UNIVERSITY OF CAMBRIDGE

NHS Cambridge University Hospitals NHS Foundation Trust

#10000 BLACK INTERNS

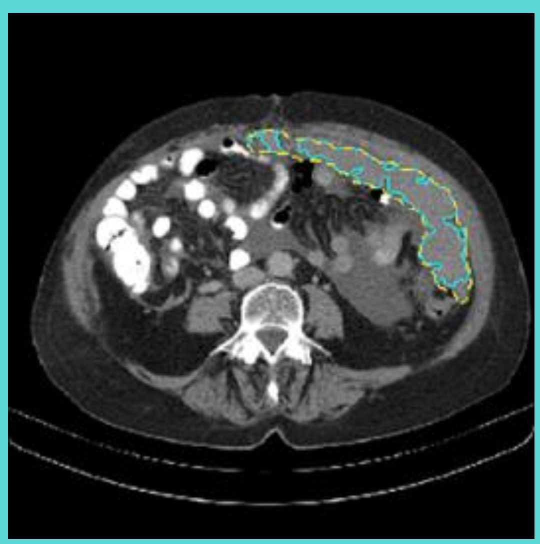
UK Health Data Research Alliance

HDRUK Health Data Research UK

Machine Learning Models To Predict Treatment Response in Ovarian Cancer

Author: Oguntade Adeniji

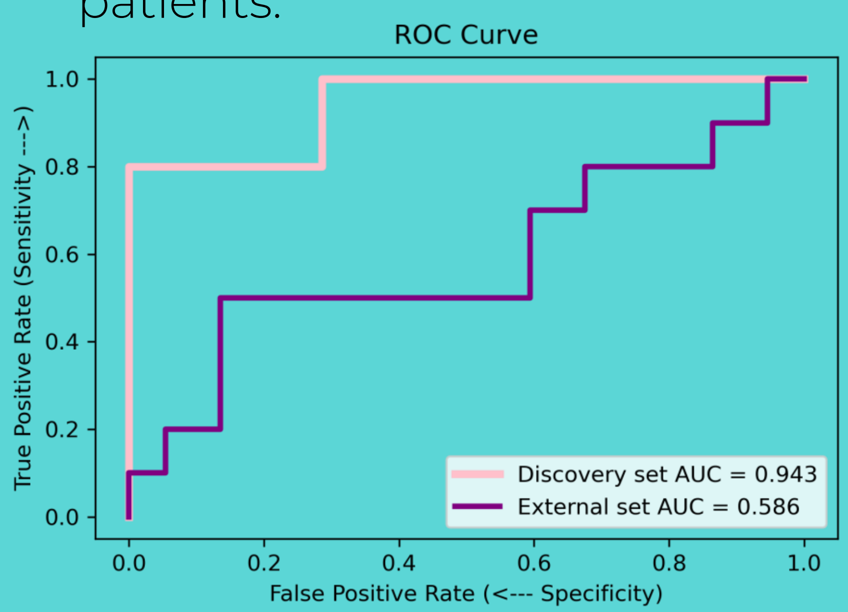
Supervisor: Lorena Escudero Sánchez, University of Cambridge



2D CT slice. Figure from [1]

Introduction

Radiomics, is a field that extracts quantitative data from medical images to provide insights into a patient's condition. This study I worked on focuses on predicting, using Computed Tomography (CT) features, the response of neoadjuvant chemotherapy (NACT) in treating ovarian cancer patients.

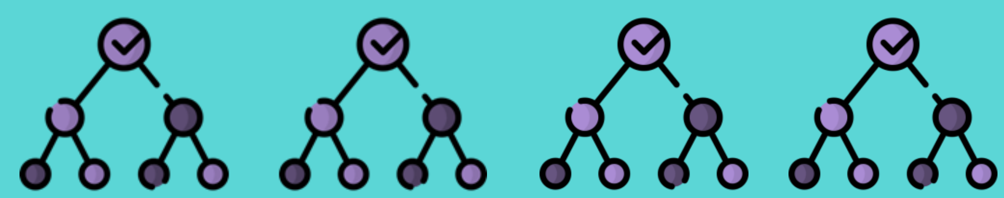


Aim

- Essentially, the idea is that if we can create a model that predicts whether NACT is effective or not depending on the features seen in the CT scans.
- If the results of the model are accurate enough, it can enable personalised treatments, avoiding NACT in some instances when it may not be most effective. Potentially saving resources and lives!

Data Collection and Preprocessing

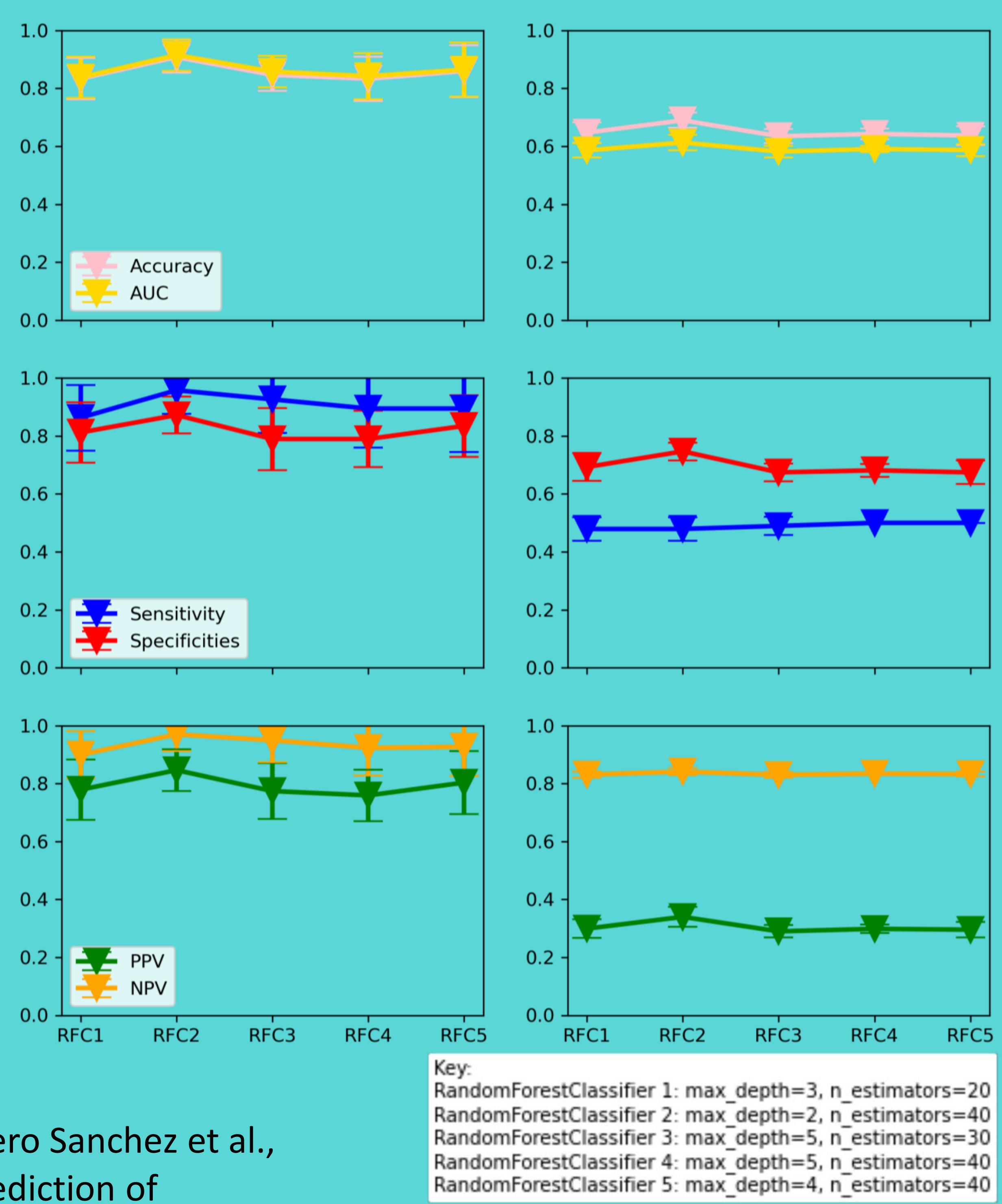
The study uses CT scans of ovarian cancer patients taken before NACT, from two different independent cohorts (discovery and external test sets). A total of 107 features were initially extracted. To avoid overcomplicating, feature selection was performed to narrow this number down to 6: **maximum 2D diameter, least axis length, elongation, inverse difference, homogeneity, difference entropy.**



Machine Learning Model

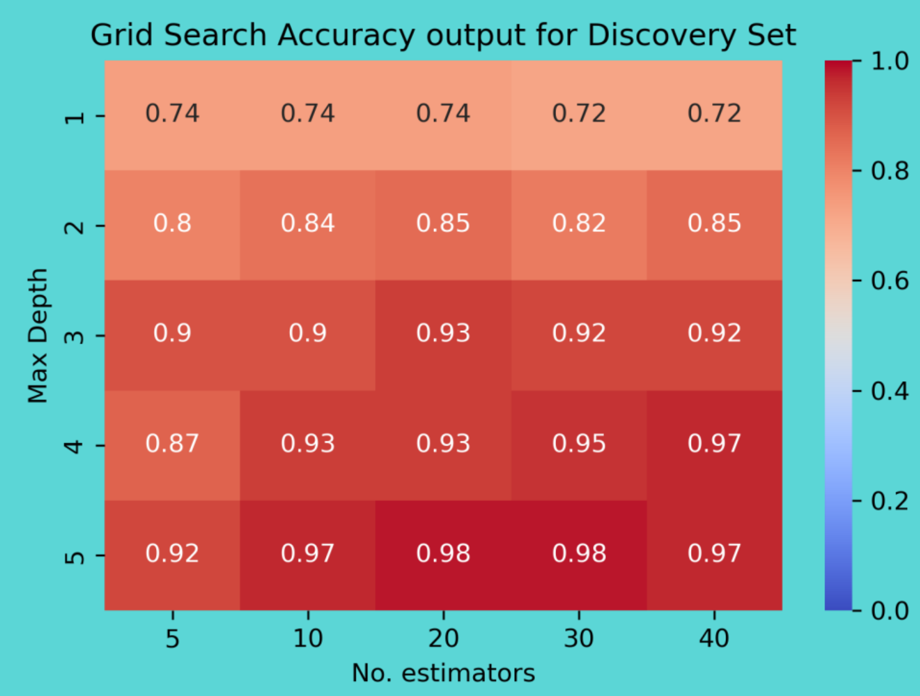
Here, **Random Forest Classifier model** was used. It integrates numerous decision tree classifiers, which are supervised learning algorithms specifically for classification. It constructs a tree-like structure, with each internal node representing a choice based on a feature and each leaf node representing the predicted class label. A decision tree is then built of a random subset of the training data, (bootstrap sampling). Once every tree is built, the Classifier predicts the class by taking a 'majority vote' from all individual trees and the class that receives the most votes becomes the final predicted class. The model training was done using the scikit-learn Python library.

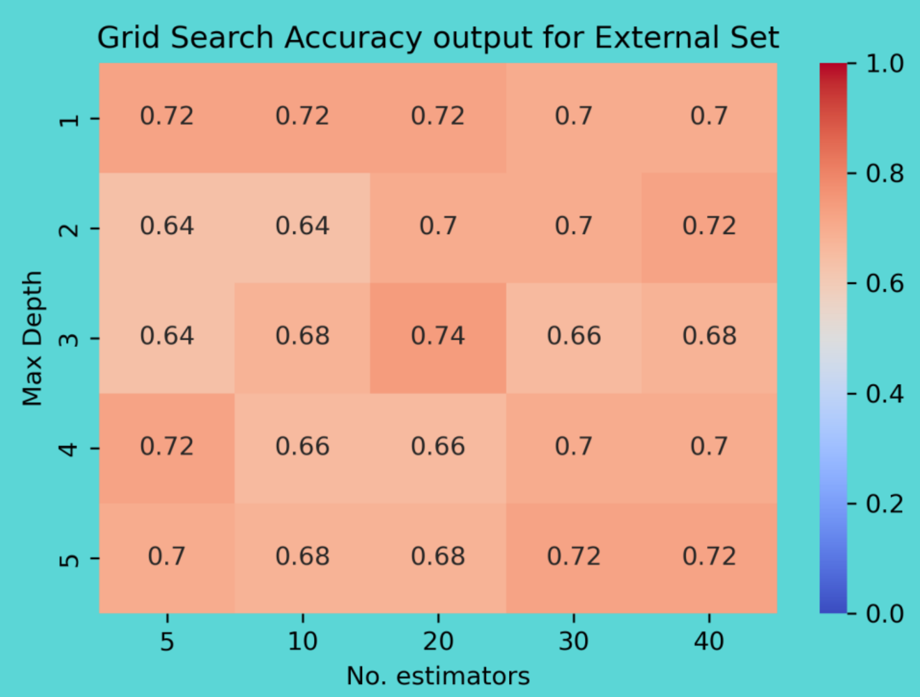
Discovery test Set [Figure 1] External test set



Optimizing the Model

Hyperparameter tuning is a method commonly used to find the best combination of model parameters. The *number of estimators* and *max depth* of the model were the parameters optimized. Using the GridSearchCV function, 25 different variations of the random forest classifier model were made. Only the 5 best performing of these were chosen and compared to one another [Figure 1].





Conclusions

The model chosen was RFC5 (*no. estimators=4, max depth=40*). We chose this model because on the discovery set the difference between sensitivity and specificity was small, and on the external set it has a relatively high accuracy of 0.72. The difference between the results for the two cohorts indicates the difficulty for the model to generalise (overtraining), which is to be expected given the small sample sizes, imbalance and differences in the two populations. Further work can be done with larger datasets in the future.

References: [1] L. Rundo, L. Beer, L. Escudero Sanchez et al., Clinically Interpretable Radiomics-Based Prediction of Histopathologic Response to Neoadjuvant Chemotherapy in High-Grade Serous Ovarian Carcinoma, Front. Oncol. 12 (2022) doi: 10.3389/fonc.2022.868265