

FACULTY OF SCIENCE & TECHNOLOGY

MSc Data Science and Artificial Intellegence

2021-2022

Evaluation of breast cancer detection techniques

by

Siddharth Kumar Negi

Faculty of Science & Technology

Department of Computing and Informatics

Individual Masters Project

**Abstract**

One of the most common form of cancer occurring in women is breast cancer, according to the National health services, about 1 in 8 women in UK are diagnosed with it at some age. Detection of breast cancer in an early stage could improve the mortality and survival rate. Breast cancer could be defined as benign (non-cancerous) and malignant (cancerous) in nature. Benign are referred as not life threatening as it does not spread outside the breast unlike malignant. There are several ways from which breast cancer could be identified from physical structure to taking out mamographs. The primary source of breast cancer was radical surgery but due to rapid advancement in the technology, new techniques were discovered for the early detection. Integration of machine learning techniques have played an important role in creation of computer aided systems, which were proven to be more accurate and precise than the human observation.

In this research paper, we are going to discuss about various supervised machine learning techniques used for the detection of breast cancer and evaluate them using various evaluation techniques

|  |
| --- |
| **Dissertation Declaration**    I agree that, should the University wish to retain it for reference purposes, a copy of my dissertation may be held by Bournemouth University normally for a period of 3 academic years. I understand that once the retention period has expired my dissertation will be destroyed.    **Confidentiality**    I confirm that this dissertation does not contain information of a commercial or confidential nature or include personal information other than that which would normally be in the public domain unless the relevant permissions have been obtained. Any information which identifies a particular individual's religious or political beliefs, information relating to their health, ethnicity, criminal history, or sex life has been anonymised unless permission has been granted for its publication from the person to whom it relates.    **Copyright**    The copyright for this dissertation remains with me.    **Requests for Information**    I agree that this dissertation may be made available as the result of a request for information under the Freedom of Information Act.    **Signed**: Siddharth Kumar Negi .    Name: Siddharth Negi    Date: 13/04/2022    Programme: Data Science and Artificial Intelligence |

**Original Work Declaration**

This dissertation and the project that it is based on are my own work, except where stated, in accordance with University regulations.

**Signed**: Siddharth Kumar Negi .

Name: Siddharth Kumar Negi

Date: 13/04/2022

**Acknowledgment**

I would like to express my gratitude towards our supervisor, Prof. Wei Koong Chai for his excellent support throughout the dissertation, whose expertise was extremely valuable in each and every step of the dissertation.

TABLE OF CONTENTS

[1 INTRODUCTION 1](#_Toc103337262)

[1.1 Breast Cancer 1](#_Toc103337263)

[1.2 Existing Breast cancer detection techniques 1](#_Toc103337264)

[1.3 Aims and objectives 2](#_Toc103337265)

[1.4 Structural Outline 2](#_Toc103337266)

[2 Literature Review 3](#_Toc103337267)

[2.1 Approaches for predicting Breast Cancer 3](#_Toc103337268)

[2.1.1 Machine learning 3](#_Toc103337269)

[2.1.2 Knowledge Discovery 3](#_Toc103337270)

[*B.* *Data Mining* 5](#_Toc103337271)

[2.2 Types of Machine Learning 6](#_Toc103337272)

[2.2.1 Supervised Learning 7](#_Toc103337273)

[2.2.2 Unsupervised Learning 7](#_Toc103337274)

[2.2.3 Semi-Supervised Learning 7](#_Toc103337275)

[2.2.4 Reinforcement Learning 8](#_Toc103337276)

[2.3 Approaches for Breast Cancer Detection 8](#_Toc103337277)

[Machine Learning 8](#_Toc103337278)

[2.3.1 Logistic Regression 9](#_Toc103337279)

[2.3.2 K-Nearest Neighbour 10](#_Toc103337280)

[2.3.3 Support Vector Machine (SVM) 12](#_Toc103337281)

[2.3.4 Decision Tree 13](#_Toc103337282)

[2.3.5 Random Forest 15](#_Toc103337283)

[2.4 Deep Learning 16](#_Toc103337284)

[2.4.1 Working of Deep Learning 17](#_Toc103337285)

[2.4.2 Deep Learning in Breast Cancer Detection 19](#_Toc103337286)

[3 Methodology 21](#_Toc103337287)

[3.1 Architecture of the Machine learning model 21](#_Toc103337288)

[3.1.1 Data Pre-processing and visualization 22](#_Toc103337289)

[3.1.2 Data splitting and training of model 23](#_Toc103337290)

[3.1.3 Implementation of Machine Learning Algorithms 24](#_Toc103337291)

[3.2 Implementation of Deep Learning 27](#_Toc103337292)

[3.3 Implementation of breast cancer using CNN 29](#_Toc103337293)

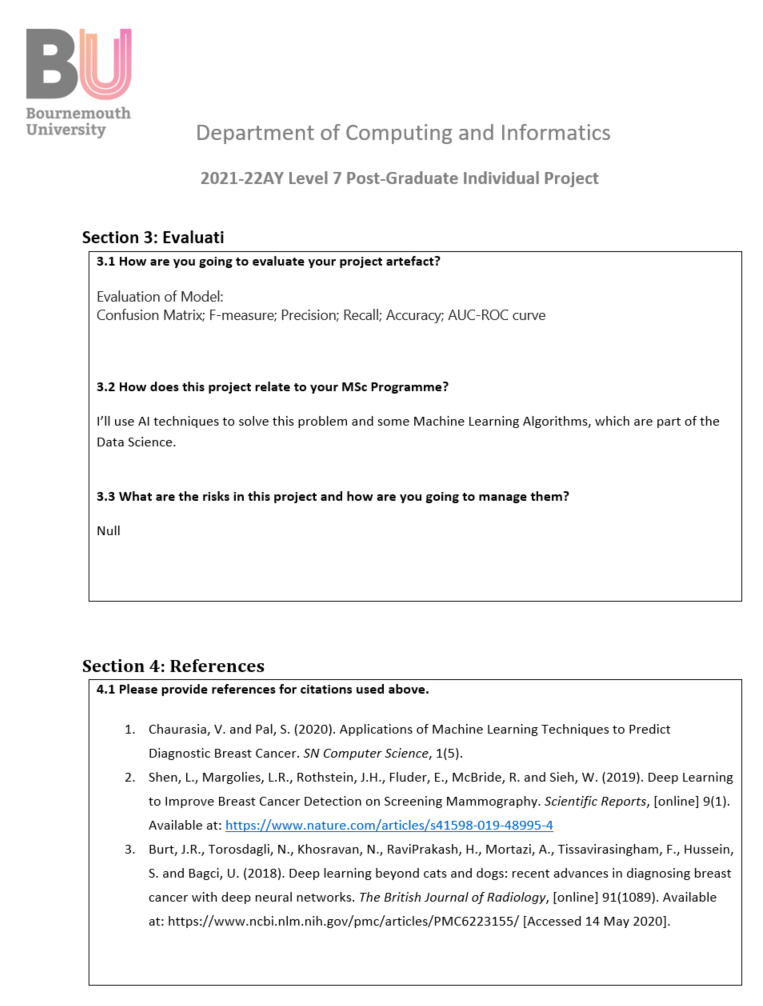
[3.3.1 Splitting the dataset for training and testing 30](#_Toc103337294)

[4 CONCLUSION & Discussion 34](#_Toc103337295)

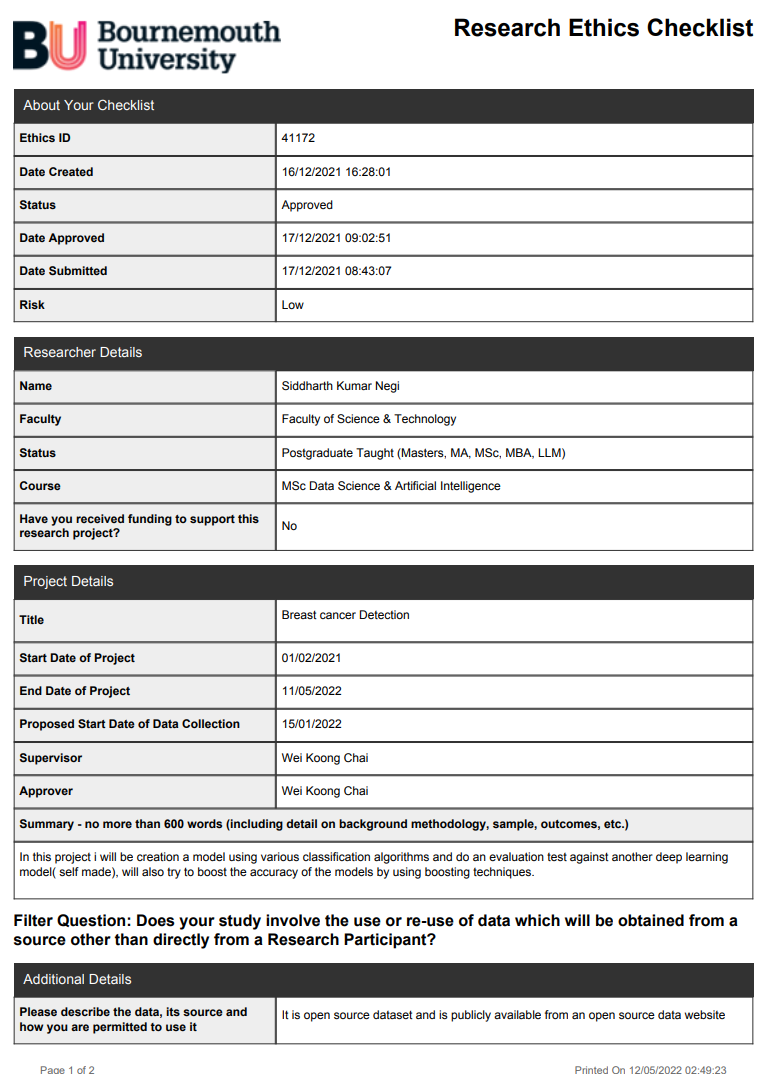
[REFERENCES 35](#_Toc103337296)

[5 PROJECT PROPOSAL 38](#_Toc103337297)

[40](#_Toc103337298)



[6 APPENDIX A PROJECT ETCIH CHECK LIST 41](#_Toc103337299)



[7 APPENDIX FIRST REVIEW 43](#_Toc103337300)

[8 APPENDIX E -Artefact 44](#_Toc103337301)

[8.1 Machine Learning model 44](#_Toc103337302)

[8.2 Deep Learning model 47](#_Toc103337303)

[8.3 Deep learning using CNN 50](#_Toc103337304)

# INTRODUCTION

## Breast Cancer

Breast cancer is one of the most common form of cancer found in women that are either middle or old aged and could be found at young ages as well, cells grow in an uncontrollable way and could be of different types depending on the cell type and region. Mankind has been familiar with this disease since ancient times and has been mentioned through different ages. Generally, the breast region is divided into three parts: ducts, lobules and connecting tissue. The lobules produce milk, ducts carry the milk to the nipples and the connecting tissue holds everything together. In 2020, over 2.3 million women were diagnosed with cancer and out of which 685,000 deaths were recorded. Cancer could be benign and malignant. Generally Benign refers to the non-cancerous cells that cannot grow and are weak to damage nearby cells while malignant are the cancerous cells that can multiply and damage the cells and could be locally invasive or metastatic.

* Locally invasive cancer: The nearby cells could be invaded by the tumour by sending out cancerous cells into the normal ones.
* Metastatic cancer: Even normal cells located away from the original tumour could be infected.

There are various symptoms of breast cancer which includes, formation of a new lump in breast or underarm region, swelling in the breast, irritation, pain, change in shape of breasts or discharge of anything other than milk. Some people may not even have these symptoms at all. The chances of survival and recovery are higher when breast cancer is detected in early stages. Breast Cancer can be distinguished into four types or stages depending upon the size of the tumour.

* Stage 0: 0 cm in size
* Stage 1: 2-5 cm in size
* Stage 2: 5-9 cm in size
* Stage 4: from 9cm and above in size

Tumour with the size more than 0 are refereed as Malignant (cancerous) otherwise Benign (non-cancerous).

## Existing Breast cancer detection techniques

The existing breast cancer detection techniques are mainly done by three triplet assessment: Clinical, where the history is checked and physical examination of the patient is done by the doctor for presence any abnormalities or any changes in the structure of body organ, that indicates cancer. Imaging, these tests helps in the diagnosis of internal body organs in non-intrusive way, these imaging test include a computerized tomography (CT), magnetic resonance imaging (MRI), ultrasound and x-ray. They have their own pros and cons. Earlier; the detection of breast cancer was done by examining the mammographs. Although this technique has been successful in reduction of the mortality rate (30%-40%) but had some drawbacks such as false positives and false negatives screening outcomes. For example, benign neoplasm may act like a breast cancer that might lead to a false detection where it isn’t, which also results in low positive predictive values. Lastly, in any case doctor finds any suspicion, pathological (biopsy)tests are required.

Different methods are used in order to get a sample of the breast tissue, that is later used for the evaluation.

* Punch biopsy: a circular blade is used to take out sample of a lesion that includes the skin and the sample is used for evaluating the abnormalities in the sample.
* Fine needle aspiration: FNA process just consists of pulling out cells from the lesion with the help of a fine needle.
* Core needle biopsy: with the help of a small needle, a thin mass of tissue is pulled out from the abnormal area, which is later used for evaluation of breast mass and this method is one of the most commonly used for obtaining the specimen.
* Excisional biopsy: If the specimen isn’t obtainable due to the depth of the mass in breast, excisional biopsy is used for the extraction.

Other than the diagnosis part, there are several risk factors that are related to increased risk of breast cancer that includes sex, age, medical history, inherited genes, obesity etc.

However, In the past few decades, rapid growth of the machine learning techniques has resulted in improvements in various fields from medical to security or in general ways, Introduction of the artificial intelligence in the medical field has led to significant developments and improvements to the existing techniques. Many new and alternative methods have been developed to improve the workflow and accuracy in detection of diseases, which is considered far better in comparison of human performance. The chances of survival and recovery are higher when breast cancer is detected in early stages. Patients diagnosed with cancer locally had the survival chances of 100%; those diagnosed regionally had an 86% survival rate while ones with distant, chances dropped to 29%.

For instance, CAD (computer assisted detection and diagnosis) software were developed in order to improve the accuracy of the predictive analysis of the mammography, but the early versions were not successful until deep learning took off. Advances machine learning and in deep learning problems like image classification, speech recognition and natural language processing were solved to a greater extent.

## Aims and objectives

The aim is to discover which of the machine learning methods has the potential to detect the type/nature of tumour (benign or malignant) and do a comparative analysis/ evaluation of the various machine learning algorithms like Logistic Regression, SVM (support machine vector), DT (decision tree) and Random Forest with the Deep Learning by the help of evaluation techniques.

## Structural Outline

The next part is the literature review, that would show about the previous researches that had been done so far by now, their outcomes and a brief discussion about the Machine learning and Deep learning techniques that have been already used for the analysis purpose and others that will be used in this research. The third part, Methodology will have an in-depth discussion about the various machine learning algorithms along with deep learning, their outcomes, evaluation and different boosting techniques that has been used for achieving higher accuracy. Fourth chapter, consists of the results from the comparison and the last chapter will have a brief conclusion and discussion related to the future work.

# Literature Review

## Approaches for predicting Breast Cancer

### Machine learning

What is Machine Learning? Machine Learning is a sub division of Artificial intelligence, which makes computer think on their own. Machine Learning refers to a branch of computer science which enables the computer to learn without the interference of the humans. The ultimate goal of machine learning is to learn about the data in order to be utilized by the models.. Until the end of 1970s, machine learning wasn’t a considered as a sub division of AI but due to rapid advancement, it was treated as a separate field on its own. With the help of Artificial Intelligence systems, humans have been able to perform tasks far beyond the human level. For instance, from detection of diseases to performing specific tasks like forming distant object images in real time. Machine learning on other hand has turned into one of the important tool used in the modern era, especially in the field of business and research. With the use of various algorithms and deep learning models, performance of the model can be improved progressively. (Keith D. Foote, December 3, 2021) Machine learning is based on various types of learning methods

### Knowledge Discovery

Knowledge discovery is the exploratory analysis and modelling of the huge data stores, where meaningful and useful patterns are extracted from the huge and complex datasets. Data mining being one of the important steps in developing a model by which can further be used in finding of new unknown patterns. Knowledge discovery Process is further divided into the following parts.

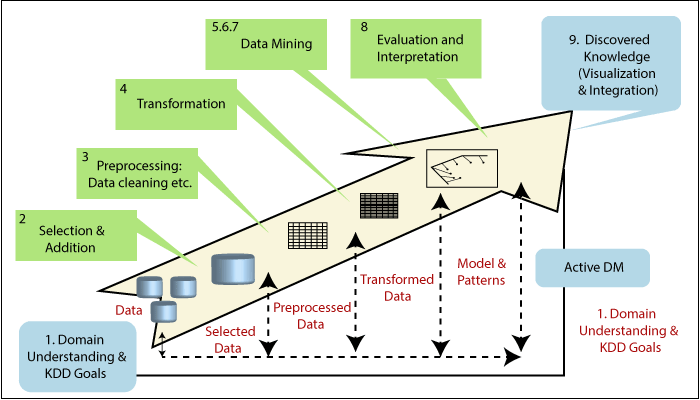


Figure 1.1 process of KDP

* Domain understanding and KDD goals
* Data selection
* Data cleaning and pre-processing
* Data Transformation
* Prediction and description
* Selection of algorithm
* Implementation of data mining algorithm
* Pattern evaluation
* Representation

1. ***Domain understanding***

Initial step for developing a model is done first by understanding the goal and the other requirements for the process that also includes setting up the environment. Once the goals are understood, only then the other steps like, data selection, pre-processing of data etc. could be started.

1. ***Data Selection***

Is considered, one of the most important step towards knowledge discovery process, this consists of finding all the valuable and accessible data and integrating the m into one dataset. It is considered as the building base for the models as the data mining techniques are applied onto the selected dataset and patterns are extracted from the dataset and later evaluated. /if in any case, nay significant data gets missing the entire research could be unsuccessful. Each and every attribute could make a huge impact in respect to knowledge discovery and modelling.

1. ***Data cleaning and pre-processing***

Even after the selection process, there’s always some amount of data in form of outliers or unrelated to prediction process or missing attributes, which causes issues during the modelling phase. For example, if any dataset has many attributes that either have some missing values, a supervised algorithm could be used for the prediction of missing values. In the case of outliers, they can be detected by the use various techniques and could be visualized for easy recognition of outliers.

* Normal distribution: empirical relations of normal distributions can be used.
* Skewed distributions: Inter-quartile range or IQR rule.
* Other distributions: Percentile based technique.

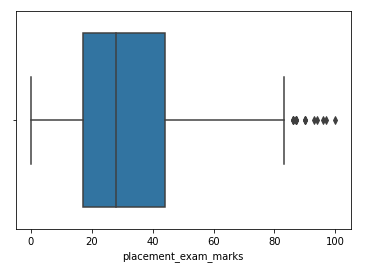


Figure 1.2 shows use of IQR method for detection of outliers

(https://lh3.googleusercontent.com/JFHr-1HWfN7RCcWAJOnTSw3ieLWvXcD2v6T4puZ5jKt5fzjzi0JfOdq2Y0e4kkEuIHCs55HJYgD4eHoQDlHfcFA-qnG-WUHzgBztgRA6xtXEgdMkQUHWJPDz3Pxelh8PQgRgrgj\_)

1. ***Data Transformation***

During this step, the data is actually turned and developed into a better version for the data mining process, which mainly consists of dimension reduction and transformation of the attributes. For dimension reduction various methods like feature selection, extraction and record sampling can be used while for transformation, functional transformation or discretization of numerical attributes could be used.

### *Data Mining*

What is data mining and why is it important? Rama Chandra discusses about the importance of the data mining in performing predictive analysis, which directly help us in gaining the insights about the dataset and making quick decisions. Data mining, means extracting valuable information from huge datasets, data mining tools helps us in finding of hidden patterns that are outside the expectations of the humans. Prediction of future probabilities is called predictive analysis and is widely used in predicting upcoming events or trends. For instance, weather forecasting, stock prediction, prediction of diseases could be done with the help of data mining tools and predictive analysis. This step consists of choosing the right task for data mining, what approach should be taken, is it solvable by regression, classification or by clustering to picking out the right data mining algorithm, which includes splitting the dataset into training and testing along with implementing it several times, till the desired outcome is shown.

1. ***Evaluation***

During this phase, we evaluate how well our model is working with respect to the goals set in the initial stage. All the changes in the parameters during pre-processing steps are being considered and the change in results as well. This step generally shows and focuses on how easy it is to understand the model and its usefulness.

1. ***Using Discovered knowledge***

This is the last step where all the feedback on the patterns are received at the user’s end, these feedbacks are used to integrate new features in the model and the effects are measured. The reports generated from the reports are visualized by use of various graphs.

## Types of Machine Learning

Machine learning is basically categorized into three broad types of learning methods, each having different functionality and usage. All of them having their positives and negatives but that too depends on the type of the data used. In machine learning, data is definded into two types: -

* Labelled: This type of data has both input and output parameters, which are in machine readable format, but requires a lot of efforts.
* Unlabelled: This type of data only has one or no parameters but in order to work in it, more complex algorithms are used.

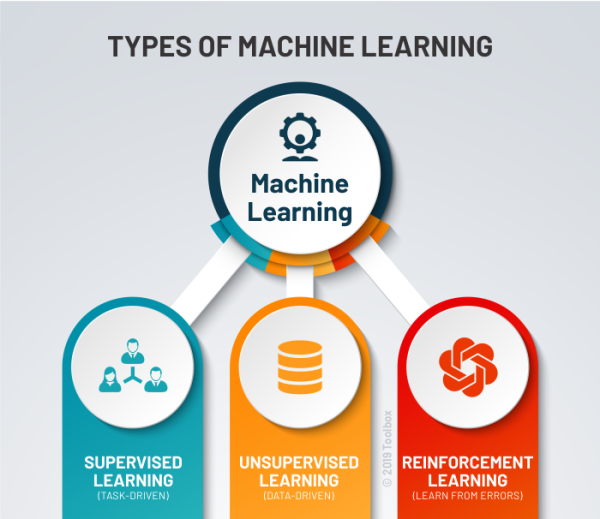


Figure 1.3 shoes differenttypes of machine learning

### Supervised Learning

Supervised machine learning is one of the most common learning methods used for classification and regression problem. It used labelled dataset for the training of algorithms which are used to classify or predicting value. For example, image recognition, predictive analysis, customer sentiment analysis or spam detection. Supervised machine learning comprises of various algorithms and computational techniques.

* Neural networks
* Naïve Bayes
* Linear Regression
* Logistic regression
* Support vector machine(SVM)
* K-nearest
* Random Forest

### Unsupervised Learning

Unsupervised learning techniques doesn’t require the users to supervised their model instead they work and discover patterns on their own. Unlike supervised learning methods, they use unlabelled data to perform complex processing tasks but are less accurate in contrast to the supervised learning algorithms. Customer segmentation, Genetics(clustering), anomaly detection etc. are some examples of unsupervised learning methods. Some of the most commonly used and popular unsupervised algorithms are listed below:

* K-means clustering
* Hierarchal clustering
* Anomaly detection
* PCA (principle component analysis)
* Neural Networks

### Semi-Supervised Learning

This learning technique is a hybrid between the supervised and unsupervised learning technique. These type of learning involves training the algorithm on a small scale of labelled data and the vast majority of unlabelled data but its applications are somehow limited. Some examples are: speech analysis, web content classification, text document classifier etc.

### Reinforcement Learning

Reinforcement learning is somehow different than all the previous learning methods, a mechanism, where desired action are rewarded whereas negative behaviours are punished. This technique gives positive values to the desired outcomes, which encourages the agent to follow a long term and maximum reward for achieving an optimal outcome and punishes for undesired outcomes. Application of reinforcement learning includes gaming, recommendations and robotics.

## Approaches for Breast Cancer Detection

### Machine Learning

Machine Learning can be defined as a sub branch of Artificial Intelligence, where machine gets the ability to work on data and learn accordingly, the data helps the machine in training in comparison to the approach of coding all the possible outcomes. With the development, detection of various diseases became easier to detect, just by feeding the data to the machine. This has led to detection of diseases at earlier stages and thus, increasing the mortality and survival rate at a huge scale. This automation process provides higher detection rates and efficiency as the new methods are far reliable than the old traditional ways. There are many algorithms like logistic regression, k-nearest neighbour, support vector machine, decision tree and random forest, which are likely to have a higher.

The cancer dataset used is the Wisconsin breast cancer, the data set has over 569 rows and 33 columns also called features. These 569 samples have 30 features describing the data and the two features perimeter\_worst and concave points\_worst were found to be the most important using the feature extraction. For evaluating different machine learning techniques, we generally use the performance metrics for evaluating the performance of the model. In order to do that, a confusion matrix is used for checking the actual and predicted values that mainly has four parameters namely TruePositive, TrueNegative, FalsePositive, FalseNegative.

* Accuracy: Accuracy of a model shows the correctness of the model in the training process and its performance. In other words, it is defined as the number of correct predictions made in contrast to the wrong ones.
* Recall: It is defined as the ratio of correctly predicted positives values to all the other values, also known as sensitivity. It could be also defined as the ability of classifier to search positive values. It ranges from 0(worst) to 1(best). Ratio of TruePositive to sum of all observations (TruePositive + FalseNegative) is recall.
* Precision: It defines the quality of the positive predictions done by the model or the degree of correctness in defining the positive results. Ratio of number of TruePositive to the total TruePositive + FalsePositive is precision.
* F1 Score: It is the mean of precision and recall, ranges from 0 representing low recall and precision while 1 showing high recall and precision.

### Logistic Regression

The early detection of the breast cancer is proved in increasing the survival chances of the patient. Many computational models have been build and tested. Although the predictions made by the clinical methods might be valuable but it is challenging in many ways. Logistic Regression is one of the most widely used statistical fitting in the medical field as it s relatively easy to build when compared to others for estimating diseases risk in many domains. Many studies in the past have shown that logistic regression useful in the diagnosis process as the ultimate goal is to integrate these computational models into the clinical practices and over the last two decades, it has been increasingly used as an analytical tool in the medical field.

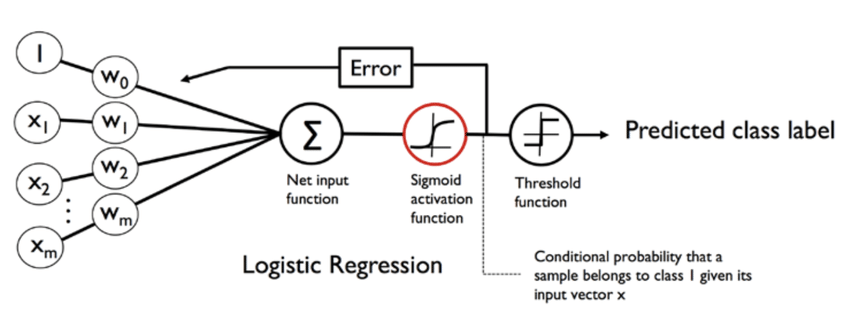


Figure .4 demonstrate working of logistic regression

(https://www.researchgate.net/publication/334575492/figure/fig7/AS:782432886349826@1563557945492/Architecture-of-a-Logistic-Regression-Model-56.ppm)

**Turgay Ayer** et al () says that logistic regression works by checking the relationship between two dependant variables by computing the possibility of something happening or already present which lies between the interval 0 and 1, then can be used to classify the data such as detecting diseases and variables that act as predictors and states that with an example of breast cancer, where chances of having breast cancer can be predicted by the knowledge of the patient’s family history, age and density of breasts. In general, logistic regression only includes the variables that are considered valuable in the prediction process and also states that if the number of samples is very large, then the predictors who have very little effect on the result will also become important. So, in order to avoid that situation, harsher measures can be used. There are various selection methods through which only more significant variable can be used, In forward selection just by putting each of predictor variables into an empty model which is basically a model with no predictor variables. In backward selection, each of the variables are checked and if proved insignificant, are removed from if proved insignificant. He also states that sometimes even clinically proved important variables can be insignificant with different selection techniques as their influence can be reduced by strong variables. During the evaluation of the model it was found that logistic regression was able to achieve an AUC of 0.963 ± 0.9. In other research lead by Cheng-Min Chao et al () proposed another model for classification of breast cancer detection and the model was successfully able to yield an accuracy of 90% + with 10-fold cross validation for identification of model. The data was split into multiple subsets with equal sizes and also stated that stratified cross validation shows displays output with lower bias and variance than regular k-fold cross validation. The logistic regression model achieved an accuracy of 95.1%. Ram MurtiRawat et al () in his evaluation of different models for breast cancer detection on the Wisconsin Diagnostic dataset uses principle component analysis for the feature extraction and the number of components were said to be 17. After the implementation of PCA, logistic regression model is used for classification process. Using test-train split function and examining the results, the model was able to achieve an accuracy of 97.70%. Another approach was taken for the detection of breast cancer by Ahmed F. Seddik et al () with reduced number of features with the employment of binary logistic regression which could discriminate between cancerous and non-cancerous tumours. The features that exhibited low correlation are used to build model. The end results turned out to be fascinating as the model had an avg classification accuracy of 98% and 98.5 %for sensitivity.

### K-Nearest Neighbour

In this article by Sayyid Ahmed Medijahed et al () states about the use of K-NN in detection of breast cancer. K-NN is a supervised learning method which is generally used for classification problems. The working of K-NN is simple, In this learning method instances does not require learning. A distance function is used for the training sample, as new elements are discovered, they are compared to the other elements for a similarity check. In other words, it assumes that every single data point that is near each other falls in the same class and the class which appears maximum number of times among the neighbours is to be classified. If the value of k is assigned as 10, it will only look for 10 nearest neighbours for the data point. He also says that for the proper functioning totally depends on the number of parameters and the type of distance used. Distance is defined as the distance between two points and in K-NN, we use the different types of distance to form the group of individual data points that are similar and separate those who don’t. Distance can be classified into six types.

* Cityblock distance or 1-distance
* Euclidean distance or 2-distance
* Minkowski distance or p-distance
* Tchebychev distance
* Cosine distance
* Correlation distance



Figure .5 shows the formation of similar data point

(https://miro.medium.com/max/1222/1\*wW8O-0xVQUFhBGexx2B6hg.png)

With the experimentation, it was found that the accuracy with the K-NN was high around 98% while using different types of distances (Euclidean and Manhattan), even increasing the value of k from 1 to 50 didn’t had any effect on the accuracy values but there were many drawbacks along with that which includes the need of huge storage as it stores all the training data, the prediction time was slow and the algorithm itself gets slower with the increase in number of observations, predictors etc. In another study led by Md. Milon Islam et al () did a comparative study between two models using supervised learning algorithms ( SVM and K-NN) on Wisconsin breast cancer. The evaluation for the models was done in terms of sensitivity, accuracy, false discovery, specificity and Mathews correlation coefficient. The dataset consisted had 11 features and discovered that the dataset had 16 missing values and for those missing values, a mean of the feature. The dataset was split into equal k block size and one was used for the training of the model using 10 k-fold validation. Performance measure indices was used for the evaluation where they found K-NN had value of 0.96 for the Mathews correlation coefficient which indicates that it was a pure binary classifier. Feature selection plays an important role in the predictive analysis of the model. Sara Laghmati et al () proposes use of neighbourhood component Analysis for the feature selection on supervised machine learning techniques. She talks about the various feature selection methods used to identify the relevant and irrelevant feature in the dataset. Neighbourhood component analysis (NCA) method which is generally used for increasing the accuracy for classification methods. NCA works similar to k-NN, finding a space where the data points having same labels are closer to each other and could be used for dimensionality reduction. It creates a loss function for its parameters and optimizes it with gradient descent. In the training and testing phase, data was split into 80 and 20% respectively. In the end it was found that k-NN had an increase in accuracy for predicting breast cancer when compare to models proposed by Puja Gupta et al () .95, Vikas Chaurasia et al () .76 and M.H Tafish et al () .76. The present K-NN model had an accuracy of 99% in classifying breast cancer with NCA as feature selection method. Kernal Polat et al () supports the idea of using least square SVM algorithm for breast cancer detection. LSSVM

### Support Vector Machine (SVM)

Support Vector Machine is another linear classification technique that also falls under the supervised learning method. Muhammad Hussain et al () explains that support vector machine has the ability to outperform every other classification technique used for the detection of breast cancer and is heavily used for pattern classification and regression. It uses input in the form of feature vector. Support Vector Machine works with the idea of creating a hyperplane that separates the data points or observations into classes. Hyperplane in SVM can be defined as an n-dimensional Euclidean space. If the ambient space is defined as n, then the hyperplane is n-1 or the subset of the ambient space.

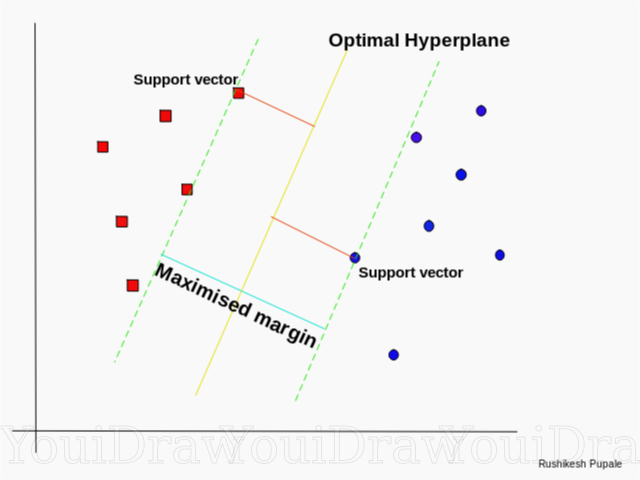


Figure 1.6 illustrates the formation of optimal hyperplane

It requires linearly separable data but sometimes the feature vectors are not linearly separable. For solving this issue, the original input space is transferred into a high-dimensional space where it could be separated linearly. Muhammad Hussain et al ()

Also states that the performance of the support machine classifier also depends on the type of kernel function used. The outcome was that the using the SVM with different kernel functions might be yielding better outcome. Mehmet Faith Akay in his research proposed a SVM model working on feature selection. For his experimentation, he uses Wisconsin Breast cancer dataset (WBCD) and the model was evaluated on the basis of performance metrics and confusion matrix. The model uses grid search technique for selection of the optimized parameters and calculates the f-score for selection of features. Mehmet Faith Akay also emphasis on the selection of kernel as it can drastically increase the classification accuracy, In his research he chooses RBF kernel. During the training and testing phase it was found that the data split with 80-20% had the highest accuracy (99.52) when compared to the other models. Kernal Polat et al () considers using least square support vector machine for breast cancer classification. Wang and Hu (2005) in their findings preferred LSSVM, as it was found to be more efficient when it comes to solve large scale problems over SVM. During the training and testing phase the data was split into 50-50% , 70-30% and 80-20% respectively, and the model at the end was able to achieve 95.89% accuracy for 50-50 split, 96.59% for 70-30% split and 97.08% for 80-20 split using RBF kernel and 98.53% using 10 k-fold validation technique. LSSVM was considered to be more accurate for examining data in short period of time and in a more detailed way.

### Decision Tree

Alaa. M. Elsayad et al () says that the decision tree has some powerful techniques for the purpose of classification and prediction. Decision Tree works by repetitively divides the data points into a tree structure (branches) for improving the accuracy of the prediction made by the model. Each tree node is further divided into either a leaf node or a decision node and following through different decision node would lead to a different outcome. Building a decision tree comprises of following steps.

* Training of the training subset and tree.
* If all the samples at the node are of same class, then node is labelled as leaf node with a label.
* If not, then the splitting attribute with the more significance in segregating the training samples into classes is selected.
* Then for each distinct value of s, a branch is created and the samples are separated accordingly and this process is continued in a loop until a desired value which can fulfil the criteria is achieved.

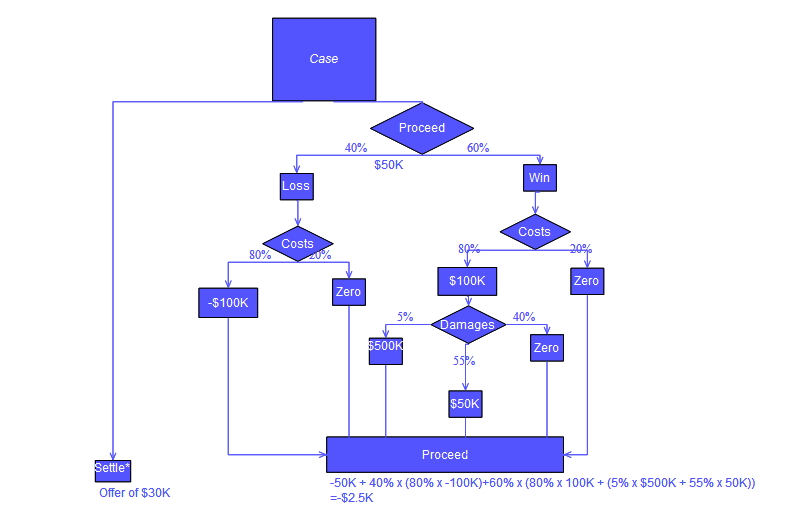


Figure 1.7 shows the flowchart for the Decision tree working

(https://upload.wikimedia.org/wikipedia/commons/4/48/DecisionCalcs.jpg)

Alaa. M. Elsayad et al () concludes with the results from his experimentation, where he shows the performance of four different models of decision tree CHAID, C&R, QUEST, C5.0 and all of the models were able to achieve almost 100% classification sensitivity with a distribution of training and testing data into a ratio of 7:3. P. Sathiyan Anarayanan et al () did a study did a comparative analysis of various machine learning techniques and wasn’t satisfied with the previous results. So, he focused mainly on the working of Decision Tree. For the experimentation, WBCD was used and during the training and testing phase data was split into 80-20%. Without using feature selection or other techniques, the model was able to predict the type of cancer was 99% while compared to K-NN. In my opinion using different types of models does gives a clear view about how the performance, accuracy, precision of each model is different from each other, isn’t sensitive to outliers, easy to understand but on the other hand, decision trees are prone to over fitting, doesn’t guarantees optimal trees and could become more complex with the increase in class variables.

### Random Forest

In this paper Bin Dai et al () talks about prediction of Breast cancer using the machine Learning techniques and their outcomes. He puts emphasis on use of Random Forest algorithm and also states that with the combination of multiple characteristics of eigenvalues and results of multiple decision trees, the outcomes could be predicted more accurately and in the past decade has made a huge impact on auxiliary medical diagnosis. TIn Kam Ho, who created the random forest algorithm believed that as long as decision tree is able to differentiate the sloping hyperplane, better accuracy could be achieved without exaggerating and the random selection of features is one of the precondition for achieving higher accuracy. Another advantage of using the random forest is that even the small batch of data is distributed among the algorithms for sampling and because of this random approach can lead to desired outcomes.

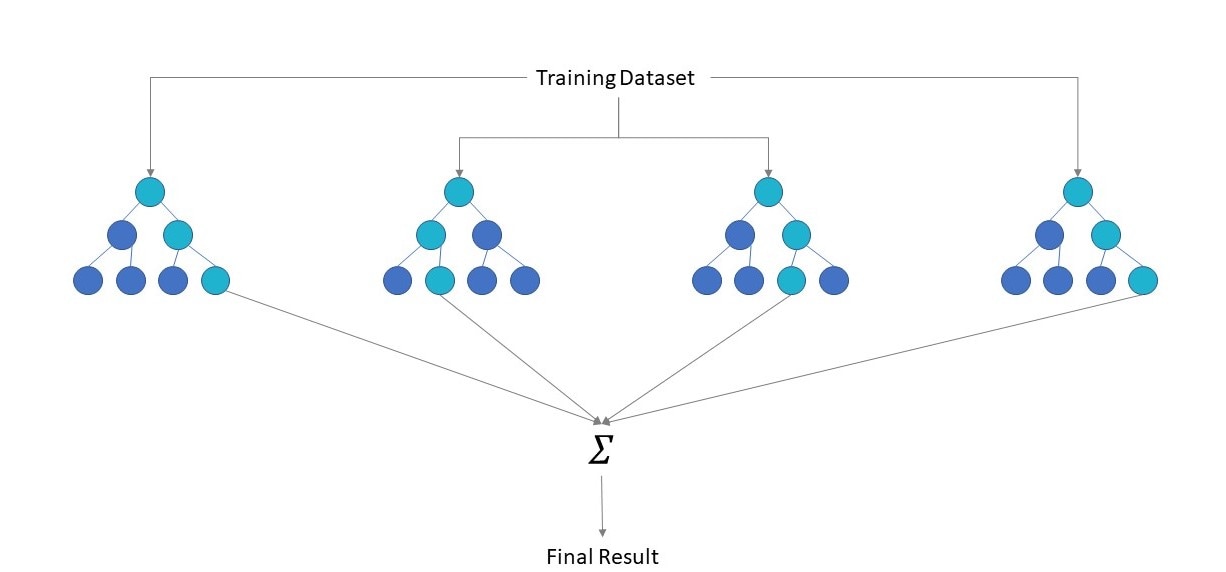


Figure 1.8 illustrates the working of Decision Tree

While building the decision tree, we simple take the benefit of the random selection of features. In simple words, if the dataset has 60 features, random features are selected and to obtain the accuracy, the dimensions shouldn’t be too short or too big as it would reduce the independence of each decision tree. The model was trained and tested with the Irvine Breast Cancer Wisconsin dataset, the model was able to achieve an average accuracy of 95% in classifying the nature of the breast cancer (benign or malignant). Another study aiming at diagnosing and predicting breast cancer using Random Forest classifier with feature selection technique by Cuong Nguyen et al (). The process was divided into two phases. The first one, training and testing was done and the feature selection and ranking was done and using Bayesian probability and feature impurity, a ranking value for features is assigned in ascending order. For checking the importance of each feature, backward elimination method is used and during the second phase, training of the classifier was done on the same set of selected features for improving the classification accuracy of the model. Due to this, number of tress formed in random forest was 25 with a set of 15 features. An accuracy of 99.82 % was achieved by the model.

## Deep Learning

In her article Priyanka et al () describes about the advancement in development of the deep learning techniques which has a huge potential in finding solutions for medical imaging problems. Deep Learning is considered as an unsupervised learning method as it has the ability to learn from the data without human interference. The data fed to the machine could be unstructured or unlabelled, unlike the supervised learning methods, which requires labelled data. Twain Taylor in his article explains about the structure of deep neural networks which generally has three types of layer.

* Input Layer: Input Layer consists of the input neurons that brings in the data for the whole neural network which is further used for processing by the other layers in the network
* Hidden Layer: The deep neural networks consist of one or many hidden layers inside the structure. These hidden layers allow us to model the complex data by processing it with the help of mathematical functions. Each layer, sometimes might have only a specified function, depending on the use and architecture. For example, for classification of objects CNN is used but a hidden layer that has the purpose of identifying a single part of an object but with the additional layers, it could identify other features of the object.
* Output Layer: The output layer is the third layer which shows contains the output layer.

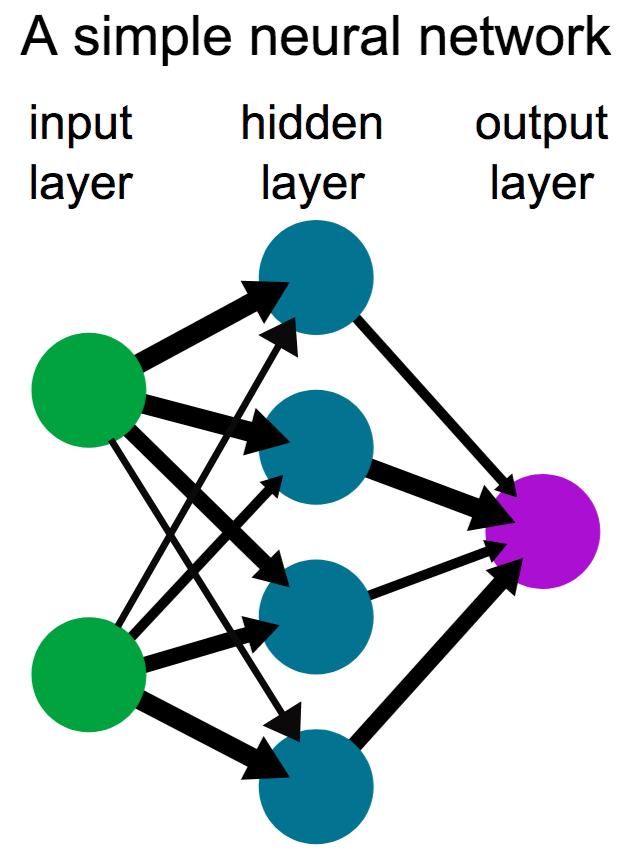


Figure 1.9 shows the structure of the deep neural network.

### Working of Deep Learning

Ed Burns et al () in his article discusses about the deep learning and how it works in a similar way as a human tries to learn. Being a part of the supervised learning, it does need labelled data and is much faster when it comes to processing it. After the data is taken as an input, nonlinear transformation is done to it and a statistical model is formed labelled as output. The same output is used as the input and this process is iterative, if a desired degree of accuracy is achieved, then it stops at the point but requires huge training data and processing power to complete the tasks. Solving the problem with neural networks is not only fast but more accurate with each iteration. For example, the computer is provided with a set of trained data which consists of images of a cat and each image has been labelled with cat or not cat. The program after processing, it receives information from the training data. A predictive model for the generated feature set of cat is created. The first time, model would have the perception of anything that has four legs and a tail should be classified as a cat. In the next step, it will learn another feature of a cat and with each iteration, the model will be able to predict more accurately but it becomes more complex in the process. Deep neural network has a large number of applications from pattern recognition to computational learning, a deep neural network is similar to an artificial neural network but the key difference between a deep neural network and artificial neural network lies in the multiple layers found. In a survey it was found that deep neural network was able to reduce the error rates to a greater extent than the traditional methods. There are so many types of deep neural network but three of them are widely used for most of the problems.

* Multi-Layer Perceptron: In her article Carolina Bento explains about perceptron which was referred as an image recognition machine, has similar functionality like a human eye, observing things and recognizing them. In 1958, Frank Rosenblatt proposed a perceptron model whose aim was to mimic the working of logic gates. After receiving the inputs, neurons select a random set of weights and they are used to form a weighted sum. The output value is defined by the activation function (RELU). The limitation with the proposed model was that it wasn’t able to represent every logic gate and then the concept of multi-layer perceptron was proposed. Unlike, the single layer perceptron, multi-layer perceptron has input, output and hidden layer full of neurons and with help of backpropagation, the weights could be adjusted in each iteration inside the network with low cost function.

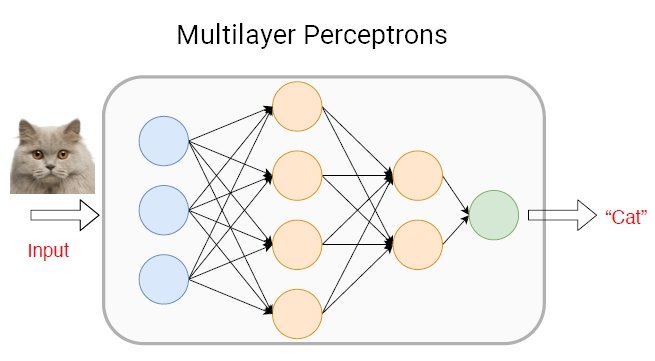


Figure 2.0 Multilayer perceptron

(https://viso.ai/wp-content/uploads/2021/04/multilayer-perceptrons-MLP-concept-1.jpg)

* Convolutional Neural Networks: CNN is another architecture used in deep learning which is heavily used for object recognition and could perform effectively against non-image data as well. With the help of CNN, feature extraction could be done automatically, has high accuracy when it comes to image recognition and prebuild models could be reused for training of new recognition tasks. There are several CNN architectures that are being used for image classification (AlexNet, VGG-16, GoogleNet, ResNet etc.) and object detection (YOLO, SSD, Fast R-CNN etc.). CNN has similar structure to neural networks when it comes to weights and biases, during the training process model learns from the values and keeps updating each time while training. The most commonly used layers are convolution, ReLu and pooling. Training is done over hundreds of layers and in each iteration, layers keep identifying different features, architecture of CNN shifts to classification.

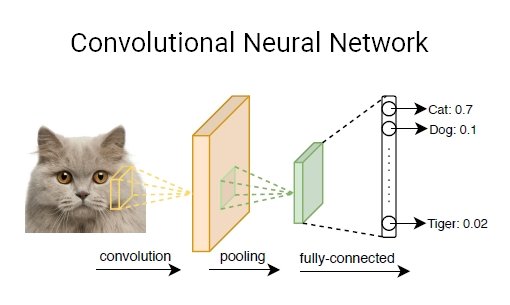


Figure 2.1 shows working of CNN

(https://viso.ai/wp-content/uploads/2021/04/convolution-neural-network-cnn-concept-1.jpg)

* Recurrent Neural Network: RNN works on sequential feeding of data and is mostly used for time series problems. RNN has similar behaviour like functioning of a human brain. The input and samples from the previous input are used as the current input which leads to formation of a directed graph between the connected nodes. RNN model has a huge usage in Natural Language Processing (NLP) due to high quality of processed data with input of no fixed length and due to presence of internal memory, which gives it an advantage for solving sequential data related problems.

### Deep Learning in Breast Cancer Detection

Computer Aided systems with the help of Deep learning architectures like RNN, CNN and DNN had been extensively applied to medical image analysis in the past few years and have been found more accurate and efficient when compared to the traditional machine learning methods. In a study led by Saad Awadh Alanazi et al () proposed a system that utilizes convolutional neural networks and its architectures for automated detection of breast cancer. The proposed CNN consists of three layers, convolutional layer, pooling layer and fully connected. Three CNN models were created for the experimentation, model 1 only having two convolutional layers with 32 and 64 kernels respectively. The activation function (ReLu) is being used for all the layers except the output layer as the Softmax function is being used. A batch size of 128 and 12 epochs is used for training. A validation loss of .69 per was detected with accuracy level low as 59%. For the second model, the number of features were increased and the number of convolutional layers were tripled but the accuracy had a slight increase and was reported to be .76. The last model used a five-layer deep convolutional neural network layer and gave out an accuracy of 887% that was the best out of the other two and similar distribution of predicted labels was shown to that of actual ones. Shuyue Guan and Murray Loew, created a single CNN model from scratch but used pre trained VGG-16, another architecture used for classification problems for feature extraction and used back propagation method for tuning the weights in the final layers. Training a CNN model from scratch is usually quite challenging as it requires lot of labelled data. Therefore, transfer learning was found to be an easy alternative solution. They found out that using transfer learning, pre trained VGG-16 with FC NN classifier could achieve an accuracy of 90.5% for fine-tuned model which was 0.008% higher than the feature extraction model but the time cost function was 5% of the fine tuning model. Ahmed Hamza Osman and Hani M. A. Aliahdali came with a new idea of using Radial Based Function Neural networks for assessing and diagnosis of chronic diseases. They used ensemble features along with boosting method for gaining more accuracy and precision in the RBF NN model. The datasets used for the proposed method consisted of WBC WBCD, BCP and BCD. RBF having fast training, easy designable, robust tolerance to the noise features and has the capabilities to calculate the data extensively. In their fining s using ensemble boosting methods were found improving the learning rate and accuracy. During the training and testing phase, features from the dataset were used as the input for RBFNN. The combination of winner takes all and most weighted strategies gives ensemble learning advantages in reducing the misdiagnosis and increase in accuracy as the weak learners acting as strong ones. The end results were quite fascinating as a comparison test was done between the models, one without ensemble boosting methods. The model with ensemble boosting achieved 97.36% and 97.4% during the training and testing phase for WBC database and other one achieved an accuracy of 95.92 and 97.6 % during training and testing. The best result for EBL- RBF NN was 97.09 and 98.4% in the entire comparison but there was a limitation found while using the model in CAD, doctor has to support the diagnostic stance which can delay the detection rate.

#### **Tensorflow**

Tenserflow is an open source library that is mostly used for computational task and training of machine learning models at a large scale. It consists of vast number of pre trained models and algorithms used in machine learning and deep learning as well which makes it easy as one doesn’t have to go through acquiring large set of data and training models with it. These pre trained models and algorithm can be used for several task which includes image classification, object recognition, natural language processing etc. There are several reasons why tensorflow is in heavy demand as the architecture itself is based upon the graph computation. The architecture of tensorflow consists of following components.

* Loaders: are regarded as the continuation point from where another dataend or an algorithm could be added.
* Servables: instances that are applicable for performing computation.
* Models: described as one or multiple servables that could make up for a section of a model.
* Batcher: reduces the cost of running data points in the model for computing a task.

# Methodology

This section deals with the planning and structure of all the models proposed for the comparison analysis between machine learning and deep learning algorithms but in order to start building the models, we need to understand the basic concepts of machine learning and deep learning. Most of the machine learning models has to go through the following steps shown in the figure 2.2

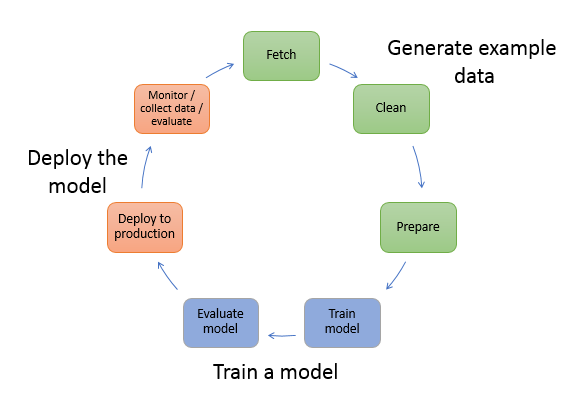


Figure 2.2 displays the basic working of a machine learning model

## Architecture of the Machine learning model

In this comparative analysis we have designed two models, one on traditional machine learning methods with breast cancer Wisconsin dataset and the other one is based on deep learning with inbuilt breast cancer Wisconsin dataset. The other two models are based on medical imaging (deep neural networks). The dataset has 569 observations and 33 features including null values, that would be eventually removed while cleaning the dataset. In the next few steps, I’m going to briefly explain the working of the models.

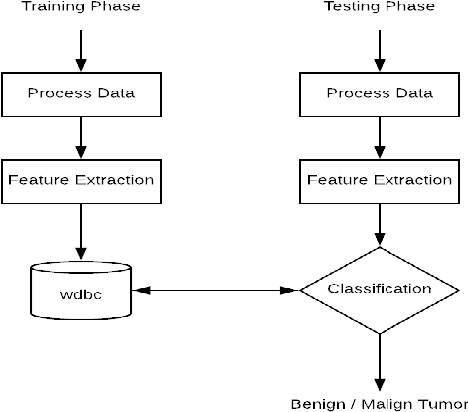


Figure 2.3 shows the global architecture used for building machine learning model

### Data Pre-processing and visualization

The first step towards building the machine learning model is to perform data cleaning where all the null values are either dropped or a mean of those value is assigned in place of the null values. Data cleaning is followed by data pre-processing, where the features in the dataset are checked for correlation. In simple words, how the features are related to each other and the features with lowest correlation values are dropped out of the dataset

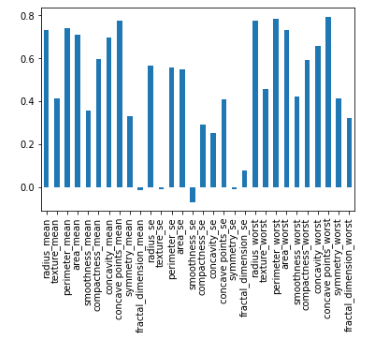


Figure 2.4 shows correlation

LabelBinarizer is used for turning all the variables into binary format inside the matrix where they are referred as the columns and with the help of the seaborn library, a heat map is generated that shows the correlation between the features/predictors.

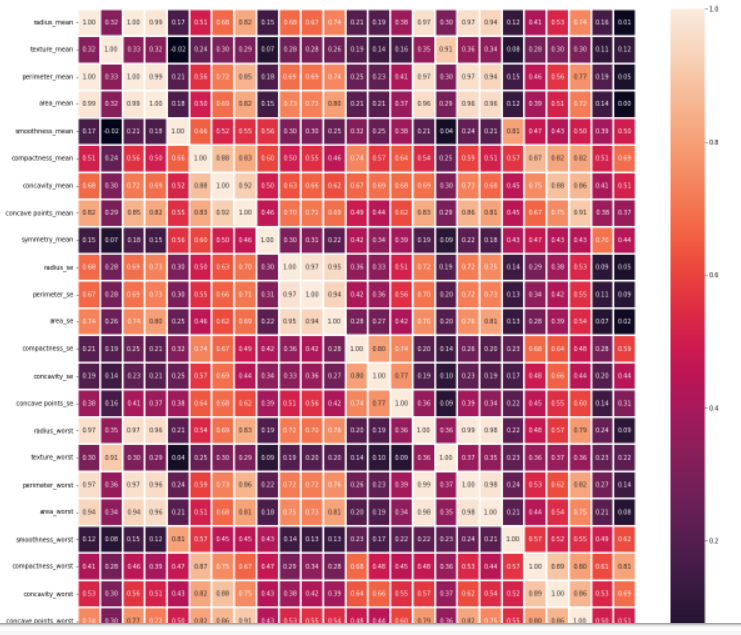


Figure 2.5 shows correlation heat map

### Data splitting and training of model

For testing and splitting the data, stratified shuffle split is being used for providing the training and testing data and is a combination of startifiedkfold and shufflesplit. Before breaking the indices for training and testing set, it shuffles the indices but there is no guarantee that it won’t give out folds in repetitive manner. To tackle this, it maintains the proportion of the classes.

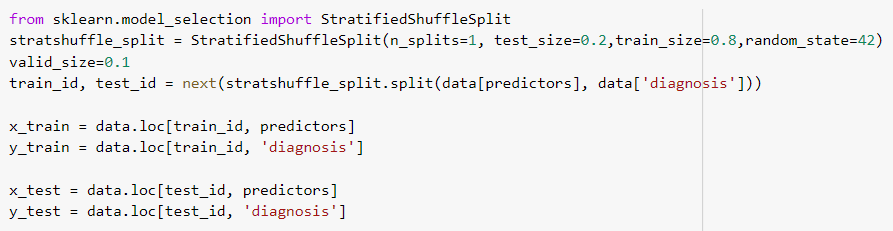


Figure 2.6 shows splitting of data by using stratified shuffle split

Standard scaling is used for elimination of mean and scaling the unit variance and is done by processing statics on the samples in the training set. Standardization is one of the requirements for the learning estimators so that the data looks standardly distributed. For example, elements uses objective function of learning algorithms and if any feature has variance of huge order which is usually of same order but then it would cause problem for the estimators as the feature might dominate the function of the algorithm. Some of the most common methods used for standardization includes: -

* Fit () – calculate the mean and standard deviation
* Fit\_transform () – fitting and transforming the data
* Transform () – cantering and scaling of data for standardization

### Implementation of Machine Learning Algorithms

The first model is based upon supervised learning methods that includes logistic regression, support vector machine, decision tree, K- Nearest neighbour and Random Forest and for the evaluation, performance metric is being used in terms of precision, recall and f1-score.

* Precision: defines the proportion of positive prediction that were actually correct.
* Recall: defines the ratio of actual positives that were predicted correctly
* F1-score: tells us the mean of the precision and recall.

1. Logistic Regression

The logistic regression model was designed as an alternative to the linear regression as linear regression was good for solving regression problems but when it came to classification, it was a total disaster as it lacked output probabilities and thus, for classification logistic regression was chosen. Instead of fitting a straight line or hyperplane like linear regression, logistic regression uses logistic function to press the output from linear equation lying from 0 to 1. Even though logistic regression can be used for multi class classification, it still has some drawbacks, the interpretation of weights is challenging as it is multiplicative and not additive, suffers from complete separation and has challenges with the restrictive expressiveness.

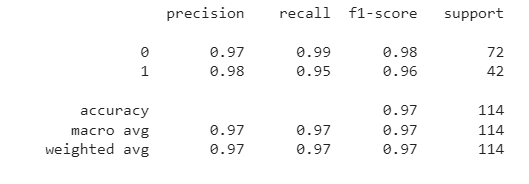


Figure 2.7 logistic regression output

The outcome from the logistic regression was quite good, it was able to get a precision of 97% and 98% for benign and malignant. Recall and f1-score was highest for classification of benign (99 and 98%) while for malignant, it was noted to be 95 and 96%.

1. K-Nearest Neighbour

K-Nearest Neighbour is one of the easiest supervised algorithm to deploy, works on the concept of similar data points similar and close to each other. The working of KJ-NN includes loading the data, initializing the value of k for the number of neighbours, calculate and add the distance along with index, sort the indices and distance in ascending order then select k entries from sorted group and if it is regression then return mean otherwise mode of k folds.

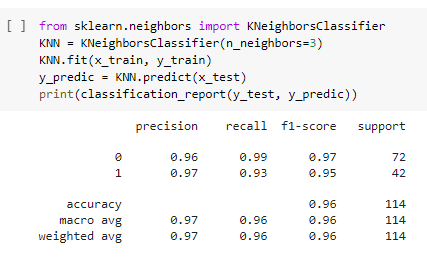


Figure 2.8 k-nearest output

Using the performance metrics, it was found that the classification of benign has the highest average score in terms of precision, recall and f1-score. I.e. 96%, 99% and 97% while for malignant it was 97%, 93% and 95% respectively. K-NN is being easy to deploy and doesn’t requires tuning of parameters and is versatile but and one big disadvantage that if the number of observations or predictors are increased, the algorithms degrades in terms of speed and in a situation when predictions are required rapidly, it slows down while there are algorithms with better accuracy when it comes to classification.

1. Support Machine Vector

Support vector machine being another supervised learning method could be used for both classification and regression. In my experimentation, Gaussian radial bias kernel is being used up which has the function to seek out any non-linear data in a linear based model and with RBF kernel, data points could be moved from one dimension to another. The other parameters gamma, which defines the influence of each single trained example over the distance it reached and C is used to set a tolerance margin for the model to lower the error rates. Using the performance metrics, it was found that the model for classification of benign was able to have a score of .97 for precision, recall and f1-score while for malignant .95.

1. Decision Tree

Another supervised learning algorithm that maps down all the possible outcomes by weighing down all the possible actions, comparing them against each other depending on their costs, advantages and probabilities forming a tree like structure.

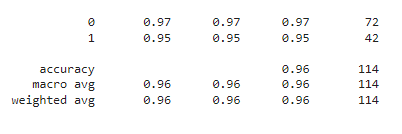


Figure 2.8 DT output

The results for decision tree and SVM were found to be same for benign and malignant (.95 and .97) for precision, recall and f1-score.

1. Random Forest

Random forest falls under the ensemble learning, which means generating multiple classifiers and combining them to solve a give problem. The difference between decision tree and random forest is that random forest does not depend on one single decision but on multiple which does makes it a bit slower but has higher accuracy than decision tree.

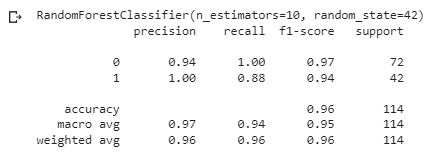


Figure Random Forest output

1. Feature Importance

It is important to know which features are important for the prediction model, feature importance consists of methods to assign a score which indicates, how important is the feature for the model.

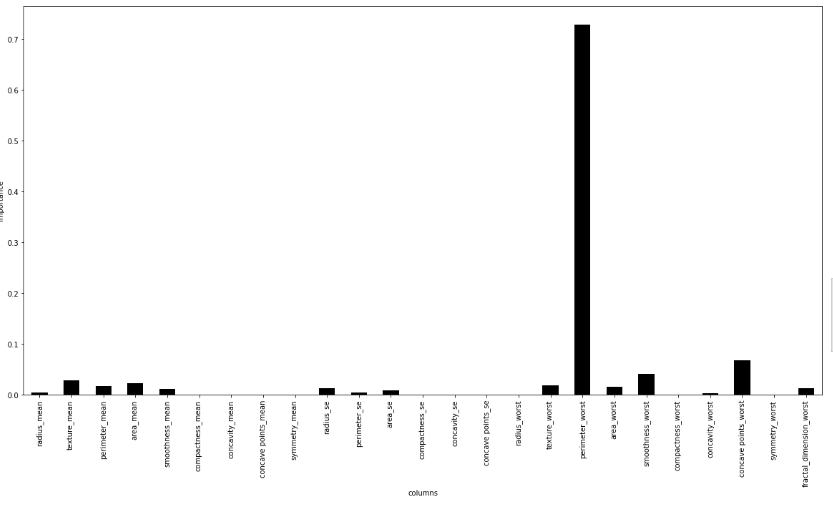


Figure 3.1 shows the feature importance chart

## Implementation of Deep Learning

The first step to build the deep learning model starts with importing important libraries along with functions and layers. We are using data that is in form of one dimension, so we import the layers in single dimension and we are using the Adam optimizer for our model which is an extension of stochastic gradient descent and is used for updating networks weights in training data. Importing train\_test\_split for splitting the data into two indices (training and testing) and standardscaler for standardization

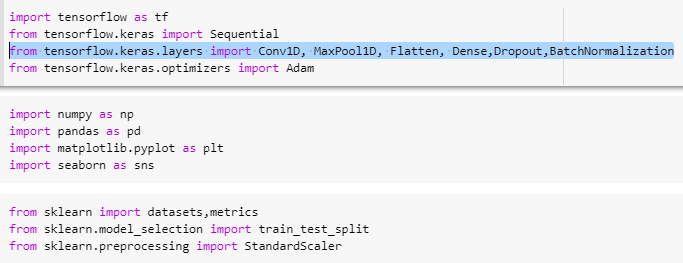


Figure 3.2 importing libraries

The next step involves initializing the sequential model along with the hidden layers.

Sequential model

* Filters: defines the dimensionality of output space
* Kernel\_size: specifies the length of convolutional window
* Activation: initializes the activation function
* Input\_shape: (30,1) shows 30 elements in the first dimension and 1 in second.
* BatchNormalization: adds more layers inside the hidden layer to make the network fast
* Dropout: is used to drop the neurons from the neural network as we are training the network on a small dataset due to overfitting problem.
* Flatten (): this function is used to transform the data into 1-dimensional array as an input for other layers

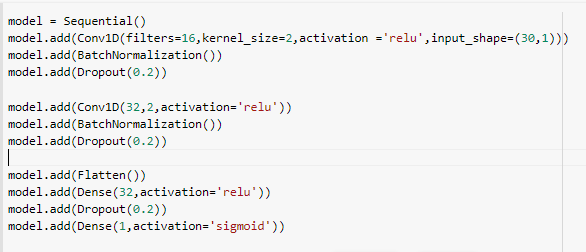


Figure 3.3 shows initializing the model and importing layers

model.summary () shows a summarized table for the model which consists a summary of layers and the output shape of it. After summarizing the model, the next step is to compile the model. While compiling, we pass the arguments like the type of optimizer we are using, learning rate at which the model learns, loss function which calculates the loss in predicted labels and actual labels. The last step is to check the history of the network while training in terms of accuracy, loss, validation accuracy and validation loss. The visualization part includes the graphs that show the performance of the model after each epoch in terms of model accuracy and model loss.

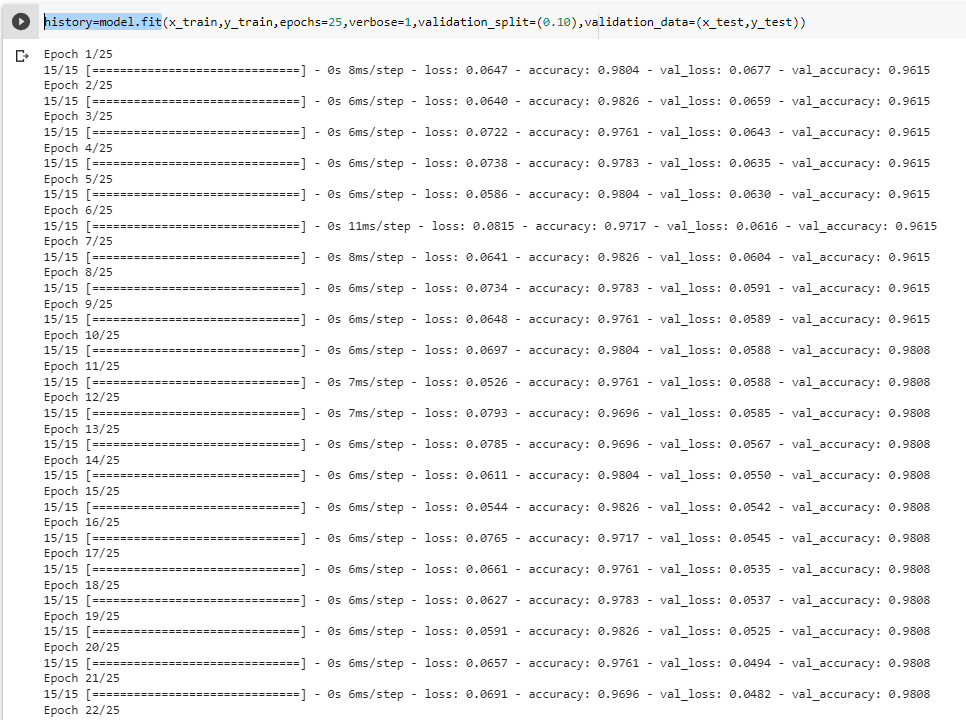


Figure 3.4 displays the history of trained model

## Implementation of breast cancer using CNN

The first step starts with importing essential libraries and uploading the dataset. The dataset being used for the model is the Breast Histopathology images dataset consists of labelled images 0 for benign and 1 for malignant. After uploading the dataset, we are going to check whether the images are being loaded or not.



Figure 3.5

Figure 3.5 displays the number of labelled images from the dataset 0 and 1 and total number of images. The next step is pre-processing the images and using random randint function random sampling in numpy for cancerous and non-cancerous images. Converting the images into the array for both types and labelling (-) for benign and (+) malignant. For both cancerous and non-cancerous images are being resized toa dimesions of (50,50)

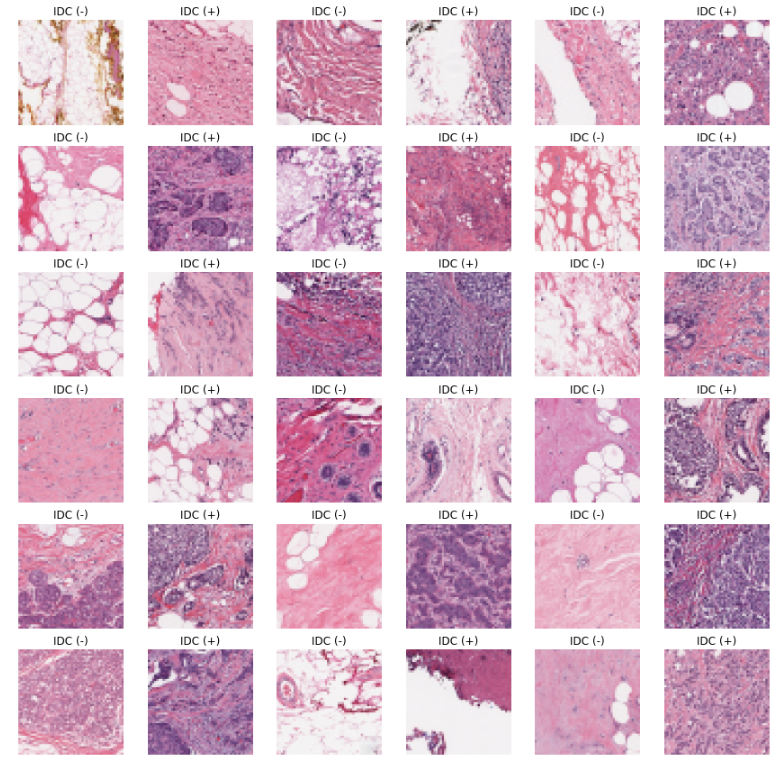


Figure 3.7 shows the bc samples with new labels

Now the array of both cancerous and non-cancerous are being added into a single array of shape (315144,50,50,3) by using random shuffle method, which is used to improve the performance and generalization of network.

### Splitting the dataset for training and testing

For splitting the dataset, train\_test\_split from skLearn is used to split the dataset into random subsets along with to\_categorical from keras is imported for converting the vector into the binary matrix. The dataset is split into a test size of 30 and rest for training.

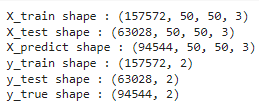


Figure 3.8

The next step involves initializing the hidden layers for the network and a sequential model, sequential model is easiest to build and as our model doesn’t have multiple inputs or outputs, it’s the best to use in these condition.

* Conv2D: convolutional layer which helps in producing the tensor of outputs
* Maxpooling2D: is used to downscale an image and replace it with a convolution for extraction of maximum and important features.
* Dropout: is used for neglecting the hidden and visible layers, for sorting the problem of overfitting.
* Flatten () : the function is used to copy of an array that is transformed into single dimension.
* Dense (): is used to generate connected layers, where the output is totally dependent on the input.
* Padding: padding = ‘same’ refers that the input is half padded, which refers to the amount of pixels added to the input image while being processed by the kernel.
* Activation function: decides whether the neuron should be activated and for the model two of them is being used ReLu and sigmoid, ReLu being the most widely used because of its computational and convergence speed.

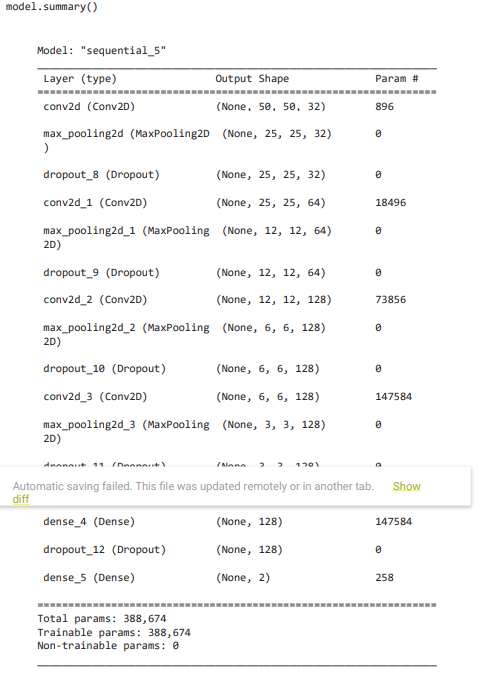
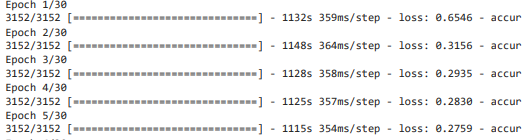


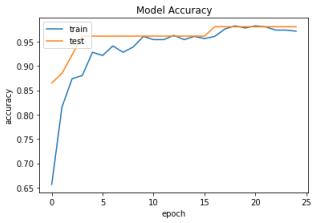
Figure 3.9

The summary of the entire network shows all the layers along with the output shape and total number of parameters.



history = model.fit(X\_train, y\_train, validation\_data = (X\_test, y\_test), epochs = 30, batch\_size = 50)

Returns the records of the training done by the network and the epochs denotes the total number of cycles, network would be trained and a batch size of 50.



Figure

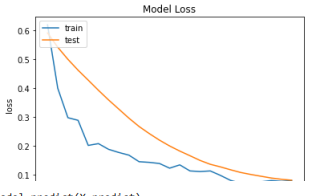


Figure 4.1

The data visualization shows two graph plotted for accuracy vs model accuracy and loss vs model loss and predicted accuracy of 90.4414 was achieved by the network.

# CONCLUSION & Discussion

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | | K-Neighbours | | Support vector machine | | Decision Tree | | Random Forest | |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Precision | .97 | .98 | .96 | .97 | .97 | .95 | .97 | .95 | .94 | 1 |
| Recall | .99 | .95 | .99 | .93 | .97 | .95 | .97 | .95 | 1 | .88 |
| F1-score | .98 | .96 | .97 | .95 | .97 | .95 | .97 | .95 | .97 | .94 |
| support | 72 | 42 | 72 | 42 | 72 | 42 | 72 | 42 | 72 | 42 |

Our main objective for this study was to predict the type of cancer correctly and which of the models had the least false negative reports and highest recall. With the comparison test between the supervised learning methods, algorithms were able to achieve different results which were being recorded with the help of performance metrics in the first model.

The dataset used for the comparison is breast cancer Wisconsin dataset which has 569 observations and 33 features. All the algorithms had good performance but the dataset has low number of observations and features. Larger datasets can increase in complexity and computational time but that too depending on the what is inside the dataset. The deep learning model had the same score with the accuracy level of .9717 with a loss of 0794 and validation loss and validation accuracy 0f .0809 and .9808 respectively, which could be also seen by the performance graph and confusion matrix. The last model was based on the breast histopathology images, had the same architecture like the second model. The image pre-processing is the extra part, the training of the model did took some time as the number of sample in the dataset consists of 277524 images with the learning rate of 0.0001. The final loss was closed down to 0.2300 approx. and accuracy to 90 percent. There are several methods that could’ve been used to improve the performance of the models, even by a smaller fraction for example hyper parameter tuning, boosting methods and using important features at the early stages for both machine learning and deep learning models.

# REFERENCES

1. CDC (2019). *CDC - What Is Breast Cancer?* [online] Centers for Disease Control and Prevention. Available at: <https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm>.
2. WHO (2021). *Breast cancer*. [online] www.who.int. Available at: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>.
3. stanfordhealthcare.org. (n.d.). *Cancer*. [online] Available at: https://stanfordhealthcare.org/medical-conditions/cancer/cancer.html#:~:text=What%20is%20the%20difference%20between.
4. ‌ Tagliaferri, L. (2017). *An Introduction to Machine Learning*. [online] Digitalocean.com. Available at: https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning [Accessed 8 Nov. 2019].
5. Susan G. Komen®. (n.d.). *Understanding Breast Cancer Survival Rates*. [online] Available at: https://www.komen.org/breast-cancer/facts-statistics/breast-cancer-statistics/survival-rates/.
6. ‌Chapter 1 INTRODUCTION TO KNOWLEDGE DISCOVERY IN DATABASES. (n.d.). [online] Available at: https://www.ise.bgu.ac.il/faculty/liorr/hbchap1.pdf.
7. ‌ Analytics Vidhya. (2021). *How to Detect and Remove Outliers | Outlier Detection And Removal*. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/>.
8. Chapter 1 INTRODUCTION TO KNOWLEDGE DISCOVERY IN DATABASES. (n.d.). [online] Available at: https://www.ise.bgu.ac.il/faculty/liorr/hbchap1.pdf.
9. ‌ www.javatpoint.com. (n.d.). *Unsupervised Machine learning - Javatpoint*. [online] Available at: https://www.javatpoint.com/unsupervised-machine-learning.
10. ‌ Foote, K. (2019). *A Brief History of Machine Learning - DATAVERSITY*. [online] DATAVERSITY. Available at: <https://www.dataversity.net/a-brief-history-of-machine-learning/#>.
11. Brownlee, J. (2021). *What Is Semi-Supervised Learning*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/what-is-semi-supervised-learning/.
12. ‌ SearchEnterpriseAI. (n.d.). *What is Reinforcement Learning? A Comprehensive Overview*. [online] Available at: https://www.techtarget.com/searchenterpriseai/definition/reinforcement-learning#:~:text=Reinforcement%20learning%20is%20a%20machine.

‌

1. Sathiyanarayanan, P., Pavithra., S., SAI SARANYA., M. and Makeswari., M. (2019). *Identification of Breast Cancer Using The Decision Tree Algorithm*. [online] IEEE Xplore. doi:10.1109/ICSCAN.2019.8878757.
2. ‌ Team, T.A. and Team, T.A. (n.d.). *Why Choose Random Forest and Not Decision Trees – Towards AI — The Best of Tech, Science, and Engineering*. [online] Available at: https://towardsai.net/p/machine-learning/why-choose-random-forest-and-not-decision-trees.

‌

1. Nguyen, C., Wang, Y. and Nguyen, H.N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, 06(05), pp.551–560. doi:10.4236/jbise.2013.65070.
2. ‌
3. Ayer, T., Chhatwal, J., Alagoz, O., Kahn, C.E., Woods, R.W. and Burnside, E.S. (2010). Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. *RadioGraphics*, 30(1), pp.13–22. doi:10.1148/rg.301095057.
4. ‌Ayer, T., Chhatwal, J., Alagoz, O., Kahn, C.E., Woods, R.W. and Burnside, E.S. (2010). Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. *RadioGraphics*, 30(1), pp.13–22. doi:10.1148/rg.301095057.

‌

1. Chao, C.-M., Yu, Y.-W., Cheng, B.-W. and Kuo, Y.-L. (2014). Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree. *Journal of Medical Systems*, 38(10). doi:10.1007/s10916-014-0106-1.
2. ‌ MurtiRawat, R., Panchal, S., Singh, V.K. and Panchal, Y. (2020). *Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning*. [online] IEEE Xplore. doi:10.1109/ICESC48915.2020.9155783.

‌

1. Harrison, O. (2019). *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. [online] Medium. Available at: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761#:~:text=KNN%20works%20by%20finding%20the.
2. ‌ Laghmati, S., Cherradi, B., Tmiri, A., Daanouni, O. and Hamida, S. (2020). Classification of Patients with Breast Cancer using Neighbourhood Component Analysis and Supervised Machine Learning Techniques. *2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet)*. doi:10.1109/commnet49926.2020.9199633.

‌

1. Akay, M.F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, [online] 36(2), pp.3240–3247. doi:10.1016/j.eswa.2008.01.009.
2. ‌ Bento, C. (2021). *Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis*. [online] Medium. Available at: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141.

‌

1. viso.ai. (2021). *Deep Neural Network: The 3 Popular Types (MLP, CNN and RNN)*. [online] Available at: https://viso.ai/deep-learning/deep-neural-network-three-popular-types/.
2. ‌ viso.ai. (2021). *Deep Neural Network: The 3 Popular Types (MLP, CNN and RNN)*. [online] Available at: https://viso.ai/deep-learning/deep-neural-network-three-popular-types/.
3. ‌ viso.ai. (2021). *Deep Neural Network: The 3 Popular Types (MLP, CNN and RNN)*. [online] Available at: https://viso.ai/deep-learning/deep-neural-network-three-popular-types/.

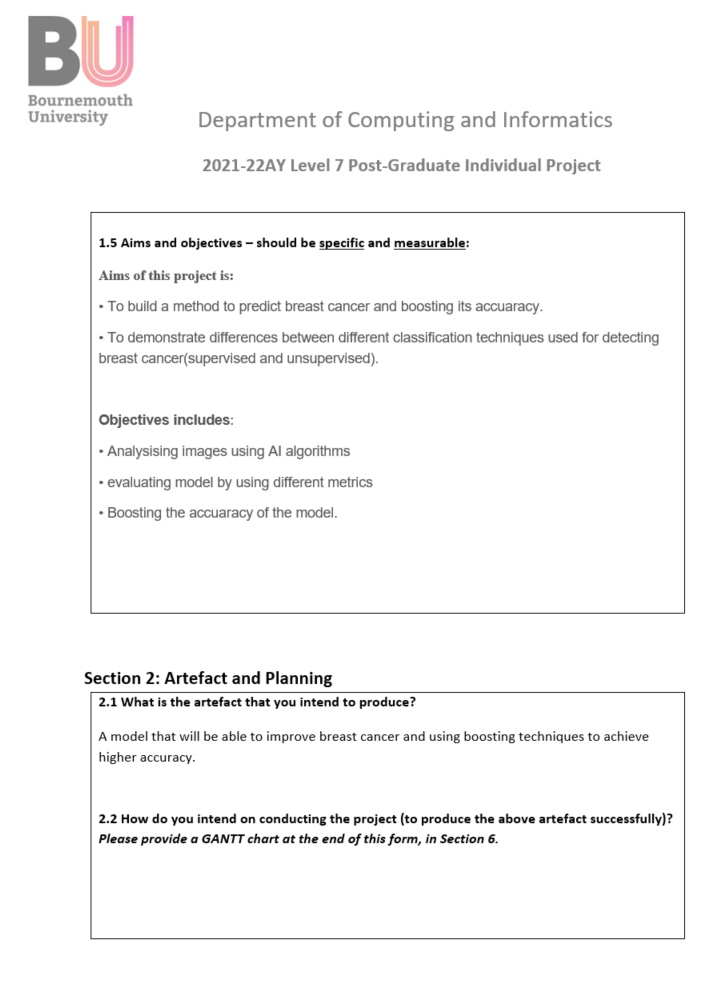
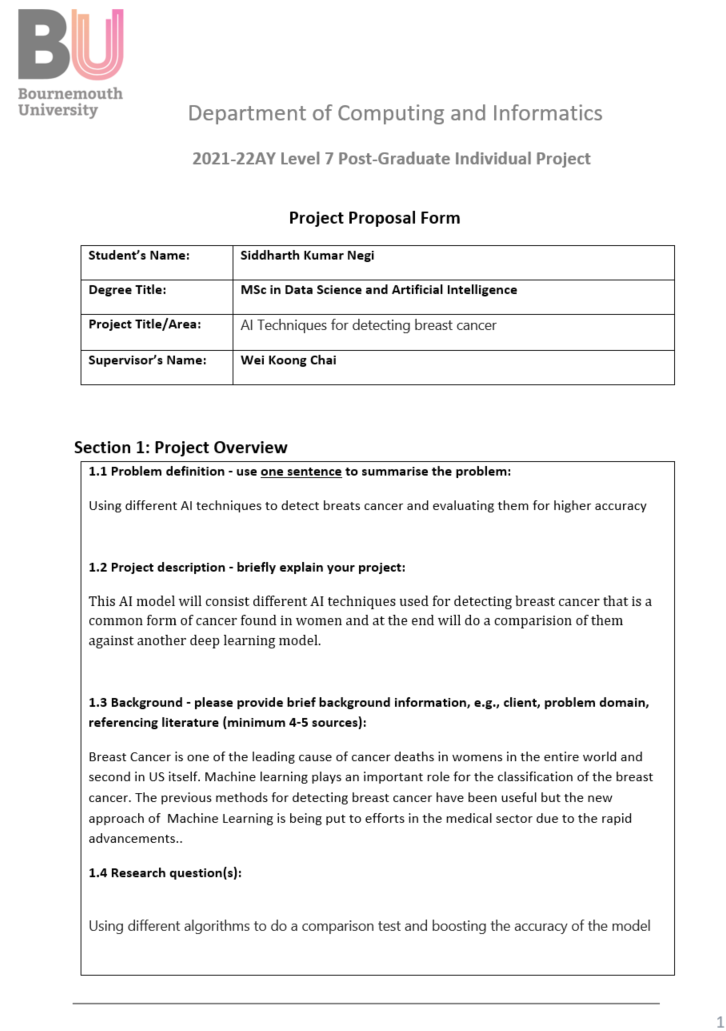
‌

1. www.xenonstack.com. (n.d.). *TensorFlow: Introduction, Architecture, Benefits | Quick Guide*. [online] Available at: https://www.xenonstack.com/blog/tensorflow.

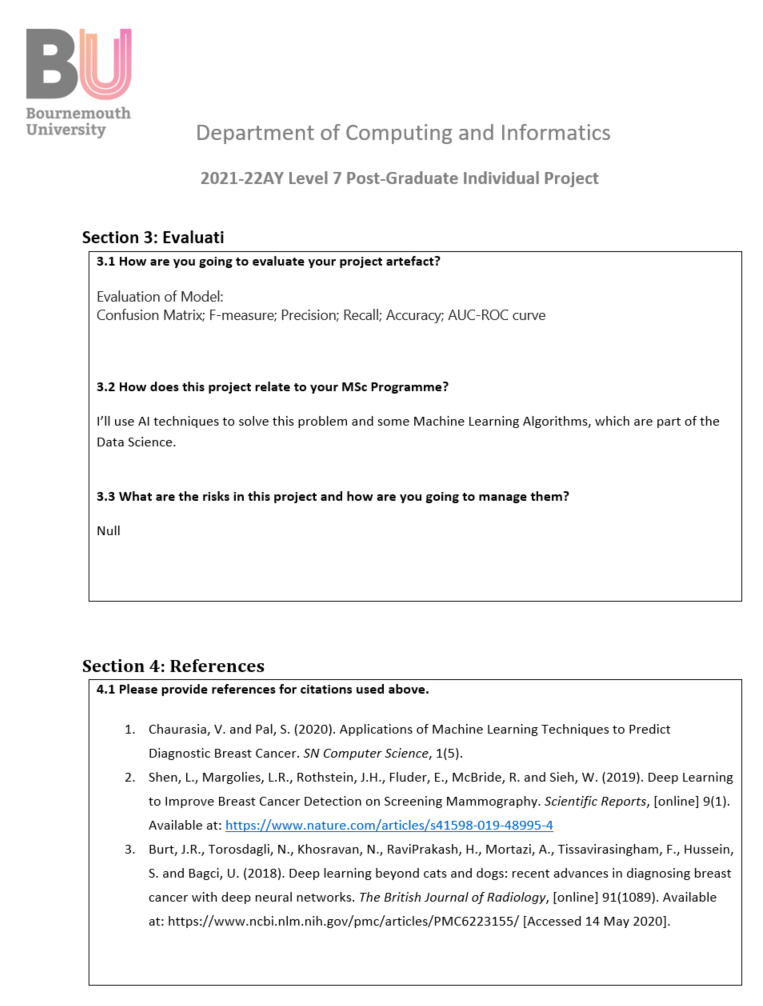
‌

‌

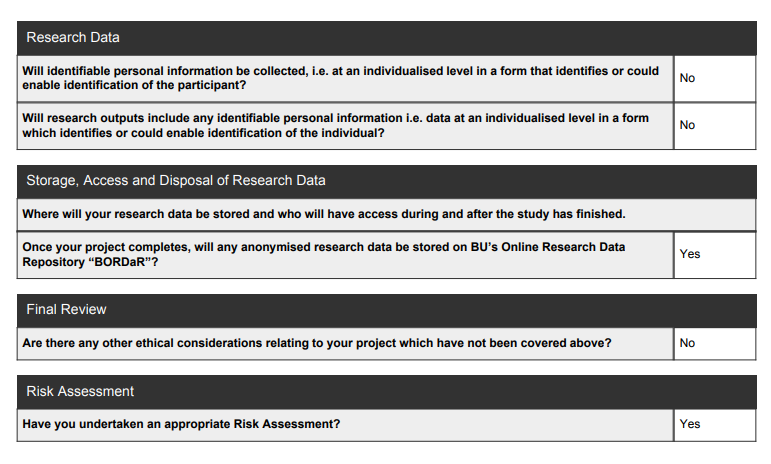
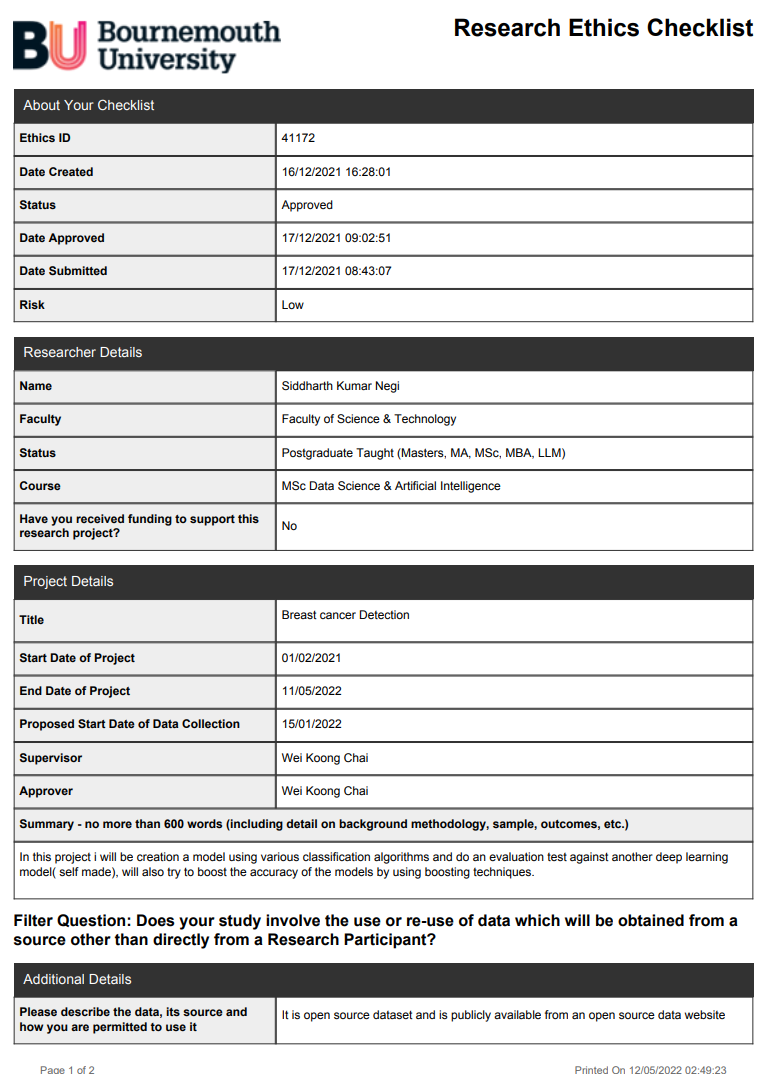
# PROJECT PROPOSAL



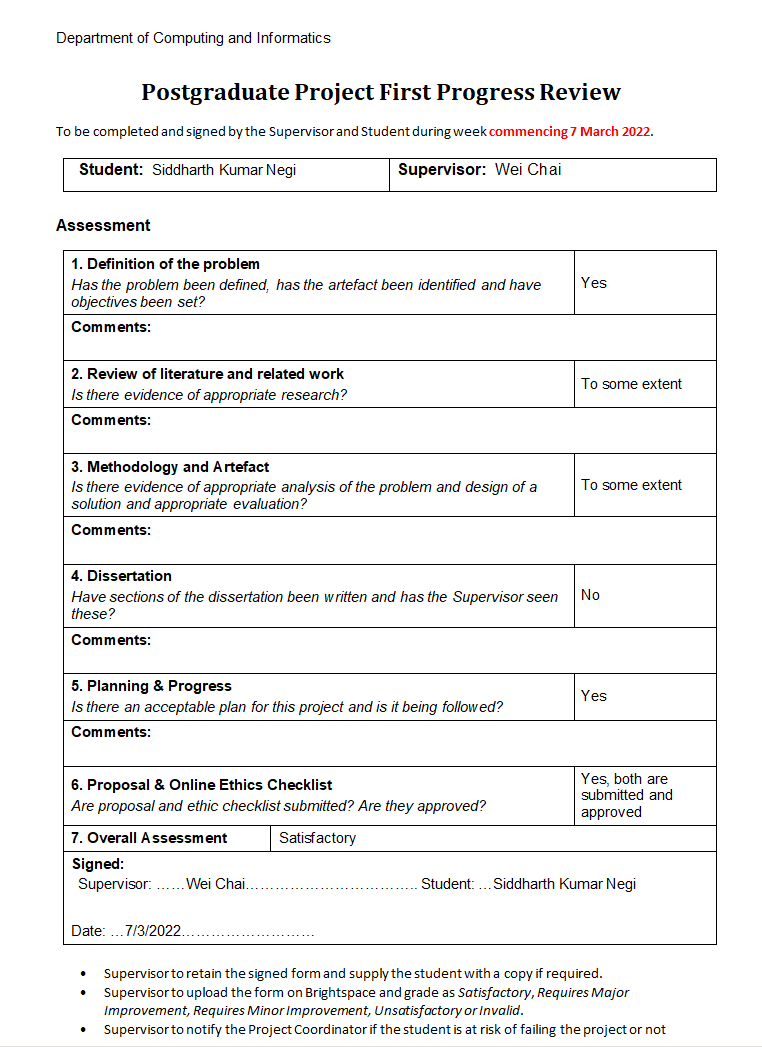
# 



# APPENDIX A PROJECT ETCIH CHECK LIST



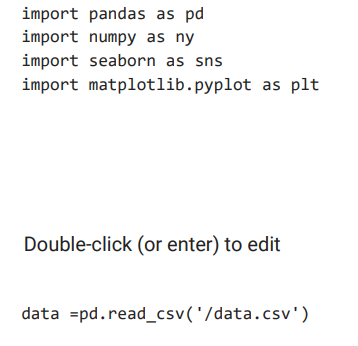
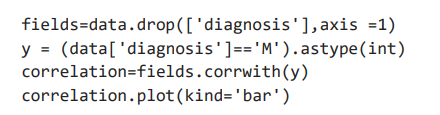
# APPENDIX FIRST REVIEW



# APPENDIX E -Artefact

The artefact consists of explanation of all three models that have been used for the comparision test. A line by line brief summary is given below

## Machine Learning model



Importing the important requirements and libraries which include pandas, numpy, seaborn for graph and matplotlib for plotting data points on the graph.



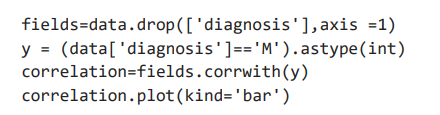
Data.head() shows the first five observations of the dataset and data.shape() shows datatype of the columns/features.

Data.desc() shows the summary of data frame columns.

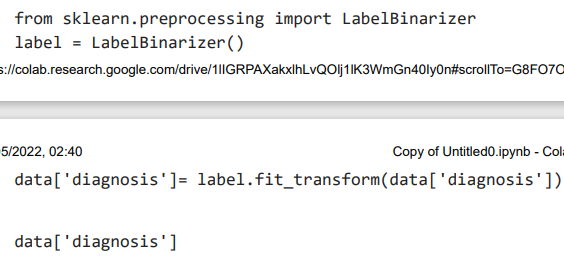
Daat.isnull.sum() is used for checking all the null values in the dataset and replace them with a mean of the sum.



Displays the number of sample count for benign and malignant.



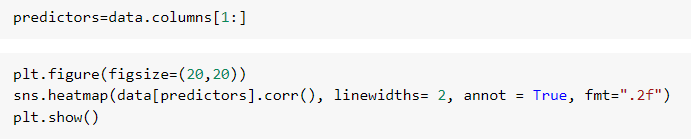
The next part is the exploratory data analysis and the next step is to check for correlation by visualization. Where we found that there were some features which were least correlated so we drop them from the dataset.



For converting the multi class labels into a single binary label



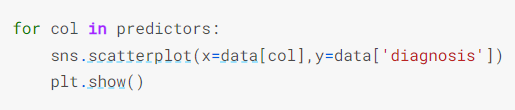
We have missing values for one column so we are going to drop it (unnamed and id)



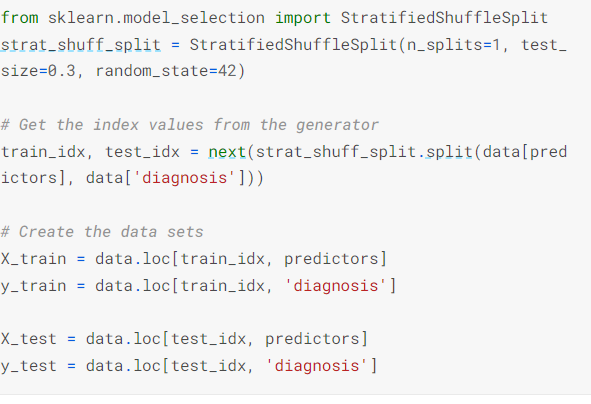
Now, we are checking the correlation among the features of the dataset



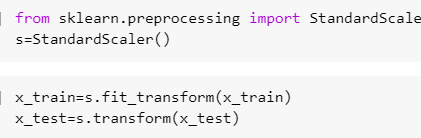
By plotting the bar graph for the features, we can check if features are left skewed or right.



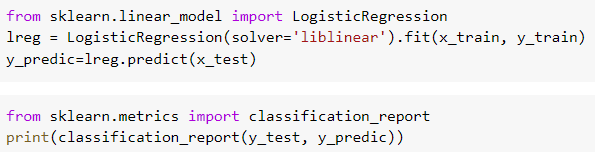
The scatterplot graph shows us how plotting data points of features could help us determine the nature of cancer.



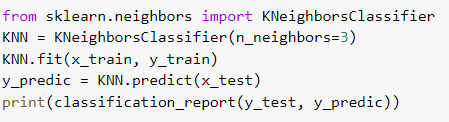
In this step we are preparing data for training and testing phase, for splitting the dataset we are using straitifiedshufflesplit and x = features while y = prediction value



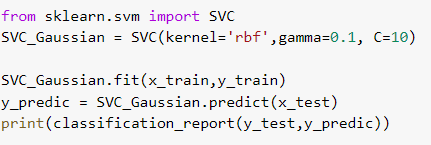
In the next two steps, we are going to perform scaling of data as in the dataset there are some points that are far from each other so to put them closer scaling is used. For training data, we use fit\_transform () and for test data we are using tansform () for calculating the mean as well as variance of the features in the dataset.



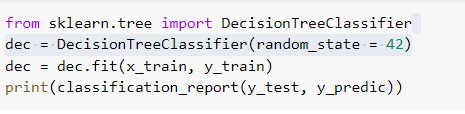
After performing the exploratory data analysis and splitting/ training and testing of data comes the implementation of algorithms. In logistic regression, we are using the liblinear library which is used for large linear classification and utilizes the coordinate descent algorithm. With the performance metrics the output is going to be displayed in terms of precision, recall and f1 score.



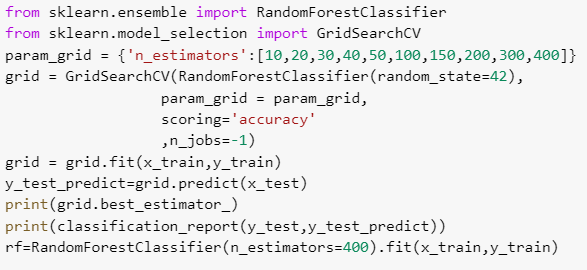
The next algorithm is K-NN which uses the Euclidean distance to calculate nearest neighbour and the ones which are similar are put in a same class. n\_neighbors defines the number of neighbours.



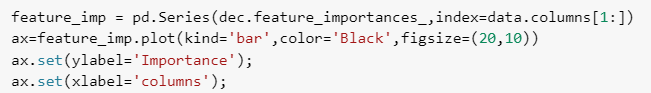
Support vector machine uses the Gaussian to find the hyperplane that distributes the best features into different fields. The Rbf kernel checks out for non-linear data in linear model. Gamma parameters shows how influential is the training example and how far can it reach. C defines the minimum correct classification limit that SVM has to reach.



Decision tree is being implemented and a random\_state is being assigned which defines the random selection of features and samples for the sub-optimal greedy algorithm.



Random forest is an extension of the decision tree which makes decision on the basis of multiple decision trees and uses averaging to improves the predictive accuracy and controls over fitting in the model. GridsearchCV is being used to provide the best suitable hyper parameters suitable to get higher accuracy and also chooses the best number of estimators (number of trees).



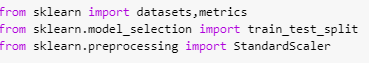
This block of code tells us the which features in the dataset are important and ranks them accordingly in for the bar graph.

## Deep Learning model

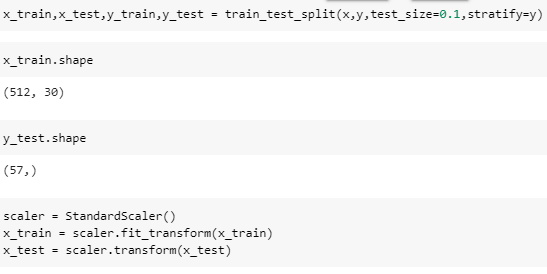
The second and third model is based on the same architecture, the difference is in the working of layers and the image pre-processing.



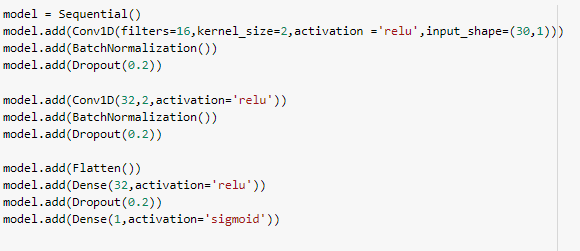
The first step is installing tensorflow and other deep learning layers which include convo1D, Maxpool1D, flatten, Dense, Dropout and BatchNormalization along with python libraries numpy, pandas, seaborn and matplotlib.



SkLearn consists of some in build datasets somewhere around 150, 50 of each class and the metrics consists of utility functions to evaluate classification model’s performance in terms of loss, score, accuracy etc. Train\_test\_split for splitting the dataset into training and testing.



We have sliced the dataset into x and y using train\_test\_split with a test\_size of 0.1 is passed which could be increased to 0.2 or 0.3 but it would add another step of passing down the exact size of array. For example, if the size of array is 15360 the new have to pass the shape of (512,30,1). For scaling the data, standard scaler is used.



The next step is preparing the model and adding layers to it. For the preparation, we are using the sequential model () which is each layer is laid upon each other and has one input and output tensor. It is a single layer perceptron model so it doesn’t have too many hidden layers.

The conv1D layer consists of:-

* Filters: defines the number of output filters in convolution.
* Kernel size: defines the length of the convolution windows/strides.
* Activation function: user has to pass which activation function is to be used ( by default no activation function is passed)
* Input shape: is the number of steps and the input dimensions which in our case is 1D.
* Dropout: means slicing down the weights at a probability if there is too much correlation among them
* Dense: Dense layers are layers of neurons, where the output of previous neuron is used as an input for the layer.

ReLu and sigmoid being the type of activation function. Relu being usedto saturates output which are lower than 0 while sigmoid is used as an classifier to classify if the output is either 1 or 0.

* BacthNormalization: used for training deep neural networks faster and increases the stability of the network.



Shows the summary of the network that includes the layers and the output shape.



Model compile defines the type of optimizer, learning rate at which the network learns while training, loss function and metrics. The optimizer used in the network is Adam optimizer, which is one of the best adaptive optimizers in classification. Binary cross entropy is used for loss function which compares the predicted values to the actual output.

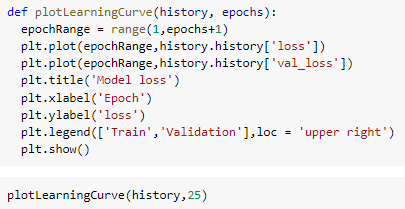


History is used as it consists of all the records during the training of the network in each epoch.

Epoch refers to a complete training cycle of the network, verbose is used to display all the output if set to 1. Validation\_split what fraction of the data isbeing used for validation (10% in our case).



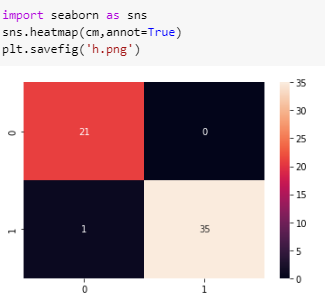
Histoy.history?? is the dictionary that holds all the training records (loss and metrics value)



In the next two steps, we are plotting two separate graphs as a part of visualization.



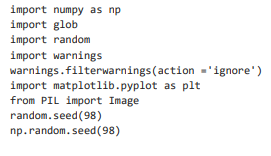
Shows the accuracy of the model which was 98.24%



The last part of our model and data visualization, initiating a confusion metrics that shows all types of predictions made during the training.

## Deep learning using CNN

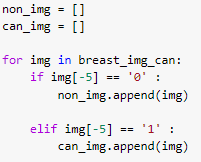
This is the last model for our comparison test, has the same architecture as previously told, Breast Histopathology Images was used which almost consisted of 555068 labelled images. The size of the dataset is quite big, tackle this problem the entire dataset was uploaded to the google drive as a zipped file and later unzipped manually using google colab.



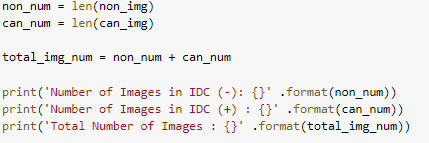
Importing all the necessary libraries which includes numpy, global, random, matplotlib and random seed which is used to get the same training and validation data.



This set of lines uploads the dataset from the google drive and prints the first three images from the dataset.



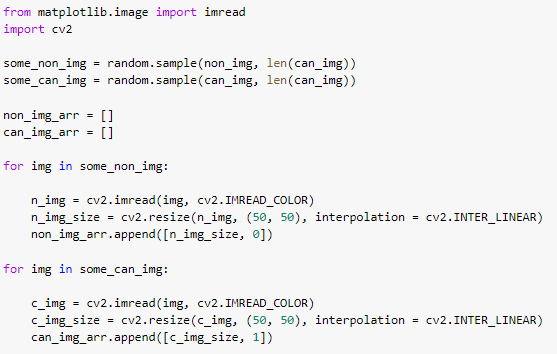
Two different dictionaries for cancerous and non-cancerous images is initialized and if the image has the label 0 then image would be append to non\_img otherwise can\_img



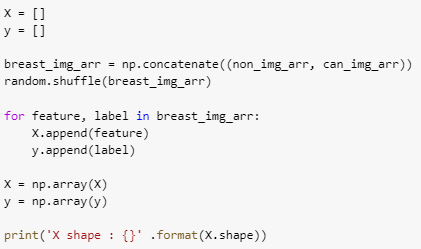
Now we are trying to find the length of the non\_num and can\_num arrays and a thirld array total\_img\_num shows the total number of images. For the visualization purpose, non\_num are labelled as IDC (-) and can\_num are labelled as IDC (+)



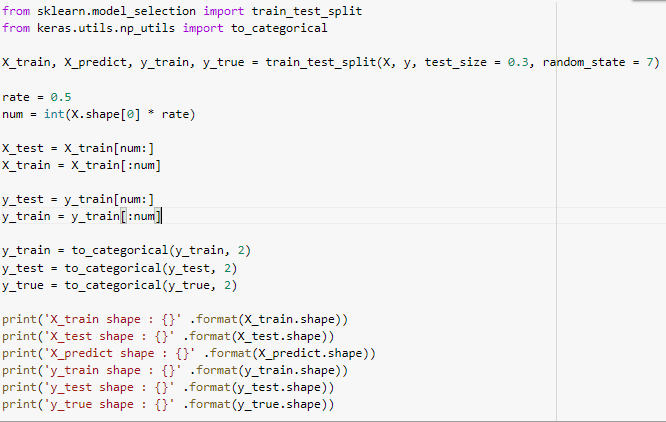
The image library from keras is being imported and the dimensions of the images for plotting is passed as (15,15). For cancerous and non-cancerous. Randint function returns the images in form of array or integers and a random number of images (18) is passed. This line of code returns the images with the IDC labels on them.



The images are being resized to the dimension of (50,50) and by using Inter\_linear as the interpolation technique we can calculate the pixel values for the new images from original ones.



Another array x and y are being initialized and are randomly shuffled, where x contains the features and y consists of the labels (0 and1)

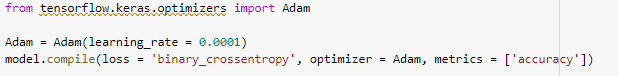


Now the image dataset is being split into training and testing using the train\_test\_split with a test size of .3 and train size of 0.7 with a random state of 7. Then the images are divided into x\_test, x\_train, y\_test and y\_train by multiplying the number of images with rate and storing them into training and testing variables.

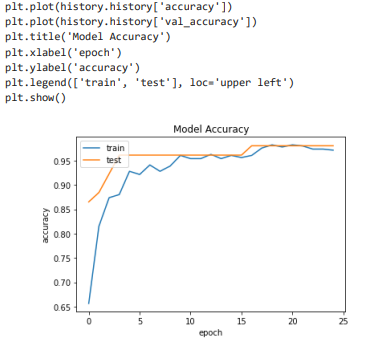


The next step is intializing the sequential model and importing the layers for the network.

The layers includes Conv2D, Maxpooling2D, Dropout, Flatten and Dense.For intializing the convo2D layer we pass the total number of filter(32), the size of the kernel (3,3) which specfies the width and height of the 2D convolutional window. The padding can handle two values either valid being the default value or same. In our case we want to preserve the dimensions of the volums so the size of output volum and input volume are same. The activation function ReLu, which returns 0 if the it receives neagtive value and 1 for positive values. All together total number of 4 convolutional layers are inserted into the network. Th flatten () layer removes all the other dimensions leaving it into a cector format (single dimension). The sigmoid function returns the output in form of 0 and 1.



Now from the optimizer library we are using adam optimizer as it is the best optimizer and the adaptive learning rate could be used for almost all types of classification datasets and there isno need to emphasis on learning rate value. The loss function again is binary cross entropy where we calculate the entropy loss between predicted and actual classes (0 and1). The total predicted accuracy is90.441% which was achieved during the14th cycle (epoch)



The last step is the data visualization for the end result of training the model for both accuarcy and the loss.



At last, we check for the actual true prediction and false predictions made by the model.