H2020-SC1-FA-DTS-2020-1
AI for Genomics and Personalized Medicine

PANCAIM

Pancreatic cancer AI for genomics and personalized Medicine

Starting date of the project: 01/01/2021
Duration: 48 months

= Deliverable D4.1 =

Open-source prediction algorithms beta-tested and made available in GitHub

Due date of deliverable: 30/06/2023
Actual submission date: 28/06/2023

Responsible WP: Tero Aittokallio, WP4, Oslo University Hospital
Responsible TL: Tero Aittokallio, T4.1, Oslo University Hospital
Revision: V1.0

Dissemination level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016851.

AUTHOR

Author	Institution	Contact (e-mail, phone)
John Zobolas	OUS	ioannisz@uio.no
Tero Aittokallio	OUS	t.a.aittokallio@medisin.uio.no

DOCUMENT CONTROL

Document version	Date	Change
V0.1	21/06/2023	First draft
V0.2	22/06/2023	Review by project manager
V0.3	27/06/2023	Review by other partners
V1.0	28/06/2023	Final version

VALIDATION

Reviewers		Validation date
Work Package Leader	Tero Aittokallio, OUS	30/06/2023
Project Manager	Kristin Aldag, AMI	30/06/2023
Coordinator	Henkjan Huisman, RUMC	30/06/2023

DOCUMENT DATA

Keywords	Survival analysis, ML benchmarking, multi-omics AI modeling
Point of Contact	Name: John Zobolas Partner: OUS Address: Ullernchausseen 70, OUS, Radiumhospitalet, 0379, Oslo E-mail: ioannisz@uio.no
Delivery date	28/06/2023

DISTRIBUTION LIST

Date	Version	Recipients
21/06/2023	V0.1	Project manager, coordinator, other partners via email
27/06/2023	V0.3	Project manager, coordinator via email
28/06/2023	V1.0	Commission via portal, partners via OwnCloud

DISCLAIMER

Any dissemination of results reflects only the authors' view and the European Commission Horizon 2020 is not responsible for any use that may be made of the information Deliverable D4.1 contains.

Executive Summary

This report describes the development, implementation and beta-testing of multi-modal AI algorithms for integrating diverse data types and predicting the survival of pancreatic cancer patients. The report outlines the key features and functionalities of these AI algorithms and the utilization of modern AI techniques for model optimization and hyperparameter tuning. Key developments include a new algorithm to perform multi-modal feature selection on high-dimensional datasets and a novel methodology to benchmark multiple AI models on multi-modal datasets. Beta-testing of these implementations is carried out on a public TCGA pancreatic dataset and key results are presented. All the methods, models, and benchmarks have been implemented in R as open-source code (https://github.com/bblodfon/survmob/). The benchmarking analyses and results, along with the code, are openly available on GitHub (https://github.com/bblodfon/tcga-survmob/). Overall, the report showcases the advancements made in developing openly-accessible and reliable survival prediction algorithms to empower translational researchers and clinicians in effectively managing patients with pancreatic cancer.

Table of contents

List of tables

List of figures

1. Introduction

Deliverable 4.1 directly relates to PANCAIM WP4 Task 4.1. The main objective of Task 4.1 is to **develop multi-modal AI algorithms for integrating diverse data types** (such as genomic, pathological, and radiological information) and constructing explainable prediction models. Task 4.1 involves the development of feature selection algorithms to identify subsets of predictors for various clinical outcomes, prioritizing clinical significance and cost-effectiveness. It also encompasses the investigation of methodologies for ranking the tested AI models and selecting the modalities that are most useful for the prediction of patient survival. These will provide the basis for the development of clinically applicable AI models that leverage routinely available and affordable data and can provide valuable insights for translational researchers and clinicians. This report outlines the main efforts in AI model development and the exploration of new modeling techniques to implement, test and compare such AI models **in the context of pancreatic cancer patients' survival prediction**.

1.1. Abbreviations

Table 1: Acronyms and abbreviations

ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
IBS	Integrated Brier Score
ERV	Explained Residual Variation
RCLL	Right-Censored Log Loss
KM	Kaplan-Meier
ML	Machine Learning
AI	Artificial Intelligence
CRAN	Comprehensive R Archive Network
TCGA	Cancer Genome Atlas
PDAC	Pancreatic Ductal Adenocarcinoma
PAAD	Pancreatic Adenocarcinoma
AFT	Accelerated Failure Time
RSF	Random Survival Forests
OOB	Out-Of-Bag error
RFE	Recursive Feature Elimination
eFS	Ensemble Feature Selection
mRNA	Gene expression data
miRNA	microRNA data
CNA	Copy Number Alterations data
RPPA	Reverse Phase Protein Array data
SVM	Support Vector Machines

2. Results and Discussion

2.1. AI models review

The project involved an extensive evaluation of various types of survival prediction models, most of which are implemented in the R programming language. To ensure comprehensive coverage, we consulted the CRAN task review on survival analysis[1], which provides a curated list of relevant survival modeling packages. We further explored various survival models described in research publications. Our focus was on selecting models that have demonstrated superior predictive performance in the literature. We specifically prioritized open-source implementations that adhere to sound software code practices and can be effectively tuned (Table 2). Several models did not meet the desired criteria during the evaluation. For instance, the survival Support Vector Machines faced challenges with optimization on a small dataset[2]. The Gradient Boosted Cox Regression model[3] suffers from limited development and lack of parallelization, making alternatives like XGBoost more favorable. Additionally, the model-based boosting family of algorithms[4], including Generalized Linear, Additive and Boosting Tree models, exhibited frequent memory errors and were very slow during training, significantly hampering their computational efficiency. Based on the survey, the *mlr3* framework was utilized [1] for effective model tuning, evaluation, and benchmarking, along with the R package *mlr3proba* [2], which offers a unified interface for survival machine learning (ML) models and performance evaluation measures (Table 3).

Table 2: Overview of survival models tested.

Model	Abbreviation	Prediction type	R package	Publication
Cox Proportional Hazards	CoxPH	risk + survival	survival	[3]
Regularized Cox Regression with elastic net penalty	CoxNet	risk	glmnet	[4]
Likelihood-based Boosting	CoxBoost	risk + survival	CoxBoost	[5]
Extreme Gradient Boosting (Cox)	XGBoost (Cox)	risk	xgboost	[6]
Extreme Gradient Boosting (AFT)	XGBoost (AFT)	risk	xgboost	[7]
Random Survival Forests (Logrank spitrule)	RSF (Logrank)	risk + survival	ranger	[8]
Random Survival Forests (C-index spitrule)	RSF (C-index)	risk + survival	ranger	[9]
Random Survival Forests (Maximally selected rank statistics splitrule)	RSF (Maxstat)	risk + survival	ranger	[10]
Random Survival Forests (Extremely Randomized Trees)	RSF (ExtraTrees)	risk + survival	ranger	[11]
Accelerated Oblique Random Survival Forests	AORSF	risk + survival	aorsf	[12]

One of the key challenges in survival modeling is assessing the performance of survival models based on their predictions. There exist two main prediction types in survival modeling: the **survival distribution** (a patient's survival probabilities over a specified time period) and the **risk score** (a single time-independent value where a higher number corresponds to a higher risk of an event taking place, e.g., a patient's death). Survival models, based on their implementation details, can either output the first prediction type, the second or both (usually via a transformation from the main prediction output). In Table 2, we list the default prediction types a user can expect by using the survival models from the corresponding R packages.

Several performance measures exist to evaluate survival models. These can either assess the **discriminatory** power of a model (i.e., how well it ranks patients - here, input is the patients' risk scores), how well a model is **calibrated** (i.e., how closely the predicted survival probabilities agree numerically with the actual survival outcomes - input is the patients' survival distributions) or the model's **overall predictive ability** (i.e., both calibration and discrimination). For example, measures such as the integrated area under time-specific ROC curve

[1] https://cran.r-project.org/web/views/Survival.html
[2] https://github.com/mlr-org/mlr3proba/issues/287
[3] https://github.com/gbm-developers/gbm
[4] https://github.com/boost-R/mboost

(ROC-AUC [13]) and the concordance index (C-index [14]) are measures of discrimination, D-calibration is a measure of calibration [15], the Integrated Brier score (IBS [16]) is used to evaluate both discrimination and calibration performance. In Table 3 we list several survival performance metrics and the prediction type they require as input.

Table 3: List of survival metrics for the assessment of discrimination and calibration performance of survival models.

	Prediction type	
Metric	**Risk**	**Survival**
C-index	YES	
ROC-AUC	YES	
Right-Censored Log Loss (RCLL)		YES
Integrated Brier Score (IBS)		YES
D-calibration		YES

It is crucial to exercise caution and consider the aforementioned factors (prediction types and performance measures) when evaluating the performance of survival ML models. What should guide such choices is primarily the specific scientific or clinical question at hand [17]. Therefore, depending on the objective, one may need to prioritize aspects such as discriminating between patients or generating more accurate and calibrated survival predictions, and as such, appropriate evaluation metrics and models should be selected. To highlight several of the aforementioned modeling challenges and provide a comprehensive guide for future survival model evaluation in the context of multi-modal data integration, we compiled **a step-by-step R tutorial**[5] using TCGA survival and omics data. Additionally, we have uploaded **a paper to the arXiv preprint server** which outlines a general workflow for survival analysis [18]. Our aim is to provide researchers with a valuable resource for survival modeling and prediction, presenting practical and effective strategies for conducting survival analysis and enabling the exploration of diverse modeling techniques.

Another challenging task is **AI model tuning**. Survival ML models incorporate hyperparameters, which greatly influence the model's behavior and require careful tuning. Achieving the desired model performance and execution time depends largely on the optimization techniques employed. The search space for hyperparameters is often vast, demanding thorough exploration and evaluation of different values and combinations. Moreover, the intricate interplay between hyperparameters and model architecture adds further complexity. Therefore, efficient optimization strategies are necessary to overcome these challenges and achieve optimal model performance. As an example of model hyperparameter tuning, a CoxNet model with 2 hyperparameters (alpha, lambda) was tuned on a gene expression dataset from TCGA's PAAD patient cohort (Figure 1). The first optimization strategy, random search [19], explores the hyperparameter space and discovers many suboptimal configurations with lower C-index values. On the other hand, **Bayesian optimization** [20] quickly converges to hyperparameter spaces with higher C-index values, resulting in faster model optimization and better tuning. Therefore, Bayesian optimization was the strategy we used for all subsequent AI model tuning.

[5] https://ocbe-uio.github.io/survomics/survomics.html

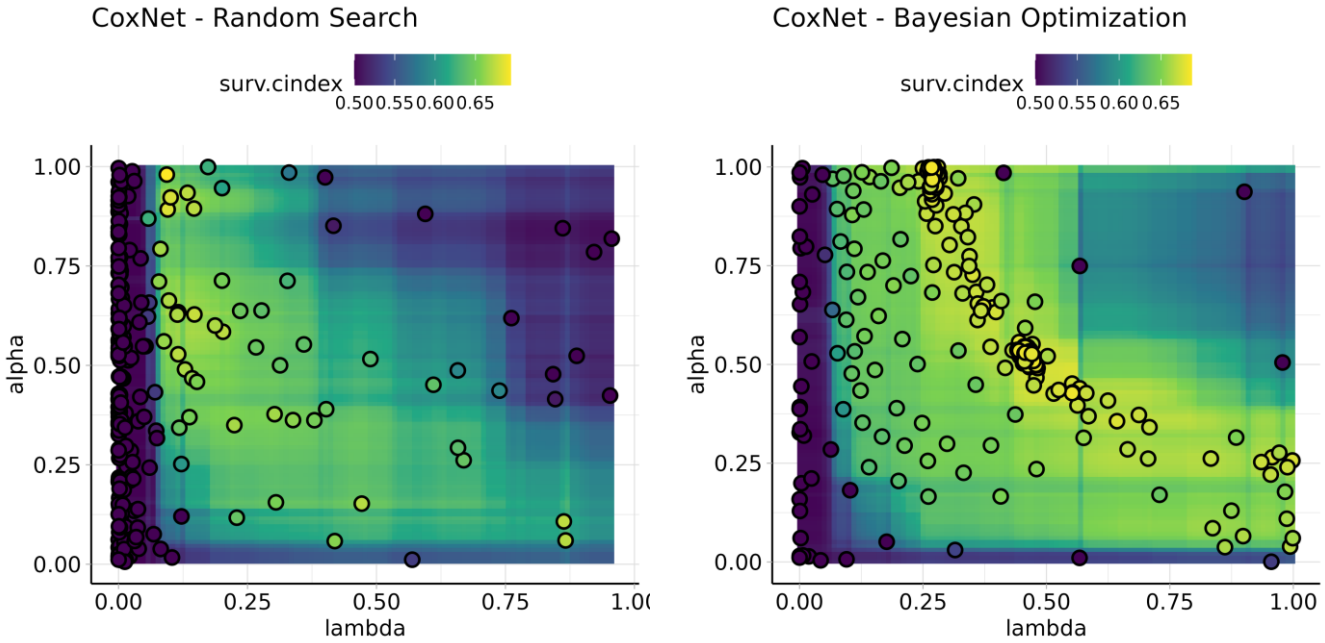

Figure 1: Comparison of two optimization strategies for AI model tuning (Random search and Bayesian optimization).
The figure illustrates the range of values considered for the two regularization hyperparameters of a CoxNet model (alpha and lambda), showcasing the exploration of various combinations by the optimization strategies.

2.2. Open-source repositories

We used TCGA's pancreatic ductal adenocarcinoma cohort (PAAD public dataset) to benchmark and test the survival AI models reviewed. The data retrieval and initial patient filtering were carried out using the *curatedTCGAData* and *TCGAutils* R packages [21]. These filtering steps were implemented to ensure a more reliable and meaningful analysis by selecting relevant patient samples, removing incomplete or heterogeneous data, reducing potential confounding effects and maintaining a sufficient number of patients and events for robust AI modeling. In total, **145 patients** were included, out of which **85** died of pancreatic cancer (censoring rate of 41%). The median follow-up time for the surviving patients was 16 months, with a standard deviation of 15 months. The multi-modal dataset for the 145 PDAC patients consisted of 5 different modalities (common features across all the patients in the cohort), including **6 clinical features** (age, gender, number of lymph nodes, pathologic stage, radiation therapy, residual tumor) and **4 omic layers** of biological profiling information. We carefully considered which omic layers to include from the ones available from the TCGA PAAD cohort. Our main objective was to incorporate a maximum number of omic layers, while ensuring a sufficiently large sample size (>100 patients) in the cohort, without significantly increasing the censoring rate. Accordingly, the following omics were selected: gene expression (mRNA), segmented somatic Copy Number Alteration calls (CNA), microRNA expression (miRNA) and methylation beta values (Methylation). Appropriate pre-processing steps were applied to each specific omics layer, the most important of which were variance filtering (keeping the 10000 most variable features due to the ultra-high dimensionality of several of the omics data), model-based imputation of missing values using XGBoost, removal of pairwise correlated features exceeding 0.95 Pearson correlation, log2-transformation and standardization of features to have a mean of 0 and a standard deviation of 1, ensuring that all features are on a comparable scale. In the end of the pre-processing step, every omic had ~10K features, except miRNA which had ~350 features.

We have made our initial code for model development, testing, and several investigations (as the one shown in Figure 1 for AI model tuning), available on a GitHub repository, serving as a comprehensive resource for more detailed information (https://github.com/bblodfon/paad-survival-bench/). In order to advance our work, we prioritized improving the code quality by incorporating rigorous testing and comprehensive documentation. Additionally, we focused on modularizing the codebase to enable seamless integration of new enhancements and functionalities while building upon the existing code and workflows. The result of this process was the development of a new R package called ***survmob*** (https://github.com/bblodfon/survmob/) for benchmarking

survival ML models on multi-omics data. Key features of the *survmob* package include preprocessing *mlr3*-compatible pipelines for various types of data (including omic datasets), a list of survival ML models that can provide both risk and survival distribution predictions by applying transparent transformations for models that do not support both prediction types (Table 2), predefined ranges for tuning the models' hyperparameters, as well as a wide range of survival measures including both discrimination and calibration measures (Table 3).

Furthermore, the *survmob* R package incorporates a novel **ensemble feature selection algorithm,** called eFS, which leverages Random Survival learners (RSFs) and a Recursive Feature Elimination (RFE) optimization strategy to identify robust features for survival modeling from high-dimensional omic datasets. Additionally, *survmob* offers a **benchmarking methodology** for evaluating the performance of different models and ranking the different omics according to their contribution to model's performance when they are used as part of a multi-modal dataset. Lastly, we provide a post-hoc Bayesian-based analysis of benchmarking results, facilitating the interpretation and comparison of survival models. With its extensive capabilities, *survmob* serves as a valuable tool for researchers and practitioners involved in survival ML analysis studies using multi-omics datasets. We applied *survmob* on several TCGA studies and made the results available on GitHub (https://github.com/bblodfon/tcga-survmob/). This repository includes the detailed benchmark on the pancreatic TCGA PAAD cohort, the findings from which are presented in this report.

2.3. Ensemble feature selection

Given that high-dimensional biological datasets often contain thousands of features, it is crucial to perform feature selection prior to utilizing machine learning models for robust and cost-effective outcome prediction. Selecting relevant features additionally helps improve model performance, reduces overfitting and results in faster model training and tuning [22]. Moreover, to achieve cost-effective and explainable models, a minimum set of interpretable features is often required [22]. Therefore, we set out to implement a novel **ensemble feature selection algorithm (eFS)**, that generates a ranked list of predictive features on a given high-dimensional dataset. The eFS algorithm employs a wrapper-based feature selection approach called **Recursive Feature Elimination (RFE)**, which selects the most predictive features by evaluating the performance of a random survival forest model (RSF) trained on subsets of features. RFE is a backward selection method that starts with all features and iteratively eliminates the least important ones based on their impact on the model's performance. The best feature subset is the one that achieved optimal performance according to the chosen survival metric (Table 3) and resampling scheme.

For executing the eFS algorithm using the different PAAD omic datasets, the **Out-Of-Bag (OOB) error** of random forests (equivalent to 1 - C-index) serves as an optimization metric. The OOB error has the advantage of being very fast to calculate and it does not require an additional resampling scheme, as all (subsampled) patients are used to train each RSF model.

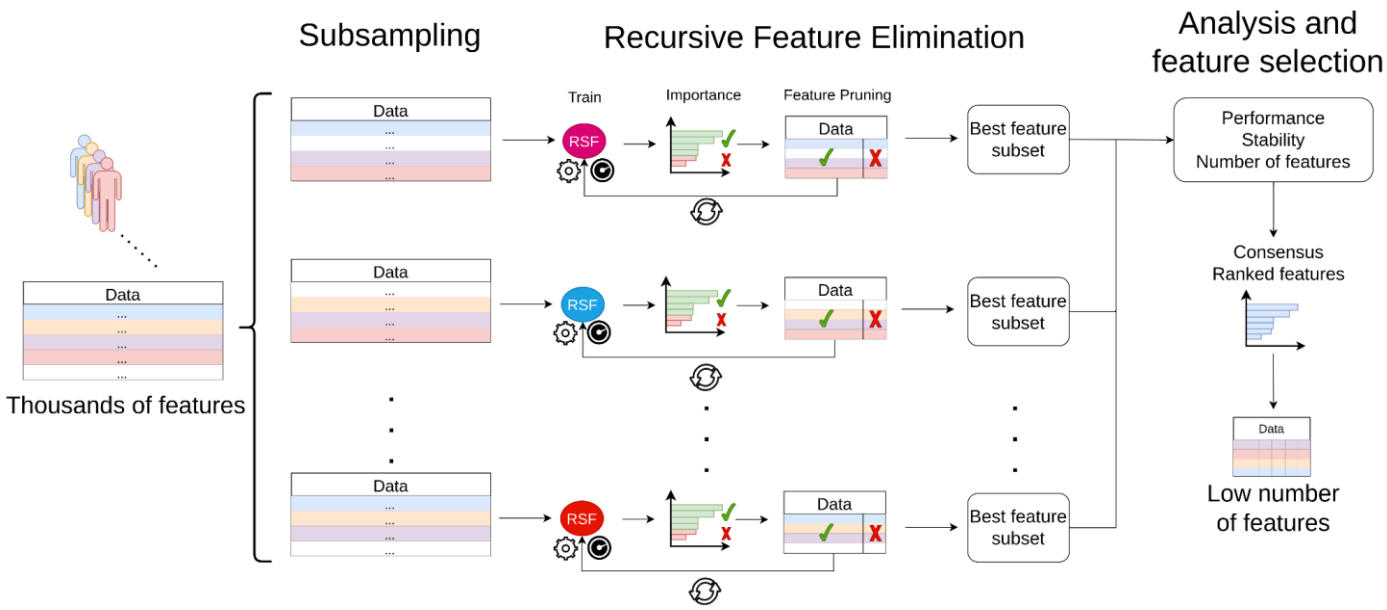

Figure 2: Ensemble feature selection workflow

Our current implementation includes five different versions of Random Survival Forests (RSFs), see the last five rows of Table 2. This approach introduces **diversity on both the algorithmic and dataset levels**. By using a heterogeneous ensemble of stochastic models, the selected features from the RFE process may vary even with the same input data. Additionally, by subsampling the data and selecting different subsets of patients, we can derive optimal feature subsets, allowing for diversity at the dataset level. Also, Random Forests in general require less tuning compared to other models, making them well-suited for our scenario. By combining these ideas in an ensemble feature selection workflow (Figure 2), we obtain a collection of the best feature subsets, each derived from a slightly different patient population and potentially utilizing a different RSF model. Aggregating these best feature subsets can result in a more **robust consensus feature list**, where features are ranked based on how frequently they appear in the collection of the best feature sets [23,24,25]. Figure 3 illustrates two examples of ranked features obtained from the eFS algorithm, when applied to the mRNA and CNA omic datasets from TCGA's PAAD cohort. Lastly, our implementation prioritizes user-friendliness by utilizing R6 classes[6] and parallelized execution, ensuring faster retrieval of results and facilitating further downstream analysis.

Using as input to the eFS algorithm the 4 high-dimensional omic datasets from the PAAD cohort, we generated a collection of optimal feature subsets for each omic data type, whose performance was evaluated based on the C-index metric (1 - OOB error). Figure 4 (left) shows that these **features are highly predictive** in terms of discrimination performance (C-index ranges from 0.70 to 0.85) and we also observed that **some omics are more predictive than others** (mRNA and Methylation have higher median C-index than miRNA and CNA). Upon closer examination of each individual RSF model's contribution, we observe that the AORSF model results in a better performance at the cost of utilizing a larger number of features (Figure 4, right). Aggregating the results across all RSF models, it seems that on average **a minimum number of features** are required to achieve the aforementioned level of performance, independent of the omic used (less than 25 features).

[6] https://adv-r.hadley.nz/r6.html, https://r6.r-lib.org/

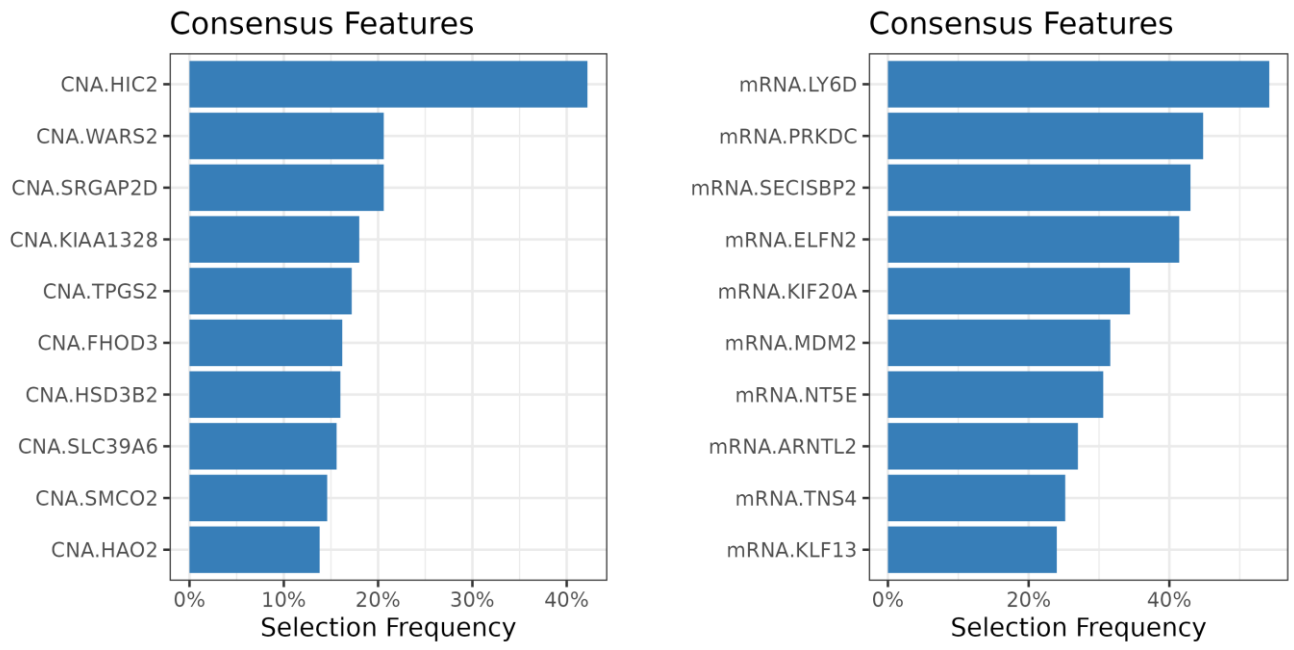

Figure 3: Ranked features associated with survival outcomes (optimizing for better patient discrimination via the C-index) on the PAAD TCGA dataset
(Left: CNA, Right: mRNA).

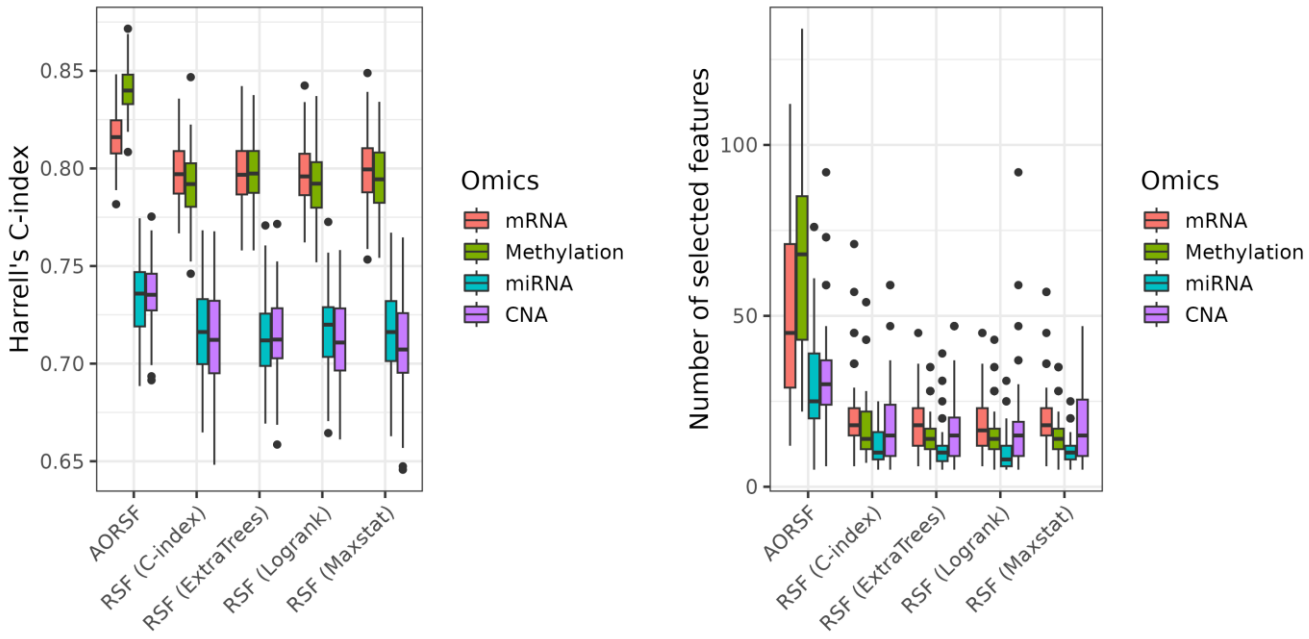

Figure 4: Discrimination performance (C-index, left) and number of selected features (right) per omic dataset and RSF model used in the eFS algorithm

We next assessed the robustness of the eFS algorithm, i.e., investigated how similar are the best feature subsets generated in each RFE run. We used the Nogueira similarity [26], which takes into account the number of features of each omic. This measure remains unaffected by randomly drawing independent feature sets of varying sizes and assessing their similarity (i.e., a chance-corrected similarity measure). We observed that all omics profiles have a Nogueira similarity score < 0.4, indicating a relatively poor agreement (best features are in general more different than are the same), suggesting that the aggregation of these sets provides a more robust consensus feature set per omic dataset (Figure 6).

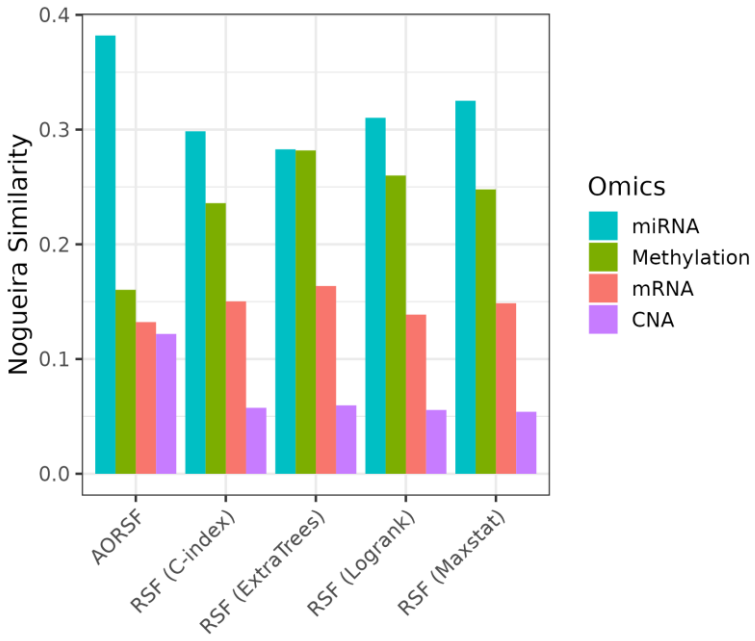

Figure 5: Stability assessment of the eFS algorithm

2.4. Multi-omics benchmarking

We have developed a novel methodology, implemented in the *survmob* R package, for benchmarking multiple survival ML models using a multi-modal dataset, such as the pancreatic TCGA PAAD patient cohort (Figure 6). The benchmarking process begins with a patient cohort containing clinical information and various data modalities, potentially high-dimensional (i.e., having more features than patients). Each data modality can encompass different aspects of the patient's biology, including **molecular factors, phenotypic characteristics (such as extracted features from radiology or pathological images)** or other relevant indicators. The cohort is divided into a training set (80% of the cohort) and a test set (20%) for model validation, ensuring that the proportion of censored patients remains consistent between the two sets. Next, the eFS algorithm is applied to each high-dimensional data modality in the training set, generating a ranked list of predictive features for each modality. The top-ranked features are then selected to subset the corresponding data modality in both the train and test sets. Continuing, these reduced datasets, along with the clinical data, serve as building blocks to create **all possible combinations of multi-modal datasets** by combining the selected features in various configurations. These generated datasets are used to tune the selected AI models, optimizing a specific survival performance metric on the training set. Finally, the trained models are evaluated on bootstrapped versions of the validation patient set to obtain a more robust estimate of their performance (variance estimation and confidence intervals) and allow for better decision making when comparing the different models or selecting the best-performing one. Lastly, the workflow enables further post-hoc processing of the benchmarking results using Bayesian methods.

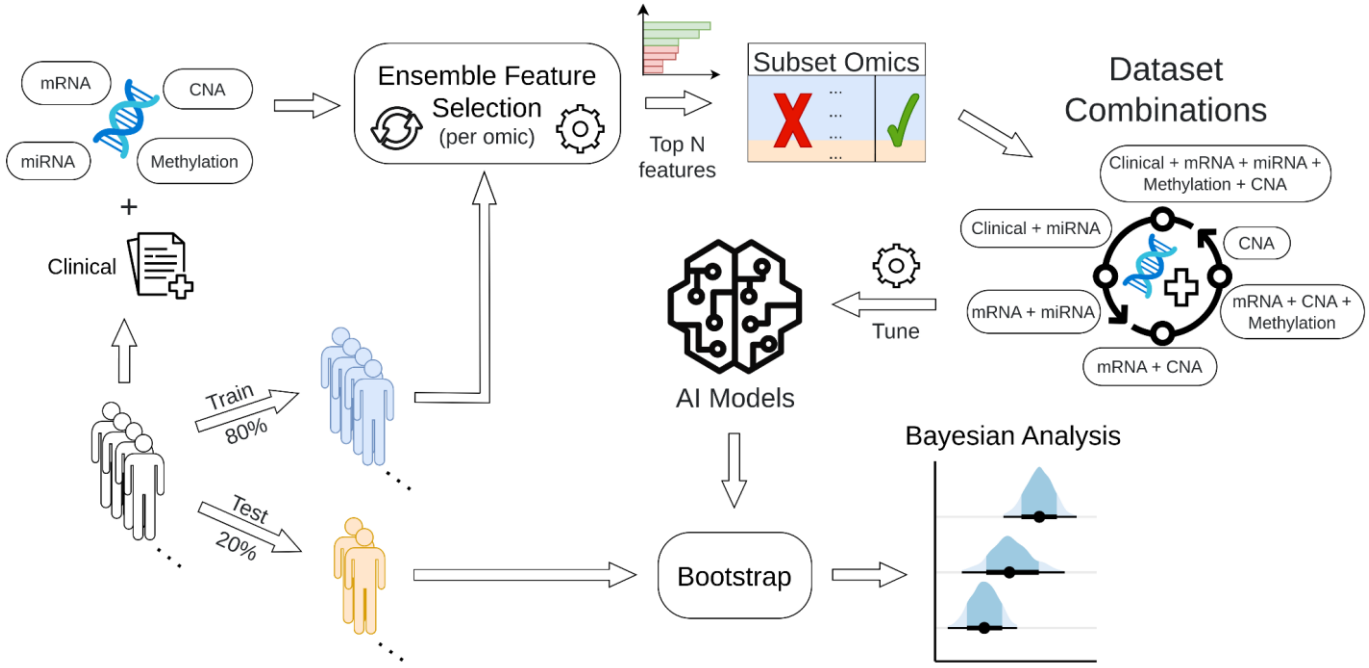

Figure 6: Multi-omics AI modeling and benchmarking workflow applied to TCGA's PAAD patient cohort

In the TCGA'S PAAD cohort, we split the patients into a training set (116 patients) and a test set (29 patients). The four high-dimensional omics (mRNA, CNA, miRNA and Methylation datasets) were used as input to the eFS algorithm and the top 20 features per omic were selected. A total of 31 multi-omics datasets were generated encompassing various combinations of omics and clinical data (5 modalities in total result in 2^5 combinations minus the empty features dataset). We used the AI models outlined in Table 2, along with predefined hyperparameter tuning ranges, both provided by the *survmob* R package. The primary clinical outcome used was patient survival, given that it was the only available outcome in the TCGA data (other clinical outcomes related to PANCAIM goals were not readily accessible). To tune the AI models, we employed 5-fold stratified cross-validation, ensuring a balanced representation of censored patients across the folds. The C-index metric was used for optimization, and Bayesian Optimization was performed with 100 evaluations to identify the optimal hyperparameter configuration for each model on a specific multi-modal dataset.

To evaluate the models' performance, we employed 100 bootstrapped resamplings on the test set (parallelized for faster execution). The survival measures used for **model evaluation** included the **C-index**, which assessed model discrimination performance, and the **Integrated Brier Score (IBS)**, which considers both discrimination and calibration performance (Table 3). Notably, we normalize the IBS scores of the ML models to those of a simple Kaplan-Meier (KM) model with no features. The resulting score is more **interpretable** as it represents the percentage increase in performance, also known as **explained residual variation or ERV**, relative to the baseline KM model. IBS-ERV values below 0 indicate models performing worse than Kaplan-Meier, while values above 0 indicate better-performing models (1 is a perfect score).
Our workflow addresses three key questions in the context of survival prediction:

- **Ranking of models**: It allows us to rank the tested survival ML models based on a chosen evaluation metric, providing insights into their relative performance from worst to best, independent of the chosen multi-modal dataset they were trained on.
- **Modality ranking**: It facilitates the ranking of different modalities (such as omic datasets and clinical data) to identify the most informative ones for patient survival prediction. This information guides the selection of modalities to be included in the training of a final predictive model.
- **Baseline model comparison**: It enables the comparison of the tested models with a baseline model that utilizes only clinical data. This comparison is performed on a separate validation set, allowing us to assess the added value of incorporating additional modalities into a predictive model.

2.5. Analysis of benchmarking results

To comprehensively evaluate the benchmarking results, address the aforementioned key questions, while accounting for the fact that the generated multi-modal datasets share common features and are therefore **interdependent** (every single data modality can be a part of many multi-modal datasets tested, see Figure 6), we adopted an alternative approach compared to conventional methods used to compare models across different datasets [27]. The benchmarking data were used to fit **linear mixed-effects Bayesian models** [12, 28, 29], from which we could derive posterior distributions of the performance scores (e.g., C-index) on the validation set, either per AI model or data modality used in the benchmark. These types of Bayesian models are also called random-effects or hierarchical models, because they offer the advantage of accounting for the correlation between datasets as well as the within-resample correlation when bootstrapping from a single validation set [30].

We employed separate Bayesian models to rank models and omics respectively, while considering the interdependence between the tested multi-modal datasets. For model ranking, the Bayesian analysis generated a posterior distribution for each AI model, representing the expected value of a specific metric (e.g., C-index). In the case of omics ranking, the Bayesian methods allowed us to draw a posterior distribution comparing the use of a specific data modality (omic dataset) in a multi-modal combination versus not including it at all. This comparison was made possible thanks to the design of our benchmarking workflow (Figure 6), which facilitated the assessment of each omic's contribution in a comprehensive manner.

2.6. Model ranking

For each survival ML model used, we summarized the respective posterior distribution using **credible intervals**. In the context of Bayesian statistics, a credible interval represents a range of values that are likely to contain the true parameter value (e.g., a performance metric such as Harrell's C-index, see Figure 7). It is a **measure of uncertainty** and provides a specific probability level, such as 95%, within which the parameter value is expected to fall. Figure 7 and Figure 8 (left) show the model ranking results using two different measures: Harrell's C-index and the IBS-ERV. We observed that the **rankings are highly measure-dependent,** e.g., the "RSF (Maxstat)" model has the best discrimination performance according to the C-index, while it scores worse than a simple Kaplan-Meier model (with no features, used as a reference model) when model calibration is also assessed using the IBS-ERV metric (Figure 8, left). In the case of the IBS-ERV metric, only the CoxBoost and AORSF models were able to lead to a combined discrimination and calibration performance better than the KM baseline model, considering that all possible multi-modal datasets from TCGA's PAAD cohort were tested (IBS-ERV > 0, Figure 8, left).

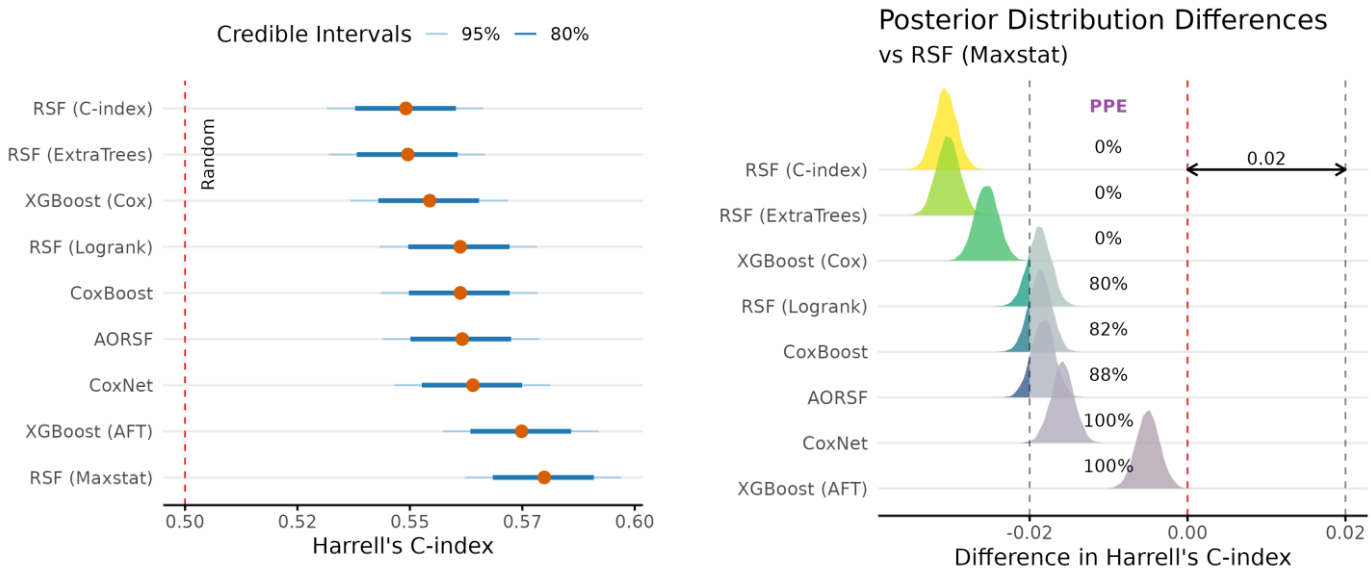

Figure 7: Model ranking according to the C-index metric
***Left**: Credible intervals for the discrimination performance (C-index).*
***Right**: Posterior distribution differences relative to the best performing model across all the possible multi-modal datasets combining different omics and clinical features from the PAAD TCGA cohort. A region of practical model equivalence (ROPE) of 2% is drawn and the Probabilities of Practical Equivalence (PPEs) between the different models are shown.*

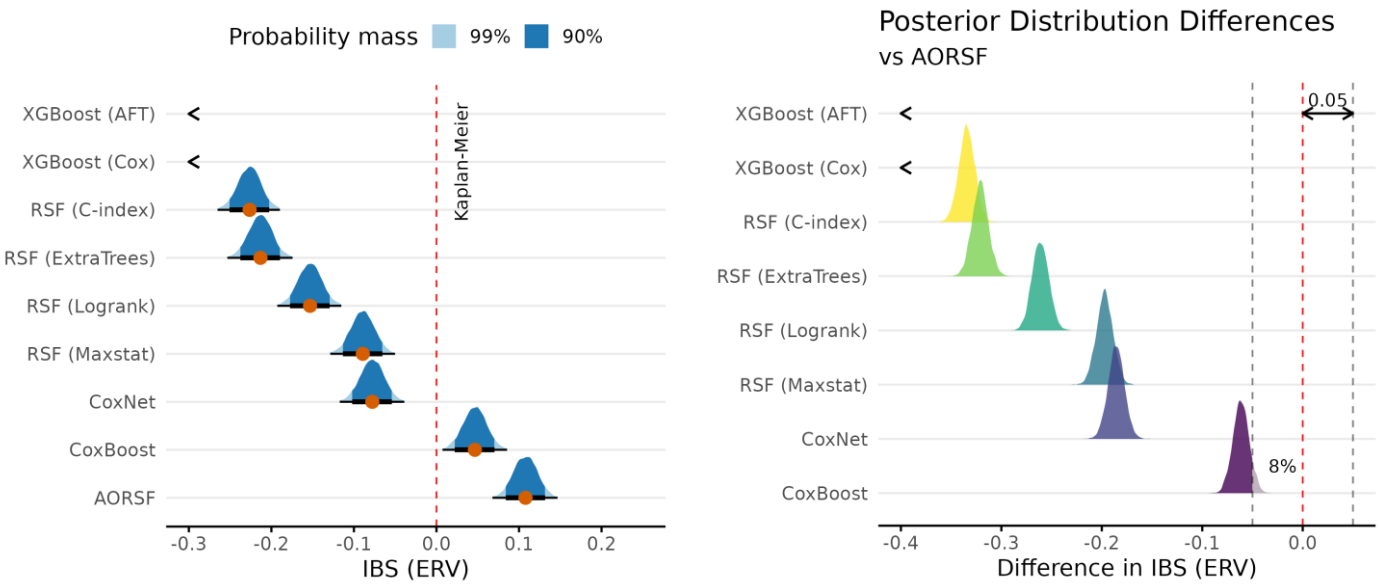

Figure 8: Model ranking according to the Integrated Brier Score (IBS).
***Left**: Posterior distributions and credible intervals for the IBS. **Right**: Posterior distribution differences relative to the best-performing model across all the possible multi-modal datasets combining different omics and clinical features from the PAAD TCGA cohort. A region of practical model equivalence (ROPE) of 5% is drawn and the Probability of Practical Equivalence (PPE) with the second best model is shown. '<' symbols are used to denote inferior performance scores, and as a result, the corresponding distributions are not visualized.*

The Bayesian approach described above offers a distinct advantage in addressing the question: "What is the probability that one model is practically equivalent to another?" To answer this question, we computed the difference between the posterior distributions of the AI models with respect to a chosen performance metric, relative to the best-performing model. To establish a practical threshold for equivalence, we define a **Region of Practical Equivalence (ROPE)** [31] encompassing a 2% difference for the C-index, where model differences falling within this range are considered practically indistinguishable. Based on the visualization of the posterior distribution differences and the ROPE area for the C-index (Figure 7, right), we observe that the top 5 models exhibit practically identical discrimination performance. In terms of the IBS-ERV metric (Figure 8, right), employing a larger 5% ROPE, the second-ranking model (CoxBoost) demonstrates only 8% equivalence relative to the best-performing model (AORSF). Consequently, based on these findings, we can confidently select **AORSF as the best-performing model** from our benchmark evaluation.

2.7. Omics ranking

Figure 9 and Figure 10 show the posterior distribution differences for the C-index and IBS-ERV performance metrics. These differences reflect the impact of including a specific data modality (omic dataset) in a multi-modal combination, regardless of the AI model employed. A larger distribution difference for a specific modality reflects its importance. We observed that out of the five different modalities used in our benchmark workflow, **only the clinical and mRNA features** have the potential to contribute significantly to a **positive practical difference** in both discrimination and calibration performance, as shown in Figure 9 and Figure 10 (C-index and IBS-ERV metric respectively). This finding also suggests that for the TCGA PAAD cohort, the most important data modalities for survival prediction remain the same regardless of the performance metric chosen.

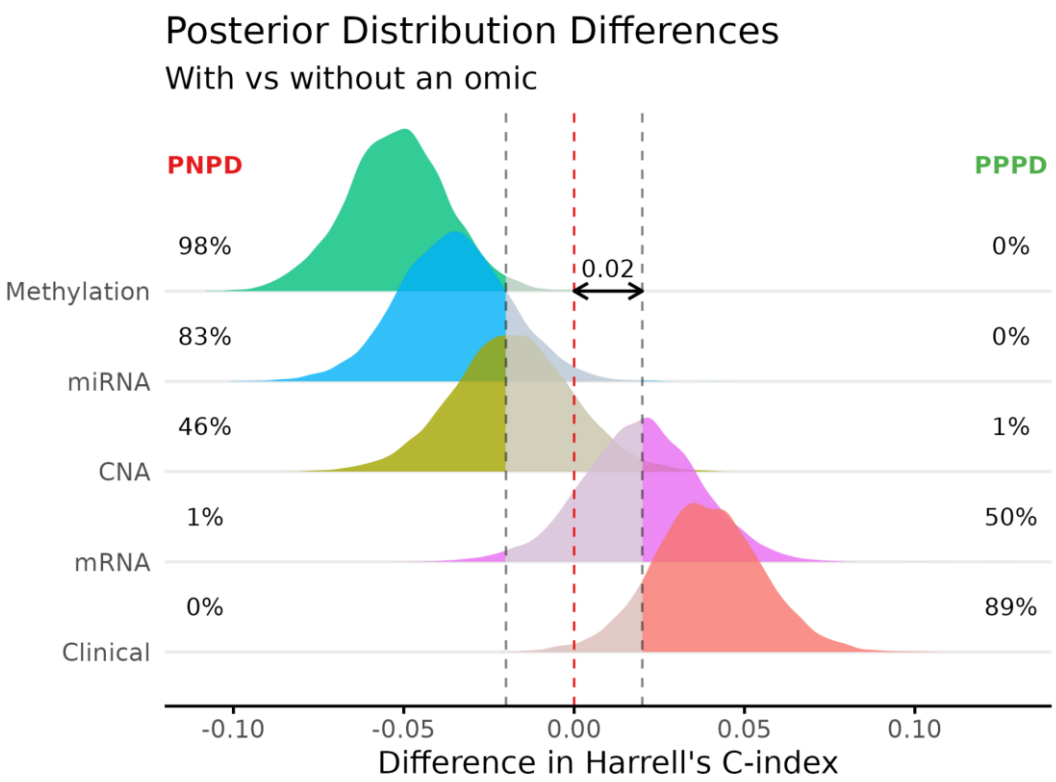

Figure 9: Ranking of omics by their contribution to a practical difference in discrimination performance (C-index) when incorporated into a multi-modal dataset.

Ranking results are aggregated across all the tested models via Bayesian methods. PPPD = Probability of Positive Practical Difference, PNPD = Probability of Negative Practical Difference. The size of the ROPE (Region of Practical Equivalence) is 2%.

Using a ROPE of 2%, we can assess the **probabilities of different omics contributing to a positive, negative, or no practical difference in performance**. When incorporating clinical features into a multi-modal dataset, there is a high probability (89% for C-index, 85% for IBS-ERV) of observing a 2% or greater increase in performance, regardless of the AI model used for training. In contrast, the use of mRNA data shows a lower probability (50% for C-index, 78% for IBS-ERV), indicating a smaller but still positive performance impact. On the other hand, the inclusion of other omics (Methylation, miRNA, and CNA) was associated with higher probabilities of negative differences in performance, suggesting that their incorporation may actually lead to a decline in performance in the validation set, as assessed by both the C-index and IBS-ERV metrics. These findings highlight the **consistent importance of clinical and mRNA features** and the varying impact of different omics modalities on survival prediction.

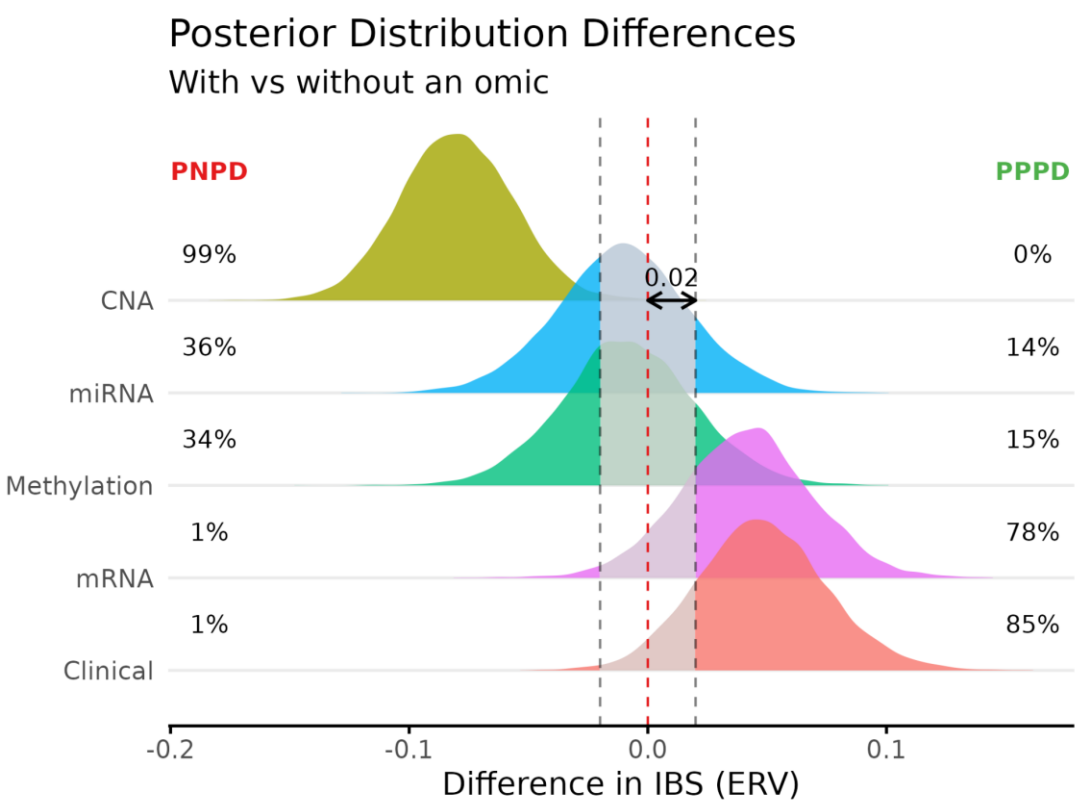

Figure 10: Ranking of omics by their contribution to a practical difference in both discrimination and calibration performance (Integrated Brier Score) when incorporated into a multi-modal dataset

Ranking results are aggregated across all the tested models via Bayesian methods. PPPD = Probability of Positive Practical Difference, PNPD = Probability of Negative Practical Difference. The size of the ROPE (Region of Practical Equivalence) is 2%.

2.8. Comparison with Baseline Cox model

We conducted a comprehensive comparison of the tested AI models with a baseline model that exclusively utilizes clinical data. Based on our benchmarking workflow (Figure 6), this comparison is conducted on a separate validation set, allowing us to address two important questions: Can an AI model trained on multi-modal data outperform a **baseline Cox Proportional Hazards (CoxPH) model** [3] **trained on clinical data alone**? And if so, which combinations of models and omics demonstrate superior performance? **Bootstrap confidence intervals** were employed to summarize and assess the empirical results on the validation set from TCGA's PAAD dataset, without the use of post-hoc Bayesian analysis. Results here refer to both the discrimination and calibration performance (measured by the C-index and IBS-ERV metrics) on bootstrapped test sets from the validation patient cohort.

The baseline CoxPH model stands out in terms of **discrimination performance**, as indicated by the C-index, surpassing almost all other combinations of omics, clinical data, and AI models (Figure 11). Interestingly, regardless of the AI model employed, the **exclusive use of clinical data consistently outperforms alternative multi-modal configurations**, highlighting its predictive efficacy in accurately ranking predicted patient risks in the TCGA PAAD cohort. These findings underscore the predictive power of clinical data in survival prediction tasks, which was expected, emphasizing its significance in achieving optimal discrimination performance.

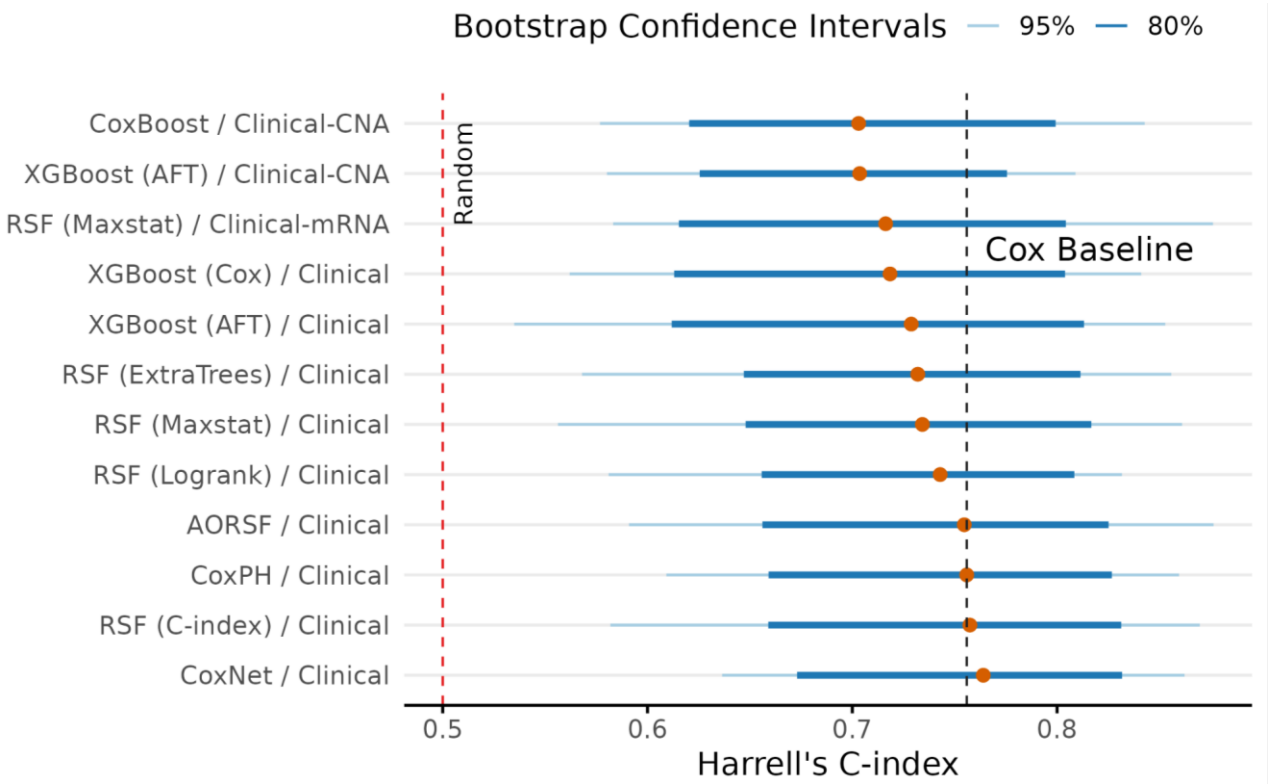

Figure 11: Bootstrap confidence intervals for assessing discrimination performance (C-index) on the test TCGA PAAD patient cohort. *The top 12 combinations of survival ML models / multi-modal datasets are shown on the y-axis. The "Cox Baseline" (dotted vertical line) indicates the median C-index test set performance using a Cox Proportional Hazards model trained on clinical features only (used as a baseline model here).*

On the other hand, when evaluating **both the calibration and discrimination performance** using the IBS-ERV metric (Figure 12), we observed that the CoxBoost and AORSF models showcased significantly improved performance compared to the baseline CoxPH model. This suggests that achieving such levels of overall predictive performance requires a combination of advanced ML models and a wider range of omics/modalities. Notably, among the top 12 combinations of models and multi-modal datasets, various modalities were included (all expect CNA, in accordance with the omics ranking for the IBS-ERV metric in Figure 10), highlighting the **need for a diverse set of data sources to achieve optimal overall performance**.

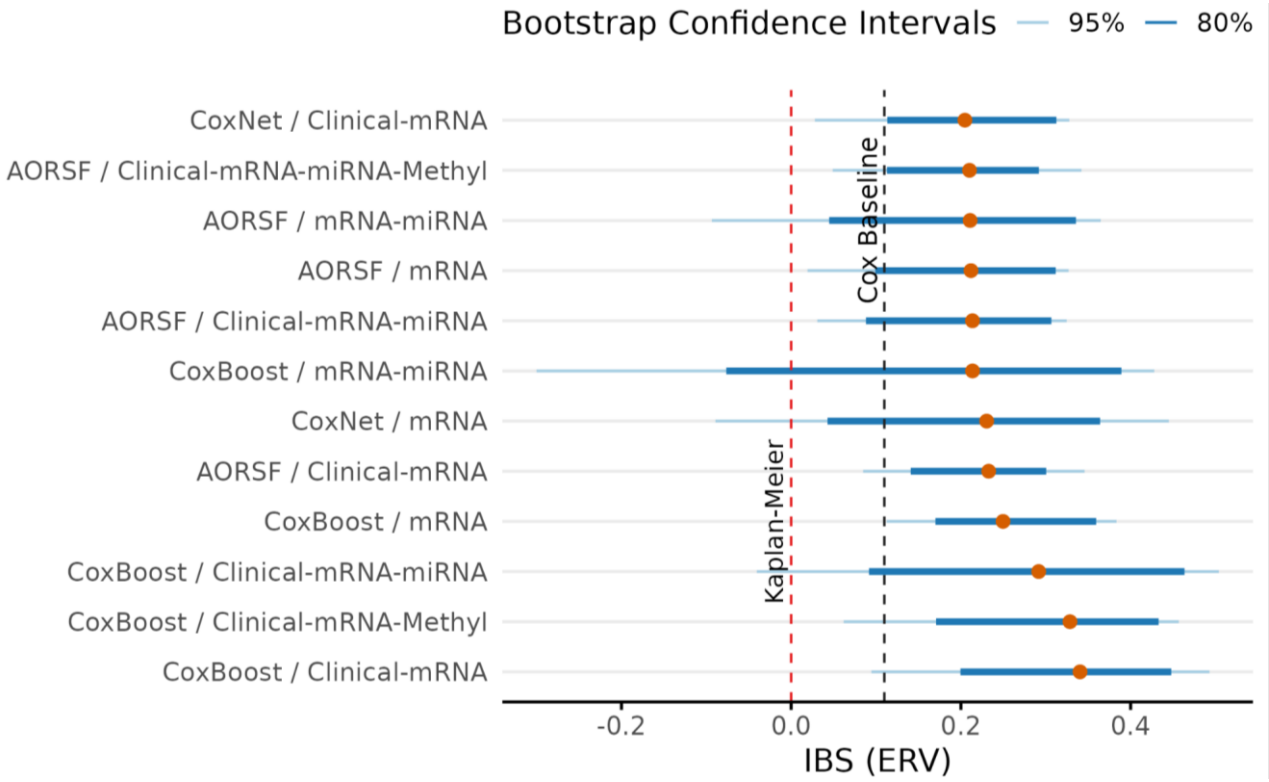

Figure 12: Bootstrap confidence intervals for assessing both discrimination and calibration performance (Integrated Brier Score) on the test TCGA PAAD patient cohort

The top 12 combinations of survival ML models / multi-modal datasets are shown on the y-axis. The "Cox Baseline" (dotted vertical line) indicates the median IBS test set performance using a Cox Proportional Hazards model trained on clinical features only (used as a baseline model here).

3. Conclusions

The report outlines the development of new methodologies for feature selection and AI model benchmarking on a multi-modal dataset. The methods developed are applicable to any data modality, regardless of the data source. All AI models, performance metrics, methodologies, source code, and generated results are openly accessible through online GitHub repositories (https://github.com/bblodfon/survmob/,https://github.com/bblodfon/tcga-survmob/). The TCGA pancreatic ductal adenocarcinoma (PAAD) dataset was used to develop, apply and test these methodologies, as the PANCAIM project multi-modal data was not accessible at the time. Four high-dimensional biological omic datasets were selected, while features extracted from radiology and pathological images were excluded due to insufficient number of patients with enough information across all data modalities. Moreover, due to the unavailability of other clinical outcomes aligned with PANCAIM goals, the primary clinical outcome under investigation was patient survival in this report.

In the context of developing clinically applicable AI models, our benchmarking results on the PAAD pancreatic cohort highlighted the importance of considering the specific clinical question at hand. For patient stratification purposes, utilizing clinical data alone proves to be sufficient in TCGA PAAD cohort, regardless of the AI model employed. However, when aiming for more accurate and well-calibrated survival predictions that can significantly impact clinical decision making and patient management in pancreatic cancer, the incorporation of multiple data modalities and the use of advanced ML models are essential. Among the models analyzed in the benchmark study, the Accelerated Oblique Random Survival Forests (AORSF) emerged as the recommended AI model, achieving optimal overall discrimination and calibration performance. Furthermore, the incorporation of optimally selected mRNA features, in addition to clinical data, resulted in further improvement in the overall predictive capability of the tested AI models.

To further enhance the effectiveness of our framework, future applications on larger and higher-quality multi-modal and multi-cohort datasets, such as those available in PANCAIM, will be invaluable in driving improvements in pancreatic cancer health care and facilitating more comprehensive analyses.

4. Degree of Progress

100%

5. Dissemination Level

The Deliverable 4.1 is public.

6. References

[1] Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), 1903. https://doi.org/10.21105/JOSS.01903

[2] Sonabend, R., Király, F. J., Bender, A., Bischl, B., & Lang, M. (2021). mlr3proba: an R package for machine learning in survival analysis. *Bioinformatics*, *37*(17), 2789–2791. https://doi.org/10.1093/BIOINFORMATICS/BTAB039

[3] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://www.jstor.org/stable/2985181

[4] Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, *39*(5), 1–13. https://doi.org/10.18637/JSS.V039.I05

[5] Binder, H., & Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, *9*(1), 1–10. https://doi.org/10.1186/1471-2105-9-14

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[7] Barnwal, A., Cho, H., & Hocking, T. (2022). Survival Regression with Accelerated Failure Time Model in XGBoost. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2022.2067548

[8] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). *Random survival forests*. *2*(3), 841–860. https://doi.org/10.1214/08-AOAS169

[9] Schmid, M., Wright, M. N., & Ziegler, A. (2016). On the use of Harrell's C for clinical risk prediction via random survival forests. *Expert Systems with Applications*, *63*, 450–459. https://doi.org/10.1016/J.ESWA.2016.07.018

[10] Wright, M. N., Dankowski, T., & Ziegler, A. (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in Medicine*, *36*(8), 1272–1284. https://doi.org/10.1002/sim.7212

[11] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning 2006 63:1*, *63*(1), 3–42. https://doi.org/10.1007/S10994-006-6226-1

[12] Jaeger, B. C., Welden, S., Lenoir, K., Speiser, J. L., Segar, M. W., Pandey, A., & Pajewski, N. M. (2022). *Accelerated and interpretable oblique random survival forests*. https://doi.org/10.48550/arxiv.2208.01129

[13] Heagerty, P. J., & Zheng, Y. (2005). Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, *61*(1), 92–105. https://doi.org/10.1111/J.0006-341X.2005.030814.X

[14] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the Yield of Medical Tests. *JAMA: The Journal of the American Medical Association*, *247*(18), 2543–2546. https://doi.org/10.1001/jama.1982.03320430047030

[15] Haider, H., Hoehn, B., Davis, S., & Greiner, R. (2020). Effective Ways to Build and Evaluate Individual Survival Distributions. *Journal of Machine Learning Research*, *21*(85), 1–63. http://jmlr.org/papers/v21/18-772.html

[16] Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, *18*, 2529–2545. https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18%3C2529::aid-sim274%3E3.0.co;2-5

[17] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, *21*(1), 128–138. https://doi.org/10.1097/EDE.0B013E3181C30FB2
[18] Zhao, Z., Zobolas, J., Zucknick, M., & Aittokallio, T. (2023). *Tutorial on survival modelling with omics data* (pp. 1–13). https://doi.org/10.48550/arxiv.2302.12542
[19] Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*(10), 281–305. http://jmlr.org/papers/v13/bergstra12a.html
[20] Turner, R., Eriksson, D., Mccourt, M., Kiili, J., Com Valohai, J., Laaksonen, E., Xu, V. Z., & Guyon, I. (2021). *Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020*. https://doi.org/10.48550/arxiv.2104.10201
[21] Ramos, M., Geistlinger, L., Oh, S., Schiffer, L., Azhar, R., Kodali, H., de Bruijn, I., Gao, J., Carey, V. J., Morgan, M., & Waldron, L. (2020). Multiomic Integration of Public Oncology Databases in Bioconductor. *JCO Clinical Cancer Informatics*, *4*, 958–971. https://doi.org/10.1200/cci.19.00119
[22] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. https://doi.org/10.1093/BIOINFORMATICS/BTM344
[23] Awada, W., Khoshgoftaar, T. M., Dittman, D., Wald, R., & Napolitano, A. (2012). A review of the stability of feature selection techniques for bioinformatics data. *Proceedings of the 2012 IEEE 13th International Conference on Information Reuse and Integration, IRI 2012*, 356–363. https://doi.org/10.1109/IRI.2012.6303031
[24] Saeys, Y., Abeel, T., & Van De Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. *Machine Learning and Knowledge Discovery in Databases*, *5212 LNAI*, 313–325. https://doi.org/10.1007/978-3-540-87481-2_21
[25] Pes, B. (2020). Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. *Neural Computing and Applications*, *32*(10), 5951–5973. https://doi.org/10.1007/s00521-019-04082-3
[26] Nogueira, S., Sechidis, K., & Brown, G. (2018). On the Stability of Feature Selection Algorithms. *Journal of Machine Learning Research*, *18*(174), 1–54. http://jmlr.org/papers/v18/17-514.html
[27] Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*, 1–30. https://doi.org/10.5555/1248547.1248548
[28] Benavoli, A., Corani, G., Demšar, J., & Zaffalon, M. (2017). Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis. *Journal of Machine Learning Research*, *18*(77), 1–36. http://jmlr.org/papers/v18/16-305.html
[29] Goodrich et al. (2023). *rstanarm: Bayesian applied regression modeling via Stan*. R package version 2.21.4 https://mc-stan.org/rstanarm
[30] Kuhn (2022). *tidyposterior: Bayesian Analysis to Compare Models using Resampling Statistics*. R package version 1.0.0 https://tidyposterior.tidymodels.org
[31] Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, JAGS, and stan*. Academic Press.