Classification of Breast Cancer Tumours

Bomin Kim, Laszlo Morgan, Paria Naderi Nasab

Abstract:
Correct diagnosis of benign or malignant tumours is crucial in saving money, time and especially lives; thereby at the centre of interest of everyone. At present, machine learning algorithms have been utilised to do this with minimal intrusiveness (e.g., fine needle method) while also achieving a high performance accuracy and accurate diagnostic capabilities compared to a physician analysing the data. In this paper, we present a logistic regression binary classifier built based on key breast cancer biomarkers. Our results show that our model correctly classifies tumour cells with the accuracy of 81%, alongside a recall score of 70% (i.e., Type II error rate). Additionally, we implement a variety of optimisation techniques such as k-fold cross validation and two distinct feature selection methods to improve our model's performance. In conclusion, our model performed with adequate (i.e., not optimal) accuracy and recall, which are especially important when classifying tumour cells. Future works include combining different classification algorithms and exploring different existing techniques of optimisation (e.g., Lasso, electric net) to not only further increase accuracy but also recall score.

1 Introduction

After lung cancer, breast cancer is one of the leading causes of death among women in both developed and underdeveloped countries. Due to late diagnosis and inefficient preventative measures, approximately half of the women diagnosed with breast cancer die [5]. Consequently, early and accurate detection of malignant cancer cells are vital to increase the chance for full recovery.

Cancer is caused by mutation in genes, which begins with unchecked growth of cells that ultimately leads to palpable lumps called tumours. Whereas benign tumours are harmless, malignant tumour cells grow uncontrollably and expand to neighbouring cells. Due to lack of clear physical symptoms in the early stage, it is difficult for patients to notice whether they have cancer or not. After some clinical tests (e.g., biopsy, mammogram, MRI), an accurate diagnosis can be concluded. Individual cells have many distinct features (e.g., texture and size of cell lump) that provide physicians with the useful information they need to distinguish between malignant and benign tumours. However, the process inevitably introduces bias and subjectivity potentially yielding erroneous results (i.e., false positives and negatives). Therefore, to increase the precision and objectivity of this procedure, machine learning is utilised to systematically classify tumour cells and other malignancies. In this paper, we built a supervised learning classifier leveraging a Logistic Regression algorithm to determine whether a patient has benign or malignant cells. The performance of our model is evaluated in terms of accuracy, recall, precision, k-fold cross validation, along with area under the receiver operating characteristic (AUC-ROC curve). To optimise our model, we implemented a wrapper and filter feature selection method which reduces irrelevant attributes and only attains key underlying features.

2 Methods

2.1 Data Description

The Wisconsin Diagnosis Breast Cancer (WDBC) dataset was taken from the UC Irvine repository [3] and developed by Wolberg, Street and Mangasarian [7] at the University of Wisconsin. Each sample was taken from breast masses with the images generated through a camera on a microscope and analysed using the Xcyt program [8]. The data consisted of 569 clinical cases and ten characteristics of the mass: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. For each feature, the mean, standard error, and highest value were additionally measured, resulting in thirty total features. Other information included the patient's ID and a binary value of diagnosis (Malignant (M) =212; Benign (B) = 357).

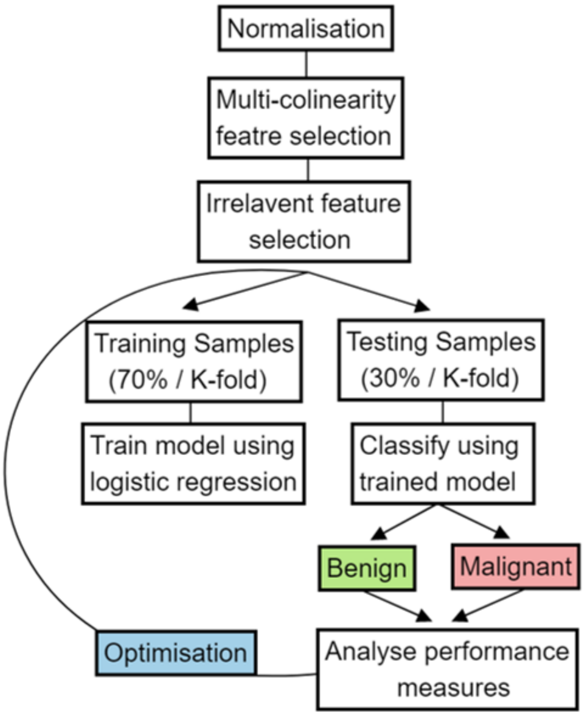
Normalisation
Multi-colinearity featre selection
Irrelavent feature selection
Training Samples (70% / K-fold)
Testing Samples (30% / K-fold)
Train model using logistic regression
Classify using trained model
Benign
Malignant
Optimisation
Analyse performance measures

Fig. 1: Schematics design for breast cancer detection with supervised Machine Learning.

2.2 Pre-processing

The attribute "ID Number" was excluded as it had no relevance to our classification model. Moreover, we were only interested in analysing the mean as it is the measurement that best represents each feature, thereby eliminated standard deviation and the highest value. Next, a correlation matrix was generated which illustrated the relationship of one feature to another. Multi-collinearity is a common problem in linear models with numerous covariates, as it can potentially yield unstable estimates with high variance [1]. Therefore, in our data, perimeter (r = .99 with radius), area (r = .99 with radius), concavity (r = .88$f$omde with compactness) and concave points (r =.83 with compactness) were also removed, yielding a total of six distinct features.

2.3 Logistic Regression Model

Logistic regression is a ML predictive algorithm for binary classification. This algorithm transforms the output using the logistic sigmoid function ($S(x) = 1/1 + e^{-x}$) to return a probability value that an observation belongs to one of two classes. Where the predicted value falls on a sigmoid curve determines if it will be classified as benign or malignant. Any values less than 0.5 are classified as benign, whereas values equal or above 0.5 are indicated as malignant.

The shuffled data was normalised to ensure that each feature with different measurement units is now in a similar range of values to each other, allowing for numerous statistical analysis to be done. The data was split into two sets: a training set (70%) and a testing set (30%), which is a common ratio that is used throughout machine learning. Subsequently, the model was trained to minimise the loss over multiple iterations and resulted in a model that will as accurately as possible categorise unseen examples as either benign or malignant. The loss function used for the model was mean squared error, which aggregates the individual losses for each training data point and calculates the average squared loss for the entire data set. The training procedure entails deducing optimal weights for the features from the labelled data.

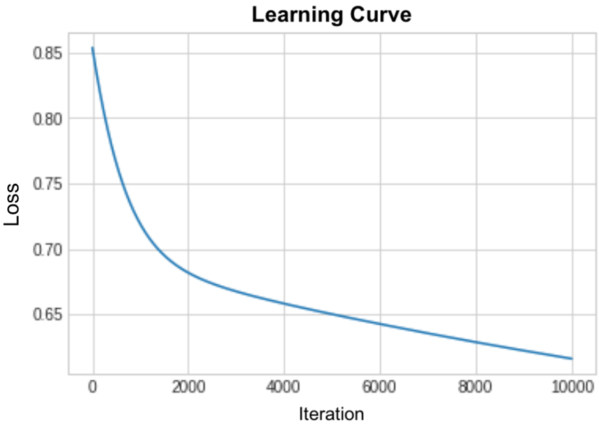

Fig. 2: The learning curve for logistic regression with the training set. The line depicts diminishing loss with more experience, ultimately reaching 0 after training.

3 Results

3.1 Model Performance

The confusion matrix depicts the number of correctly and incorrectly classified malignant and benign tumours (see figure 2). Using these values, we calculated accuracy (i.e., proportion of correctly identifying B as B, M as M), recall (i.e., proportion of true positives on all actual positives), and precision (i.e., proportion of true positives on all positive predictions). Our model reported an accuracy of 74%, a precision of 66%, recall of 62%, and a F1 score of 64% (see table 1); where the F1 score describes an equal weight of precision and recall. The precision score revealed that it predicted malignant tumours correctly on average 66% of the time. The recall score revealed that 62% of the observed malignant tumours were correctly identified, meaning that 38% of malignant tumours were misclassified as benign; this is the Type II error.

Table 1 Performance Metrics of YOLOv3

	Performance		
Measure	Obstacle	Right	Features Selection
Accuracy	0.74	0.73	0.81
Precision	0.66	0.65	0.77
Recall	0.62	0.67	0.70
F1 Score	0.64	0.66	0.73
AUC	0.72	0.72	0.84
Pseudo R^2	0.53	0.55	0.69

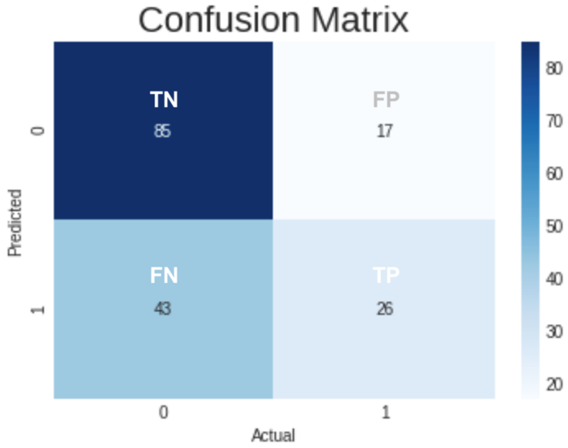

Fig. 3: Confusion matrix of logistic regression classifier. Accuracy= 0.65; F1 score= 0.46.

3.2 ROC & AUC

To assess the discrimination ability of our classifier, we calculated the area under the curve (AUC) and the receiver-operating characteristics (ROC). The ROC curve (see figure 4) represents the relationship between true positive rate (TPR) (i.e., hits) and false positive rate (FPR) (i.e., false alarms) at different threshold levels. An AUC of 0.5 means that the model separates the two classes by chance; and of 1 means that it always predicts each class with 100% accuracy. Our logistic regression model achieved an AUC score of 0.72, meaning that it has a 72% chance of correctly distinguishing between the two classes. In other words, our AUC score suggests that our model classifies the malignant and benign tumours better than chance level but not optimally.

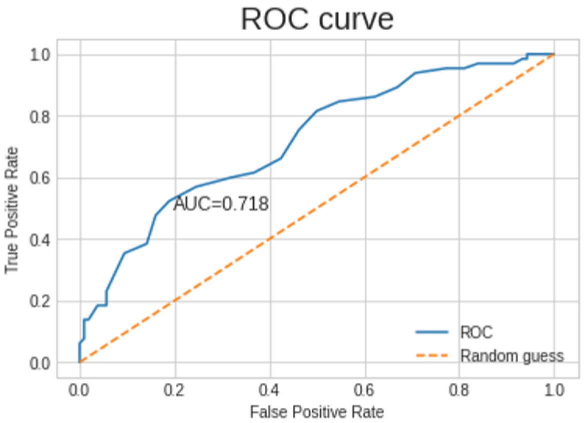

Fig. 4: ROC curve of logistic regression shows the trade-off between TPR and FPR. The curve closer to the random guess line indicates not very good performance at discriminating between malignant and benign tumours.

3.3 Pseudo R^2

R^2 from Ordinary Least Squares (OLS) regression cannot be used for logistic regression. This is because logistic regression uses a categorical dependent variable and continuous independent variables, therefore, their differences cannot be easily compared and interpreted. Hence, we used Efron's Pseudo R^2 as a statistical measure of goodness-of-fit, to understand how well the model fits the data. Our model's R^2 value was 0.53 which indicates that the model does not explain variability of the data very well. These results indicate that our model is likely over-fitting the training data and not generalising well to new data.

$$R^2_{Efron} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{N}(y_i - \hat{y})^2},$$

where $\hat{\pi}$=model predicted probabilities.

3.4 K-Fold Cross Validation

k-fold cross validation is an important re-sampling procedure which evaluates learning algorithms by partitioning the shuffled input data into *k* bins (i.e., folds), and then successively repeating the hold-out method *k* times. For each iteration, the *k*th bin is the testing set, whereas the remaining bins are aggregated as the training set. That is, each instance is used as a hold out at least once. Subsequently, the model calculates the performance measure and averages the values computed for each loop, validating the performance of the model on multiple folds. Though this procedure is computationally expensive, it has many advantages including detecting over-fitting (i.e., failing to generalise), utilising all of our data-set and reducing variance [4]. Conventionally, setting *k* as 5 or 10 is considered a good choice, but there is no formal rule. As k gets larger, the training set also increases, lowering the bias towards estimation of generalisation error (i.e., the true expected error). Using 10-fold cross validation, we achieved 73% accuracy and an F1 score of 66

3.5 Optimisation

The goal of optimisation is to produce a "better" model than the prior model. Different optimisation techniques of supervised learning (e.g., feature selection, Principal Component Analysis) are employed to find the best combination of the feature subsets that could reduce the complexity of the data and ultimately improve performance measures.

Feature selection is the procedure of choosing a subset of significant input variables to avoid overfitting, improve accuracy by storing only relevant attributes, and reduce computation time. There are three main forms of feature selection: filter (e.g., correlation, Chi-square, ANOVA), wrapper (e.g., feed-forward selection, stepwise selection), and embedded method (e.g., LASSO, electric net). Specifically, the wrapper method is a much more sensitive process than filter method, in which different combinations of the features are successively fitted to a model and the variables that fail to meet a predetermined threshold are removed.

In our model, we implement the feed-forward selection method by fitting logistic models (significance level of 5%) for each feature and then sorting p-values for each feature. The feature that does not reach the significance level is eventually eliminated from the dataset. In our model, fractal dimension was the feature with the lowest p-value (did not reach the significance level) and thereby removed. Subsequently, we reran our model with the 10-fold cross validation and the accuracy of our model increased to 81% (5 features) from 73% (6 features). Improvement in our accuracy without fractal dimension feature shows that it is not relevant in classifying breast cancer cells, and can decrease the performance of our model if included.

4 Discussion

In this paper, we built a supervised learning classifier to predict benign and malignant tumours based on six different features. The logistic regression model with 70/30 split achieved 74% accuracy, 66% precision and 62% recall. R^2 value showed that the model was over-fitted. In order to prevent over-fitting, we analysed our data, selecting features that were most relevant to our model using feed-forward selection. Removing the feature 'fracture dimension' increased the accuracy of the model to 81%, precision to 77%, and recall to 70%. Although high accuracy, precision and recall are all important in producing an optimal classification algorithm, in the case of our model, we care particularly about a high recall percentage. The recall value reveals the percentage of false negatives, where a high value reveals lower false negatives and better classifications of malignant tumours. There is a bigger risk associated with misclassifying malignant tumours as benign tumours, than the opposite misclassification, therefore we would require a recall close to 1. In other words, we especially care about having a high recall percentage which our optimisation method did improve, however, 70% recall is still not high enough.

Subsequently, there are some limitations and caveats to consider about our proposed model. One limitation in our study is the limited amount of training data, potentially resulting in the model to adjust excessively to the training data. Therefore, adding more clinical cases could fix the problem of overfitting. Furthermore, including regularisation techniques, such as Least Absolute Shrinkage and Selection operator (Lasso) and ridge regressions could also help in improving generalisability and classification accuracy of our proposed model. Specifically, they both will perform feature selection by shrinking coefficients of insignificant features to zero and preserving significant features. Further research should integrate different algorithms (e.g., Decision Trees, Naive Bayes, and Support Vector Machine) to examine which method is most precise and reliable in binary classification of breast cancer. Moreover, usage of different feature reduction transformation methods such as Principal Component Analysis and autoencoders could provide further insight into more sensitive reduction methods. Another interesting future research will be adding different features to our current data-set to examine whether the accuracy of our classifier improves. Previous studies on breast cancer show that mitotic cell count and tubule formation are good biomarkers in classifying tumour cells [2]. That said, adding key features that could improve the accuracy of our model is a good next direction.

In conclusion, our logistic regression model is not yet ready for use in the real-world as it did not perform as optimally as other existing breast cancer classifiers [6]. This type of classification for illness screening highly prioritises the recall score, and so a score close to 1 is desirable. This is to ensure that malignant tumours are correctly classified.

References

1 Arthur E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970.

2 Chao Li, X. Wang, W. Liu, and L. J. Latecki. Deepmitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical Image Analysis*, 45:121 – 133, 2018.

3 D. Dua and C. Graff. UCI machine learning repository, 2017.

4 Gaoxia Jiang and W. Wang. Error estimation based on variance analysis of k-fold cross-validation. *Pattern Recognition*, 69, 2017.

5 M. Althuis, J. Dozier, W. Anderson, S. Devesa, and L.Brinton. Global trends in breast cancer incidence and mortality 1973–1997. *International Journal of Epidemiology*, 34(2):405–412, 02 2005.

6 M. Amrane, S. Oukid, I. Gagaoua, and T. Ensarİ. Breast cancer classification using machine learning. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pages 1–4, 2018.

7 Olvi L. Mangasarian, W. N. Street, and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.

8 W. Nick Street. Xcyt: A system for remote cytological diagnosis and prognosis of breast cancer. *SERIES IN MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE*, 39:297–326, 2000.