COMP 4948 Assignment 2 Report

# Problem Definition

The goal is to create a stacked model comprised of 5 Artificial Neural Networks (ANNs) that can predict whether a tumor is benign or malignant, given different physical measurements of tumor cells.

# Exploratory Data Analysis

The chosen dataset is “Cancer Data”, uploaded onto Kaggle by Eredem Taha (<https://www.kaggle.com/datasets/erdemtaha/cancer-data>). The dataset currently looks at 570 cancer cells, each classified as either benign (B) or malignant (M). The columns are physical measurements taken of each tumor:

Graphical user interface

Description automatically generated with low confidence

I’ve split the dataset into train.csv and test.csv, with train.csv being comprised of 513 entries. No data is missing from any column. As the 'diagnosis' column has String values, they must be converted into a binary 'diagnosis\_M' column, where a 1 indicates that a tumor is malignant.

Using both Recursive Feature Elimination and Forward Feature Selection, 14 features were determined to be significant:

['compactness\_worst',

'concavity\_worst',

'radius\_worst',

'radius\_se',

'radius\_mean',

'symmetry\_worst',

'concavity\_mean',

'concavity\_se',

'compactness\_se',

'concave points\_mean',

'concave points\_worst',

'perimeter\_worst',

'perimeter\_se',

'compactness\_mean']

From these significant features, we can see that there are 6 major categories of tumor measurement used to predict whether a tumor is benign or malignant:

1. Compactness
2. Concavity
3. Concave points
4. Radius
5. Symmetry
6. Perimeter

As changes in the tumor were recorded over time, most categories have been separated into ‘worst’, ‘mean’, and ‘se’. ‘worst’ refers to the worst (i.e., largest) measurement, ‘mean’ refers to the average of all measurements, and ‘se’ refers to the standard error. Let’s examine each category, and how they compare to the target variable.

Compactness refers to the ratio between the volume and surface area of the tumor; in other words, the ‘smoothness’ of a tumor. From the following boxplot and scatterplots, we can see that both a low worst and average compactness is highly correlated to the tumor being benign. If the tumor’s worst compactness is high (over 0.6), the tumor is always malignant.

'compactness\_se' the standard error observed among multiple measurements of a tumor’s compactness. The average SE is lower for benign tumors.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

'concavity\_worst' refers to the worst concave portion of the contour of the tumor’s cell nuclei. Again, a low concavity correlates to tumor’s being benign, while tumors with concavity over 0.8 are almost always malignant (except one outlier). The average concavity is higher for malignant tumors.

'concavity\_mean' refers to the average concavity of a tumor. Benign tumors have a much lower average concavity, although there are a few outliers.

'concavity\_se' refers to the standard error of concavity. It’s low for both malignant and benign tumors, although benign tumors have extremely high-value outliers.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart

Description automatically generated

Concave points refer to the concave portions of the tumor cell nuclei. Both average and worst values for concave points are higher for malignant tumors.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Radius refers to the highest radius recorded for a tumor. We can see that all benign tumors have a 'radius\_worst' value under 20, while malignant tumors can reach up to 35. The spread of malignant tumor radiuses is also much larger than that of benign tumors.

'radius\_mean' refers to the average radius recorded for a given tumor. The correlations are like those found for 'radius\_worst'.

'radius\_se' refers to the standard error of the tumor radius. Like the previous features, benign tumors are correlated with low values, but most malignant tumors are not much larger. However, if the ‘radius\_se’ exceeds 1.0, the tumor is always malignant.

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

'symmetry\_worst' refers to the degree of symmetry of the tumor cell nuclei. Again, benign tumor values are lower, while malignant tumor values have a much larger spread of symmetry. The more asymmetric tumor’s are usually malignant.

Chart, box and whisker chart

Description automatically generated

Finally, perimeter refers to the length of the boundary of a cross-section of a tumor. Like pretty much all of the other features, perimeters are larger for malignant tumors versus benign ones.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

# Data Treatment

As discussed earlier, the String ‘diagnosis’ column was converted into a numerical binary ‘diagnosis\_M’ column. No values were missing, so no values were imputed. No other treatment methods, such as binning or scaling, were used; the significant features were selected, and the model was tested before any treatment methods were used and already performed highly.

# Model Development, Tuning, and Comparison

Recursive feature elimination and forward feature selection were used to identify the most significant features. Features common to both methods were used for the model.

|  |  |  |
| --- | --- | --- |
| **Recursive Feature Elimination** | **Forward Feature Selection** | **Common Significant Features** |
| ['radius\_mean',  'compactness\_mean',  'concavity\_mean',  'concave points\_mean',  'radius\_se',  'texture\_se',  'perimeter\_se',  'compactness\_se',  'concavity\_se',  'radius\_worst',  'perimeter\_worst',  'compactness\_worst',  'concavity\_worst',  'concave points\_worst',  'symmetry\_worst'] | ['radius\_mean', 'texture\_mean', 'perimeter\_mean', 'area\_mean', 'smoothness\_mean', 'compactness\_mean', 'concavity\_mean',  'concave points\_mean', 'symmetry\_mean', 'radius\_se',  'perimeter\_se',  'area\_se', 'compactness\_se', 'concavity\_se',  'concave points\_se', 'radius\_worst', 'texture\_worst', 'perimeter\_worst', 'area\_worst', 'smoothness\_worst', 'compactness\_worst', 'concavity\_worst',  'concave points\_worst', 'symmetry\_worst'] | ['concavity\_worst', 'compactness\_mean', 'compactness\_worst', 'symmetry\_worst', 'perimeter\_se', 'radius\_worst',  'concave points\_mean', 'concavity\_mean', 'compactness\_se', 'radius\_se',  'radius\_mean', 'perimeter\_worst', 'concavity\_se',  'concave points\_worst'] |

These significant variables were used to build Artificial Neural Networks (ANNs). The number of nodes, learning rate, number of layers, activation function and kernel initializer for the network model were all optimized using GridSearchCV. The whole search took about 2 hours. These were the best parameters:

**{'activation': 'linear', 'initializer': 'normal', 'learning\_rate': 0.01, 'numHiddenLayers': 2, 'numNeurons': 25}**

A stacked model was created using 5 ANNs, each using the best parameters shown above. Here is the loss/accuracy of training one ANN:

Chart, histogram

Description automatically generated

We can see that just a single ANN has a very high accuracy (over 90% accurate). Each individual ANN was evaluated, and then the stacked model:

|  |
| --- |
| **Evaluating single models with test data.**  Model 1 Accuracy: 0.987  Model 2 Accuracy: 0.974  Model 3Accuracy: 0.974  Model 4 Accuracy: 0.961  Model 5 Accuracy: 0.987  **Evaluating stacked model with test data.**  Stacked Test Accuracy: 0.987 |

The final stacked model has a 98.7% accuracy when determining whether a tumor is cancerous or not, which is very accurate.