Evaluating the Efficacy of Models in Radiomics: A Study with NSCLC Patients

Hervé Hiu Fai CHOI, Ph.D, CertMedPhy (HKIPM)

JULY 12–16 2020 VIRTUAL
JOINT AAPM | COMP MEETING
EASTERN TIME [GMT-4]

AIM

As the number of readily available machine-learning algorithms increases, it may be getting increasingly difficult to select which model to use for prognosis prediction in radiomics. A highly interpretable model with a slightly lower accuracy score may still be advantageous over a more obscure but slightly more accurate model.

This study hence aims to compare the various machine-learning algorithms in their efficacy.

METHOD

Dataset

A dataset consisting of CT images of 422 non-small-cell lung carcinoma (NSCLC) patients was used for the study. The dataset is publicly available through the Cancer Imaging Archive. All CT images were contoured. Two of the 422 CT scans were found to have GTV contours not matching the images and were excluded from subsequent analysis.

Feature extraction

Tumour features were extracted from the region lying within the GTV contour using the Pyradiomics library. For each tumour mask, 107 features were extracted. Some highly collinear variables were removed from the analysis.

The dataset was then partitioned into test and training datasets. An exploratory data analysis was performed on the training dataset. For optimal performance of subsequent machine-learning algorithms, features having highly skewed distributions were transformed with a logarithmic or a Box-Cox transform.

Models

The models attempting to find out if the patients can survive past 1 year, 3 years and 5 years based on the tumour features were:

- logistic regression
- decision tree
- random forest
- gradient boosted decision tree (XGBoost)
- support vector machine
- quadratic discriminant
- neural network

Hyperparameters were optimized by a 5-fold validation, with number of principal components also optimized.

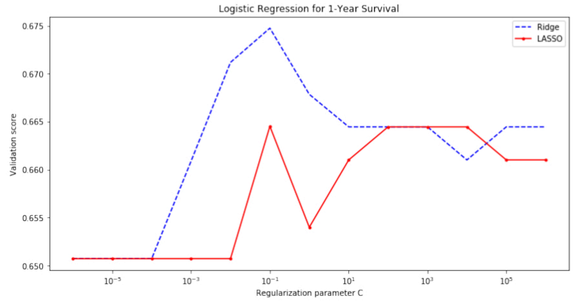
RESULTS

Different models have different hyperparameters to be tuned. Depending on the value of the hyperparameters, the efficacy of the model can vary greatly. It is then paramount to evaluate the accuracy based on the optimal model trained and validated with the training data. Furthermore, regularization of certain models is needed to prevent overtraining of the model and to partially remove multicollinearity between features.

Figure 1 shows the semilogarithmic plot of the validation score from a 5-fold validation in the training set for the 1-year survival classification problem as a function of the regularization parameter C. As an L2-regularization (Ridge regression) results in a higher accuracy score than an L1-regularization (LASSO regression), a Ridge regression model with C = 0.1 was used to calculate the accuracy score on the test data set.

The number of principal components used for the model was treated as a hyper-parameter to be tuned and was found to vary from one model to the other.

Figure 1: Validation score as a function of the regularization parameter and the type of regularization (Ridge and LASSO) for the logistic regression on classifying the 1-year survival



The metric to measure the efficacy of the model used in this study was the accuracy score. The accuracy score is the proportion of patients in the test set where the prognosis was correctly determined by the model. The accuracy scores for the respective models are summarized in Figure 2.

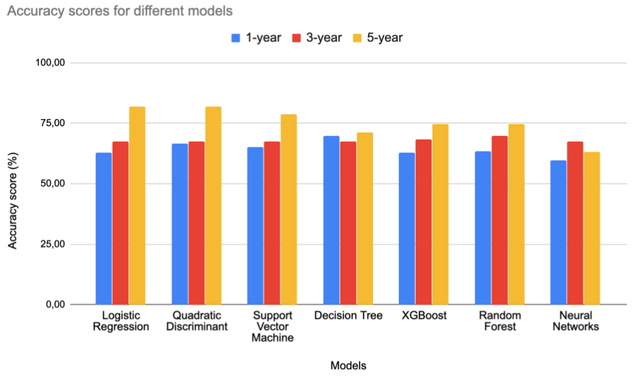
For the 1-year classification problem, all models yield an accuracy score of approximately 60%, with the decision tree yielding the highest accuracy of 69.8%.

For the 3-year classification problem, random forest yielded the best accuracy score of 69.8%.

The accuracy scores for the 5-year classification problems were calculated for completeness. Although the accuracy scores for logistic regression and quadratic discriminant exceeded 80%, these results should be interpreted with caution. The high accuracy scores were mainly attributed to the class imbalance, where 82% of the patients in the training set did not survive after 5 years. This was seen subsequently in the confusion matrix, as all patients in the test set were simply predicted to expire at the 5-year mark regardless of the actual prognosis outcome.

To attempt to elucidate problems arising from the class imbalance, the smaller class in each of the classification problems was oversampled. This was done with SMOTE, an algorithm that interpolated data points within the smaller class in the feature space. Even after oversampling the smaller class in the training sample, the results in the test sample remained unchanged.

Figure 2: Accuracy scores for different models



CONCLUSIONS

The machine-learning models in this study showed comparable accuracy scores for the 1-year and 3-year survival classification problems. Models for the 5-year survival may need to be further optimized before conclusions can be drawn. When training the models, oversampling alone may not be able to elucidate the class imbalance. If all machine-learning models are yielding comparable accuracy scores, then clinicians may consider a more interpretable model in estimating prognosis.

REFERENCES

1. **van Griethuysen JJM, Fedorov A, Parmar C et al.** Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 77(21): e104-e107. URL: https://doi.org/10.1158/0008-5472.CAN-17-0339
2. **Aerts HJWL, Wee L, Rios Velazquez E et al.** Data From NSCLC-Radiomics [Data set]. The Cancer Imaging Archive. URL: https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI
3. **Aerts HJWL, Velazquez ER, Leijenaar RTH et al.** Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5, 4006 (2014). URL: http://doi.org/10.1038/ncomms5006
4. **Clark K, Vendt B, Smith K et al.** The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging 26(6): 1045-57.

CONTACT INFORMATION

Hervé H.F. Choi Ph.D. hhfchoi@gmail.com