Predicting Survival Time of LGG Patients: Refining Feature Selection Methods

Sanchayan Bhowal[1], Rosalie Daniels[2], Neo Kok[3], Mackenzie Mueller[4]

[1]Indian Statistical Institute, [2]University of Pennsylvania, [3]University of Michigan, and [4]Pacific Lutheran University

Data

The research was conducted using data from 61 samples, all of which were obtained from individuals diagnosed with Low-Grade Glioma (LGG). The dataset consisted of three main sources of information. Firstly, gene pathway data provided 1283 genetic pathway scores derived from four pathway collections (Hallmark, KEGG, C4, and C6) for each of the 61 LGG samples. These scores are computed using the gene-set variation analysis (GSVA) procedure in [iii].

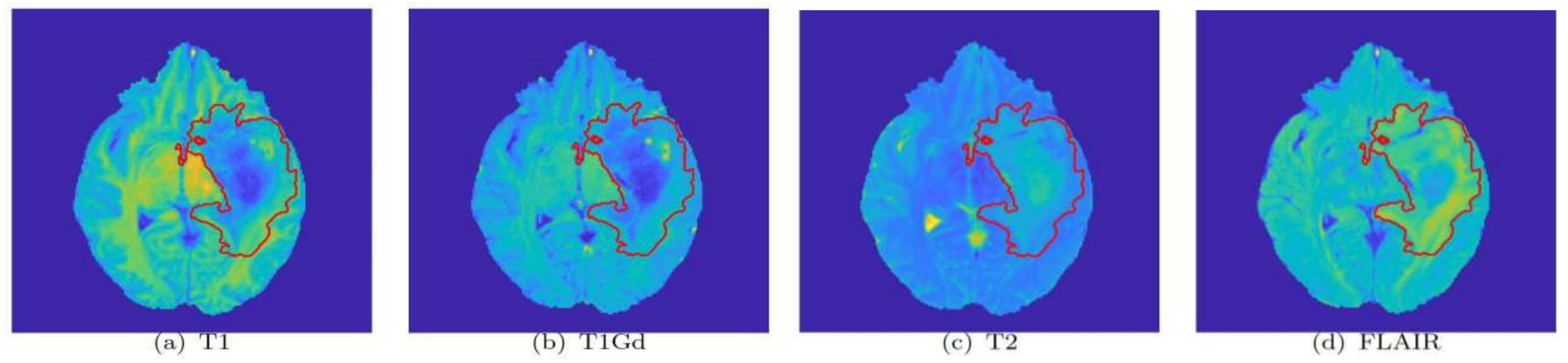

(a) T1 (b) T1Gd (c) T2 (d) FLAIR

Figure 1: Axial slice from the MRI scan of an LGG patient.

Secondly, imaging data encompassed 143 measurements derived[vii] from medical imaging scans of the LGG patients, offering valuable insights into their physical characteristics. Four types of MRI sequences (T1, T1Gd, T2, and FLAIR) are considered. Each of these sequences displays tissues with varying contrasts based on the tissue characteristics, as shown in Figure 1. Lastly, survival time represents the period of time that individuals have survived since being diagnosed with cancer.

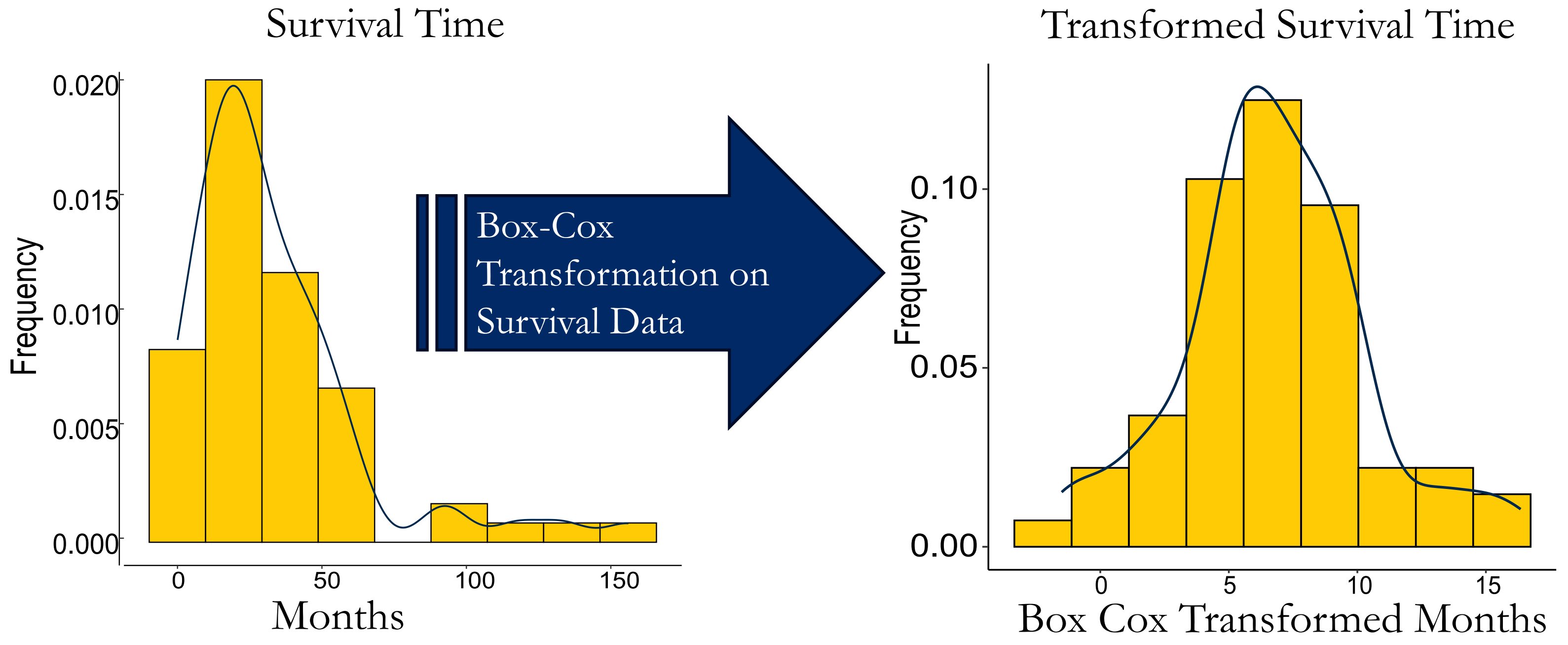


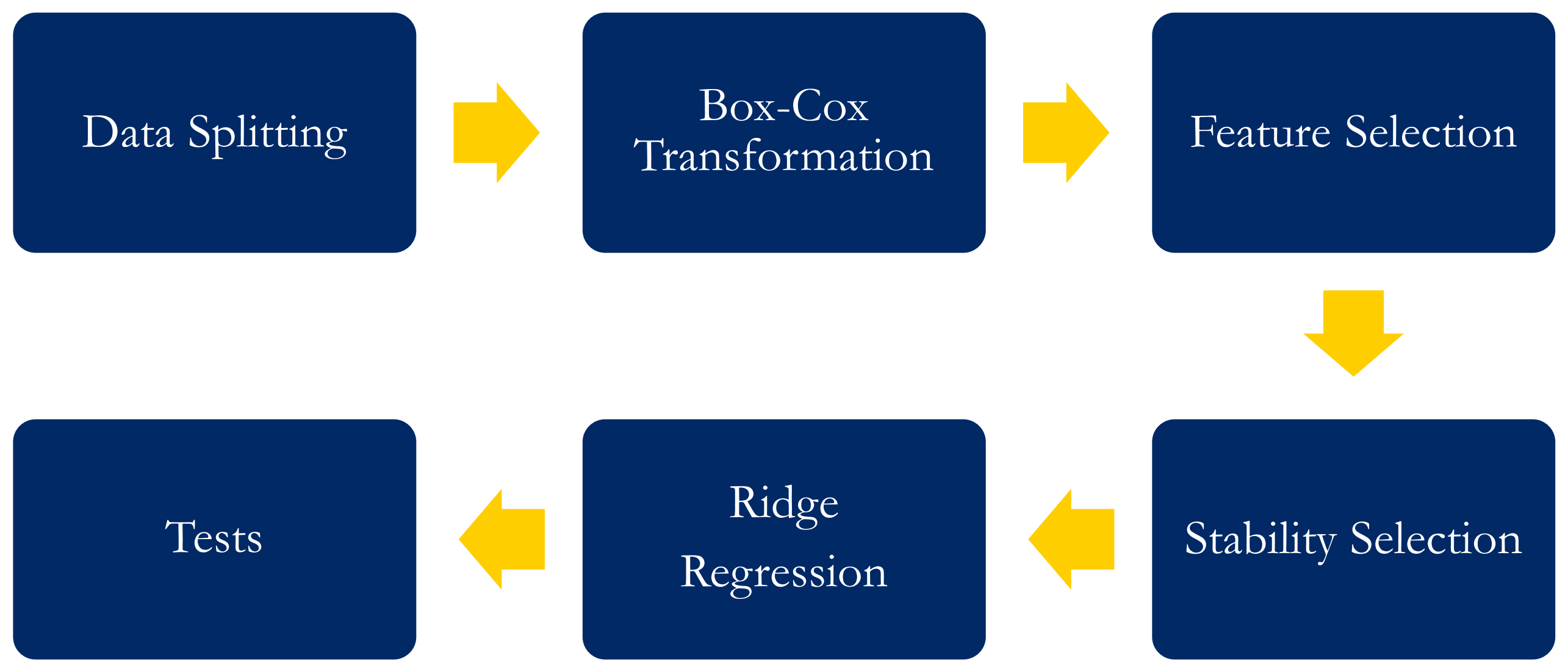
Figure 2: The survival time data does not have a normal distribution and hence it is transformed to a distribution close to normal using Box Cox Transformation. Here the transformation is $(x^{\lambda}-1)/\lambda$, where $\lambda = 0.4$

Objective

Predicting survival times in Low-Grade Glioma (LGG) patients is crucial for optimizing treatment decisions and improving patient outcomes. Our aim is to:

- Compare different feature selection techniques to identify the one that yields the lowest Root Mean Squared Error (RMSE) in predicting survival time
- Develop a model that can be effectively generalized to newer data, mitigating the risk of overfitting and ensuring robust performance
- Address multicollinearity issues in the dataset by implementing appropriate methods to manage and control the impact of highly correlated predictor variables

Process



Data Splitting → Box-Cox Transformation → Feature Selection → Stability Selection → Ridge Regression → Tests

Feature Selection Methods

- **LASSO:** Linear regression method that adds penalty term to the loss function to encourage sparsity, resulting in feature selection and regularization
- **Boruta:** Identifies important variables by comparing importance of original features with that of shadow variables, facilitating robust and accurate model building [vi]
- **Recursive Feature Elimination:** Recursively removes less important features from a model to improve its performance and reduce overfitting
  - ❖ **Random Forests:** Removes features that increase mean error of random forest regressor [ii, v]
  - ❖ **LM:** Orders according to magnitude of coefficients of features
  - ❖ **Bagged Trees:** Removes features that increase mean error of the bagging predictor [i]

Stability Selection

1. Let N be the subsample number
2. Start with the full dataset Z. For each i in $1, \ldots, N$ do:
   a) Subsample from Z without replacement to generate a smaller dataset of size, given by Z_i
   b) Run the feature selection algorithm on dataset Z_i with parameter λ to obtain a selection set S_i
3. Given the selection sets from each subsample, calculate the empirical selection probability for each model component:

$$\Pi_k = \frac{1}{N}\sum_{i=1}^{n} \mathbb{I}\{k \in S_i\}$$

The selection probability for feature k is its probability of being selected by the algorithm

4. Construct the stable set[iv] according to the following definition:

$$S^{stable} = \{k : \Pi_k > 0.8\}$$

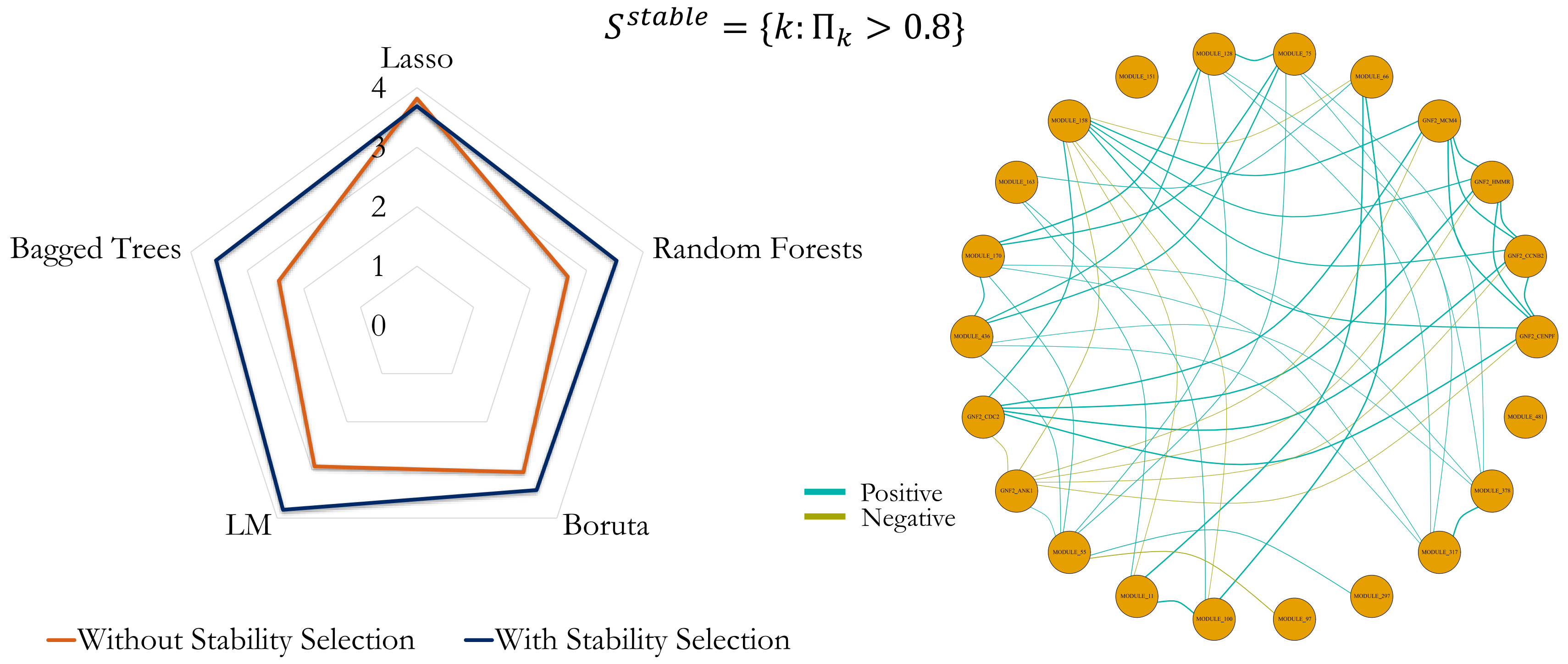


Figure 3: The plot on the left shows that stability selection increases RMSE on genomics data. The plot on the right shows the high correlation between variables of genomics.

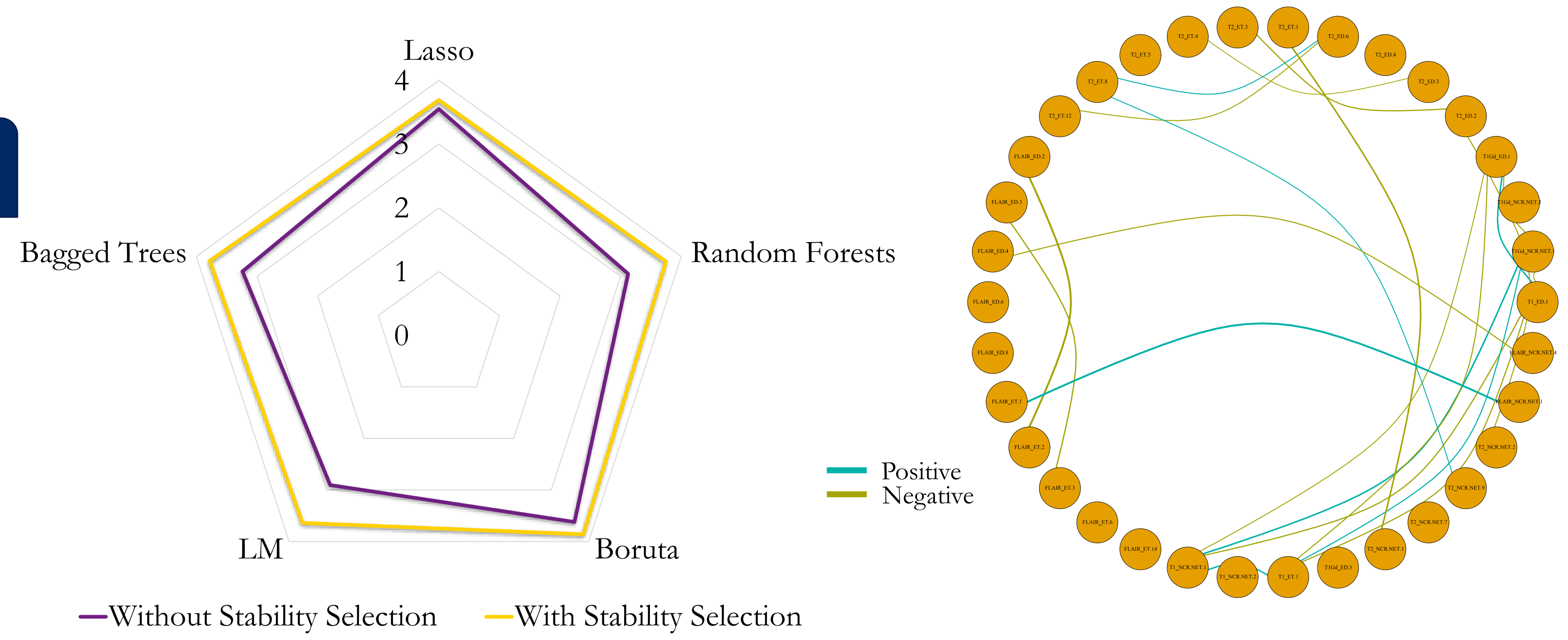


Figure 4: The plot on the left shows that stability selection increases RMSE on imaging data. The plot on the right shows the high correlation between variables of imaging data.

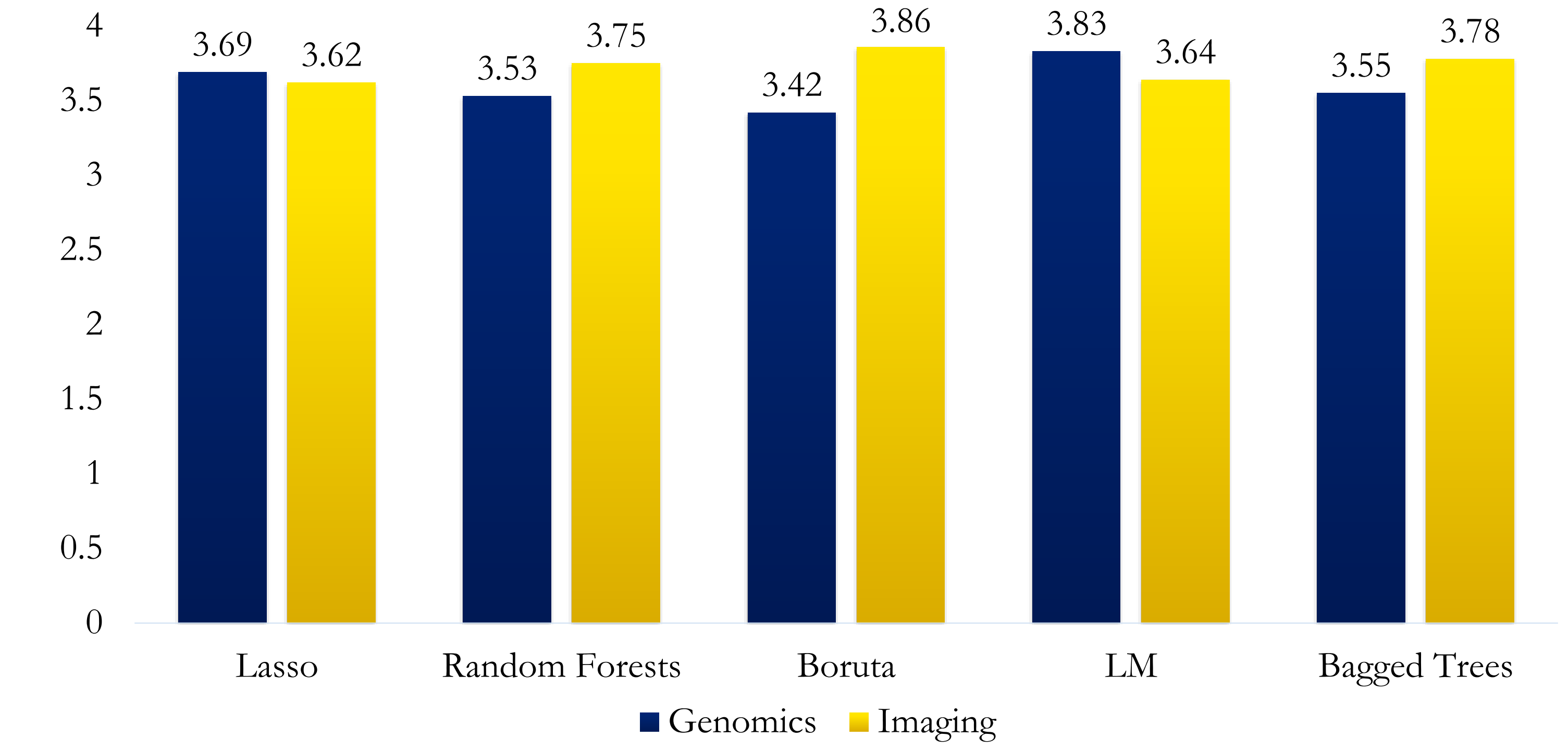


Figure 5: The plot shows the RMSE after stability selection for each of the feature selection methods.

Results

- The Boruta method, when applied to the genomic data, demonstrated the best overall performance among the feature selection methods tested.
- The LASSO method proved to be consistently reliable in its feature selection capabilities. The results consistently showed its ability to perform well across various scenarios. There was a relatively small RMSE difference of 0.13 between the pre- and post-stability models for LASSO.
- Stability selection was particularly beneficial in reducing the impact of random seeds, especially when dealing with low n (sample size) situations.
- The techniques did not yield much improvements over the mean predictor (which gives an RMSE 3.8)

In summary, the research underscores the efficacy of the Boruta method when applied to genomic data, the consistent performance of LASSO, and the importance of stability selection in ensuring reliability and robustness in feature selection processes, especially in the context of high-dimensional data and low sample sizes.

Limitations

- The relatively small sample size of the dataset might affect the model's ability to generalize to larger populations.
- Certain algorithms were sensitive to the random seed used, leading to variability in results.
- Dealing with high-dimensional data ($p \gg n$) and highly multicollinear variables presented challenges in building precise and interpretable models.
- Some feature selection methods and model-building techniques were computationally expensive
- The dataset had limited information on disease severity and progression in patients, which might have impacted the model's accuracy.
- Due to time constraints, we were limited in exploring an extensive range of feature selection methods, and with more time, we would have attempted additional approaches for better prediction

Acknowledgments

We sincerely thank our mentors, Johann Gagnon-Bartsch, Ph. D., Snigdha Panigrahi, Ph. D., and co-mentor Josh Wasserman, for their invaluable guidance and support throughout the project. We would also like to extend our heartfelt acknowledgement to Prof. Bhramar Mukherjee for organizing BDSI, which provided us with this incredible learning and research experience.

References

i. Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.
ii. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
iii. Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC bioinformatics, 14, 1-15.
iv. Huang, S., & Marchetti-Bowick, M. (2014). Summary and discussion of: "Stability Selection" Statistics Journal Club, 36-825.
v. Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. R news, 2(3), 18-22.
vi. Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11
vii. Saha, A., Banerjee, S., Kurtek, S., Narang, S., Lee, J., Rao, G., ... & Baladandayuthapani, V. (2016). DEMARCATE: Density-based magnetic resonance image clustering for assessing tumor heterogeneity in cancer. NeuroImage: Clinical, 12, 132-143.