Design and Development of Application by using Classification Algorithm for Breast Cancer Diagnostics

Karan Gupta 10315210075 B.Tech-Department of Computer Science and Engineering 8th Semester-B

5th February,2019

1. Abstract

Breast Cancer is a serious threat and one of the largest causes of death of women throughout the world. The identification of cancer largely depends on digital biomedical photography analysis such as histopathological images by doctors and physicians. Analysing histopathological images is a nontrivial task, and decisions from investigation of these kinds of images always require specialised knowledge. However, Computer Aided Diagnosis (CAD) techniques can help the doctor make more reliable decisions. The state of-the-art Deep Neural Network (DNN) has been recently introduced for biomedical image analysis. Normally each image contains structural and statistical information. This project classifies a set of biomedical breast cancer images (BreakHis dataset) using novel DNN and Machine learning techniques guided by structural and statistical information derived from the images. Specifically, a Convolutional Neural Network (CNN) and one Machine Learning Algorithm will be used, connection between them will be known as transfer learning and will be proposed for breast cancer image classification. In this experiment the best Accuracy value model will be selected using performance measurement models.

2. Literature Review

The unwanted growth of cells causes cancer which is a serious threat to humans. Statistics show that millions of people all over the world suffer various cancer diseases. As an example, below mentioned Table summarises the statistics concerning the recent cancer situation in Australia. These statistics reveal the number of newly cancer-affected people diagnosed in Australia and also the number of people who died in 2017 in Australia. These statistics also divulge that the number of females affected and the number of females dying due to breast cancer are more than the numbers for males. This indicates that females are more vulnerable to breast cancer (BC) than males. Although these statistics are for Australia they might be representative of what is happening throughout the world.

Table 1 Cancer Statistics for Australia 2017

	Female	Male	Total
Estimated number of new diagnoses (all cancers)	62005	72169	134174
Estimated number of deaths	20677	27076	47753
Estimated new cases of diagnosis	17586	144	17730
Deaths due to breast cancer	3087	57	3114

Below figure shows the number of females newly facing BC as well as the number of females dying since the year 2007 in Australia. This figure shows that more and more females are newly facing BC, and the number of females dying of it has also increased in each year This is the situation of Australia (population 20–25 million), but it can be used as a symbol of the BC situation of the whole world.

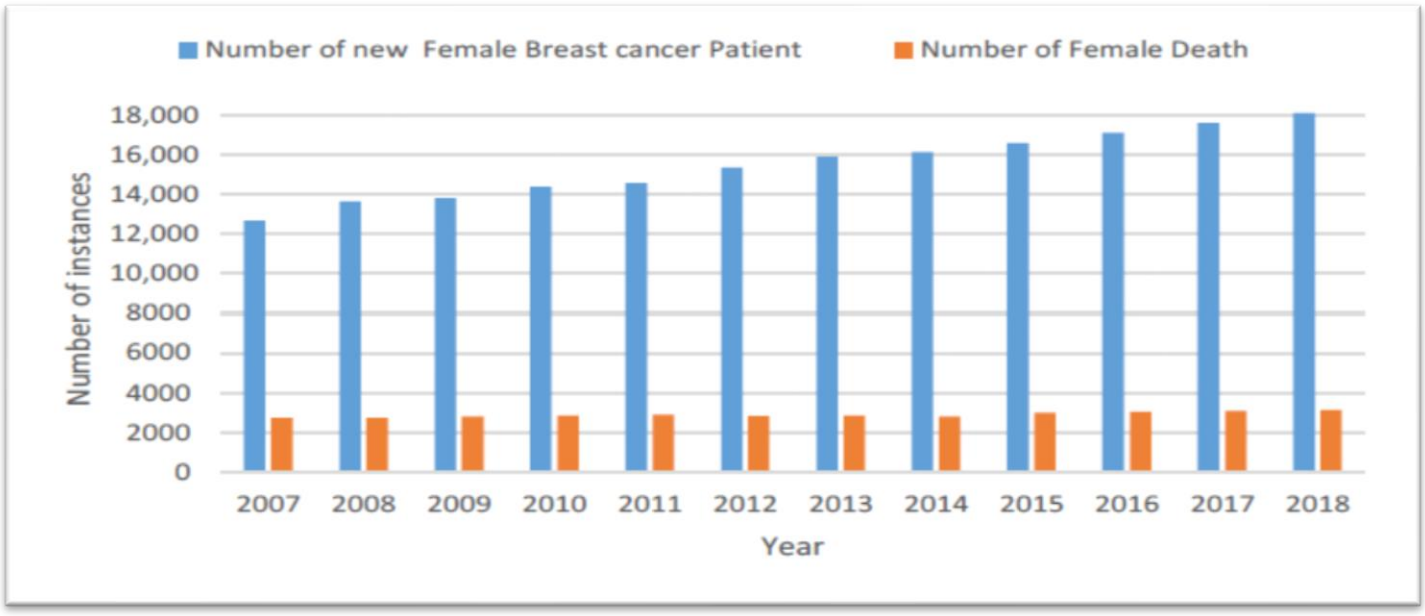


Figure 1Analysis of BC over the past year

Detecting macrocalcification in dense breast tissue can be a difficult task as both tends to depict white pixel on the mammogram. The number of false positive cases on dense breast tissue are higher. Indicators of cancer symptom are generally, masses and microcalcifications. Detecting masses are more challenging task than detection of micro-calcifications. As their size and shape varies in large variation and they often exhibit poor image contrast. The utilization of grouping frameworks in classification and pattern recognition system, in medical diagnosis, specially cancer diagnosis is growing rapidly. Evaluation and decision making based on machine learning for medical diagnosis is a key factor. Intelligent classification algorithm may help doctor in identifying symptoms that may not be possible through traditional approaches.

Any Image processing and analysis applications would require a unique function for alignment of feature for classification and segmentation. Mainly texture features and statistical features are of more suitable in pattern recognition area to find this alignment.

Screening Mammography is the easiest and affordable way to diagnosis for breast cancer. The mammography image is checked through several techniques like finding edges, smoothing border, finding structures & shapes among matrixes. Finally finding the size distribution of tissues in an Image without explicitly segmenting each object.

Digital mammography is the standard procedure for breast cancer diagnosis, various classification problem is applied on the digital mammography image. Various features are extracted as per standard procedure for breast cancer diagnosis. These features are calculated from the sensitive part of the breast to avoid any unwanted features to affect the classification problem. Area of tumour is calculated by the Maximum Likelihood Estimation (MLE). All the features extraction techniques are applied on the stored database image.

This Project mainly studies the multiple image processing in deep learning and machine learning algorithms which can be extensively used for finding cancerous cells. The techniques in computer aided mammography includes image pre-processing, image segmentation, feature extraction, feature selection and classification.

Further developments are required to extract more features to find pattern in tumour to have a better understanding on them. Texture analysis method can be used to classify between benign and malignant masses by means to identify the micro-calcification in the mammography.

Research in The Field of Cancer

Many research has been done in the field of image processing to find the cancer. Yet, the accuracy rate lies between 75% - 92%. Thus, there is still a gap of 8% to 25% of accuracy to be achieved. The new research analysis and techniques to find the cancerous cells and eradication methodology to cure the cancer from any person. However, even cancer cells have evolved them to hide from drugs and medications. As cancer cells are immortal they are not affected by the immune system. There is a research for curing the cancer tumour, the methods are as follows.

1. **CRISPR**

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) also known as Cas is a simple and powerful gene editing tool. Genetic engineering has allowed cancer researchers to screen the drug to target the cancer cells in efficient manner. There is also a vast door for direct treatment of cancer through gene interference or activation.

2. **Artificial Immune System**

Artificial Immune Systems resembles the natural properties of our biological immune system. Natural immune system has the property to pattern matching which is used to distinguish between normal and abnormal cell.

3. **Nano Technology**

Nano technology can give fast and delicate location of cancer cells in the breast tissues. Empowering researchers to identify molecular changes not withstanding when they happen just in a smaller amount of cells. The nanodevices can be programmed to annihilate infected cells and kill those infected cells.

BreakHis Data Description

Cancer and its subtype

1. Benign
    a. Adenosis
    b. Fibroadenoma
    c. Phyllodes tumor
    d. Tubular adenoma

2. Malignant
    a. Carcinoma
    b. Lobular Carcinoma
    c. Mucinous Carcinoma
    d. Papillary Carcinoma

Data is collected from 82 Patients. RGB images having size 700 X 460.

3. Objectives

The research will carry out with the following objectives

1. To study various combinations of Supervised Machine Learning approaches for Breast Cancer diagnosis through their implementation.
2. To make a comparative study of the approaches for the future.
3. Use that tool for Design and Development of an Application in future.

4. Scope of the Project

1. The methodology developed to integrate sensors with biological parameters can be further improved by considering industry standard tools that are not accessible by academicians in India. Hence, with the use of Technology Computer Aided Design tools, device simulations can be accurately carried out and mathematical models can be developed.

2. Due to limitations of hardware and software environment, right now it is not possible to integrate the complete system as a unit and hence testing of the system could not be carried out. Thus, with availability of resources in near future, there is a possibility of integrating the individual blocks developed to analyse the system performances.

5. Project Requirement

5.1 Hardware

5.1.1 Minimum Requirement

Table 2 Minimum Requirement for Hardware

Component	Specification
Processor	Core i5 -5th Generation
RAM	8Gb-16Gb
HDD	1TB
Graphic Processing Unit	Nvidia above 950M (Cuda and cuDNN Enabled)

5.1.2 Recommended Requirement

Table 3 Recommended Requirement for Hardware

Component	Specification
Processor	Core i5 -8th Genration or above
RAM	8Gb-16Gb
HDD	1TB
Graphic Processing Unit	Nvidia above 1050Ti (Cuda and cuDNN enabled) Or Nvidia Tesla Family Like K40, K80, P100, V100

5.2 Software

Table 4 Requirement for Software

Programming Language	Python
Libraries	SciKit-Learn, Pandas, NumPy, Tensor-Flow, Keras
Platform	Jupyter Lab, Jupyter Notebook, Google Colab.
Library Management	PIP

6. Proposed Methodology

6.1 Database Import

The Breast Cancer Histopathological Image Classification (BreakHis) is composed of 7,909 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). To date, it contains 2,480 benign and 5,429 malignant samples (700X460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format). This database has been built in collaboration with the P&D Laboratory – Pathological Anatomy and Cytopathology, Parana, Brazil. This dataset will be uploaded to google drive as database for Google Colab.

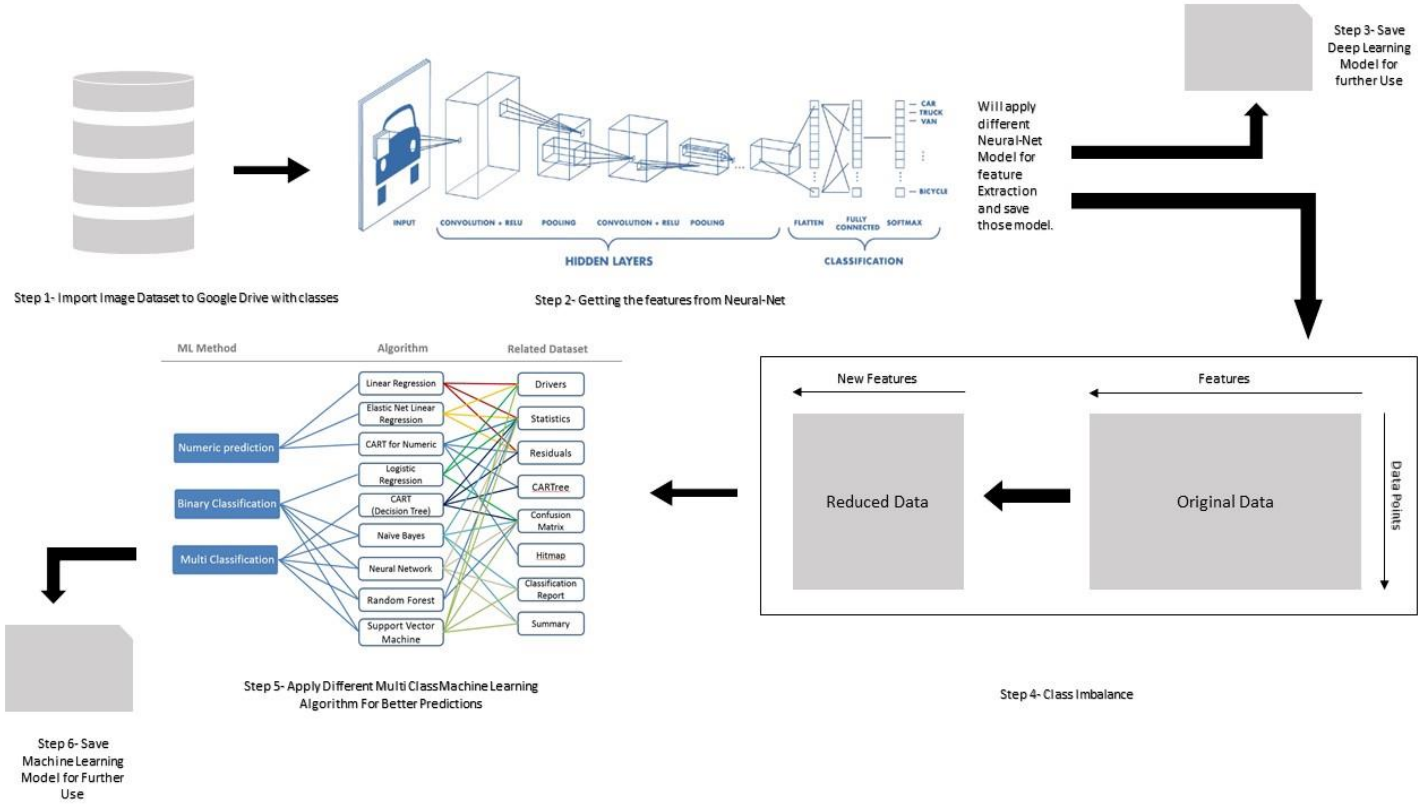


Figure 2 Work-Flow for Classification Algorithm

6.2 Convolutional Neural Network

A typical CNN has two parts:

1. **Convolutional base**, which is composed by a stack of convolutional and pooling layers. The main goal of the convolutional base is to generate features from the image.

2. **Classifier**, which is usually composed by fully connected layers. The main goal of the classifier is to classify the image based on the detected features. A fully connected layer is a layer whose neurons have full connections to all activation in the previous layer.

6.2.1 Repurposing a Pre-Trained Model

When you're repurposing a pre-trained model for your own needs, you start by removing the original classifier, then you add a new classifier that fits your purposes, and finally you have to fine-tune your model according to one of three strategies:

1. **Train the entire model.** In this case, you use the architecture of the pre-trained model and train it according to your dataset. You're learning the model from scratch, so you'll need a large dataset (and a lot of computational power).

2. **Train some layers and leave the others frozen.** As you remember, lower layers refer to general features (problem independent), while higher layers refer to specific features (problem dependent). Here, we play with that dichotomy by choosing how much we want to adjust the weights of the network (a frozen layer does not change during training). Usually, if you've a small dataset and a large number of parameters, you'll leave more layers frozen to avoid overfitting. By contrast, if the dataset is large and the number of parameters is small, you can improve your model by training more layers to the new task since overfitting is not an issue.

3. **Freeze the convolutional base.** This case corresponds to an extreme situation of the train/freeze trade-off. The main idea is to keep the convolutional base in its original form and then use its outputs to feed the

classifier. You're using the pre-trained model as a fixed feature extraction mechanism, which can be useful if you're short on computational power, your dataset is small, and/or pre-trained model solves a problem very similar to the one you want to solve.

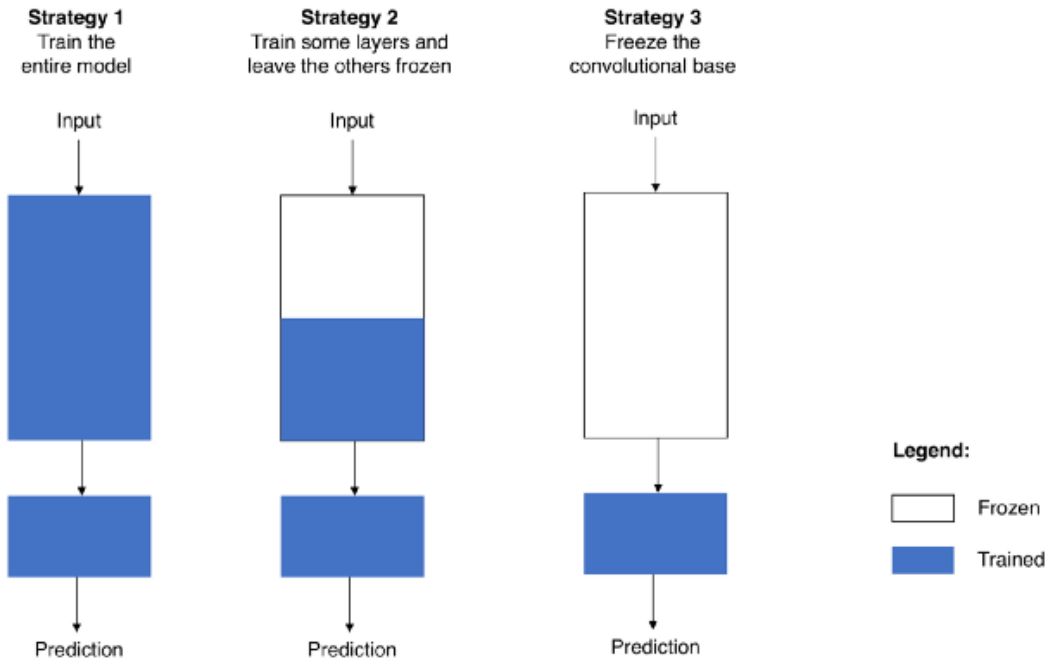


Figure 3 Representation of Above Three Stratigies in Schematic Way

Unlike **Strategy 3**, whose application is **straightforward**, **Strategy 1** and **Strategy 2** require you to **be careful** with the learning rate used in the convolutional part. The learning rate is a hyper-parameter that controls how much you adjust the weights of your network. When you're using a pre-trained model based on CNN, it's smart to use a small learning rate because high learning rates increase the risk of losing previous knowledge. Assuming that the pre-trained model has been well trained, which is a fair assumption, keeping a small learning rate will ensure that you don't distort the CNN weights too soon and too much.

6.2.2 Classifiers on Top of Deep Convolutional Neural Networks

As mentioned before, models for image classification that result from a transfer learning approach based on pre-trained convolutional neural networks are usually composed of two parts:

1. **Convolutional base**, which performs feature extraction.
2. **Classifier**, which classifies the input image based on the features extracted by the convolutional base.

Since in this section we focus on the classifier part, we must start by saying that different approaches can be followed to build the classifier. Some of the most popular are:

1. **Fully-connected layers.** For image classification problems, the standard approach is to use a stack of fully-connected layers followed by a softmax activated layer (Krizhevsky et al. 2012, Simonyan & Zisserman 2014, Zeiler & Fergus 2014). The softmax layer outputs the probability distribution over each possible class label and then we just need to classify the image according to the most probable class.
2. **Global average pooling.** A different approach, based on global average pooling, is proposed by Lin et al. (2013). In this approach, instead of adding fully connected layers on top of the convolutional base, we add

a global average pooling layer and feed its output directly into the softmax activated layer. Lin et al. (2013) provides a detailed discussion on the advantages and disadvantages of this approach.

3. **Linear support vector machines.** Linear support vector machines (SVM) is another possible approach. According to Tang (2013), we can improve classification accuracy by training a linear SVM classifier on the features extracted by the convolutional base.

6.3 Class Imbalance

The class imbalance problem occurs when the main class of interest is represented by only a few tuples. Strategies to address this problem include: -

1. Oversampling
2. Under-Sampling
3. Threshold moving
4. Ensemble Technique

6.4 Machine Learning Algorithm

Classification Report: This dataset is a tabular representation of accuracy metrics for each distinct value of target column. For ex: If the target column can have two distinct values 'Yes' and 'No' , this dataset shows accuracy metrics like F1, Precision, Recall, Support (number of rows in Training dataset with this value) for each and every distinct value of Target column.

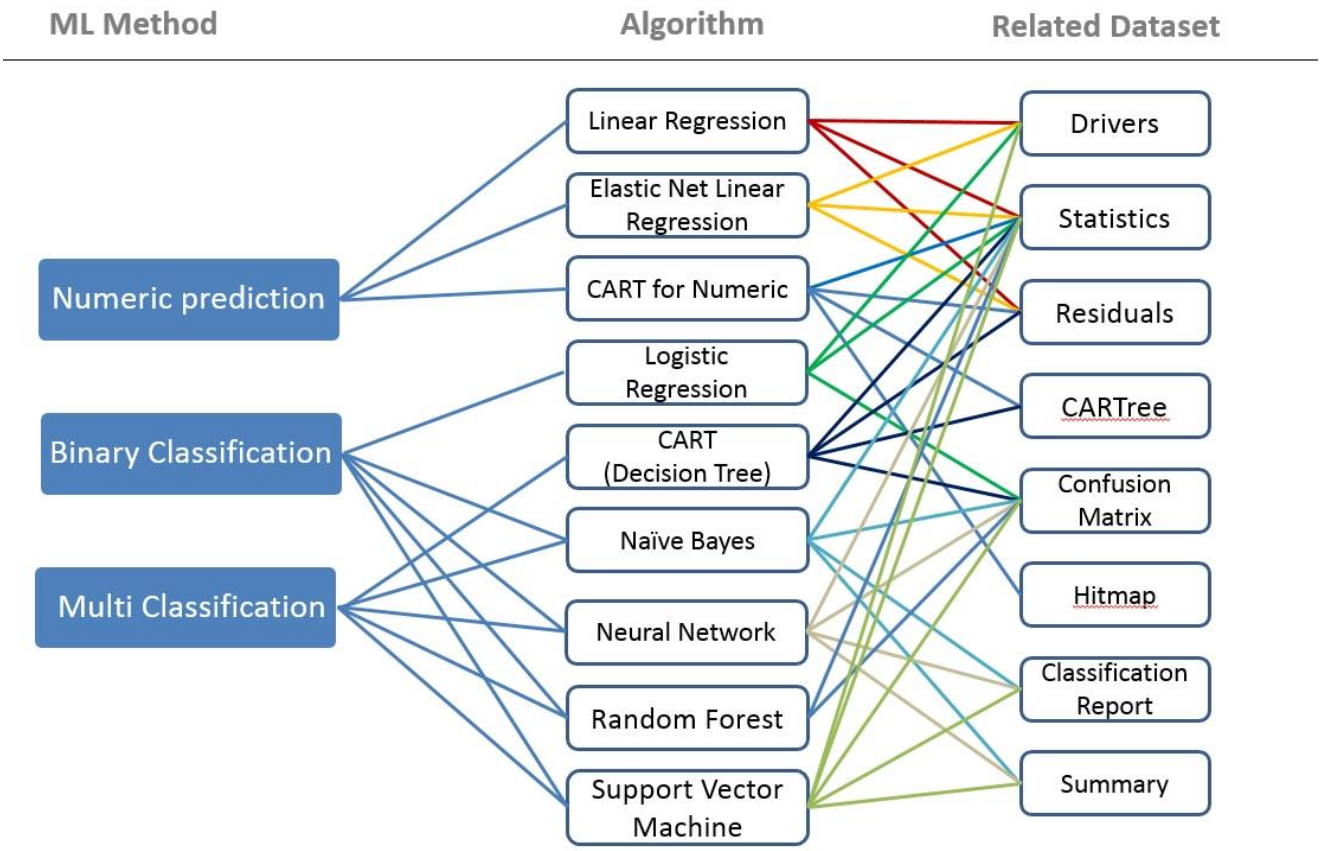
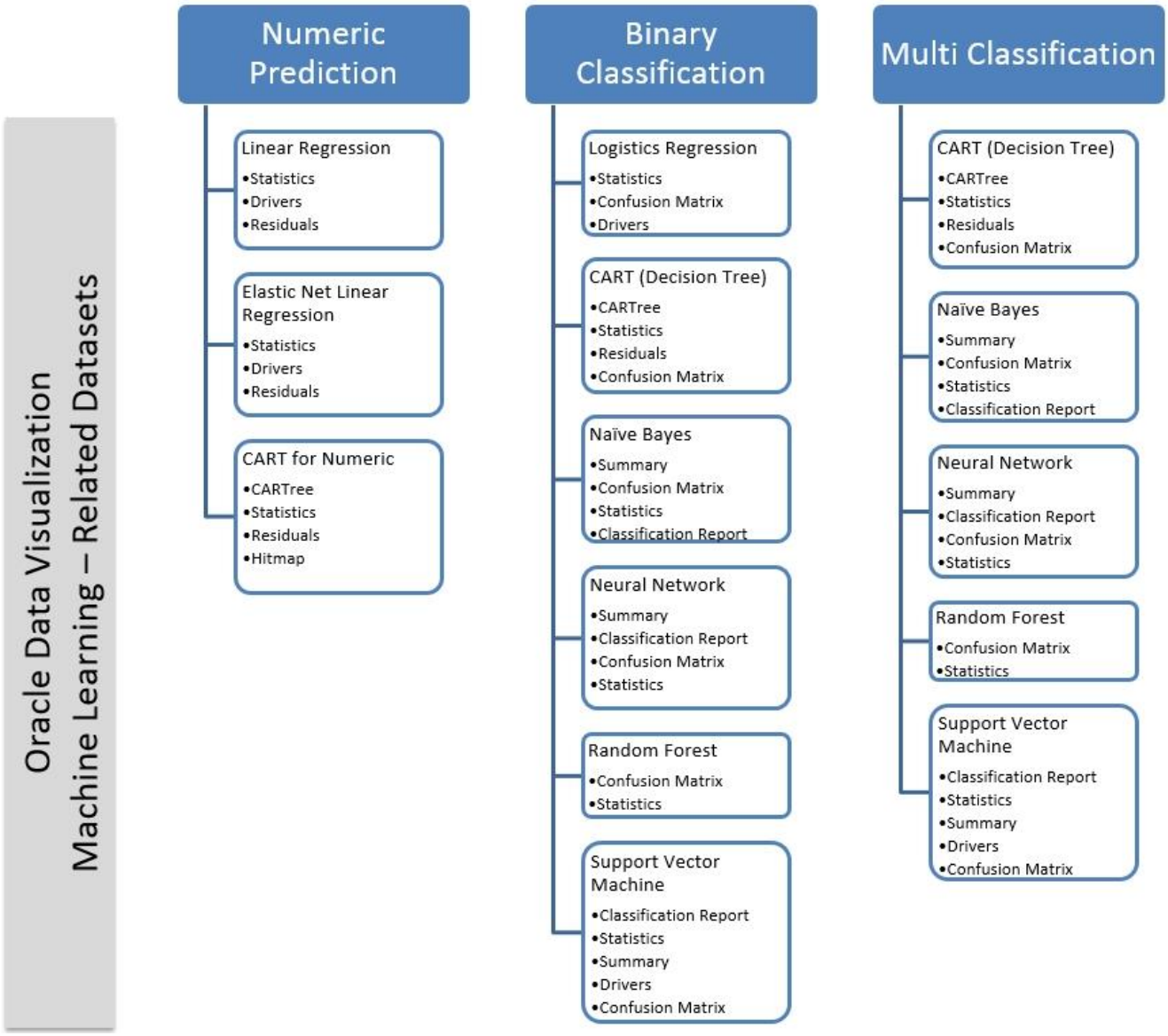


Figure 4 Combination Network for Classification Report



6.5 Designing of Web Application

Both saved Machine Learning and Deep Learning model will be used as Web API for further use

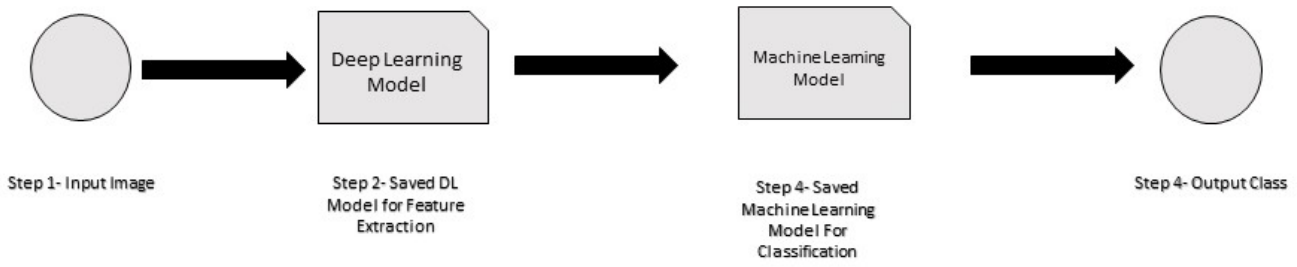


Figure 5 Checking for Best Classification Algorithm

As we can see that, we will input an image in Deep Learning Model as Web API for feature extraction and that extracted features will be passed to trained Machine Learning Model as Web API and this will predict class of that image.

7. References

1. Spanhol, F., Oliveira, L. S., Petitjean, C., Heutte, L., A Dataset for Breast Cancer Histopathological Image Classification, IEEE Transactions on Biomedical Engineering (TBME), 63(7):1455-1462, 2016.

2. Abdullah-Al Nahid, Yinan Kong," Histopathological Breast-Image Classification Using Local and Frequency Domains by Convolutional Neural Network"

3. Abdullah-Al Nahid, Yinan Kong, Mohamad Ali Mehrabi," Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering".

4. Prannoy Giri , K. Saravanakumar,"Breast Cancer Detection using Image Processing Techniques".

5. *Sebastian Raschka, Vahid Mirjalili, "Python Machine Learning ",2nd Edition, Pakt Publications.*

6. *J. Han, M. Kamber, J. Pei, "DATA MINING-Concepts and Techniques",3rd Edition, MK Morgan Kaufmann.*

7. *I. Goodfellow, Y. Bengio, A. Courville,"DEEP LEARNING", The MIT Press, Cambridge, Massachusetts, London, England.*