Benchmarking machine learning strategies for phosphoproteomic data in clear cell renal carcinoma diagnosis and staging

Nazrath Nawaz[1], Weronika E. Borek[1], Josie A. Christopher[1], Pedro Moreno-Cardoso[1], David N. Perkins[1], Amy E. Campbell[1], Andrew Williamson[1], Pedro R. Cutillas[1,2], Arran D. Dokal[1]

(1) Kinomica Ltd, Alderley Park, Macclesfield, United Kingdom; (2) Barts Cancer Institute, Queen Mary University of London, London, United Kingdom | * COI: Authors affiliated with Kinomica Ltd are employees, co-founders and/or shareholders of the company

Background & Experimental cohort

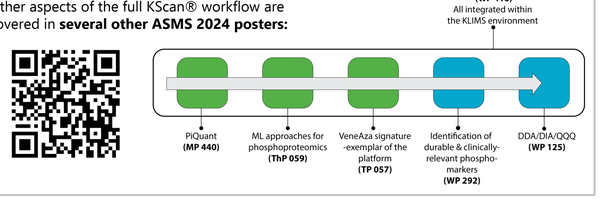
- We identified challenges in machine learning (ML) using mass spectrometry data and performed an evaluation of data pre-processing, feature selection and model building algorithms to determine which methods best address these challenges
- Dataset and sample type: Phosphoproteomic dataset from CPTAC (PCD000412) consisting of 195 clear cell renal carcinoma samples
- Labels to predict: Tumour staging (**T**=size and extent of the main tumour, **N**=degree of spread to nearby lymph nodes , **M**=presence of metastasis)
- Challenges in phosphoproteomics for ML: High feature-to-sample ratio, missing values, data complexity and technical variation
- Orbitrap Fusion Lumos on DIA acquisition mode (phosphopeptides enriched using immobilized metal affinity chromatography)
- MSFragger and Philosopher pipeline for peptide identification. DIA-NN software utilised with an experimentally-derived spectral library from both DDA and DIA data for label-free MS2-based quantification

labels	training	test
M0, M1	34	8
T3, T1, T2	81	10
N0, NX, N1	83	10

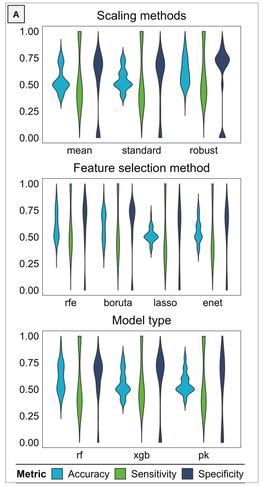
Machine learning tools and approaches

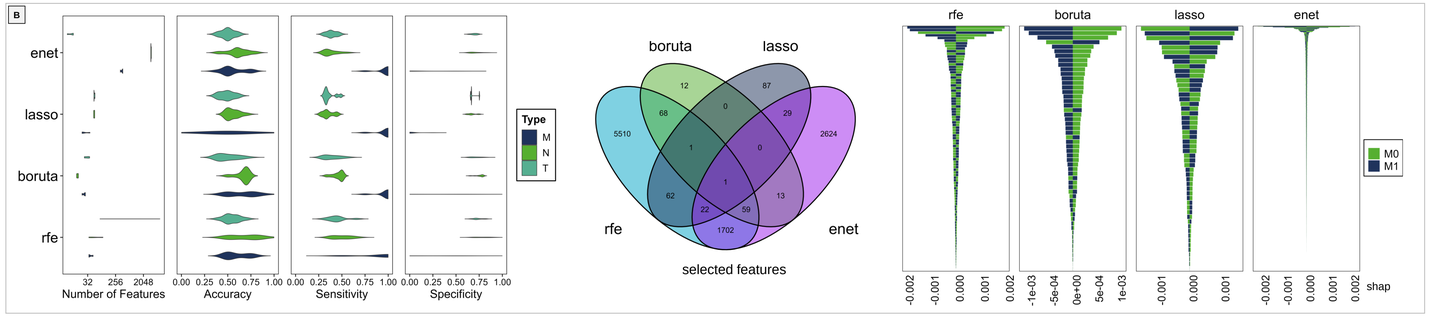
- Scaling Techniques:
  - Mean scaling to normalise feature distributions
  - Robust scaling to reduce the influence of outliers
  - Standard scaling for feature variance normalization
- Imputation: Substitution of missing values with a low constant to preserve non-detected signal information
- Feature selection algorithms:
  - Boruta for identifying all relevant features
  - LASSO for feature shrinkage and selection
  - Recursive Feature Elimination (RFE) for iteratively pruning least important features
  - Elastic Net as a combined L1 and L2 penalty to maintain model interpretability and handle correlated features
- Machine learning models:
  - Random Forest (RF) for its robustness to overfitting and handling of unbalanced data
  - XGBoost (XGB) for high-performance gradient boosting
  - Conditional Random Forest (Cforest) for conditional inference trees
- Validation: We used a holdout set for unbiased model validation and repeated 10-fold cross-validation to ensure stability of model performance

Other aspects of the full KScan® workflow are covered in **several other ASMS 2024 posters:**

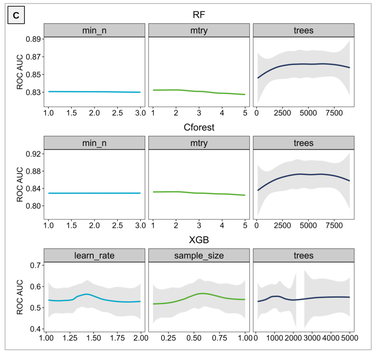


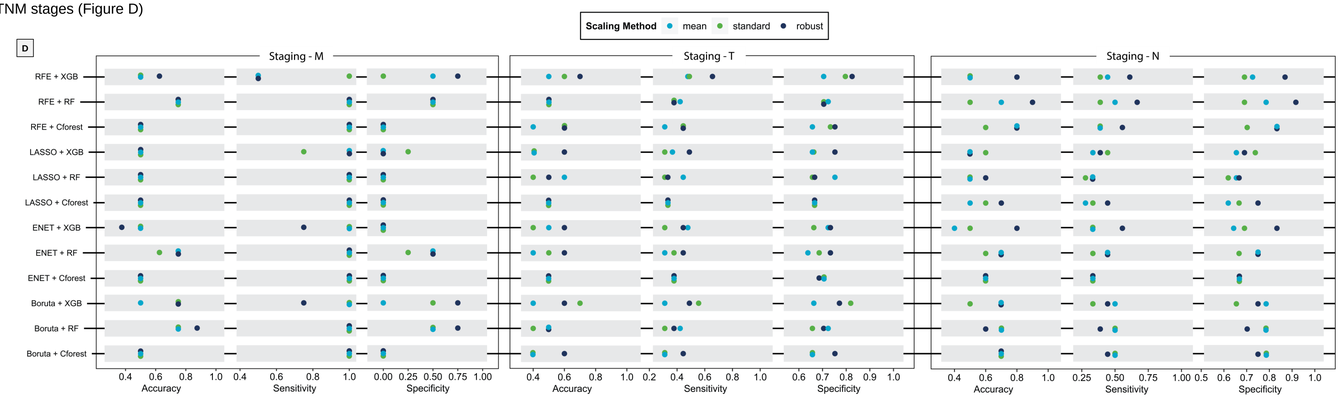
Results





- There is no single best feature selection algorithm (Figure A middle and Figure B left) or modelling type (Figure A bottom). Robust scaling dominated amongst the scaling methods for T and N (Figure A top).
- Each feature selection method finds a different set of predictive features with some overlap and Elastic Net identifies the largest number of features (Figure B middle). Shapley values, the average weighted marginal contribution, shows that each set of features have a range of importances to their classification in isolation and in combination with others (Figure B right). Boruta identifies features that predominantly explain one class (shown in the example for M in Figure B right)
- Certain parameters have a larger effect on the model accuracy (through cross validations) compared to others (Figure C)
- Predictions on a holdout set shows us that the combination of scaling, feature selection and model types for the best model / predictions in terms of accuracy, sensitivity and specificity are different between the TNM stages (Figure D)





Conclusions

- As shown in examples within the same dataset, the best methods depend on the data and specific objective. Thorough evaluations on data cleaning and pre-processing help in producing more accurate models
- Selecting the best features for the model is crucial as each selection method has its own strengths, weaknesses and features it identifies to be predictive
- Model parameter tuning and architecture can greatly affect the ability to predict on unseen data

References

- Dokal, A. (2023). Phosphoproteomics accurately predicts responses to midostaurin plus chemotherapy in two independent cohorts of FLT3 mutant-positive acute myeloid leukaemia.
- KScan: personalizing medicine with phosphoproteomics. https://www.nature.com/articles/d43747-022-00013-9].
- Li Y., Lih TM., Dhanasekaran SM., Mannan R., Chen L, Cieslik M, et al. (2022). Histopathologic and proteogenomic heterogeneity reveals features of clear cell renal cell carcinoma aggressiveness. Cancer Cell 45, 1-25.