IT1244 Report: Breast Cancer Prediction

Team 5

1 Introduction

1.1 The Issue

Breast cancer is one of the most frequently diagnosed cancers in women worldwide (Goel, 2018). In 2020 alone, there were approximately 2.3 million females who contracted breast cancer and 685,000 of those women did not survive it (World Health Organization: WHO, 2021), leading it to become a globally challenging health issue. The prevalence of breast cancer and its fatality rate has led to the questioning of the effectiveness of the current identification and classification of breast cancer, a problem that we want to solve by making use of artificial intelligence (AI), specifically machine learning (ML).

1.2 The importance of solving this issue

The effectiveness of the identification and classification of breast cancer is of utmost importance for a number of reasons. Early detection and screening acts as a secondary prevention for breast cancer as it offers the most feasible, effective and pragmatic intercessions for women globally (Coleman, 2017). Proper classification of breast cancer can also help to determine the most effective personalised treatment for eradicating the cancer specific to the patient (*Breast Cancer - Types of Treatment*, 2023). Furthermore, accurate classification of breast cancer can aid in research efforts to develop new treatments and improve existing ones. With precise classification, researchers can identify specific biological and genetic markers associated with different types of breast cancer, which can inform the development of targeted therapies. Last but not least, with breast cancer being a major global health concern, accurate classification can aid in public health efforts to track the incidence and prevalence of different types of breast cancer. This information can inform policies and interventions aimed at reducing the burden of breast cancer on society.

1.3 Methodology

The first ML technique we used is the K-Nearest Neighbours algorithm, also known as the KNN algorithm. It is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of a singular data point (IBM, n.d.). KNN is particularly suitable for image classification tasks because it works well with high-dimensional data, such as the pixel values of an image. In the case of breast cancer classification, the attributes provided in the dataset describe the characteristics of the cell nuclei present in the image, which can be considered as high-dimensional data. Additionally, KNN is a simple and easy-to-understand algorithm, which makes it a good choice for beginners in machine learning, and does not require any training, which means that the algorithm can be applied directly to new data.

The second ML method we chose to use is Logistic Regression, which is commonly used for classification problems, especially when the target variable is binary or dichotomous (Banoula, 2020). In the context of breast cancer classification, Logistic Regression can be used to predict whether a breast mass is malignant (cancerous) or benign (non-cancerous) based on the values of the input features. Logistic Regression is a simple and interpretable model that can provide insights into which features are most important for predicting the outcome. It can also be easily implemented and trained on large datasets. In addition, Logistic Regression is less prone to overfitting than more complex models, such as deep neural networks, and can perform well even with a relatively small number of input features.

The last method we chose is the Neural Network model. A Neural Network is a type of machine learning algorithm that is modelled after the structure and function of the human brain. It is composed of layers of interconnected nodes, called neurons, which process and transmit information to other neurons in the network. Neural Networks are used for a variety of tasks, including image recognition, natural language processing, and predictive modelling. The basic function of a Neural Network is to learn patterns and relationships in data. To do this, the network is trained on a set of input data and corresponding output values. The network adjusts the connections between neurons through a process called backpropagation, which involves

calculating the error between the predicted output and the actual output and updating the weights and biases of the neurons accordingly. Once the network has been trained, it can be used to make predictions on new data that it has not seen before.

2 Dataset

2.1 Information about the dataset used

The dataset we are using is a multivariate dataset with numerical attributes, with a size of 122 KB. It contains 569 instances with 32 attributes, where there are no missing values. The dataset is associated with two tasks: classification and regression, with a total of two classes for classification. The area of the dataset is medical, and specifically, it provides attributes of images of a fine needle aspirate (FNA) of a breast mass, which describe the characteristics of the cell nuclei present in the image. The dataset is commonly used in machine learning research for breast cancer classification tasks, and it is a relevant dataset for studying the performance of different machine learning algorithms for medical image analysis.

2.2 Issues with dataset

It was difficult to determine the relevance of each individual factor to the accuracy of the identification and classification of the data. In order to counter this, we created 2 separate codes for this dataset, one that used all the data in the dataset and another that filtered out relevant data according to a journal article titled “Breast Cancer Cell Type Classifier” (Qassim, 2018).

3 Methods

3.1 K-Nearest Neighbours Algorithm

It is generally quite simple to make use of the K-Nearest Neighbours (KNN) algorithm as it follows a fixed procedure. We start the implementation of this algorithm by determining the k value, which is equivalent to the number of nearest neighbours. Following that, we calculate the distance between the new data and all the training data used. Then, we sort the distance and determine the nearest neighbours based on the k minimum distances, and determine the class value of these k-nearest neighbours. Lastly, we use the simple majority of the class of nearest neighbours as the predicted class of the new data. We can find the nearest neighbours by using similarity measures between two points or Euclidean Distance. KNN algorithm also requires feature scaling, which can be done through standardisation or normalisation.

3.2 Logistic Regression

The Logistic Regression algorithm is based on the sigmoid function. This algorithm is used for solving classification problems by fitting the sigmoid function, which predicts two maximum values (0 or 1). The sigmoid function is a mathematical function used to map the predicted values to probabilities. It maps any real value into another value within a range of 0 and 1. In Logistic Regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Values above the threshold value tend to 1, and values below the threshold values tend to 0 (Javatpoint, n.d.).

3.3 Neural Network

The first step in implementing a Neural Network involves hypothesising the optimal hyperparameters for it, followed by creating the actual network and then using the model to predict the test dataset and test its accuracy. Hyperparameters indicate how many hidden layers the neural network will have and determine the type of activation function that should be used for each of these hidden layers.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_2 (Dense)             (None, 30)                930

 dense_3 (Dense)             (None, 1)                 31

=================================================================
Total params: 961
Trainable params: 961
Non-trainable params: 0
```

[1]

3.4 Problem Formulation

The gist of the problem is to find out whether a tumour is benign or malignant based on its characteristics. At the core of it, it is straightforward that the way to go about achieving that is to weigh the different features and at the end calculate the probability of each possible outcome.

[1] Figure 1:Summary of neural network model

Thus, we chose to use Logistic Regression, which is the go-to ML technique for solving classification problems. This allows us to have the weightage of each of the features and then use it to predict the nature of a new tumour.

We also chose the K-Nearest Neighbours algorithm, another popular technique for classification problems. By having a large enough dataset, we can predict the nature of a new tumour by dealing with the majority and through comparisons.

The last method that we decided to implement was the Neural Network. Neural Networks have a very broad use and are able to fit into many different types of problems. Since it is highly customisable by changing the hyperparameters, we are able to adjust them to achieve the highest accuracy.

3.5 Novelty of Methods used

What we did differently from MIDHUN DAS L, who did a similar research paper, is that we included code to optimise the recall value. We have decided that a higher False Positive Rate (Type-1 Error Rate) is better than a higher False Negative Rate (Type-2 Error Rate) . Thus we will need to find a threshold that will give us a higher recall value while also making sure the accuracy is still above an acceptable rate.

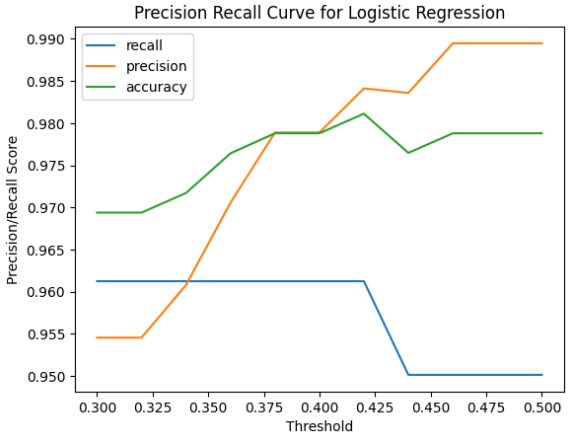
4 Results & Discussions

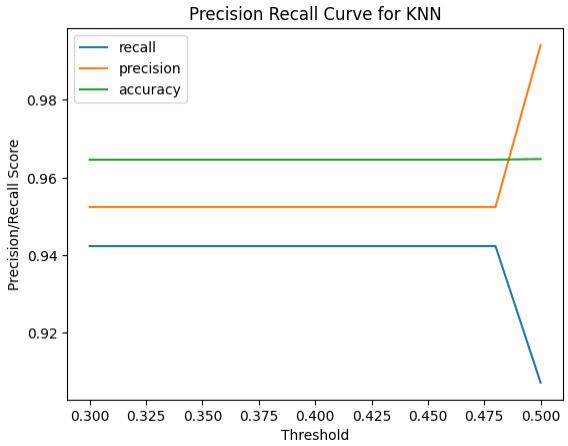
4.1 Fine-tuning Results

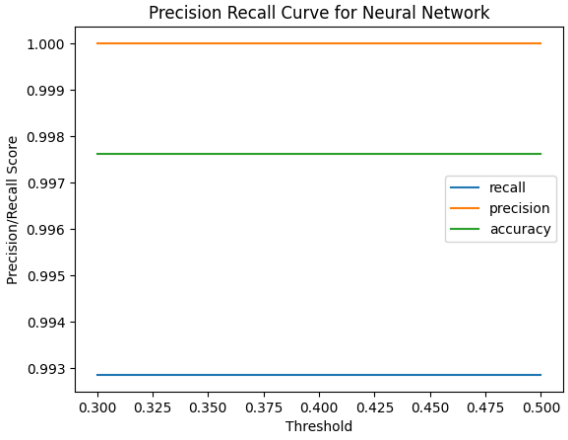
Method	Accuracy	
	9 Features	All features
K-Nearest Neighbour	0.9318936877	0.9647840531
Logistic Regression	0.9342746400	0.9787929125
Neural Network	0.9771303258	0.9952934662

After several experiments, we find that the accuracy for KNN, Logistic Regression and Neural Network using 9 features is less than when using all the features as attributes. Hence, we decided to fine-tune our models by including all the features in the dataset to increase the accuracy.

4.2 Evaluation of results







The performance metrics we used are accuracy, recall scores and precision.

- Accuracy is the ratio of the sum of true positives and true negatives out of all the predictions.
  - Accuracy Score = (TP + TN)/ (TP + FN + TN + FP)
- Recall score represents the model's ability to correctly predict the positives out of actual positives. It is the ratio of true positive to the sum of true positive and false negative.
  - Recall Score = TP / (FN + TP)
- Precision represents the ratio of true positive to the sum of true positive and false positive.
  - Precision Score = TP / (FP + TP)

Since accuracy metrics only consider the number of correct predictions, which are true positives and true negatives, it does not consider the relative importance of errors like false positives and false negatives. Hence, using only accuracy as the main evaluation metric is not good enough to provide a clear picture of the model's performance.

When classifying the images of malignant and benign breast tumours, the recall score should be higher while the precision should be lower. If there is a greater number of false negatives, it would prove to be fatal to the lives of patients. In this case, a high recall score is important because a model with a low recall score may miss some cases of breast cancer, leading to delayed diagnosis and potentially worse outcomes for patients. Hence, it is better to label patients who are negative as falsely positive rather than to label patients who are positive as falsely negative.

When the recall score is higher, the precision should be lower. This is due to the fact that precision is the measure of the proportion of true positive predictions (i.e., the number of correct positive predictions divided by the total number of positive predictions). Precision and recall are inversely related. By improving one, it will typically result in a decrease in the other. For example, a model with high precision will make few false positive predictions, but it may also miss some true positive cases and vice versa (Kumar, 2023).

4.3 Discussions

For this particular case study, we concluded that higher false positives are better than having higher false negatives. In the context of classifying breast cancer, a false positive occurs when the model predicts that a patient has cancer when they actually do not, while a false negative occurs when the model predicts that a patient does not have cancer when they actually do.

In general, false negatives are considered to be more serious than false positives in the context of classifying breast cancer because a false negative means that a patient who actually has cancer is being classified wrongly, which can lead to doctors not ordering further testing, and thus resulting in cancer going undetected for the affected patients. This can have serious consequences because early detection and treatment of breast cancer are vital in improving the chances of survival.

5 Conclusion

Neural Network is the best model to use out of the three that we have shortlisted possibly due to the non-linear nature of the activation function. By using a rectified linear unit (RELU) as the activation function, it will be able to more accurately weigh the features. After running the 10-fold cross validation, we found out that the average accuracy score of the Neural Network is much higher than the rest. Another observation that we made was that the predictions given by the Neural Network are very absolute, meaning the prediction values that it returns is often either 1 or near 0, with very few outliers. As a result, changing the threshold does not seem to work with Neural Network as it did with Logistic Regression and k-nearest Neighbours. This made optimising recall/precision difficult but we decided that since its accuracy is much higher, this tradeoff was acceptable.

```
[1.00000000e+00]
[1.00000000e+00]
[3.96293098e-22]
[1.00000000e+00]
[1.00000000e+00]
[1.00000000e+00]
[9.99966562e-01]
[1.00000000e+00]
[6.78437733e-24]
[5.30018459e-19]
```

[2] Figure 3: Values that the neural network model returned

References

Breast Cancer - Types of Treatment. (2023, March 2). Cancer.Net. https://www.cancer.net/cancer-types/breast-cancer/types-treatment

Banoula, M. (2020, July 6). An introduction to logistic regression in python. *Simplilearn*. https://www.simplilearn.com/tutorials/machine-learning-tutorial/logistic-regression-in-python

Coleman, C. (2017). Early Detection and Screening for Breast Cancer. Seminars in Oncology Nursing, 33(2), 141–155. https://doi.org/10.1016/j.soncn.2017.02.009

Goel, V. (2018, October 12). *Building a Simple Machine Learning Model on Breast Cancer Data*. Medium. https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3

Kumar, A. (2023, March 17). Accuracy, precision, recall & f1-score - Python examples. Data Analytics. https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/#What_is_Accuracy_Score

Logistic regression in machine learning - Javatpoint. (n.d.). Www.Javatpoint.Com. Retrieved April 1, 2023, from https://www.javatpoint.com/logistic-regression-in-machine-learning

Qassim, A. (2018, October 17). Breast cancer cell type classifier. *Towards Data Science*. https://towardsdatascience.com/breast-cancer-cell-type-classifier-ace4e82f9a79

What is the k-nearest neighbors algorithm? (n.d.). IBM. https://www.ibm.com/sg-en/topics/knn

World Health Organization: WHO. (2021, March 26). *Breast cancer*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer