

Classification Model for Detecting Cancerous Tumours

The aim of this report is to describe the decision making, implementation of the SVM classifying model as well as the hyperparameter tuning conducted to boost performance of the model.

The method of creating this model started by comparing untuned models; SVM and Logistic Regression, with both Leave One Out (LOO) and 70%-30% training splits. The library used to facilitate this was sklearn. **Figure 1** shows the results for this preliminary check of model performance and displays marginally better performance for Logistic Regression when using LOO.

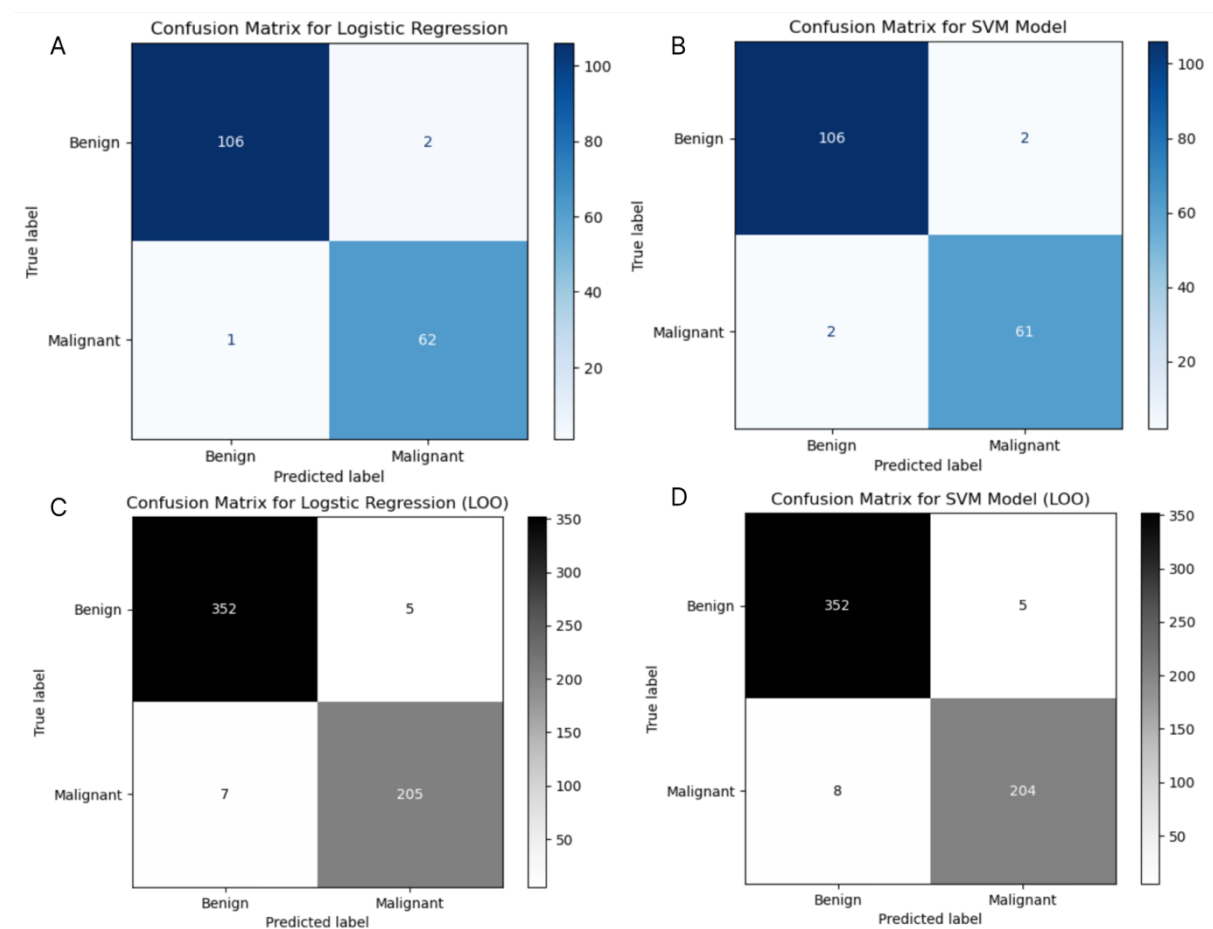


Figure 1 – four confusion matrices with graphs A and B implementing 70%-30% training, test data splits, graphs C and D implementing Leave One Out instead. Graphs A and C implement a Logistic Regression model, graphs B and D implement a SVM model.

Since this is a human centric model if the model was to return a high number of false negatives, then patients presenting with cancerous tumours would not get necessary medical treatment. With this in mind, the decision to use and improve the Logistic Regression LOO implementation was made, the reasons for this were, “off the shelf”

favouring of recall without costing on precision. **Figure 2** shows the increased recall value for negative values, further justifying the decision to use this model over others. The main drawback to this approach is the use of LOO, which is computationally intensive, the rational for its use however comes back to the human centric nature of the data as well as the small data set which the model is being trained on; we need to extract every possible piece of information for it to be useful to us.

	precision	recall
0	0.98	0.99
1	0.98	0.97

Figure 2 – Precision and recall values for the Logistic Regression model implementing LOO. On the left values 0 and 1 indicate a benign and malignant tumour respectively.

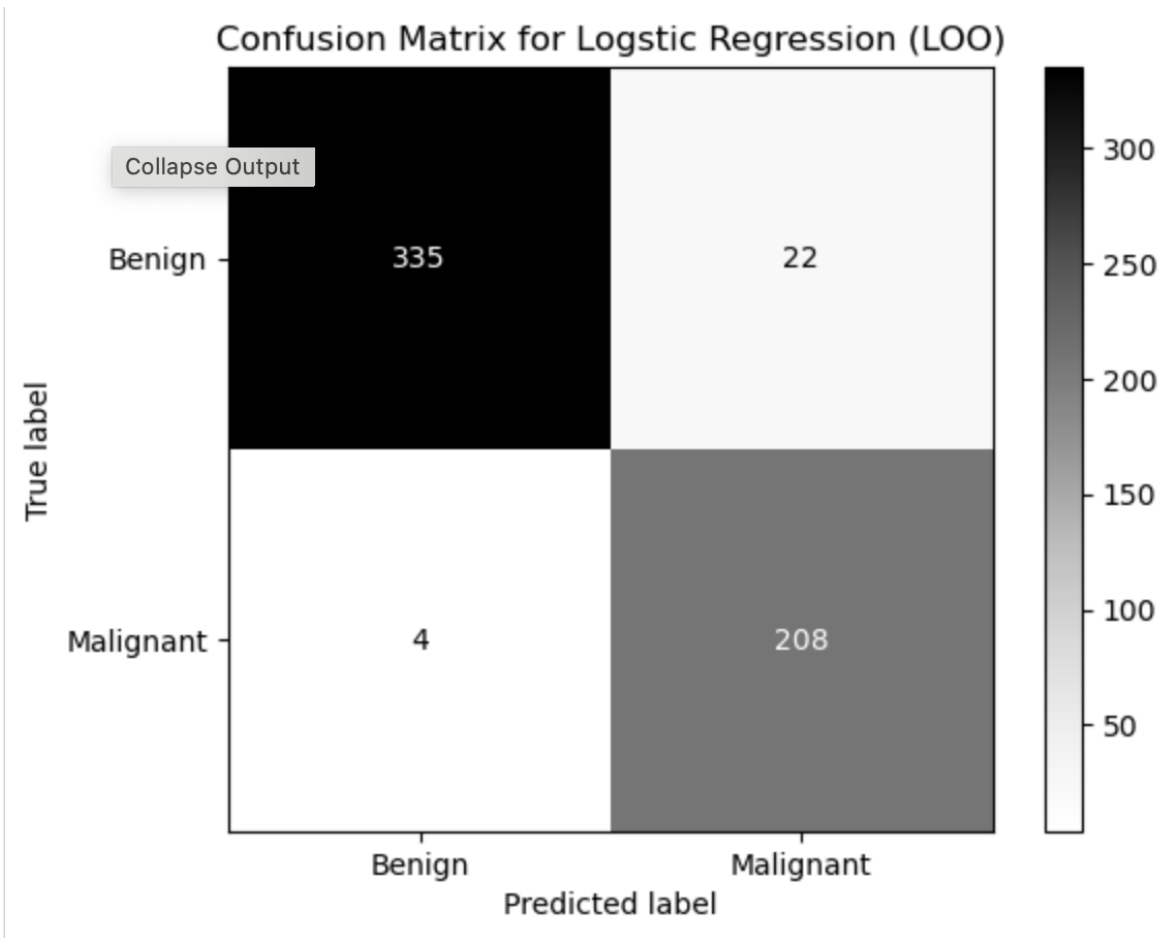


Figure 3 – Results of changing both C and class_weight parameters inside the Logistic Regression function.

	precision	recall
0	0.99	0.94
1	0.90	0.98

Figure 4 – Shows the precision and recall values achieved through changing the C and class_weight values.

Both **Figure 3** and **Figure 4**, highlights the improvements made by simply adjusting thresholds of the model to favour false positives more than false negatives. With precision increase on the benign class and increase recall value for malignant class shown empirically in **Figure 3**. The C value being decreased from the default 1 to 0.17 and making class_weight equal to {0: 0.2, 1: 1} will allow for the observed improvement.

Further marginal improvements can be achieved by changing the solver type and regularisation method used. Changing solver to “saga” and regularisation method to “l1” decreases the number of false negatives by one and the number false positives increases by ten, a drawback to this is the noticeable slowdown when training the model. This has me wondering about whether this is in-fact an improvement or not.

The use of further improvements by change of solver and regularisation method would need to be implemented if being used in a production setting. Considering this would mean that one less person would be missed when diagnosing the slowdown and false positive results are justified. In a production setting three missed diagnosed patients would also not be deemed a passible model, further changes to the threshold to ensure that no false negatives are present would be necessary. A model that implements more harsh thresholds would still be beneficial as a diagnosis tool as it would remove unnecessary use of clinician’s time in much naiver cases. In this implementation it would be more sensible to classify the tumours as “needs professional insight” and “does not need profession insight” where we class tumours close to our classification border as the former.