Deep learning-based feature importance ranking for DNA methylation data in breast cancer risk stratification

Advisor: Prof. **Francesca Ieva**
Co-advisor: Dott. **Michela Carlotta Massi**

Author: **Lorenzo Dominoni**

Mathematical Engineering- Academic year 2020-2021

Background

DNA methylation

- Epigenetic mechanism at CpG sites
- Alterations are known to be correlated with cancer

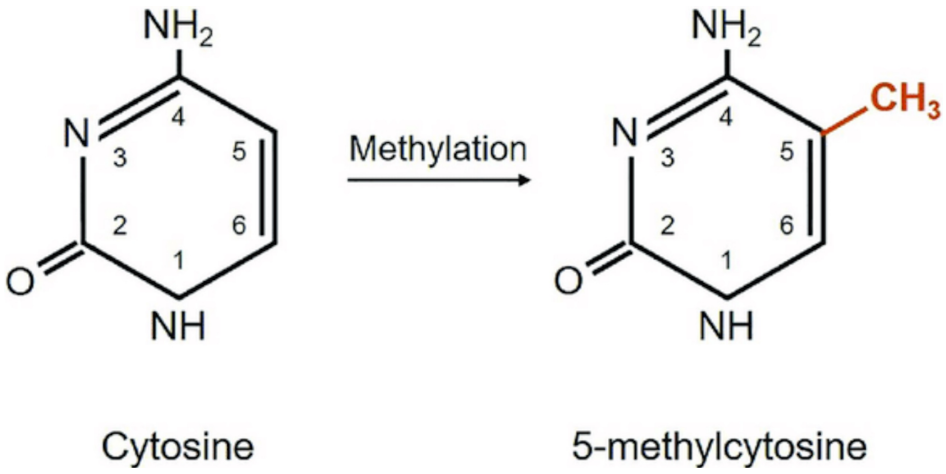


Fig. 1: DNA methylation mechanism

Background

DNA methylation

- Epigenetic mechanism at CpG sites
- Alterations are known to be correlated with cancer

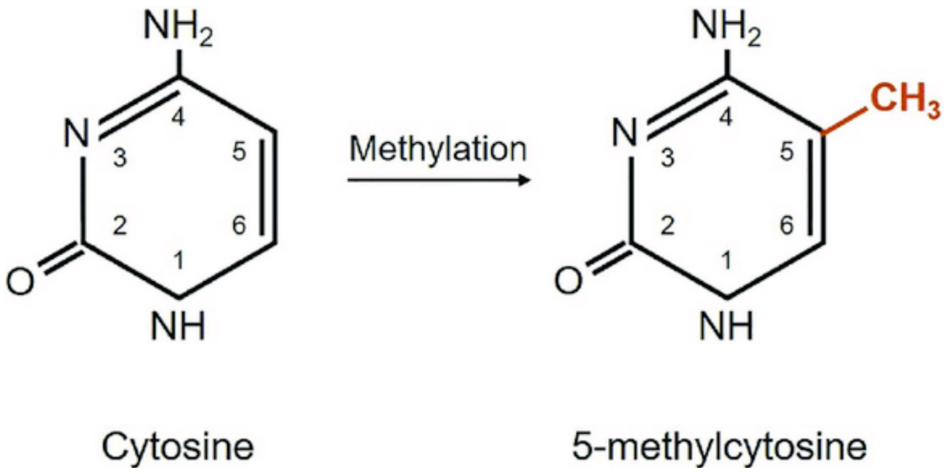


Fig. 1: DNA methylation mechanism

Breast cancer

- Most common cancer worldwide
- Early diagnosis often leads to better prognosis

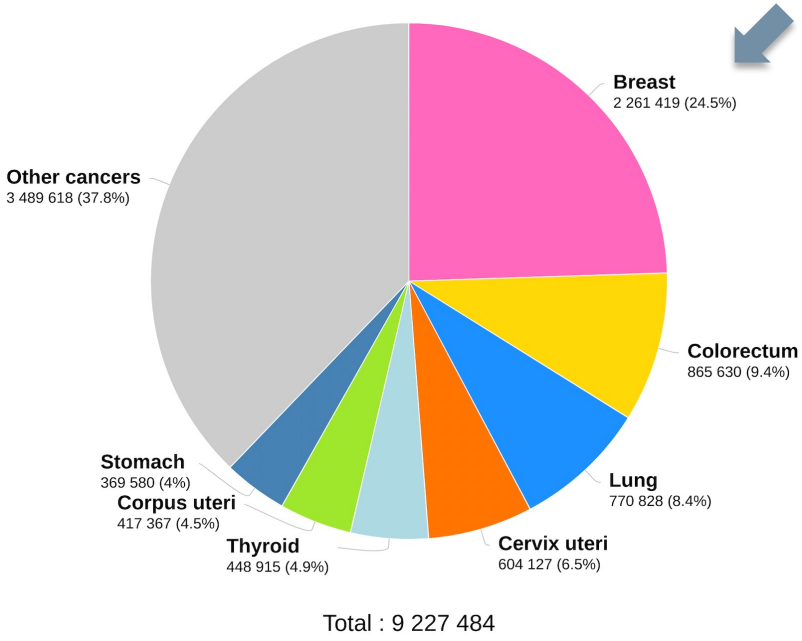


Fig. 2: Number of cancer cases among females in 2020

Objective

Goal

Discovery of the most important methylation features for detecting breast cancer risk several years before the diagnosis

Objective

Goal

Discovery of the most important methylation features for detecting breast cancer risk several years before the diagnosis

Proposal

Production of a feature importance ranking of methylation features according to their relevance in a model predicting breast cancer risk

Objective

Goal

Discovery of the most important methylation features for detecting breast cancer risk several years before the diagnosis

Proposal

Production of a feature importance ranking of methylation features according to their relevance in a model predicting breast cancer risk

Individual regions

Objective

Goal

Discovery of the most important methylation features for detecting breast cancer risk several years before the diagnosis

Proposal

Production of a feature importance ranking of methylation features according to their relevance in a model predicting breast cancer risk

Individual regions

Survival model

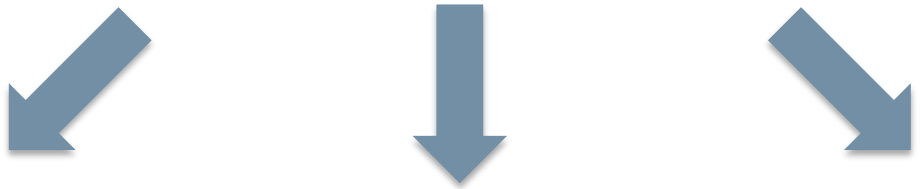
Objective

Goal

Discovery of the most important methylation features for detecting breast cancer risk several years before the diagnosis

Proposal

Production of a feature importance ranking of methylation features according to their relevance in a model predicting breast cancer risk



Individual regions | Survival model | Cases-controls use

Challenges

High dimensionality

Challenges

High dimensionality

Non-linear interactions

Challenges

High dimensionality

Non-linear interactions

Interpretability

Challenges

High dimensionality

Non-linear interactions

Methodology based on a dimensionality reduction technique, followed by a regularized deep survival pre-trained model, and then an adequate feature importance algorithm

Interpretability

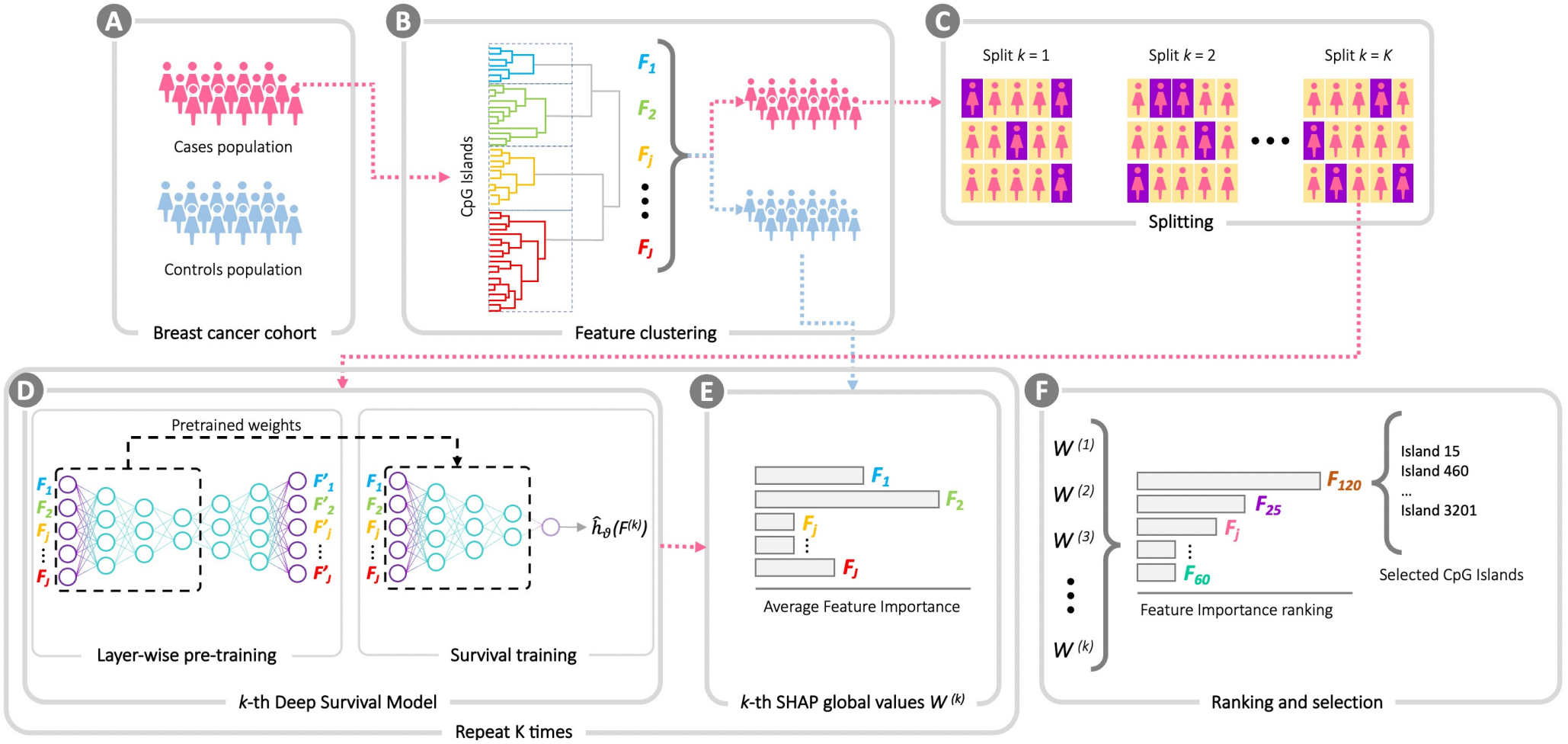
EPIC dataset (Riboli et al., 2002)

- Prospective case study with methylation data from blood samples
- Both cancer cases and matched controls
- Subjects characterized by time to diagnosis, cohort and beta values
- Group CpG sites in CpG islands

Patient	Time to disease	Study	cg03725447	cg25215298	cg03256938	cg18297246
200109360008_R01C01	6.986995	Breast	0.192584	0.277721	0.881829	0.449754
7766130100_R06C01	NaN	Breast	0.084782	0.343562	0.848235	0.481278
6042316165_R01C01	8.966461	Colon	0.216812	0.320498	0.898258	0.450233
7668610146_R06C01	NaN	Colon	0.178374	0.311865	0.900772	0.412366
3999875083_R05C02	16.169747	Lung	0.180652	0.324830	0.866643	0.533926
3999875048_R03C02	NaN	Lung	0.210455	0.318499	0.864264	0.415800

Fig. 3: Example of observations in the EPIC dataset

Methodology pipeline



Feature clustering

Biological information is shared among the CpG islands

Feature clustering used to reduce dimensionality

Feature clustering

Biological information is shared among the CpG islands

Feature clustering used to reduce dimensionality

- It works as hierarchical clustering but grouping features instead of samples
- Using the Euclidean distance and the Ward linkage

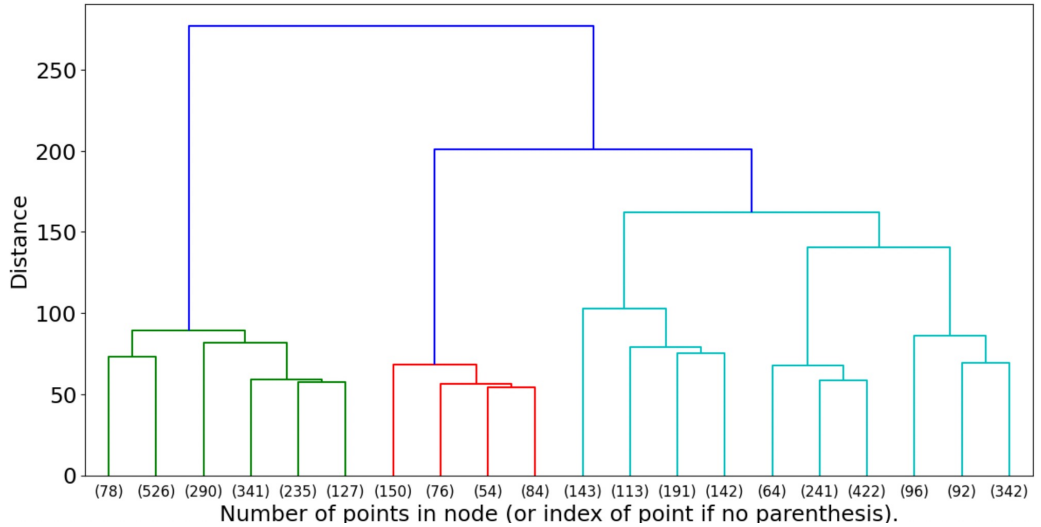


Fig. 4: Feature clustering dendrogram

Deep survival modelling

Survival training

Inspired by DeepSurv (Katzman et al., 2018)

$$\mathcal{L}(\theta) = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} (\hat{h}_\theta(\mathbf{X}_i) - \log(\sum_{j \in \mathcal{R}(T_i)} e^{\hat{h}_\theta(\mathbf{X}_j)})) + \lambda \|\theta\|_2^2$$

Models interactions

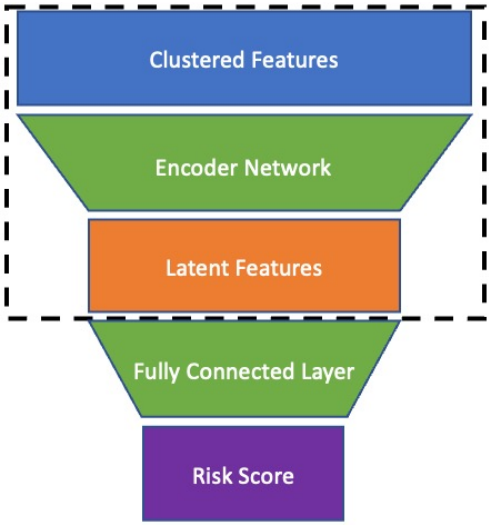


Fig. 6: Deep survival model

Deep survival modelling

Autoencoder pre-training

Layer-wise procedure

$$\mathcal{L}(\mathbf{x}, \mathbf{x}') = \frac{\sum_{i=1}^{n} |x_i - x_i'|}{n}$$

Regularization and generalization

Survival training

Inspired by DeepSurv (Katzman et al., 2018)

$$\mathcal{L}(\theta) = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} (\hat{h}_\theta(\mathbf{X}_i) - \log(\sum_{j \in \mathcal{R}(T_i)} e^{\hat{h}_\theta(\mathbf{X}_j)})) + \lambda \|\theta\|_2^2$$

Models interactions

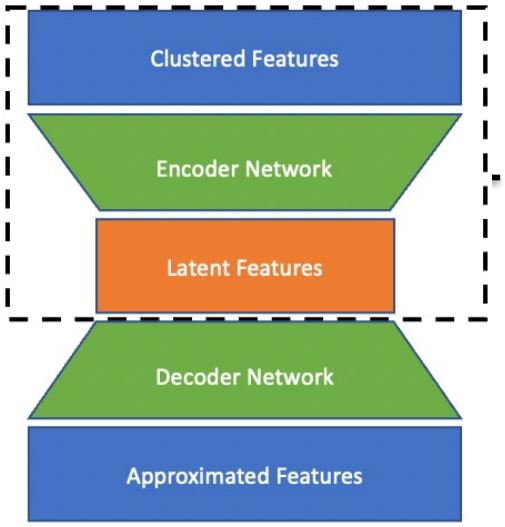


Fig. 5: Autoencoder model

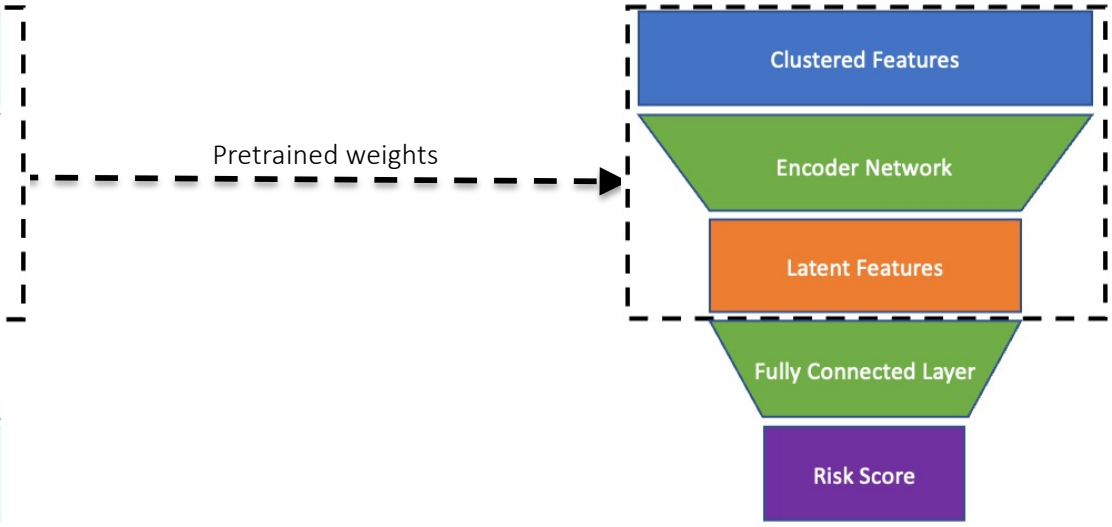


Fig. 6: Deep survival model

Feature importance ranking

Kernel SHAP (Lundberg et al., 2017)

$$\phi_j = \sum_{S \subseteq \{x_1,\dots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} (v_{\mathbf{x}}(S \cup \{x_j\}) - v_{\mathbf{x}}(S))$$

- Deals with interacting features
- Uses a reference dataset

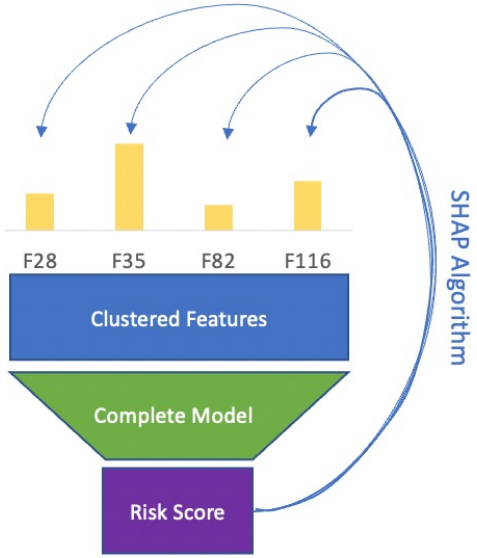


Fig. 7: SHAP values computation

Feature importance ranking

Kernel SHAP (Lundberg et al., 2017)

$$\phi_j = \sum_{S \subseteq \{x_1,\dots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} (v_{\mathbf{x}}(S \cup \{x_j\}) - v_{\mathbf{x}}(S))$$

- Deals with interacting features
- Uses a reference dataset

Fig. 7: SHAP values computation

Averaging, ordering and selection

- Exploit an ensemble approach
- Sort features according to their importance
- Select the first features in the ranking

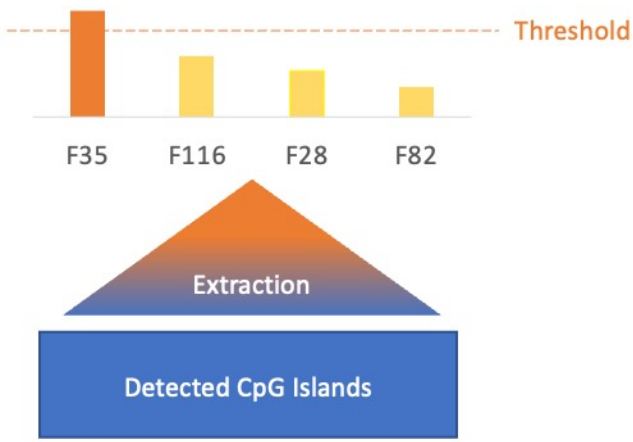
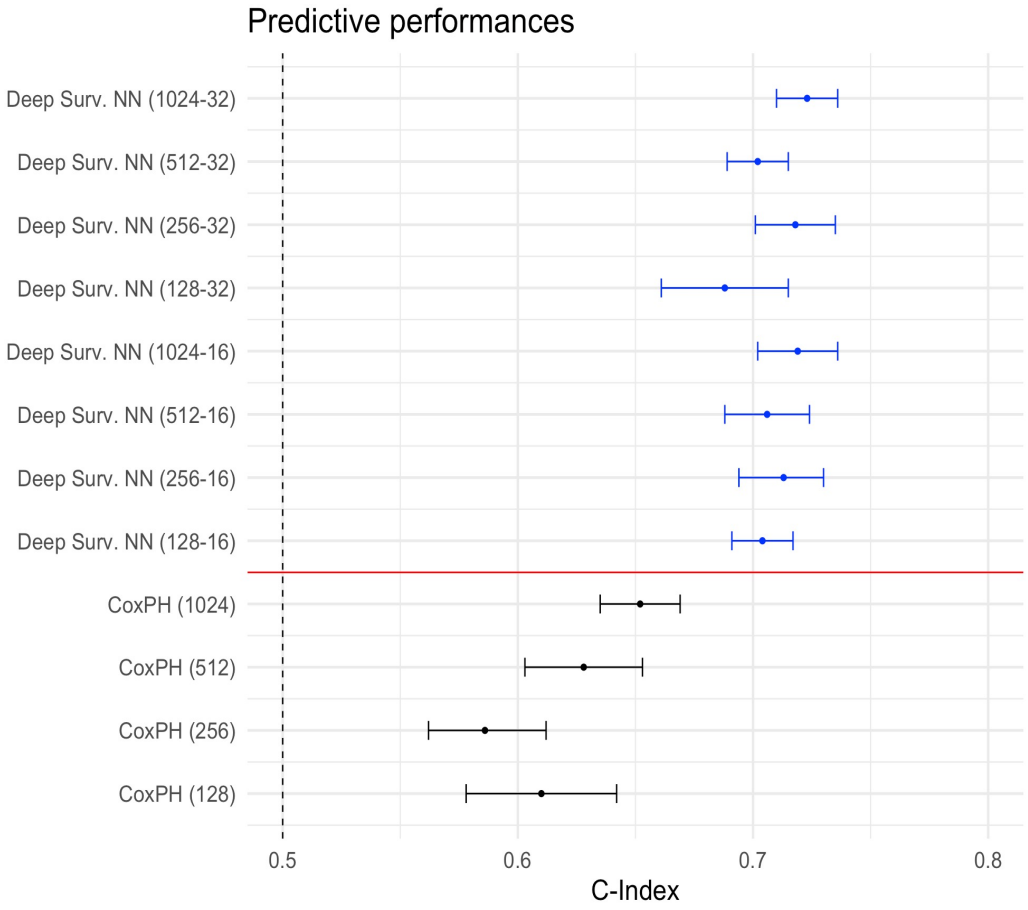
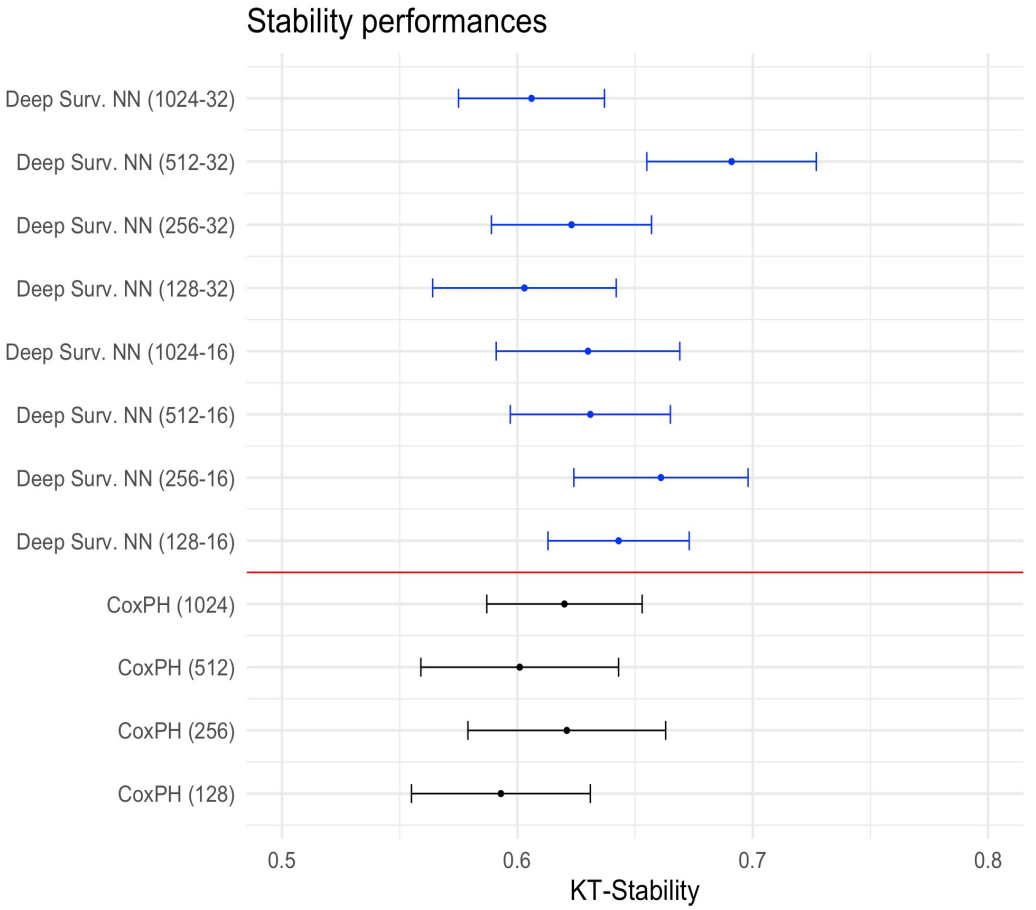


Fig. 8: Selection of the most important CpG islands

Performances of the methods





Ranking and selection

- Feature 120, composed of 20 CpG islands, selected as ranked first

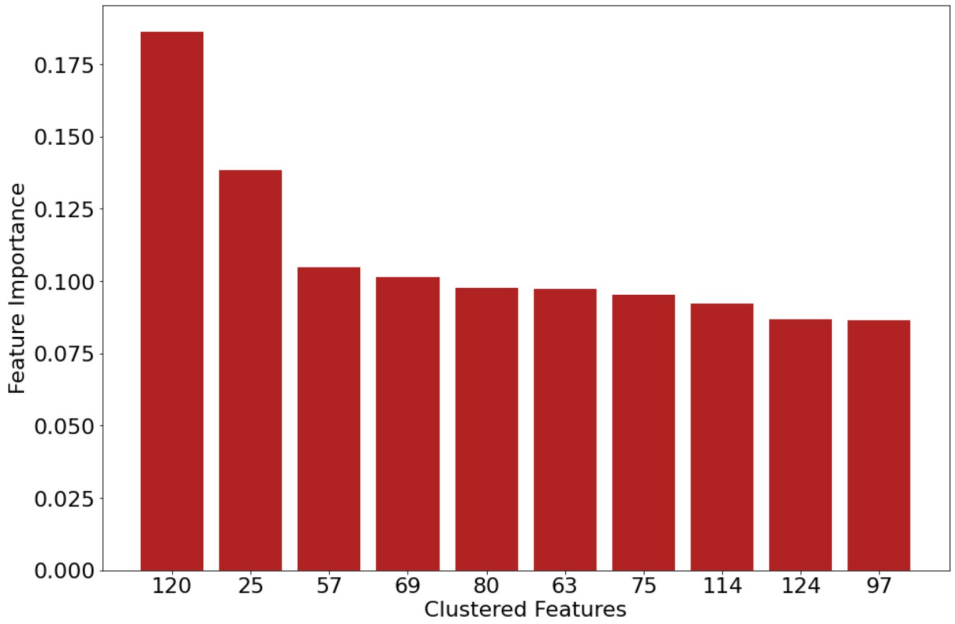


Fig. 9: Importance of the first 10 ranked features

Ranking and selection

- Feature 120, composed of 20 CpG islands, selected as ranked first
- Clear decreasing trend of methylation with respect to the time to diagnosis

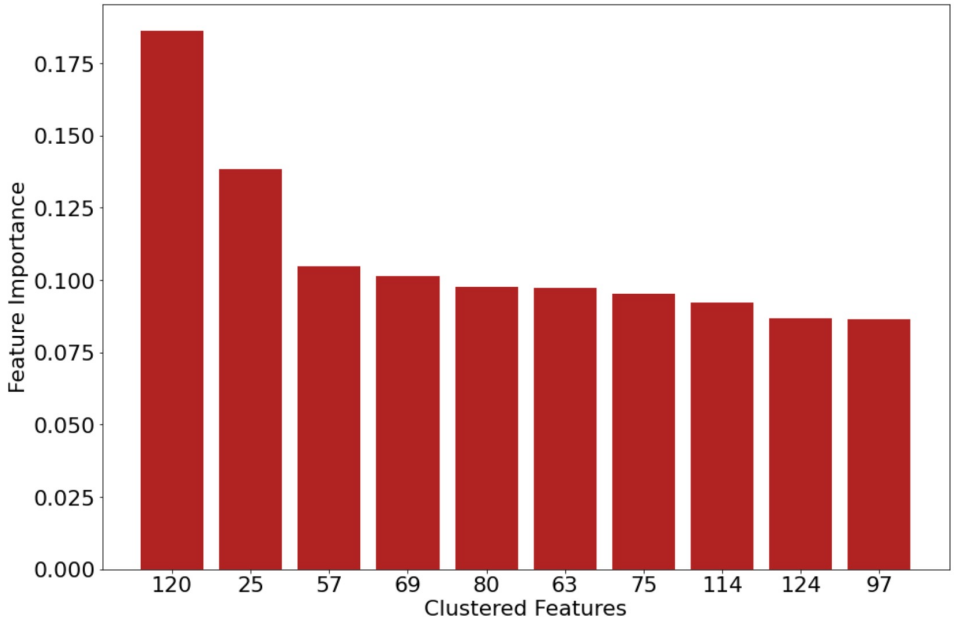


Fig. 9: Importance of the first 10 ranked features

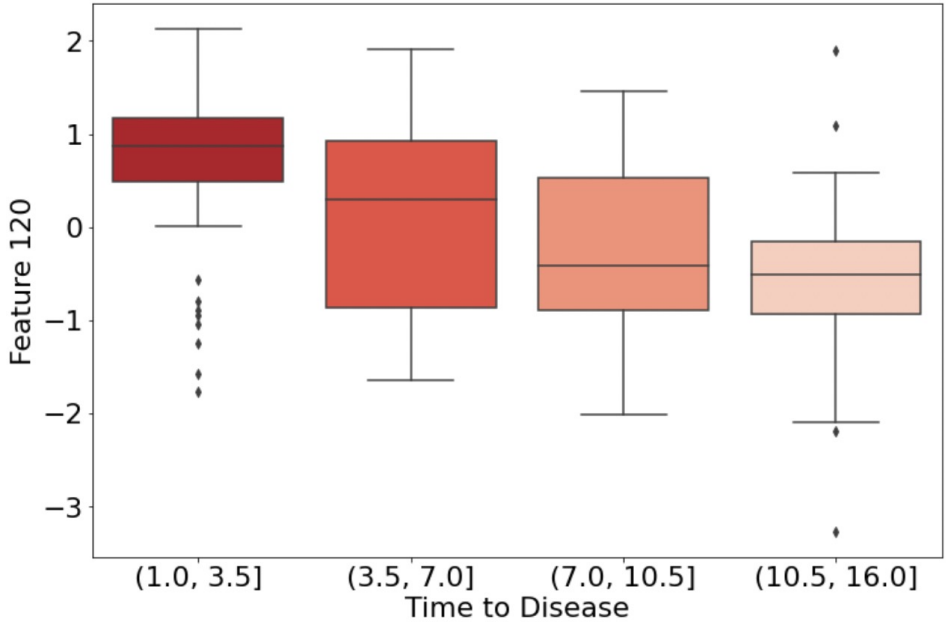


Fig. 10: Boxplot of Feature 120 divided by time categories

Post-hoc analysis

- Significant difference in the survival profiles of low and high methylated subjects
- High methylation is a risk factor

CpG island	Log-rank p-value	Hazard ratio [95% interval]
1:90945518-90945656	3.88e-08	2.03 [1.57, 2.63]
1:158090642-158091676	1.26e-03	1.51 [1.17, 1.94]
2:100086548-100088317	3.36e-04	1.58 [1.23, 2.03]
4:149584089-149584799	3.18e-07	1.93 [1.49, 2.49]
6:1570179-1570756	5.62e-07	1.89 [1.47, 2.43]
6:43530362-43531683	1.16e-05	1.75 [1.36, 2.25]
6:166137998-166138866	3.77e-03	1.45 [1.12, 1.86]
8:21701267-21701566	1.85e-08	2.05 [1.59, 2.65]
8:145119282-145120028	5.20e-06	1.82 [1.40, 2.36]
9:34618796-34619343	2.30e-12	2.49 [1.92, 3.24]
10:102493904-102494072	3.66e-03	1.45 [1.13, 1.87]
10:119294070-119294143	3.04e-04	1.58 [1.23, 2.04]
14:87862626-87863008	7.93e-05	1.65 [1.28, 2.12]
16:85096322-85097146	2.39e-07	1.96 [1.51, 2.53]
18:75811758-75814395	2.23e-08	2.04 [1.58, 2.63]
19:1704275-1706659	2.51e-11	2.35 [1.82, 3.05]
19:13070446-13070515	1.78e-02	1.36 [1.05, 1.75]
20:21438169-21438255	4.91e-07	1.90 [1.47, 2.44]
20:21449303-21449404	1.04e-02	1.39 [1.08, 1.78]
22:37180713-37182260	3.10e-09	2.17 [1.67, 2.82]

Fig. 11: Comparison of subjects with different methylation levels

Post-hoc analysis

- Significant difference in the survival profiles of low and high methylated subjects
- High methylation is a risk factor

CpG island	Log-rank p-value	Hazard ratio [95% interval]
1:90945518-90945656	3.88e-08	2.03 [1.57, 2.63]
1:158090642-158091676	1.26e-03	1.51 [1.17, 1.94]
2:100086548-100088317	3.36e-04	1.58 [1.23, 2.03]
4:149584089-149584799	3.18e-07	1.93 [1.49, 2.49]
6:1570179-1570756	5.62e-07	1.89 [1.47, 2.43]
6:43530362-43531683	1.16e-05	1.75 [1.36, 2.25]
6:166137998-166138866	3.77e-03	1.45 [1.12, 1.86]
8:21701267-21701566	1.85e-08	2.05 [1.59, 2.65]
8:145119282-145120028	5.20e-06	1.82 [1.40, 2.36]
9:34618796-34619343	2.30e-12	2.49 [1.92, 3.24]
10:102493904-102494072	3.66e-03	1.45 [1.13, 1.87]
10:119294070-119294143	3.04e-04	1.58 [1.23, 2.04]
14:87862626-87863008	7.93e-05	1.65 [1.28, 2.12]
16:85096322-85097146	2.39e-07	1.96 [1.51, 2.53]
18:75811758-75814395	2.23e-08	2.04 [1.58, 2.63]
19:1704275-1706659	2.51e-11	2.35 [1.82, 3.05]
19:13070446-13070515	1.78e-02	1.36 [1.05, 1.75]
20:21438169-21438255	4.91e-07	1.90 [1.47, 2.44]
20:21449303-21449404	1.04e-02	1.39 [1.08, 1.78]
22:37180713-37182260	3.10e-09	2.17 [1.67, 2.82]

Fig. 11: Comparison of subjects with different methylation levels

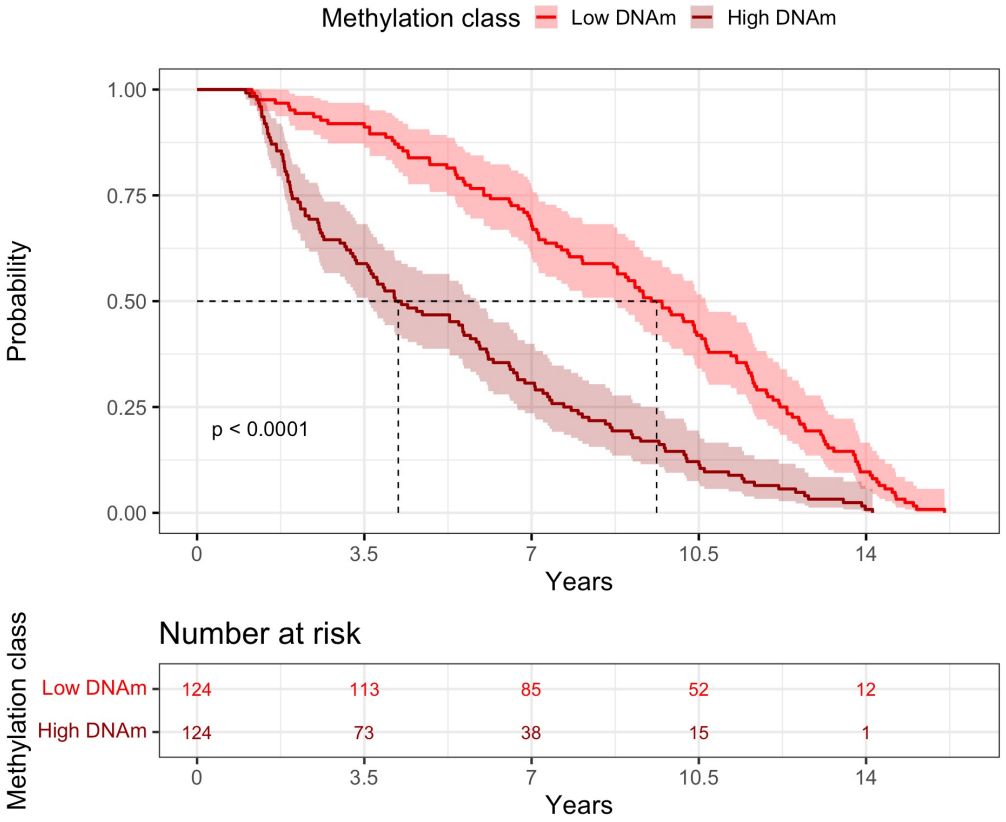


Fig. 12: Kaplan-Meier curves of island 9:34618796-34619343

Weighted Kolmogorov-Smirnoff test (Charmpi et al., 2015)

Obtains meaningful biological results and validates the methodology

Weighted Kolmogorov-Smirnoff test (Charmpi et al., 2015)

Obtains meaningful biological results and validates the methodology

All 8 pathways are correlated with cancer, 5 specifically with breast cancer

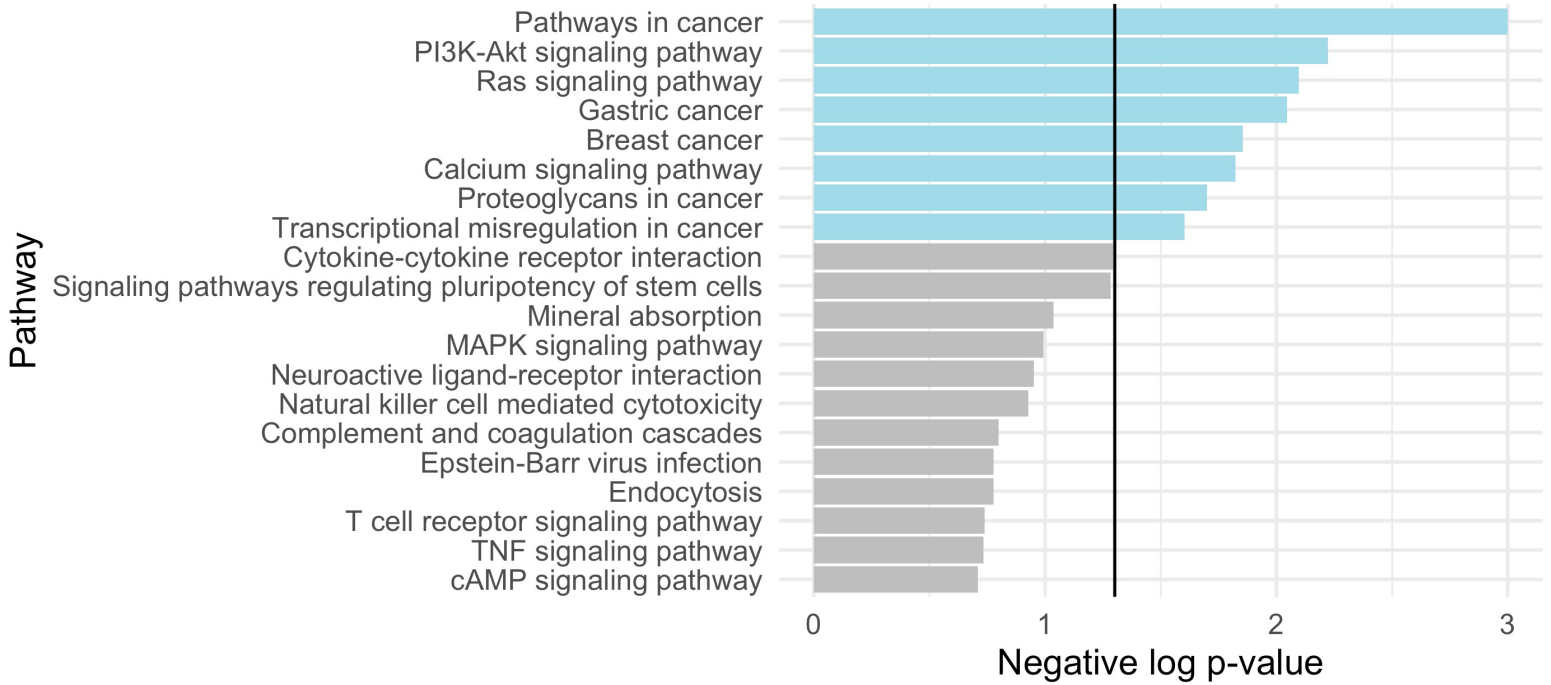


Fig. 13: Barplot of the significant pathways

Conclusions

- The novel methodology managed the motivating problems, outperforming the Cox model

Conclusions

- The novel methodology managed the motivating problems, outperforming the Cox model
- The selected CpG islands' methylation values are correlated with the time to diagnosis

Conclusions

- The novel methodology managed the motivating problems, outperforming the Cox model
- The selected CpG islands' methylation values are correlated with the time to diagnosis
- Highly methylated subjects have a higher breast cancer risk

Conclusions

- The novel methodology managed the motivating problems, outperforming the Cox model
- The selected CpG islands' methylation values are correlated with the time to diagnosis
- Highly methylated subjects have a higher breast cancer risk
- The biological results are coherent with the literature and validate the methodology

Conclusions

- The novel methodology managed the motivating problems, outperforming the Cox model
- The selected CpG islands' methylation values are correlated with the time to diagnosis
- Highly methylated subjects have a higher breast cancer risk
- The biological results are coherent with the literature and validate the methodology

To improve on some aspects of the methodology it is possible to:

- Further validate the proposal on external datasets

Conclusions

- The novel methodology managed the motivating problems, outperforming the Cox model
- The selected CpG islands' methylation values are correlated with the time to diagnosis
- Highly methylated subjects have a higher breast cancer risk
- The biological results are coherent with the literature and validate the methodology

To improve on some aspects of the methodology it is possible to:

- Further validate the proposal on external datasets
- Employ variational and adversarial autoencoders

Conclusions

- The novel methodology managed the motivating problems, outperforming the Cox model
- The selected CpG islands' methylation values are correlated with the time to diagnosis
- Highly methylated subjects have a higher breast cancer risk
- The biological results are coherent with the literature and validate the methodology

To improve on some aspects of the methodology it is possible to:

- Further validate the proposal on external datasets
- Employ variational and adversarial autoencoders
- Apply the Deep SHAP algorithm

Bibliography

- M. Massi, G. Fiorito, L. Dominoni and F. Ieva. 'A Deep Survival EWAS approach to blood-based DNA methylation global effect profile estimation for breast cancer'. *In progress,* 2021.
- A. Gagliardi, P. Dugué, T. Nøst, M. Southey, D. Buchanan, D. Schmidt, E. Makalic, A. Hodge, D. English, N. Doo, et al. 'Stochastic epigenetic mutations are associated with risk of breast cancer, lung cancer, and mature b-cell neoplasms.' *Cancer Epidemiology and Prevention Biomarkers*, 29(10):2026–2037, 2020.
- E. Riboli, K. Hunt, N. Slimani, P. Ferrari, T. Norat, M. Fahey, U. Charrondiere, B. Hemon, C. Casagrande, J. Vignat, et al. 'European prospective investigation into cancer and nutrition (epic): study populations and data collection.' *Public health nutrition,* 5(6b):1113–1124, 2002.
- J. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. 'Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network.' *BMC medical research methodology*, 18(1):1–12, 2018.
- S. Lundberg and S. Lee. 'A unified approach to interpreting model predictions.' In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.
- K. Charmpi and B. Ycart, 'Weighted kolmogorov smirnov testing: an alternative for gene set enrichment analysis.' *Statistical applications in genetics and molecular biology,* vol. 14, no. 3, pp. 279–293, 2015.