Breast Cancer Prediction

ISM 6136 - Data Mining Project

Varun Krishna Ramakrishnan

Harishma Parthiban

Background:

Breast cancer is a disease in which the cells of the breast get uncontrollably large. Normally, cells divide (create) only when new cells are required. Cells can sometimes grow and divide out of control, resulting in a mass of tissue known as a tumour. After skin cancer, it is the second most frequent cancer in women. According to global statistics, it is still the second greatest cause of cancer death among women in general and the first among Hispanic women. The type of breast cancer is determined by which cells in the breast become cancerous. The following are the most prevalent kinds of breast cancer:

- **Infiltrating (invasive) ductal carcinoma** - This is the most common form of breast cancer, accounting for 80% of cases.
- **Ductal carcinoma in situ** - It is almost always curable.
- **Infiltrating (invasive) lobular carcinoma** - It accounts for 10 to 15% of breast cancers.
- **Lobular carcinoma in situ** - It isn't a true cancer but serves as a marker for the increased risk of developing breast cancer later.

The anatomic staging system of breast cancer consists of five stages. They are

- **Stage 0** breast disease is when the disease is localized to the milk ducts.
- **Stage I** breast cancer is smaller than 2 cm across and hasn't spread anywhere.
- **Stage II** breast cancer is when the tumour has started to spread to the lymph nodes.
- **Stage III** breast cancer is when the tumour is any size with cancerous lymph nodes that adhere to one another or to surrounding tissue or that has spread to the skin, chest wall, or internal mammary lymph nodes.
- **Stage IV** breast cancer is defined as a tumour, regardless of size, that has spread to areas away from the breast, such as bones, lungs, liver or brain.

Breast cancer can start in a variety of places in the breast. Lobules, ducts, and connective tissue are the three primary components of a breast. The glands that generate milk are known as lobules. The ducts are tubes that transport milk from the breast to the nipple. Everything is held together by connective tissue, which is made up of fibrous and fatty tissue. Breast cancer usually starts in the ducts or lobules.

Breast cancer can spread to other parts of the body via blood and lymph vessels. Breast cancer is said to have metastasized when it spreads to other regions of the body.

Studies have shown that your risk for breast cancer is due to a combination of factors. Being a woman and being older are two major factors that increase your risk. The majority of breast cancers are diagnosed in women over the age of 50. There are risk factors that are under your control and risk factors that are beyond your control.

Risk Factors that are beyond control

- Getting older
- Genetic mutations
- Reproductive history
- Having dense breasts
- Personal history of breast cancer or certain non-cancerous breast diseases
- Family history of breast or ovarian cancer

- Previous treatment using radiation therapy
- Women who took the drug diethylstilbestrol (DES)

<u>Risk Factors that are under our control</u>

- Not being physically active
- Being overweight or obese after menopause
- Taking hormones
- Reproductive history
- Drinking alcohol

Motivation for solving the problem:

An early diagnosis of breast cancer is very much important as it can improve the chance of survival significantly. The most important screening test for breast cancer is mammogram. It is advisable that women aged above 45 should get mammograms every year. This helps in the early diagnosis of breast cancer.

In general, cells in the body normally divide (reproduce) only when new cells are needed. If the cells that are growing out of control are normal cells, the tumour is called benign (not cancerous). They tend to grow slowly and do not spread. If, however, the cells that are growing out of control are abnormal and don't function like the body's normal cells, the tumour is called malignant (cancerous). Malignant tumours grow rapidly, damage the tissues, and spread throughout the body.

So here we try to build a model which helps in identifying whether the breast cancer is benign (non-cancerous) or malignant (cancerous). This is very important because the classification helps in identifying the women at high risk.

For the study, we have taken our dataset from the UCI Machine Learning repository. The dataset contains 11 attributes, and they are further used to predict whether the cancer is benign or malignant.

Dataset description:

The dataset we have taken from the UCI Machine Learning repository consists of 699 observations and it has 11 attributes. The dataset was created by Dr.William.H.Wolberg from General Surgery Dept., University of Wisconsin Clinical Service Center at Madison, Wisconsin.

The attributes are id,clumpThickness,uniformityCellSize, uniformityCellShape,marginalAdhesion,singleEpithelialCellsize,bareNuclei,blandChromatin, normalNucleoli, mitoses and class.

These attributes indicate the characteristics of a cell nuclei obtained from a digitized image of Fine Needle Aspiration (FNA) of a breast mass. The attributes are shown in Figure 1.

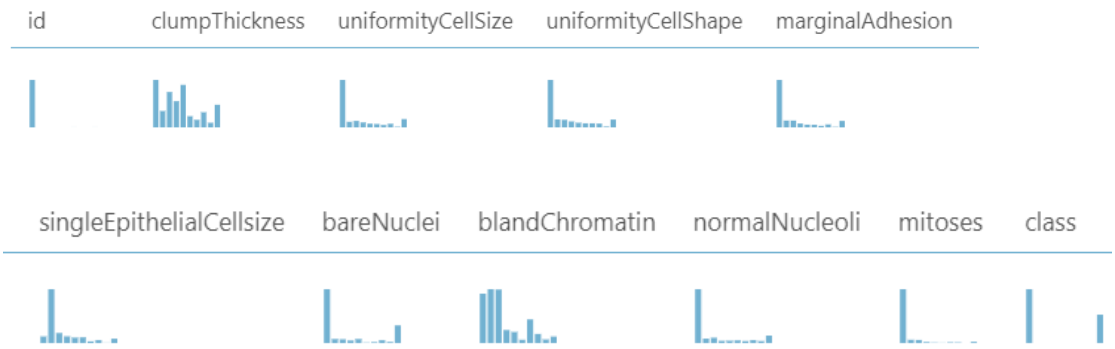


Figure 1. Attributes of the dataset

Description of the attributes:

1. **id** – id is a unique sample code number for each of the patient.
2. **clumpThickness** - Benign cells tend to be grouped in monolayers, while the mailgnant cells are often grouped in multilayers.
3. **uniformityCellSize** - Cancer cells tend to vary in size. That is why these parameters are valuable in determining whether the cells are cancerous or not.
4. **uniformityCellShape** - Cancer cells tend to vary in size as well.
5. **marginalAdhesion** – Normal non-cancerous cells tend to stick together. Cancer cells tend to lose this ability. So, the loss of adhesion is a sign of malignancy.
6. **singleEpithelialCellsize** - Epithelial cells that are significantly enlarged may be a malignant cell.
7. **bareNuclei** – It is a term used for nuclei not surrounded by cytoplasm (the rest of the cell). They are typically seen in benign tumours.
8. **blandChromatin** – It describes a uniform texture of the nucleus seen in benign cells. In Cancer cells, the chromatin tends to be coarser.
9. **normalNucleoli** - Nucleoli are small structures seen in the nucleus. In normal cells, the nucleolus is usually very small if visible at all. In cancer cells, the nucleoli become more prominent, and sometimes there are more of them.
10. **mitoses** – This describes how quickly the cancer cells are multiplying. The higher the rate, more quickly it's multiplying.
11. **class** – It indicates whether the tumour is benign or malignant.

Solution methodology and Evaluation metrics:

We are trying to build a model which helps in identifying whether the breast cancer is benign (non-cancerous) or malignant (cancerous). Class is our target variable. It indicates whether the tumour is benign or malignant.

We have used the following binary classification algorithms – Two class Boosted Decision Tree, Two class Logistic Regression and Two class Neural Network.

Model:

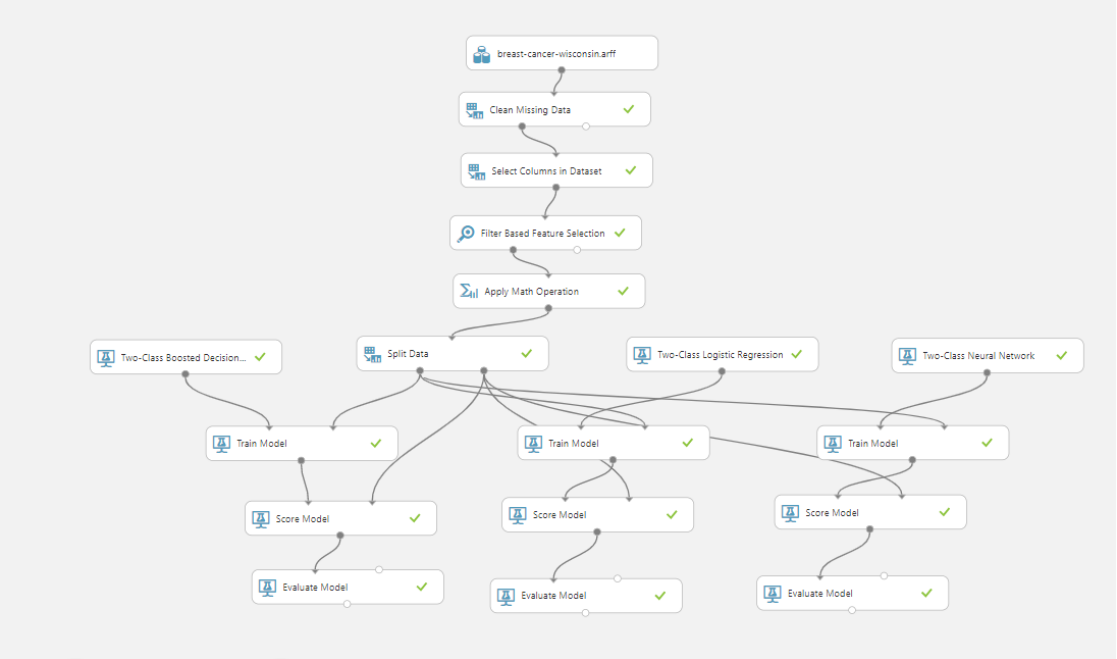


Figure 2. Machine Learning Model

Cleaning missing data:

In the dataset, we had 16 missing values for the attribute bareNuclei. So, we have used Clean Missing Data module to replace the missing values for the attribute. We have chosen Replace using MICE as the Cleaning Mode. In this cleaning mode, each missing value of the attribute is assigned a new value which is calculated by Multivariate imputations using Chained Equations. Each of the missing value is modelled with the other values of the attribute in the dataset.

***Filter Based Feature Selection*:**

We have used Filter Based Feature Selection using Pearson correlation to identify the most important attributes in our dataset. BareNuclei was identified as the most important attribute. When we replaced the missing values in bareNuclei using Replace using MICE as the cleaning mode, the strength of correlation between bareNuclei and the target variable was identified to be 0.82. But when we replaced the missing values using Replace with Mode as the cleaning mode, the strength of correlation between bareNuclei and the target variable was identified to be 0.81.

Using the Pearson Correlation Filter based feature selection, we selected the top 6 features out of the 9 available features. The top 6 features that were selected are *bareNuclei, uniformityCellShape, uniformityCellSize, blandChromatin, clumpThickness and normalNucleoli.*

Figure 3 shows the Filter based feature selection and the filtered dataset with the top 6 features.

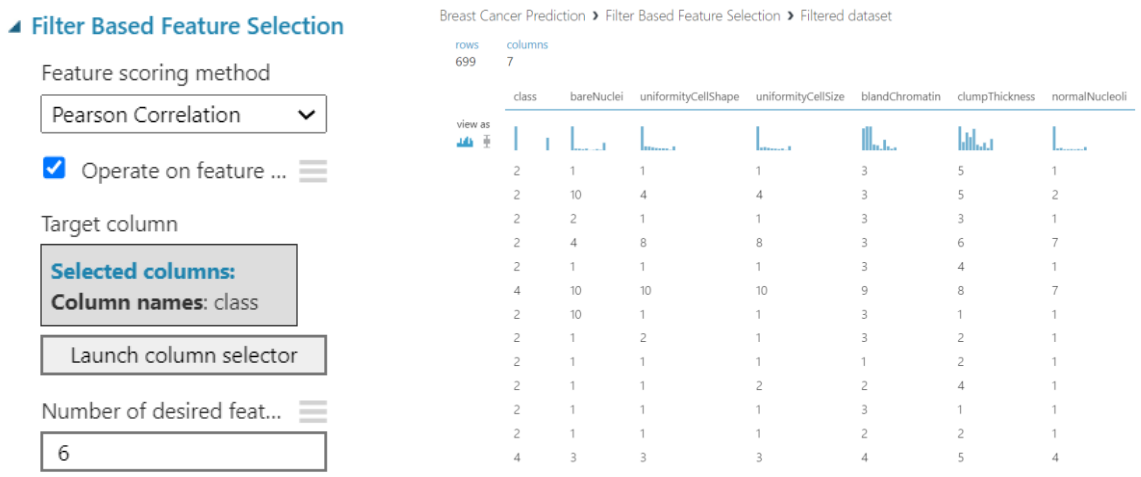


Figure 3. Feature Selection and Filtered Dataset

Figure 4 indicates the strength of the correlation between the attributes and the target variable using Pearson correlation

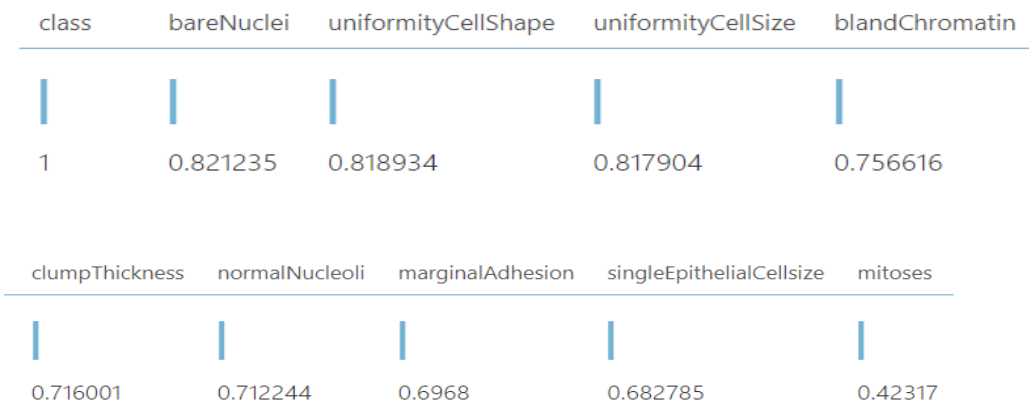


Figure 4. Correlation with Target Variable

Applying Math Operation:

In our dataset, class is our target variable. It has values 2 and 4 indicating benign and malignant respectively. So, we have used Apply Math Operation module to replace the values for class attribute as True for malignant and False for benign.

Algorithm comparison:

We are making a prediction whether the tumour is benign or malignant. So, we can identify from the dataset that class is our target variable which indicates whether the tumour present in the patient is benign or malignant. So, we are considering the following binary classification algorithms to make the predictions.

- Two class Boosted Decision Tree
- Two class Logistic Regression

- Two class Neural Network

Starting with the Two class boosted Decision Tree, which is the binary classifier, we tried to tune the default parameters Maximum number of leaves per tree to 10, Minimum number of training instances required to form a leaf to 5, Learning rate to 0.1, Total number of trees constructed to 20. This gave an accuracy of 96.4%, False positive as 2, False negative as 3 and an AUC of 99.1%.

Next, when running the Two class Logistic Regression with the default parameters L1 regularization weight as 1, L2 regularization weight as 1, we got the accuracy as 97.1%, False positive as 2, False negative as 2 and an AUC of 99.6%. Further, we tried tuning the default parameter L2 regularization weight to 50. This gave an accuracy of 95%, False positive as 1, False negative as 6 and AUC of 99.4%.

Finally, we experimented by running the Two class Neural Network with the default parameters. This gave us an accuracy of 97.9%, False positive as 2, False negative as 1 and AUC of 99.4%. We also tried tuning the default parameters Number of hidden nodes to 400, Learning rate to 0.05, Number of learning iterations to 200. This resulted in an accuracy of 97.1%, False positive of 2, False negative of 2 and AUC of 99.5%.

Confusion matrix:

In our model, class attribute as True indicates malignant tumour and False indicates benign tumour.

Figure 5 shows the confusion matrix of Two class Neural Network model:

True Positive	False Negative	Accuracy	Precision
45	1	0.979	0.957
False Positive	**True Negative**	**Recall**	**F1 Score**
2	92	0.978	0.968

Figure 5. Confusion matrix of Two Class Neural Network Model

Description:

- True Positive indicates the cases where the tumour is malignant, and our model has predicted correctly.
- True Negative indicates the cases where the tumour is benign, and our model has predicted correctly.
- False Negative indicates the cases where the tumour is malignant, but our model has predicted incorrectly.
- False Positive indicates the cases where the tumour is benign, but our model has predicted incorrectly.

Results:

The following are the results of modelling:

Two class boosted Decision Tree:

True Positive	False Negative	Accuracy	Precision
43	3	0.964	0.956
False Positive	**True Negative**	**Recall**	**F1 Score**
2	92	0.935	0.945

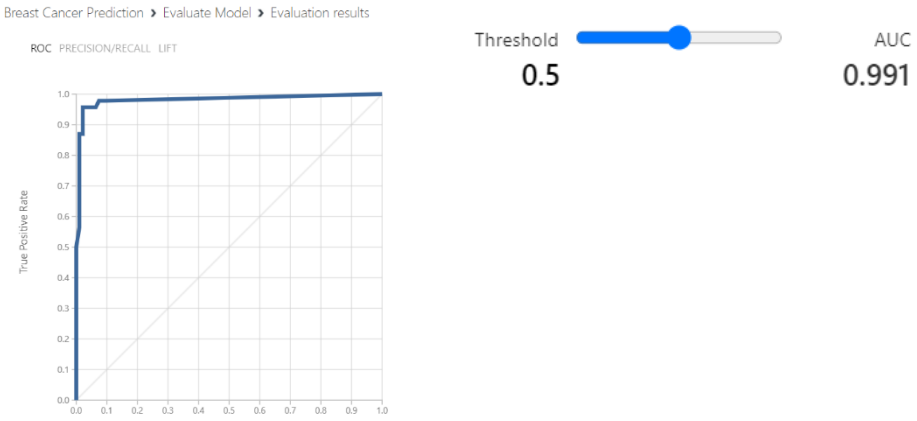
Two class Logistic Regression:

True Positive	False Negative	Accuracy	Precision
40	6	0.950	0.976
False Positive	**True Negative**	**Recall**	**F1 Score**
1	93	0.870	0.920

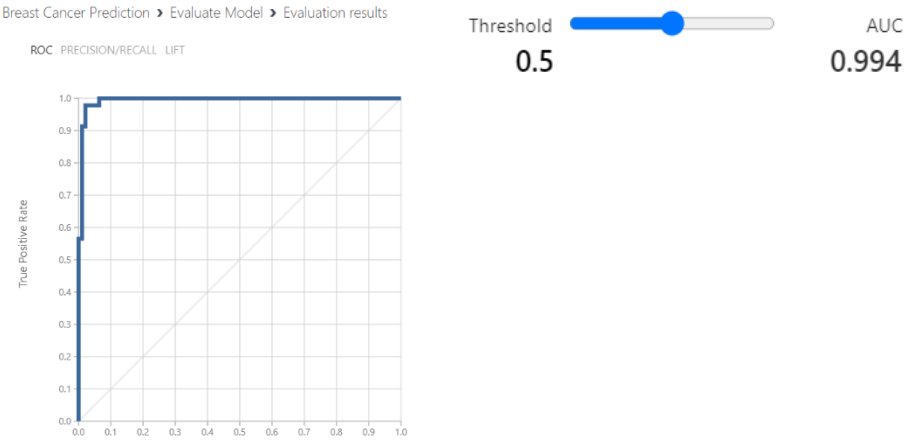
Two class Neural Network:

True Positive	False Negative	Accuracy	Precision
45	1	0.979	0.957
False Positive	**True Negative**	**Recall**	**F1 Score**
2	92	0.978	0.968

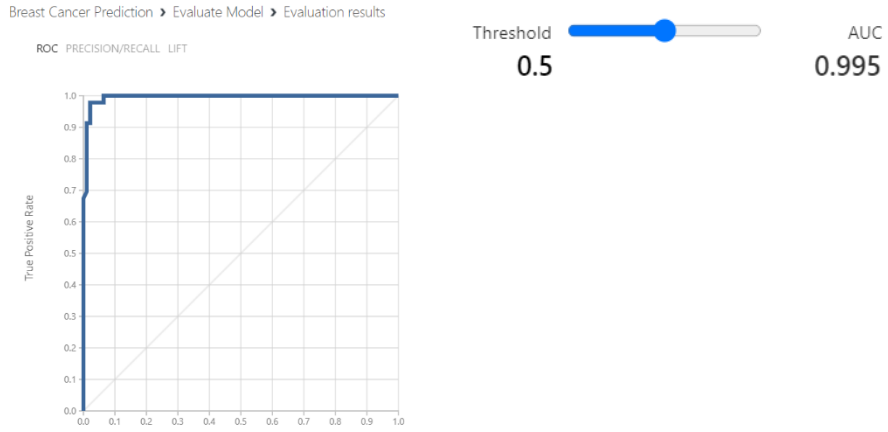
Two class boosted Decision Tree:



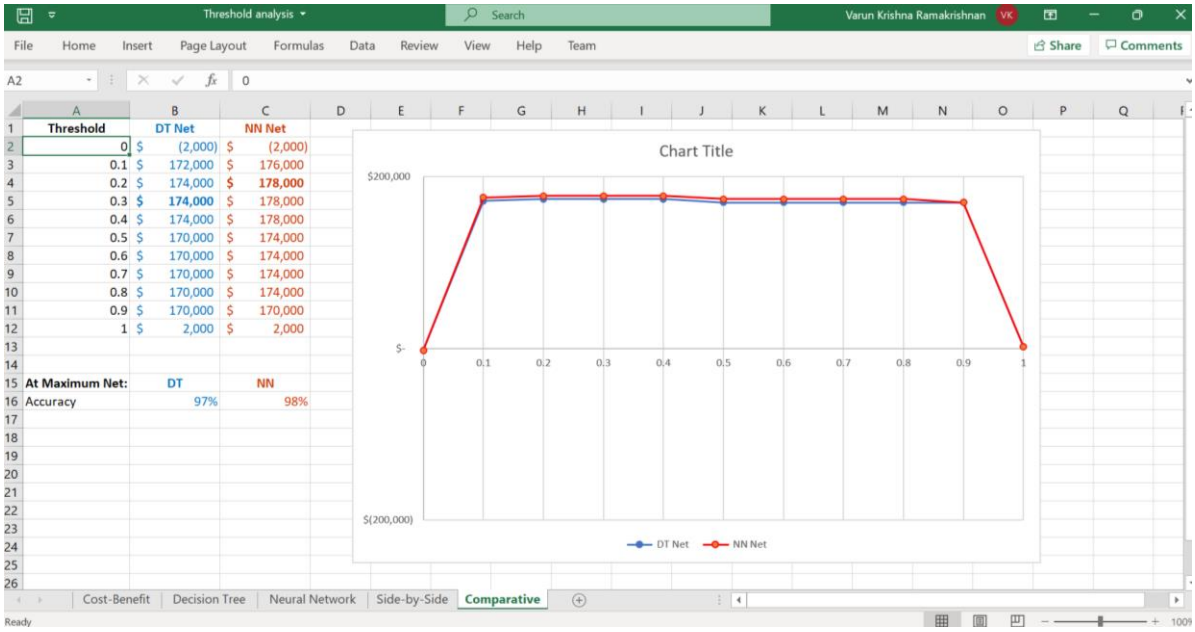
Two class Logistic Regression:



Two class Neural Network:



Threshold Analysis:



Threshold	DT Net	NN Net
0	$ (2,000)	$ (2,000)
0.1	$ 172,000	$ 176,000
0.2	$ 174,000	$ 178,000
0.3	$ 174,000	$ 178,000
0.4	$ 174,000	$ 178,000
0.5	$ 170,000	$ 174,000
0.6	$ 170,000	$ 174,000
0.7	$ 170,000	$ 174,000
0.8	$ 170,000	$ 174,000
0.9	$ 170,000	$ 170,000
1	$ 2,000	$ 2,000
At Maximum Net:	DT	NN
Accuracy	97%	98%

Cost Benefit Analysis:

In the current scenario, the deaths of women are increasing due to breast cancer. To increase the survival rate, early detection of breast cancer is very much important. The early prediction not only helps in identifying the important factors leading to breast cancer but also, it plays a significant role in preventing the unnecessary treatments given to patients, thereby saving human life, and cutting down the cost. This is indeed important because we know how expensive the cancer treatments are around the world.

Conclusion:

Summing up, we can see that there is not much difference between the three models, but Two Class Neural Network has performed better with an accuracy of 97.9%. Taking into account the problem we are dealing with, for identifying the best algorithm, we have considered accuracy, misclassification rate and area under the curve as the important features. Even though neural network may be time consuming, it has the lesser number of False negative prediction which is very much important for this problem statement. Therefore, based on the observations and interpretation, we can conclude that neural network model can be used to predict the tumour severity in the future.