**Improving Risk Stratification in TaLG Bladder Cancer Objective**

Link to repository: [Farbod-Moghaddam/BIOF520\_A4](https://github.com/Farbod-Moghaddam/BIOF520_A4)

**The figures could fit in this report and are instead in the Jupyter notebook in the repository. The important figures are indicated in the text.**

**Data Cleaning and Processing**

To prepare the data for Machine learning, only the common variables between both datasets were kept, and some variable names were adjusted to match between both datasets. The datasets were then combined to one hot encode the categorical variables, KNN impute any missing variables, and standardize all the columns. Progression status, PFS time, and the number of follow-up days were removed since these are events that happen in the future. The datasets were then separated into the original datasets.

For the classification target, the recurrence time was binned into less than 6 months, between 6 and 18 months, and greater than 18 months, as well as separating the patients who did not have a recurrence. These bins were chosen because they are clinically interpretable, as well as balancing the training dataset to have about the same number of data points for each target variable. The distribution of the target variables for both datasets can be seen in the GitHub repository. A Kaplan-Meier plot was also created to show the recurrence rate of the patients in each dataset and can also be seen in the repository.

**Feature Selection**

The training dataset was split into 80% for training and 20% for testing. Feature selection was used on the testing set since there were too many features in the dataset compared to the number of patients. Two different methods were used: a recursive feature elimination (RFE) method using a random forest, as well as using principal component analysis (PCA). A Shap plot was created of the top 20 features selected, which will help with the interpretability of the model and is shown in the GitHub Repository.

**Model Selection**

An XGboost model and a neural network (NN) model were trained on the dataset. The XGboost model was trained using 1000 estimators, 20 features from the RFE, and 20 features from PCA. The model was then tested on both the internal and external validation sets, and the weighted performance parameters were recorded for each set. The neural networks were trained via a grid search to optimize the number of features selected, the number of nodes in layers 1, 2 and 3, the rate of the dropout variable, and the rate of L2 regularization. These models were also weighted based on the distribution of each class. The model with the highest weighted accuracy on the internal validation set was chosen as the best model. The optimal hyperparameters can be seen in the Jupyter notebook. Here is a table summarizing the weighted performance of each model for the internal validation set:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | accuracy | precision | recall | F1 | AUC |
| XGboost RFE | 26.9% | 29.1% | 28.6% | 27.9% | 56.0% |
| XGboost PCA | 23.8% | 25.5% | 25.0% | 23.9% | 54.6% |
| NN RFE | 49.5% | 50.2% | 50.0% | 50.0% | 63.0% |
| NN PCA | 41.9% | 48.0% | 42.9% | 42.9% | 61.6% |

Here is a table summarizing the weighted performance for the external validation set:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | accuracy | precision | recall | F1 | AUC |
| XGboost RFE | 25% | 3.8% | 19.4% | 6.4% | 46.5% |
| XGboost PCA | 25% | 37.3% | 61.1% | 46.3% | 54.0% |
| NN RFE | 24.8% | 49.7% | 25.0% | 20.3% | 45.3% |
| NN PCA | 25% | 0.2% | 41.7% | 0.3% | 60.1% |

Overall, the NN RFE method had the best results when it came to the internal validation set, though all models failed to generalize to the external set and were heavily biased on one class. Confusion matrices were generated for each model in both the internal and external dataset and can be seen in the GitHub repository.

**Clinical Interpretability**

The best model, which was the NN RFE model, has advantages and disadvantages when it comes to clinical interpretation. One advantage is that the features selected are actual features and not features in a latent space, which means that their effect can be visualized via the Shap plot, and we can see how each feature is represented in each group. Another advantage is the way the target variables are stratified since it is now very clear to a clinician whether to expect a recurrence and, if so, when to expect this recurrence. This is an advantage over the current EAU risk stratification, since this method only stratifies the patients into low, medium, and high risk and provides no information about whether they will have a recurrence or when1. One disadvantage of the model is that it is a black box model, so it is not clear how each variable is affecting the final output.

**Conclusion and Future Directions**

All the models presented for this report failed to generalize on the external dataset. Some methods to improve the performance of the model include separating the clinical variables from the expression data and training a model for each data type separately. This is important since it will compartmentalize each data type, ensuring that the model is not dominated by one because of the higher number of features in one. The output of these models could then be combined by using a mixture-of-experts model2. Another approach is to train a classifier to classify patients based on recurrence only and then have a separate classifier that can classify the patients based on time of recurrence. This second model can either be a survival analysis model like a coxPH model or another classifier that bins patients into separate buckets. This also might improve performance since different features could be used for each model.

**Reference**

1. Sylvester, R. J. *et al.* European Association of Urology (EAU) Prognostic Factor Risk Groups for Non–muscle-invasive Bladder Cancer (NMIBC) Incorporating the WHO 2004/2016 and WHO 1973 Classification Systems for Grade: An Update from the EAU NMIBC Guidelines Panel. *Eur. Urol.* **79**, 480–488 (2021).

2. Cai, W. *et al.* A Survey on Mixture of Experts. Preprint at https://doi.org/10.48550/arXiv.2407.06204 (2024).