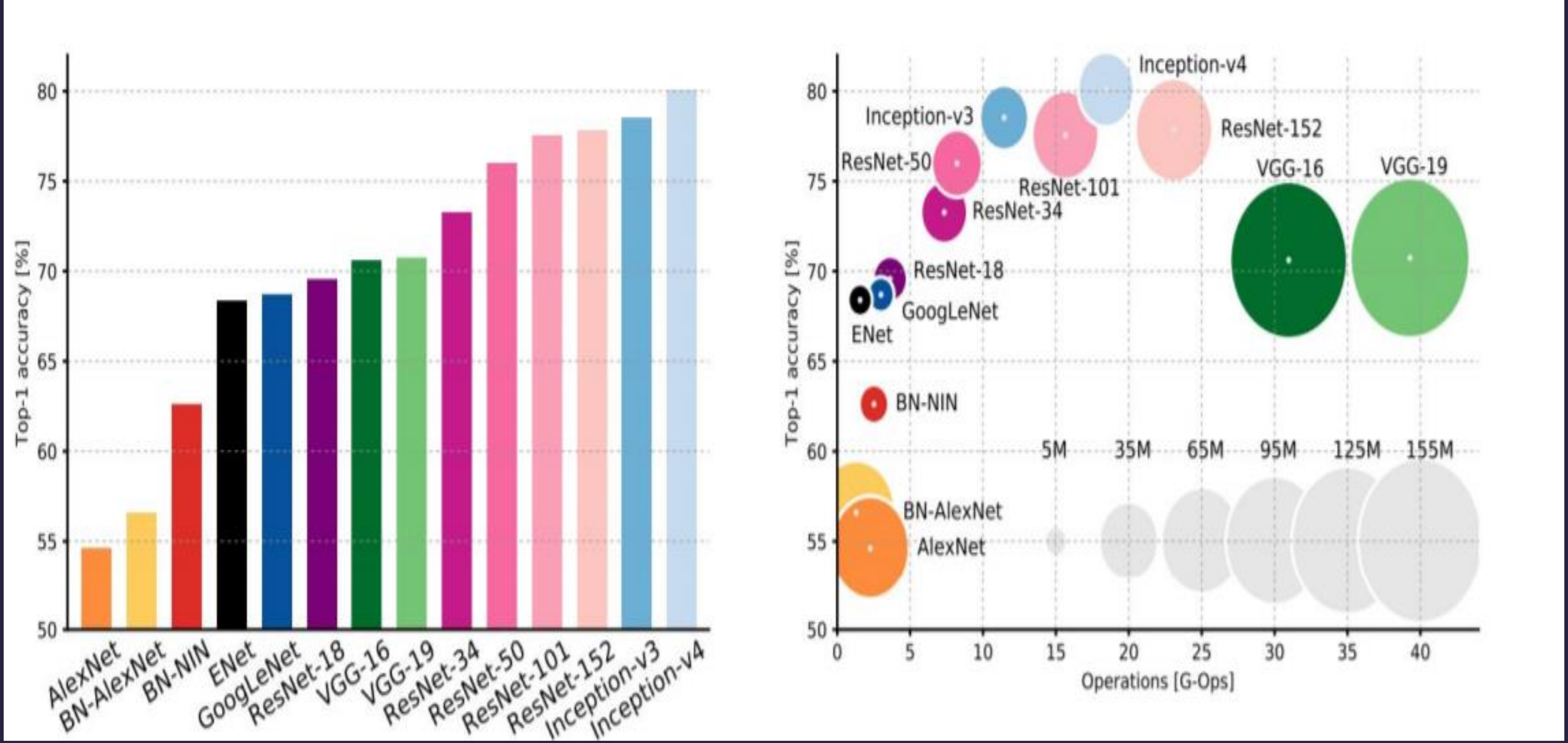


```
Confusion matrix, without normalization
[[7087  792]
 [1204 6675]]
```

Total Testing Images : 15758

True Positive : 7087 → accurate IDC prediction
False Positive : 792 → NON-IDC predicted as IDC
False Negative : 1204 → IDC predicted as NON-IDC
True Negative : 6675 → accurate NON-IDC prediction

SELECTION METHODOLOGY



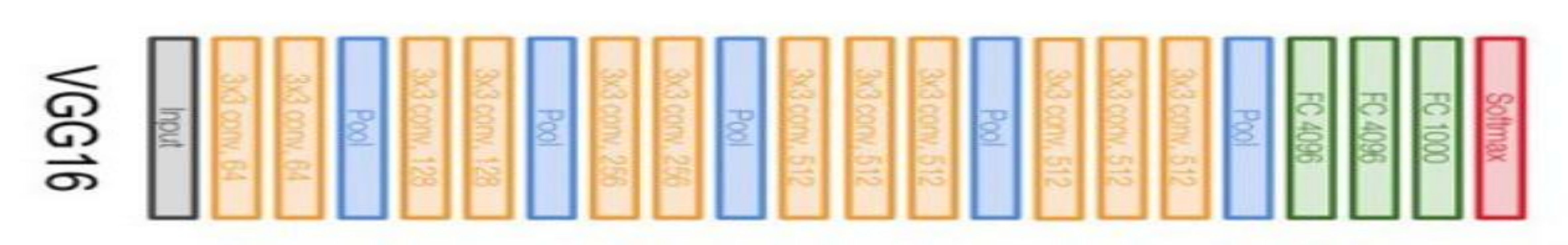
Resource credit (Stanford University) : -
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture9.pdf

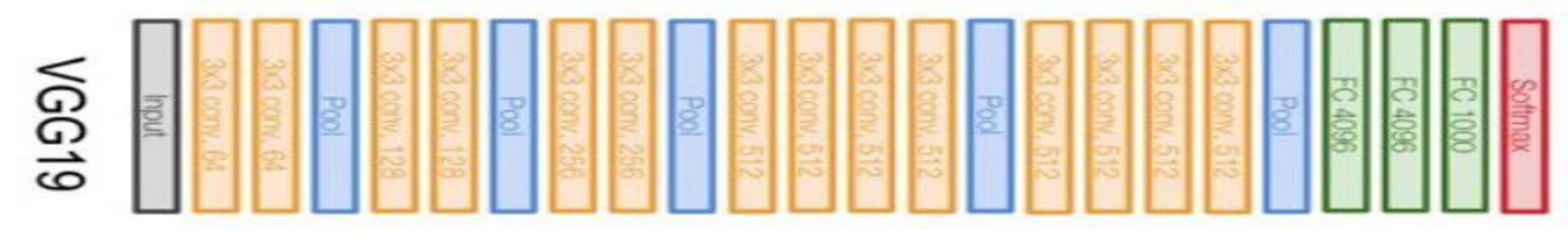
Visual Geometry Group(VGG) - Models

VGG-16 has 13 convolutional and 3 fully-connected layers, it is carrying ReLU activation function with smaller size filters (2×2 and 3×3) in use.

It consists of **138M parameters** and takes up about 500MB of storage space.

VGG-19 is deeper variant of VGG-16.





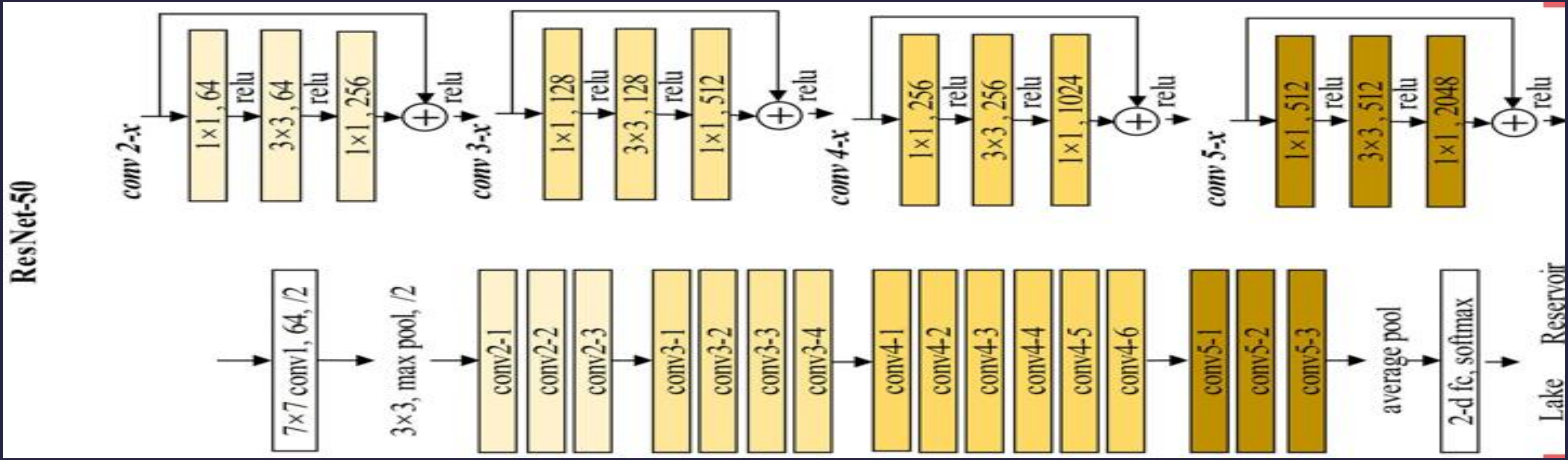
ResNet50

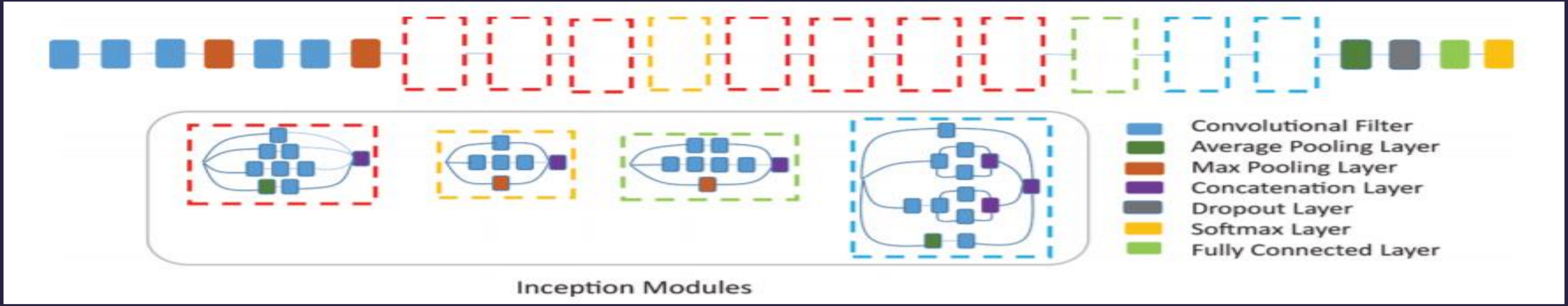
- ResNet was the winner of the 2015 ILSVRC, it is a very deep network with 34–152 layers. This deep CNN model not only avoids the problem of model degradation but also achieves better accuracy.

- In Bottleneck architecture, there are three layers which are 1 × 1, 3 × 3, and 1 × 1 convolution, where the two 1 × 1 layers play the role of reducing and then increasing dimensions, which gives the 3 × 3 layers the smallest input/output dimensions. However, ResNet-50 has fewer filters and less complexity as compared to VGG.

- Many researchers have applied this model and received trustworthy results.

- ResNet50 50 layers consist of 23 million parameters.





Inception & ResNet Combination - Models

- Inception-v3 is a successor to Inception-v1, with **24M** parameters. Inception-v2 was an earlier prototype of v3 hence it’s very similar to v3 but not commonly used.

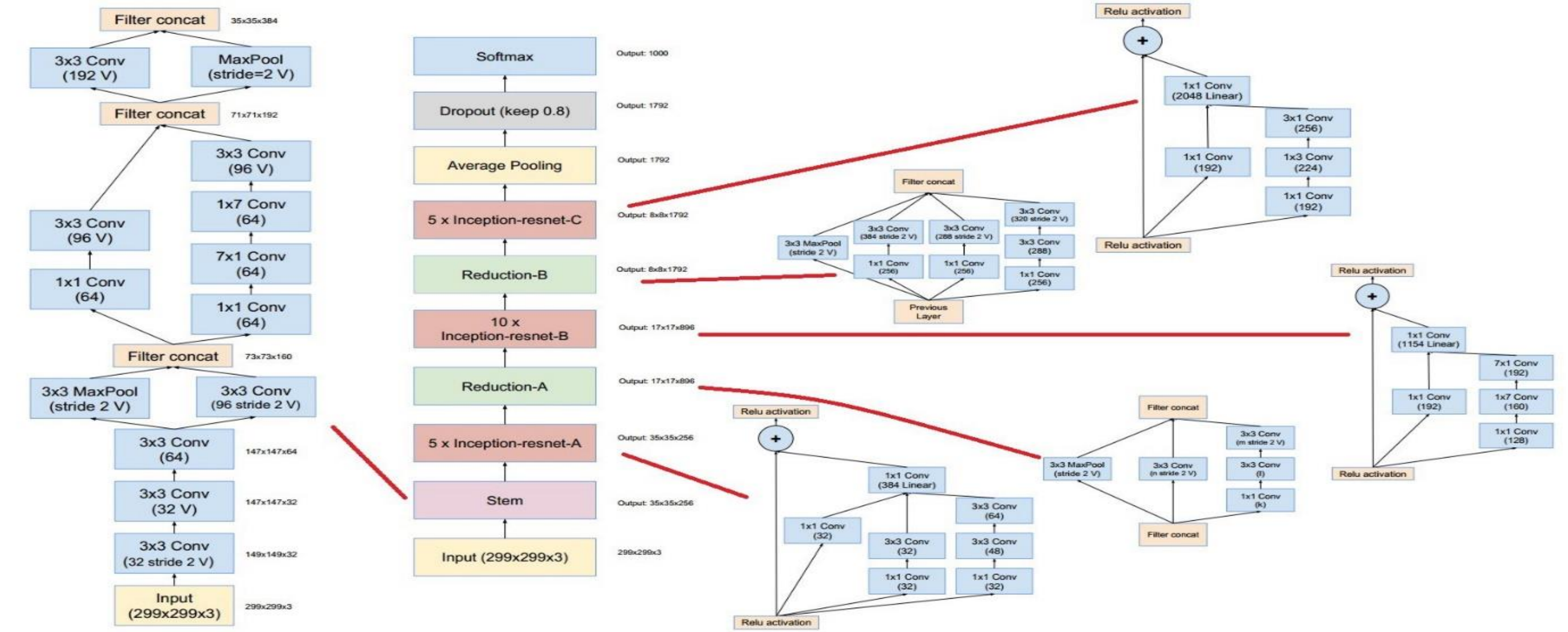
- Inception-v3 is the network that use the optimizer, loss function and adding batch normalization to the auxiliary layers in the auxiliary network

- The motivation for Inception-v2 and Inception-v3 is to drastically reducing the input dimensions of the next layer) and have more efficient computations by using factorization methods.

- Improved from the previous version, Inception-v3.

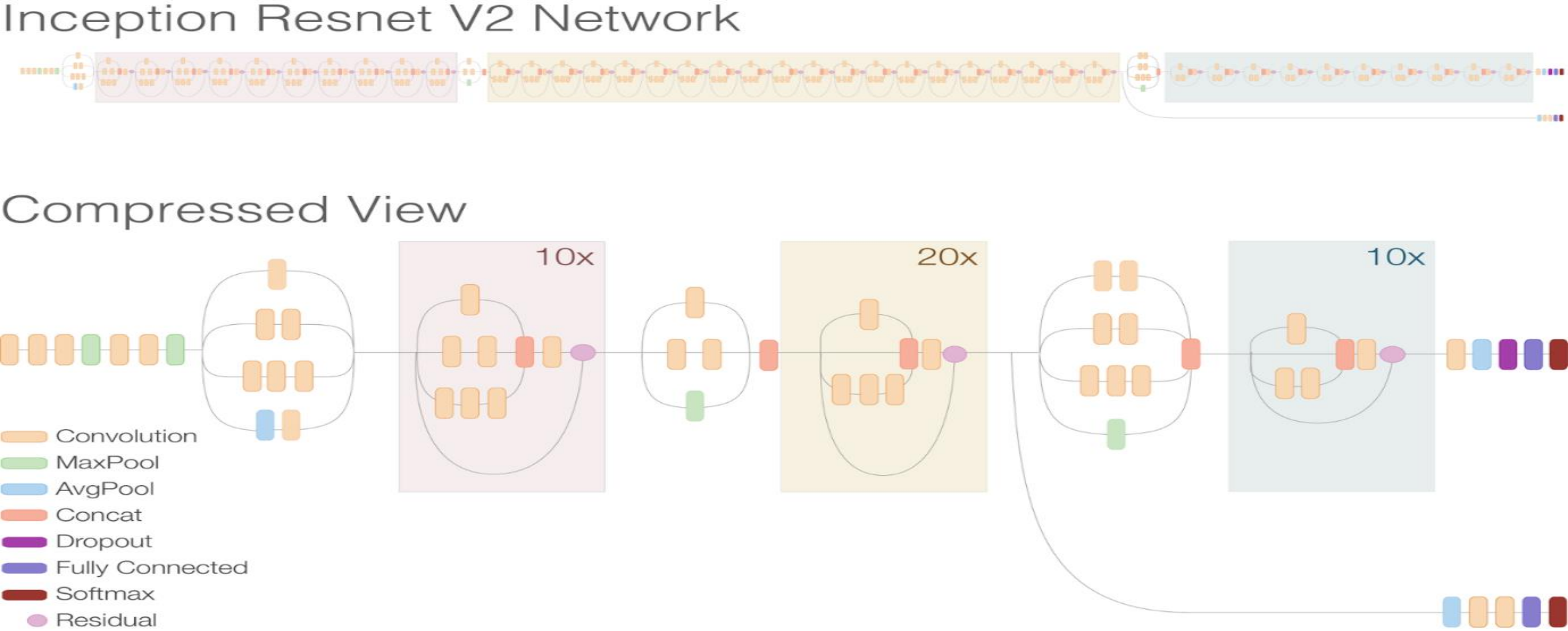
- It has total 48 layers.

Hybrid of Inception net and Residual net



- InceptionResNet-V2 is a type of earlier version of Inception-V3 model which contains some algorithm methods from Microsoft's ResNet. It has 164 total layers due to which it consists of 54 million parameters

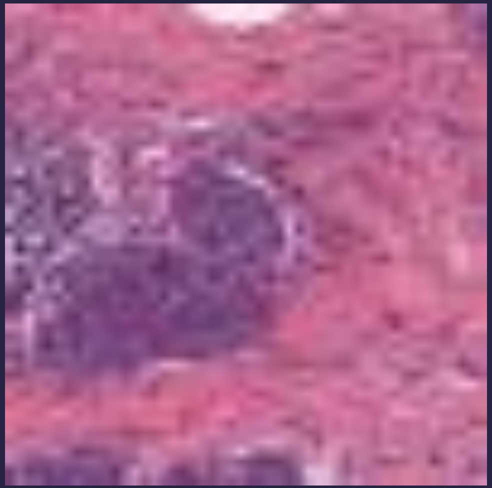
- Each Inception block is followed by a filter expansion layer of 1×1 convolution without RELU activation which is used for increasing the dimensionality of the filters before the addition so that the depth of the input become similar to the output of the previous layers.



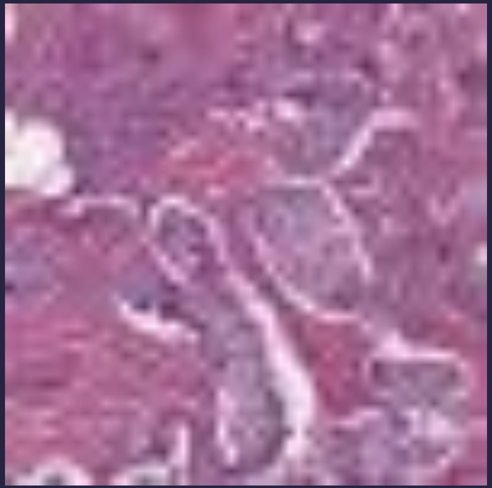
Architecture of InceptionResNetV2

CLASSIFICATION ON

BIOPSY IMAGE OF BREAST IMAGES



Benign



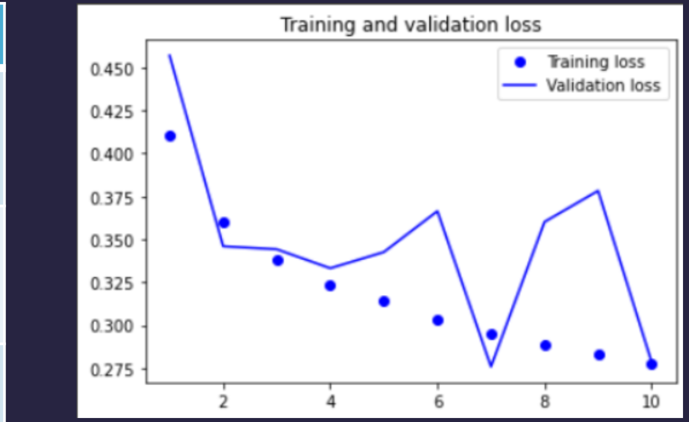
Malignant

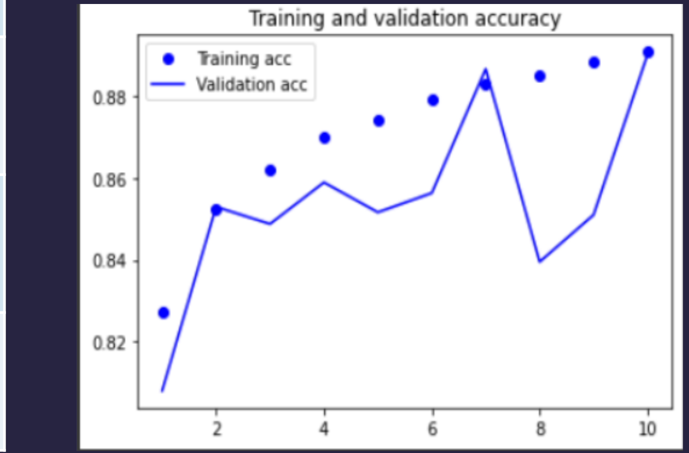
RESULTS ON
BIOPSY IMAGE OF BREAST IMAGES

Models	Recall	Precision	F1-Score	Accuracy	Selection
9- Layer CNN	B-0.90 & M-0.85	B-0.85 & M-0.89	B-0.88 & M-0.87	0.87 ~ 87%	
VGG-16	B-0.84 & M-0.88	B-0.88 & M-0.84	B-0.86 & M-0.86	0.86 ~ 86%	
VGG-19	B-0.93 & M-0.62	B-0.71 & M-0.90	B-0.81 & M-0.73	0.78 ~ 78%	
ResNet50	B-0.94 & M-0.71	B-0.76 & M-0.92	B-0.84 & M-0.80	0.82 ~ 82%	
Iception-V3	B-0.83 & M-0.92	B-0.91 & M-0.84	B-0.87 & M-0.88	0.87 ~ 87%	
InceptionResNet-V2	B-0.87 & M-0.91	B-0.91 & M-0.87	B-0.89 & M-0.89	0.89 ~ 89%	Finalized

Comparisons and Selection of Models on the Basis of F1-Score

(B: Benign & M: Malignant)

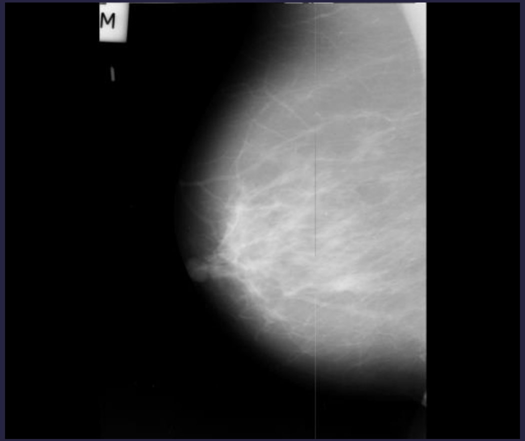




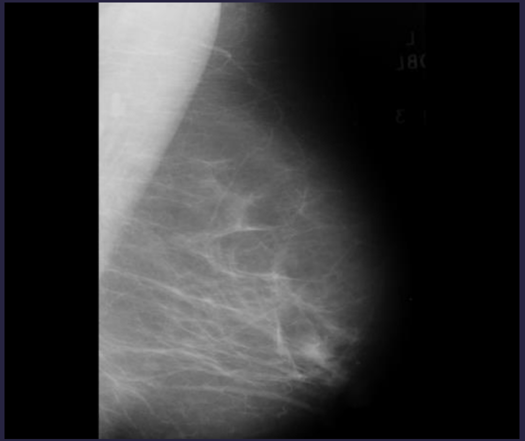
InceptionResNet-V2 provide the 95.56% ROC Curve on the final result

CLASSIFICATION ON

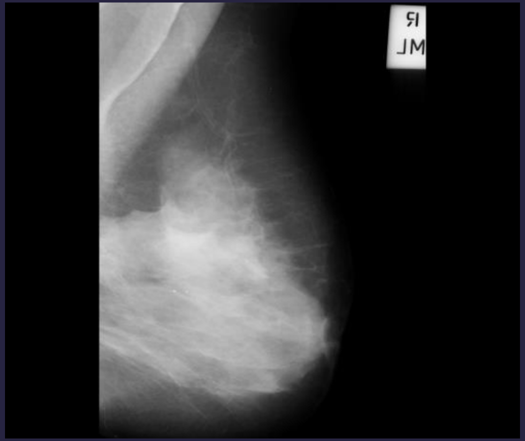
MAMMOGRAPHY IMAGE OF BREAST IMAGES



Fatty-Glandular / Benign



Fatty / Normal



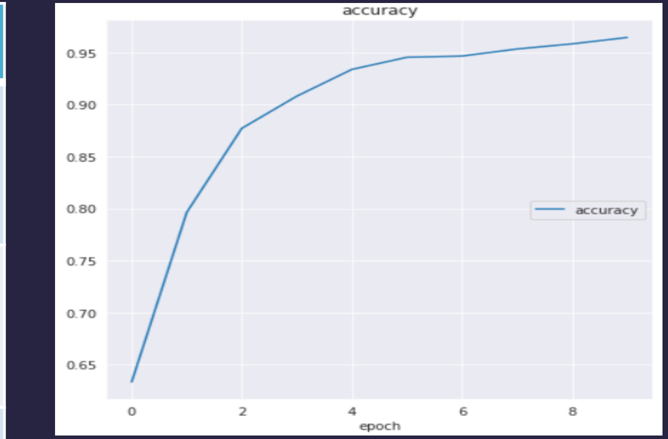
Dense-Glandular / Malignant

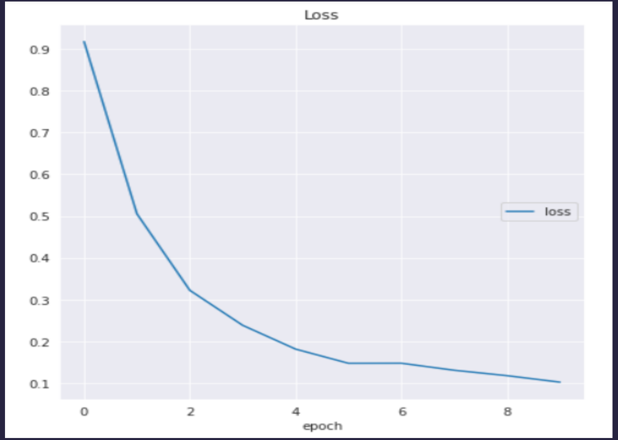
RESULTS ON MAMMOGRAPHY BASED IMAGES (NORMAL VS BENIGN VS MALIGNANT) MODEL

Models	Recall	Precision	F1-Score	Accuracy	Selection
VGG-16	N: 0.85, B: 0.84 & M: 0.92	N: 0.82, B: 0.88 & M: 0.93	N: 0.83, B: 0.86 & M: 0.93	0.90 ~ 90%	
VGG-19	N: 0.82, B: 0.87 & M: 0.93	N: 0.85, B: 0.88 & M: 0.92	N: 0.83, B: 0.87 & M: 0.93	0.90 ~ 90%	Finalized
ResNet50	N: 0.84, B: 0.82 & M: 0.92	N: 0.81, B: 0.84 & M: 0.92	N: 0.83, B: 0.83 & M: 0.92	0.89 ~ 89%	
Iception-V3	N: 0.39, B: 0.46 & M: 0.88	N: 0.57, B: 0.64 & M: 0.75	N: 0.47, B: 0.54 & M: 0.81	0.71 ~ 71%	
InceptionResNet-V2	N: 0.07, B: 0.15 & M: 0.97	N: 0.61, B: 0.59 & M: 0.66	N: 0.12, B: 0.24 & M: 0.79	0.66 ~ 66%	

Comparisons and Selection of Models on the Basis of F1-Score

(N: Normal, B: Benign & M: Malignant)





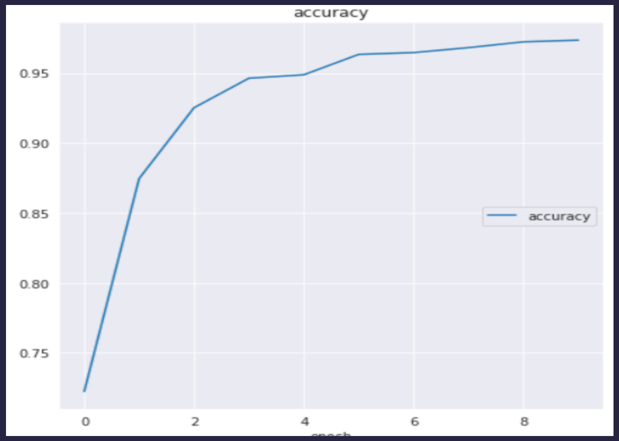
VGG-19 provide the 96.51% ROC Curve on the final result

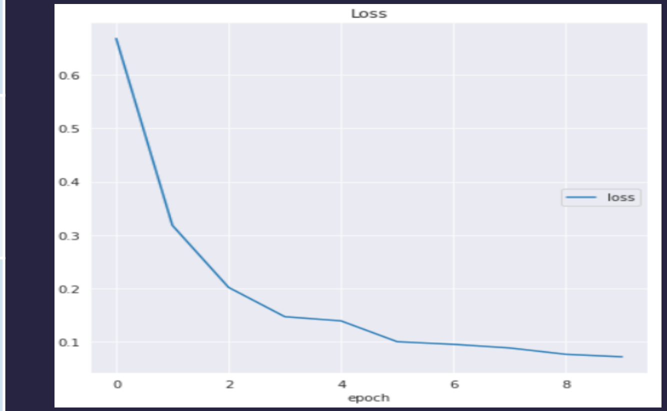
RESULTS ON **MAMMOGRAPHY BASED IMAGES (FATTY VS DENSE-GLANDULAR FATTY-GLANDULAR) MODEL**

Models	Recall	Precision	F1-Score	Accuracy	Selection
VGG-16	F: 0.93, D: 0.95 & G: 0.97	F: 0.98, D: 0.97 & G: 0.91	F: 0.95, D: 0.96 & G: 0.94	0.95 ~ 95%	
VGG-19	F: 0.95, D: 0.94 & G: 0.94	F: 0.96, D: 0.91 & G: 0.96	F: 0.95, D: 0.93 & G: 0.95	0.94 ~ 94%	
ResNet50	F: 0.98, D: 0.97 & G: 0.92	F: 0.96, D: 0.95 & G: 0.96	F: 0.97, D: 0.96 & G: 0.94	0.96 ~ 96%	Finalized
Iception-V3	F: 0.89, D: 0.83 & G: 0.69	F: 0.83, D: 0.80 & G: 0.78	F: 0.86, D: 0.81 & G: 0.73	0.80 ~ 80%	
InceptionResNet-V2	F: 0.70, D: 0.78 & G: 0.65	F: 0.74, D: 0.74 & G: 0.67	F: 0.72, D: 0.76 & G: 0.66	0.71 ~ 71%	

Comparisons and Selection of Models on the Basis of F1-Score

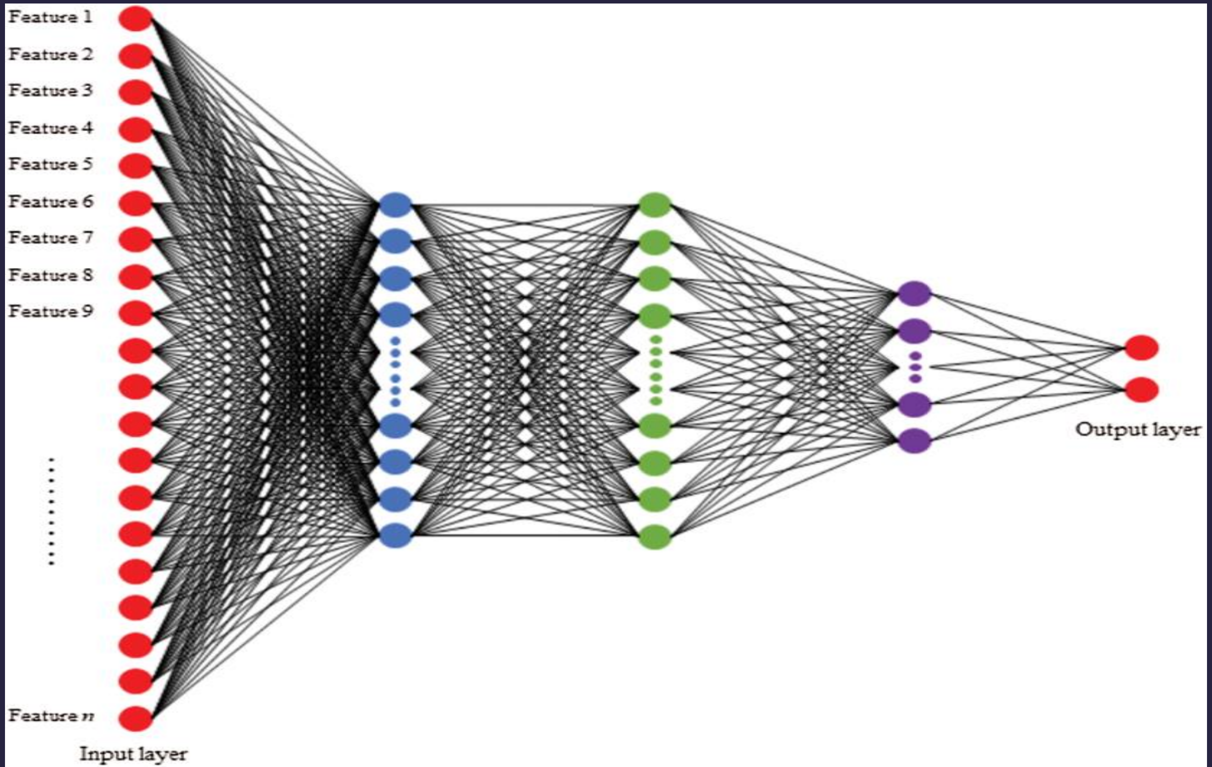
(F: Fatty, D: Dense-Glandular & G: Fatty-Glandular)





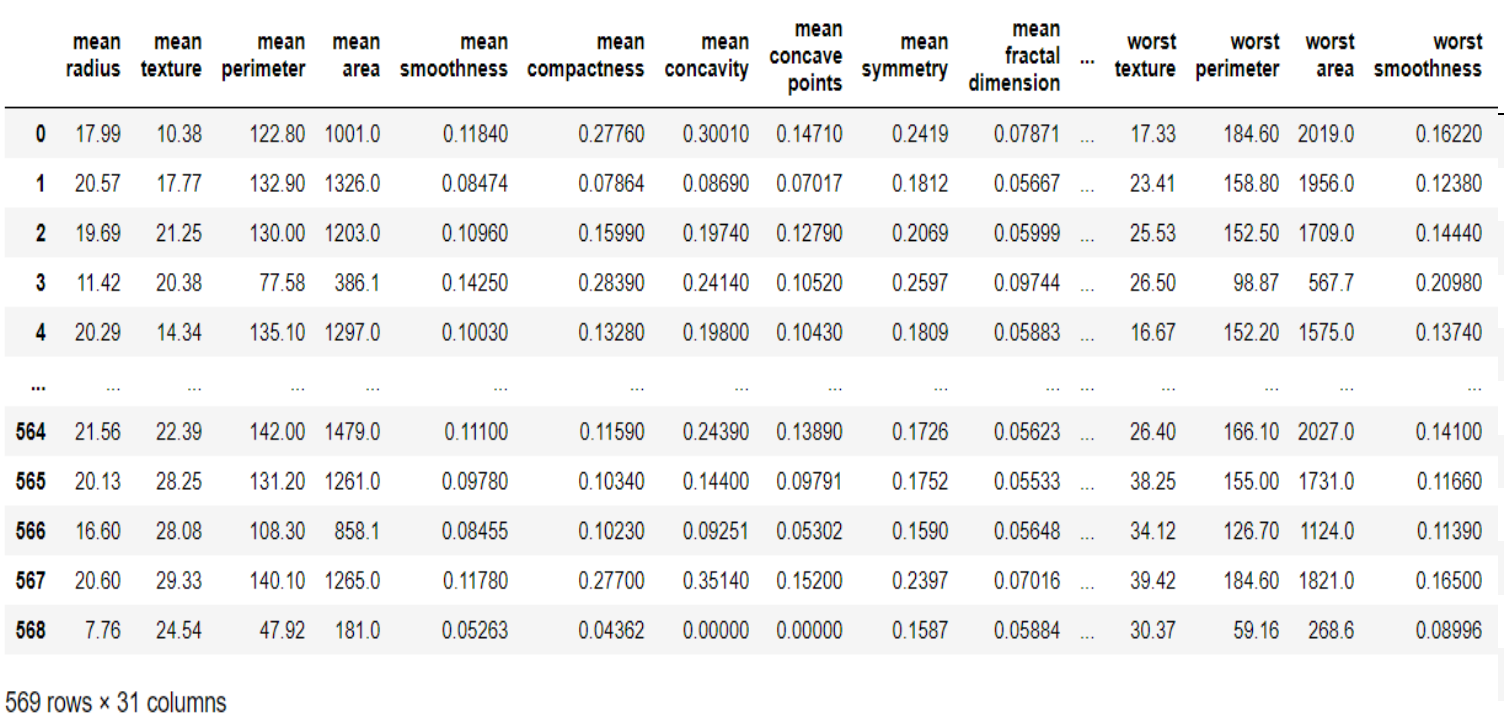
ResNet50 provide the 99.54% ROC Curve on the final result

ANN- ARCHITECTURE



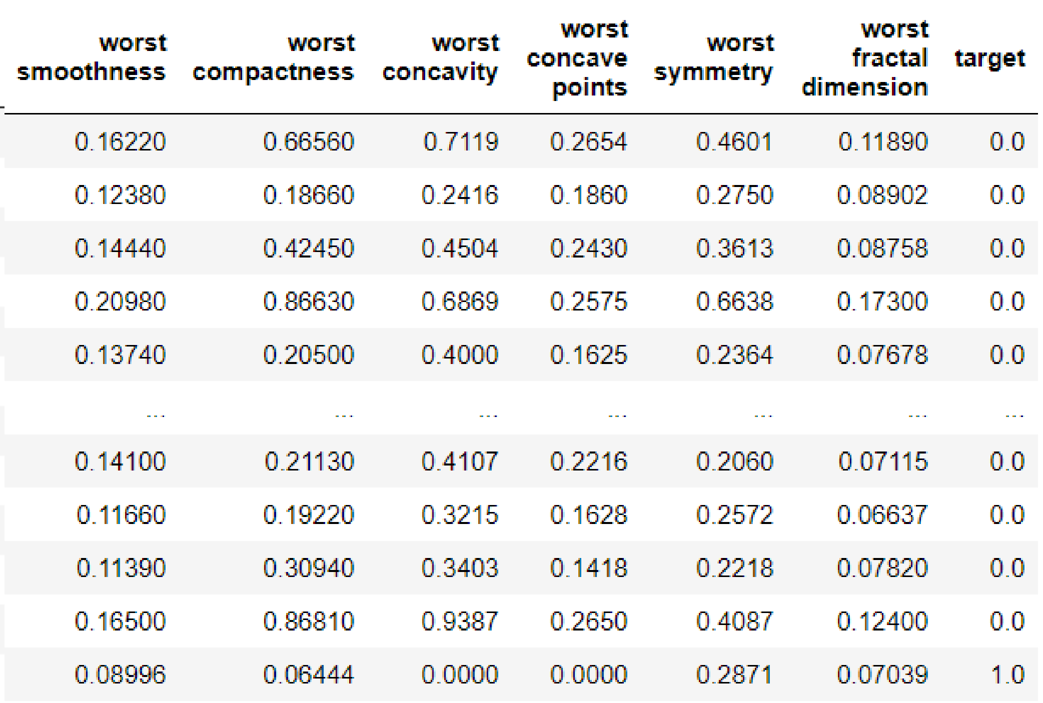
CLASSIFICATION ON

NUMERIC BASED BREAST DETECTION



	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.16220
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.12380
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.14440
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.20980
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.13740
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	26.40	166.10	2027.0	0.14100
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	38.25	155.00	1731.0	0.11660
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	34.12	126.70	1124.0	0.11390
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	...	39.42	184.60	1821.0	0.16500
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	...	30.37	59.16	268.6	0.08996

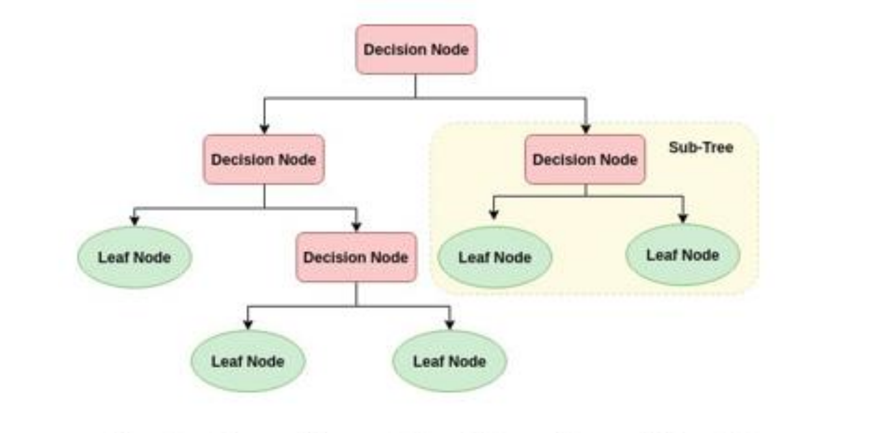
569 rows × 31 columns



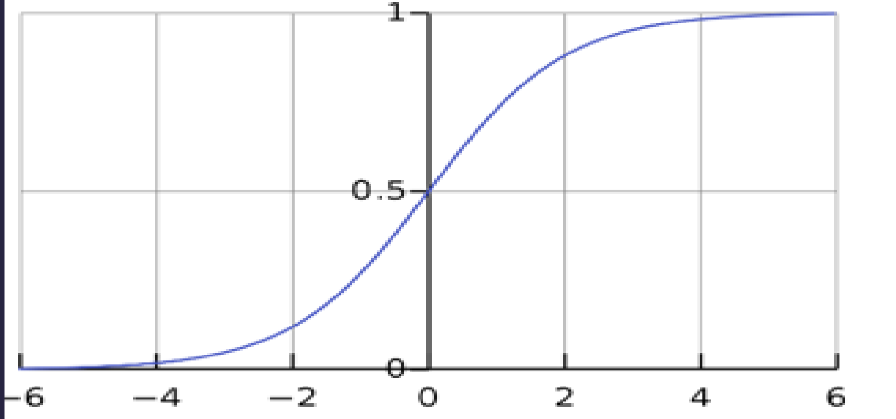
worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	target
0.16220	0.66560	0.7119	0.2654	0.4601	0.11890	0.0
0.12380	0.18660	0.2416	0.1860	0.2750	0.08902	0.0
0.14440	0.42450	0.4504	0.2430	0.3613	0.08758	0.0
0.20980	0.86630	0.6869	0.2575	0.6638	0.17300	0.0
0.13740	0.20500	0.4000	0.1625	0.2364	0.07678	0.0
...	...	...	...	...	...	...
0.14100	0.21130	0.4107	0.2216	0.2060	0.07115	0.0
0.11660	0.19220	0.3215	0.1628	0.2572	0.06637	0.0
0.11390	0.30940	0.3403	0.1418	0.2218	0.07820	0.0
0.16500	0.86810	0.9387	0.2650	0.4087	0.12400	0.0
0.08996	0.06444	0.0000	0.0000	0.2871	0.07039	1.0

SUPERVISED LEARNING ALGORITHMS

- It is a classification algorithm. It is used to estimate discrete values like Binary (0/1, yes/no, true/false) values. In this set of independent variables or features are inputted. In simple, it predicts the probability of occurrence of an event by fitting variables to a logit function or sigmoidal function. It is also known as logit regression. Since, it predicts the probability, whose output values lie between 0 and 1



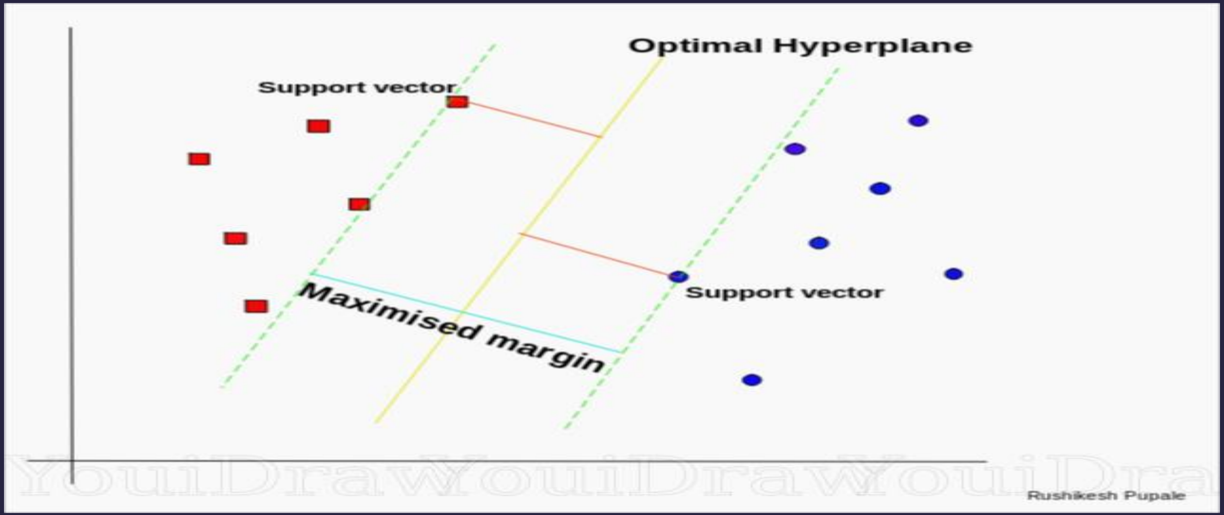
Basic Overview of Decision Tree Classifier



Logistic Regression Function

- Decision Trees is a Supervised Machine Learning in which inputs with is corresponding output of the training data are explained from the beginning. The data gets split according to a certain parameter continuously. The tree can be properly explained with the help of two entities, decision nodes and decision leaves.

SUPPORT VECTOR MACHINES



Rushikesh Pupale

KERNELS: -

i. LINEAR
ii. RBF

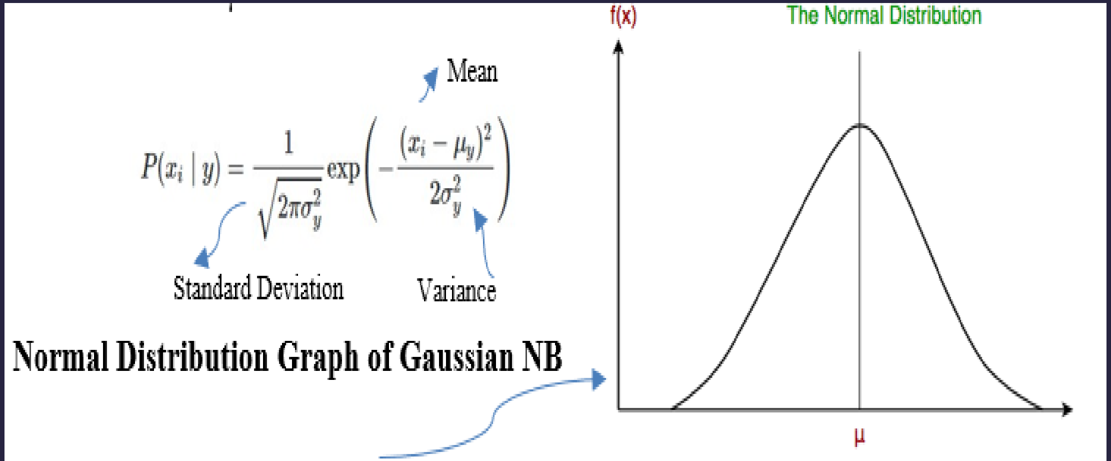
- In SVM Linear algorithm we need to find the points closest to the line from both the classes. These points are known as support vectors. After this we need to compute the distance between the line and the support vectors as shown in the figure. This distance is known as margin. Our main goal is to increase the margin. The optimal hyperplane is decided on the basis of the maximum margin. SVM tries to make a decision boundary in such a way that the separation between the two classes is as wide as possible.

- RBF (**Radial Basis Function)** is the type of SVM kernel which default form of SVM classification algorithm and can be described with the help of following formula: $K(x, x') = e^{-\gamma||x-x'||^2}$

- A **kernel is a type of a function** which takes the original non-linear task or problems and converts it into a linear form within the space of higher dimensions.

- where gamma can be set manually and has to be >0. In sklearn, the default value of gamma in SVM classification algorithm is shown below $\gamma = \frac{1}{n\,features * \sigma^2}$

- A Gaussian Naïve Bayes algorithm is a special type of NB algorithm. It’s specifically used when the features have continuous values. In this we find out the normal distribution (bell curve) on the basis of the mean, variance and standard deviation we can classify the output of a input to a specific class. The graph and formula shown below whose answer gets stored into “µ”.

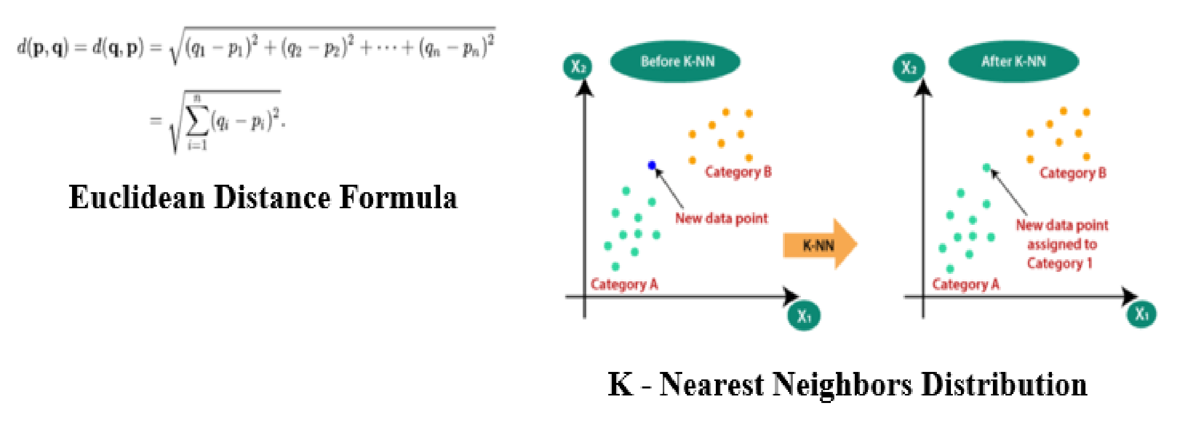


$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

Normal Distribution Graph of Gaussian NB

- K-Nearest Neighbors finds intense application in pattern recognition.

- It is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).



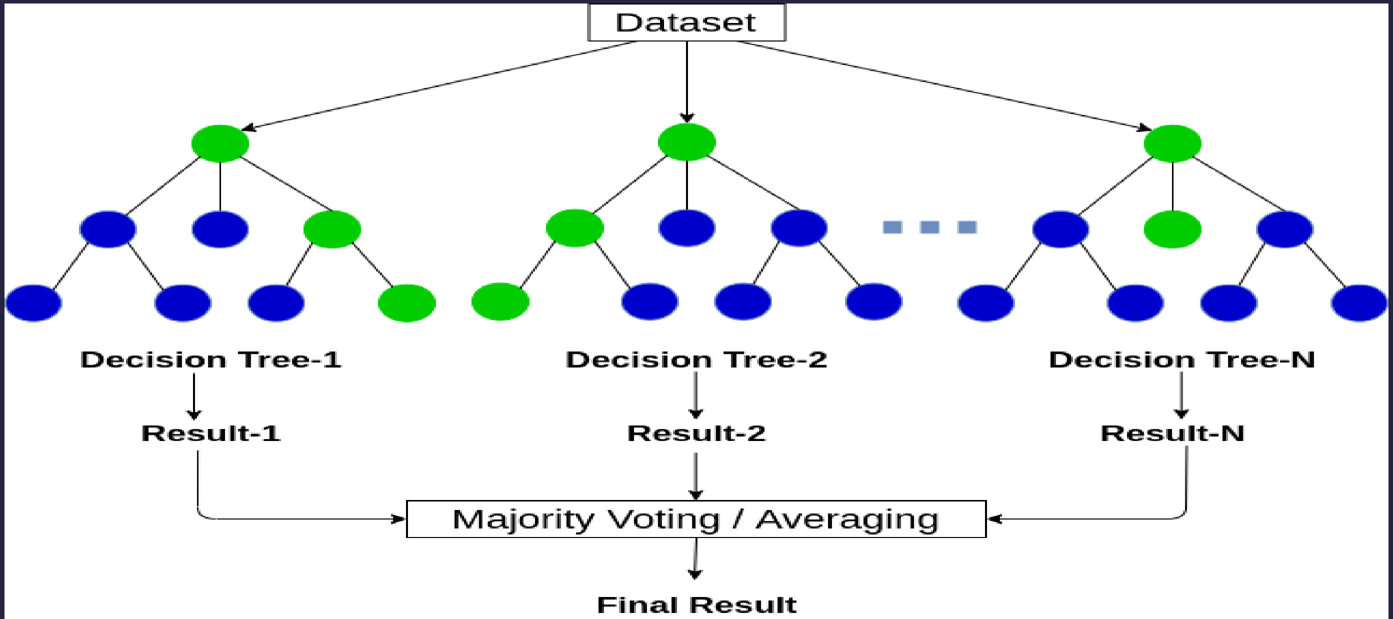
$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Euclidean Distance Formula

K - Nearest Neighbors Distribution

- The prediction is done on the basis of the distribution, this distribution is done with the help of Euclidean formula which is shown below

- **Random** forest **classifier** creates a set of decision trees from **randomly** selected subset of training set. It then aggregates every vote from various decision trees to finalized the final class of the test object for the final output.
- Each node of the decision tree works on a random subset of features or data to calculate the output. The random forest then combines all the output of individual decision trees to generate the final output.
- This can be easily understood with the help of following diagram:



RESULTS ON NUMERIC BASED DATA (BENIGN VS MALIGNANT) MODEL

Models	Recall	Precision	F1-Score	Accuracy	Selection
Logistic Regression	B: 0.89 & M: 0.97	B: 0.95 & M: 0.93	B: 0.92 & M: 0.95	0.94 ~ 94%	
Decision Tree Classifier	B: 0.94 & M: 0.93	B: 0.90 & M: 0.95	B: 0.92 & M: 0.94	0.93 ~ 93%	
SVM – Linear Kernel	B: 0.98 & M: 0.94	B: 0.92 & M: 0.98	B: 0.95 & M: 0.96	0.96 ~ 96%	Finalized
SVM – RBF Kernel	B: 0.85 & M: 0.99	B: 0.98 & M: 0.90	B: 0.91 & M: 0.94	0.93 ~ 93%	
Gaussian NB	B: 0.91 & M: 0.94	B: 0.91 & M: 0.94	B: 0.91 & M: 0.94	0.93 ~ 93%	
KNN	B: 0.94 & M: 0.94	B: 0.92 & M: 0.95	B: 0.93 & M: 0.95	0.94 ~ 94%	
Random Forest Classifier	B: 0.96 & M: 0.94	B: 0.92 & M: 0.97	B: 0.94 & M: 0.95	0.95 ~ 95%	

Comparisons and Selection of Models on the Basis of F1-Score

(B: Benign & M: Malignant)

FINAL MODELS SELECTED FOR BREAST CANCER DETECTION

- For Biopsy Image Classification:
    i. For Benign / Malignant : InceptionResNet-V2
- For Mammography Image Classification:
    i. For Benign / Normal / Malignant : VGG-19 CNN
    ii. For Fatty / Fatty glandular / Dense Glandular : ResNet50 CNN
- For Numeric Classification:
    i. SVM – Support Vector Machine (Kernel : - Linear)

THANK - YOU