

Machine Learning Based Benchmark for Space Flown Rodent Liver RNA Sequencing Data in Space Biology Research

Authors: Nishan Pantha¹, Iksha Gurung¹, Muthukumaran Ramasubramanian¹, James Casaletto², Lauren M. Sanders², Manil Maskey¹, Sylvain V. Costes²

¹The University of Alabama in Huntsville, ²National Aeronautics and Space Administration

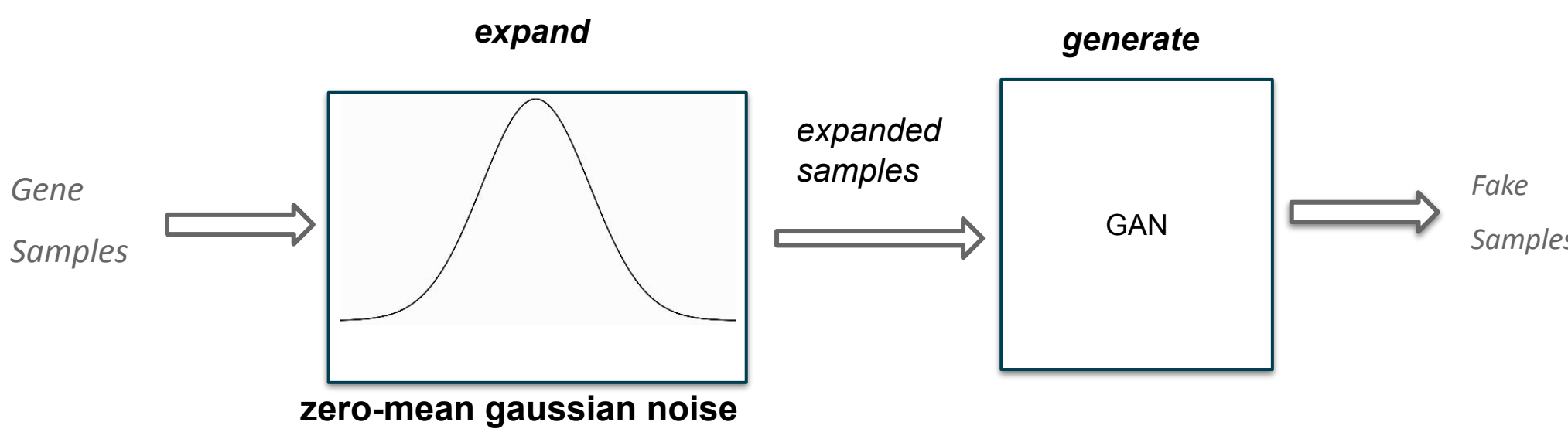


Introduction

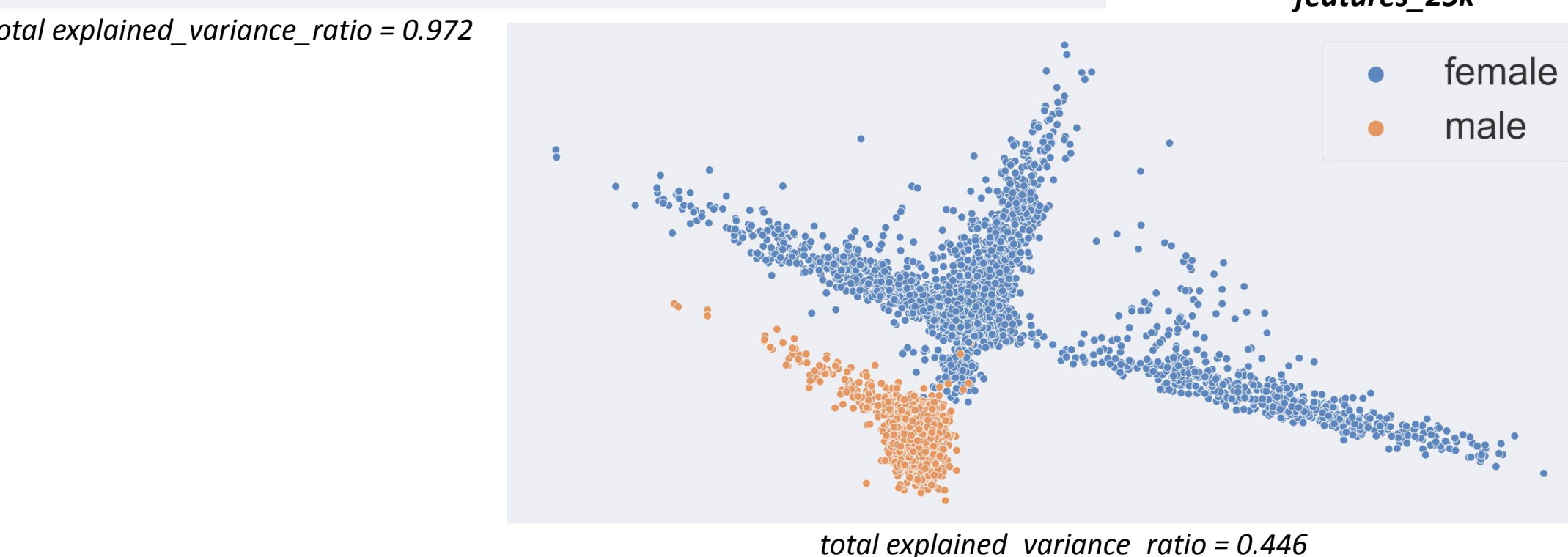
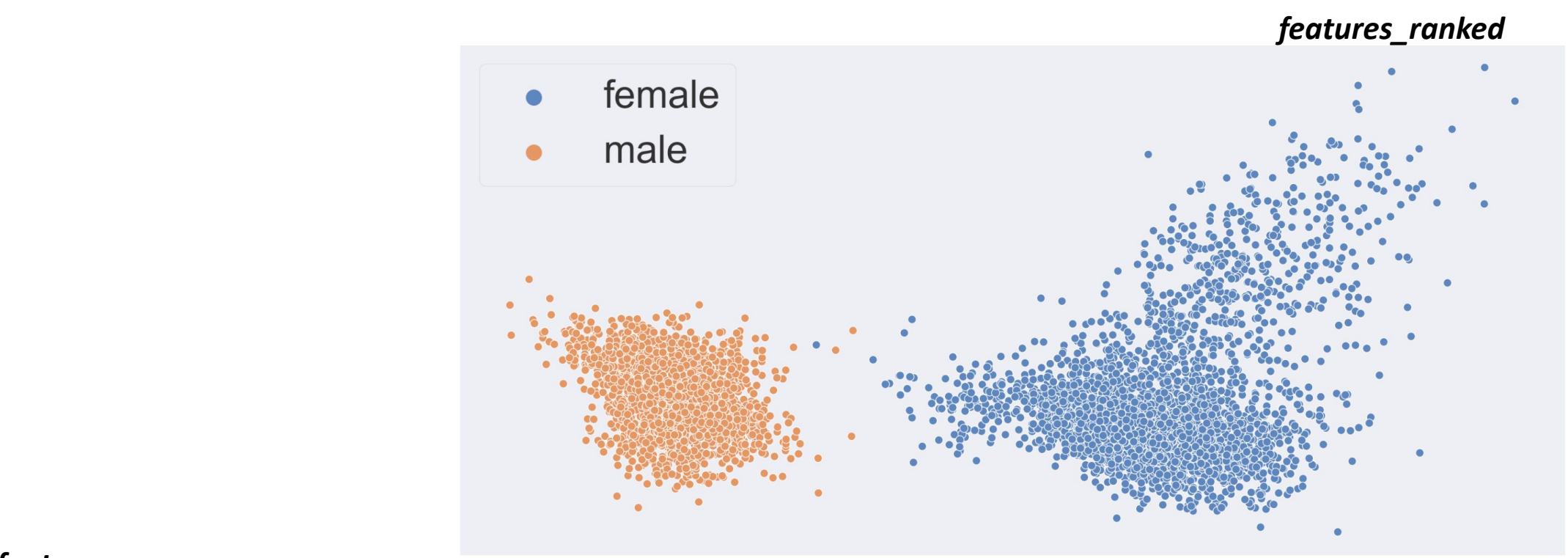
NASA's Biological and Physical Sciences division has generated a nucleic acid sequencing (RNA-sequencing) dataset from liver samples from space-flown mice. The dataset is low in sample size but has a large number of genes contributing to the characteristics of the samples. This makes it difficult to identify the genes contributing to specific attributes such as gender, strain, age, etc. Our work is to leverage machine learning algorithms to help identify important feature genes that are contributing to such characteristics.

Data Analysis

- Use zero-mean Gaussian noise to artificially expand the gene expression sample size.
- Train a Generative Adversarial Model (GAN) using expanded samples to generate new fake samples that closely match the original data distribution.
- Balance the class distribution between *male* and *female* for **gender** attribute giving the *balanced_expanded* and *balanced_fake* datasets.



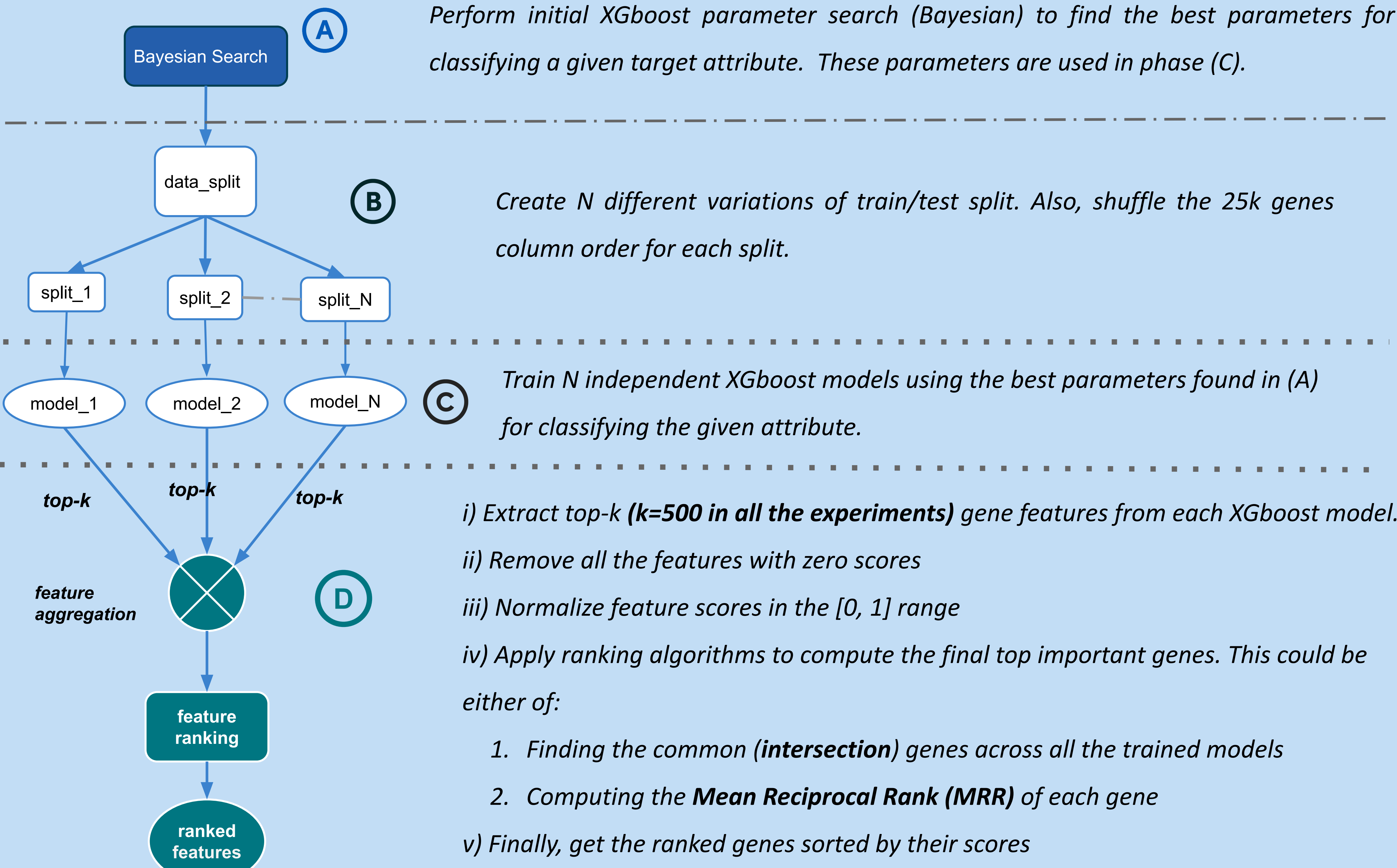
- Resulting dataset has 25000 gene expression feature columns which are highly correlated with each other.
- Perform hierarchical clustering based on absolute correlation and analyze the cluster formation across *expanded* and *fake* datasets.
- The data has a high level of redundant gene expressions, and only a subset of these gene expressions contribute to the target attributes.



PCA plots ($n_{components}=2$) of *balanced_expanded* dataset using ranked feature subsets and full 25k features for **gender**.

Gene Ranking Method

We develop an ensemble-based gene-ranking pipeline, **GeneRanker**, that uses eXtreme Gradient Boosting (XGboost) based models to identify the top *important* genes contributing to the attribute.



Results & Validation

We compare the relevancy of top-ranked feature sets by comparing the model classification performances when using ranked feature sets versus the model that utilizes all the features.

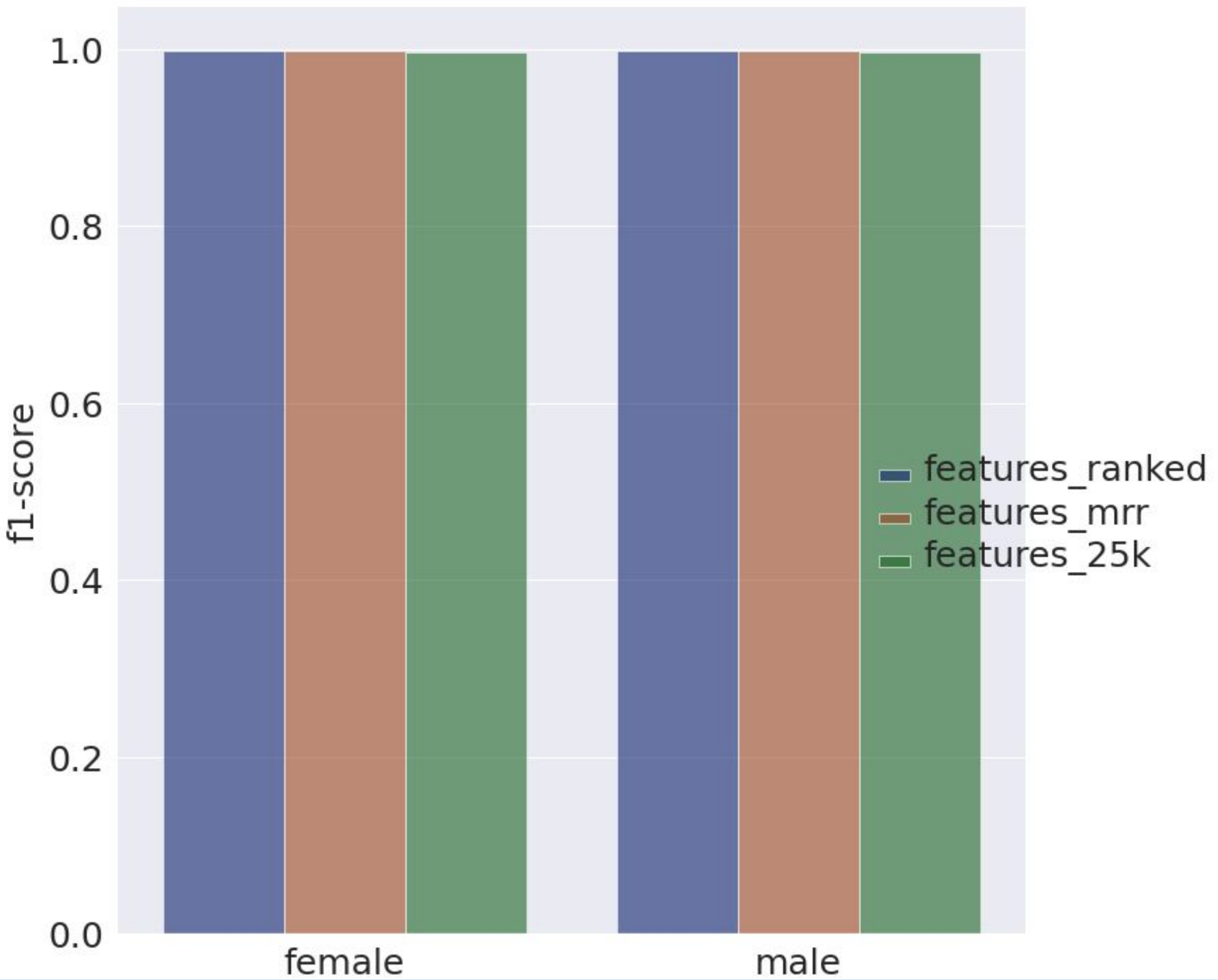
We use 3 gene feature sets for comparative analysis:

- intersection-based genes (**features_ranked**)
- MRR-based genes (**features_mrr**)
- full 25k gene features (**features_25k**).

We train the models using the fake dataset and evaluate them on the completely held-out expanded dataset. We find that using the topmost ranked features has comparable performance w.r.t original 25k feature sets.

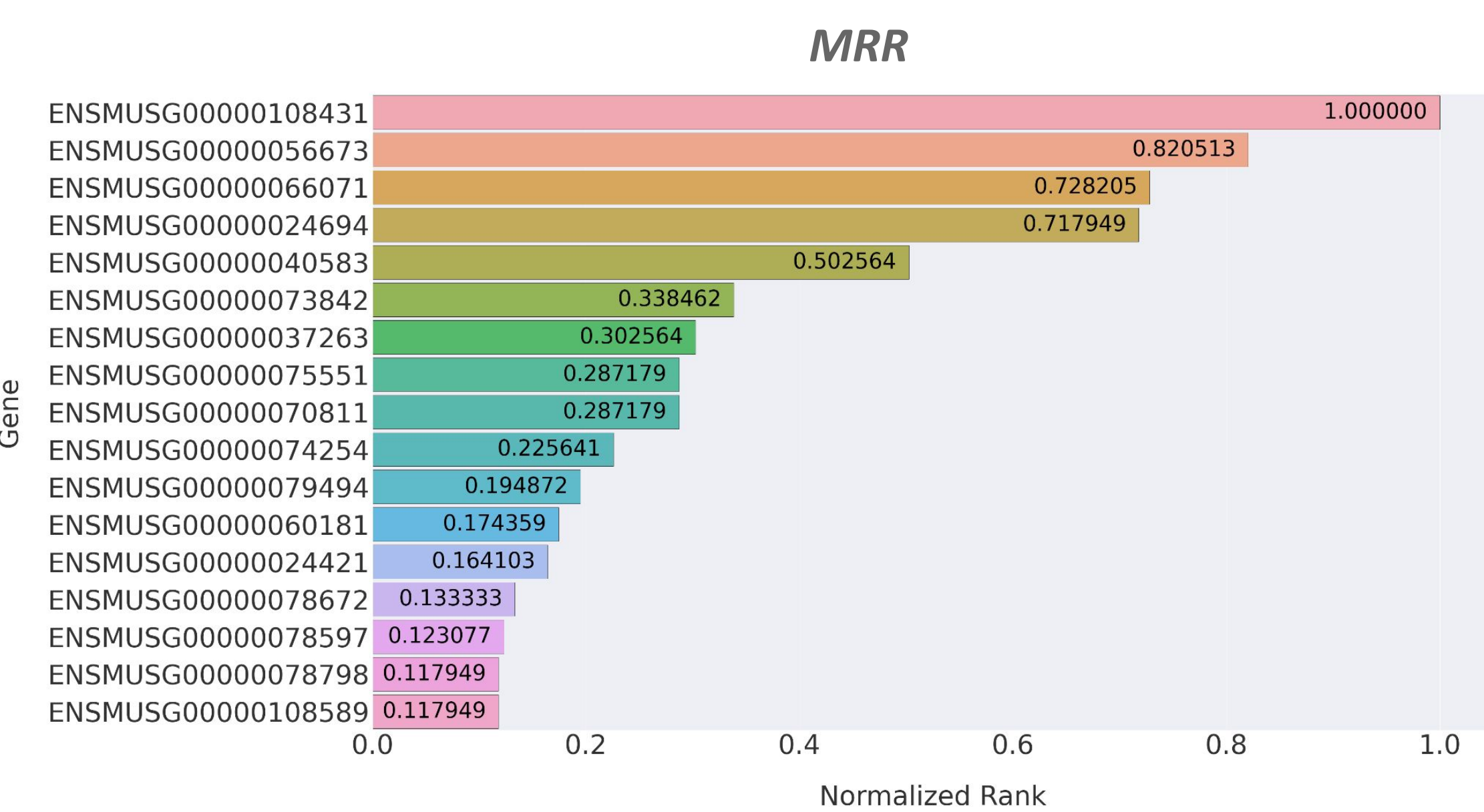
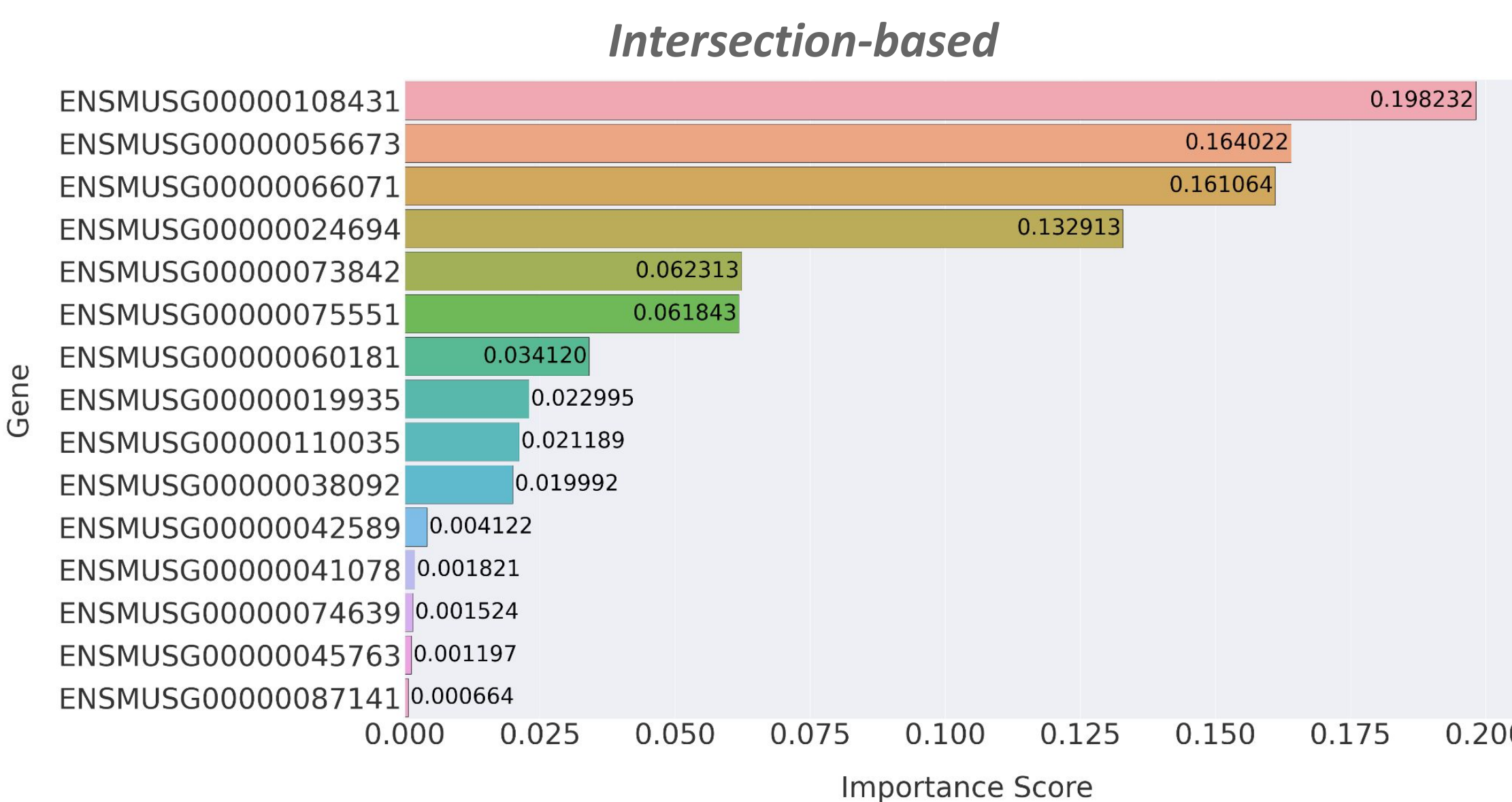
Acknowledgement

This work is supported by NASA Grant 80MSFC22M004



Model performances on *balanced_expanded* using top feature subsets (brown and blue) vs all 25k features (green) for **gender**

Top important genes for gender



Conclusion

We developed an ensemble-based feature ranking framework that takes into consideration the high-dimensionality nature of gene expressions. We used this framework to find the topmost relevant genes for some notable phenotype attributes such as **gender**, **age**, and **strain**, as well as on external factors such as **condition**. We also evaluate the performance of models on these attributes using the ranked features.

References

- Johnsen, P. V., Riemer-Sørensen, S., DeWan, A. T., Cahill, M. E., & Langaas, M. (2020). A new method for exploring gene-gene and gene-environment interactions in GWAS with tree ensemble methods and SHAP values (p. 2020.05.13.20100149). medRxiv. <https://doi.org/10.1101/2020.05.13.20100149>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. <https://doi.org/10.1145/2939672.2939785>
- Silhouette (clustering). (2022, November 5). In Wikipedia. [https://en.wikipedia.org/wiki/Silhouette_\(clustering\)](https://en.wikipedia.org/wiki/Silhouette_(clustering))
- Permutation Importance with Multicollinear or Correlated Features. (n.d.). Scikit-Learn. Retrieved November 17, 2022, from https://scikit-learn/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html
- Head, Tim, Kumar, Manoj, Nahrstaedt, Holger, Louppe, Gilles, & Shcherbatyi, Iaroslav. (2021). scikit-optimize/scikit-optimize (v0.9.0). Zenodo. <https://doi.org/10.5281/zenodo.5565057>

