A comparison of Linear and Non-linear models in separating cancer vs non cancer SNPs

Trey Stansfield and Meiheng Liang

Luddy School, Indiana University Indianapolis

INFO-B 529 Machine Learning for Bioinformatics

Dr. Juexin Wang

May 7th, 2025

Introduction

Breast cancer is a key disease facing many women around the globe. Approximately 13.1% of US women will be diagnosed with breast cancer, with 2.3% dying from the disease (American Cancer Society, 2024). Single nucleotide polymorphisms (SNPs) are point mutations in at least 1% of the human population (U.S. National Library of Medicine, 2025). SNPs are also associated with breast cancer, with many common SNPs like BRCA1 and BRCA2 being co-morbidly associated with higher breast cancer risk (Onay et al., 2006). Despite this association being known, it is very difficult to determine exactly what SNPs contribute to breast cancer and by how much due to how these SNPs interact with different types of breast cancer and other preventative treatments (Cuzick et al., 2017). SNP gene panel screenings are known to increase the quality of patient care and outcomes, but a fundamental knowledge and education gap prevents further benefits from taking hold (Lanchbury & Pederson, 2023). Machine learning offers an avenue to better study what SNPs are associated with breast cancer.

Past research has used machine learning to identify SNPs associated with disease— such as Silva et al., 2022— so a similar method could be used to identify SNPs associated with breast cancer. Past studies have used non-linear methods to uncover SNPs related to disease with success, but non-linear models tend to be complex and computationally intensive (Elgart et al., 2022). Past research into neuron networks found that, contrary to popular belief, using a linear model produced more accurate results than non-linear models, which allows for faster, more understandable, and less computationally intensive models to be used (Nozari et al., 2024). Our project will attempt a similar question but with breast cancer. Testing if linear or non-linear models better predict breast cancer in women using SNPs.

Methodology

The dataset chosen for this project was SRA project SRP162370 from the paper Fang et al., 2021. These are paired whole genome samples from breast cancer patients of healthy tissue and cancerous tissue. These samples were chosen due to having slightly smaller sizes and mostly being paired to each other. The split is even with 7 cancer samples and 7 normal samples (Table 1).

Table 1

List of Samples used

Sample Name	**Type**
SRR7890844	*Cancer*
SRR7890845	*Normal*
SRR7890850	*Cancer*
SRR7890851	*Normal*
SRR7890852	*Cancer*
SRR7890853	*Normal*
SRR7890854	*Cancer*
SRR7910003	*Normal*
SRR8955955	*Cancer*
SRR8955956	*Normal*

SRR8955957	*Cancer*
SRR8955958	*Normal*
SRR8955959	*Cancer*
SRR8955960	*Normal*

The datasets were downloaded using the SRA-toolkit on IU's Quartz Supercomputer onto the Slate storage system. They were initially processed using fastqc to check for primers and sequencing errors before being sent to trim-galore for trimming. BBMap Repair.sh was used to ensure that each strand of the split reads remained consistent. Bowtie2 was used to index the human reference genome GRCh38.p14/GCF_000001405.40 for alignment to the reads, and samtools was used to convert the sam files into bam files. GATK4 was used to identify variants. Bcftools was used to annotate the vcf files using dbsnp. PLINK2 was used to separate out multi-allelic SNPs into a bed file. These factors were then read into sklearn for dimension reduction using PCA and UMAP. For linear regression, the lasso model in sklearn was used, and for the non-linear model, the random forest module of sklearn was used. Leave-one-out crossfold validation was used due to our small sample size using the sklearn module.

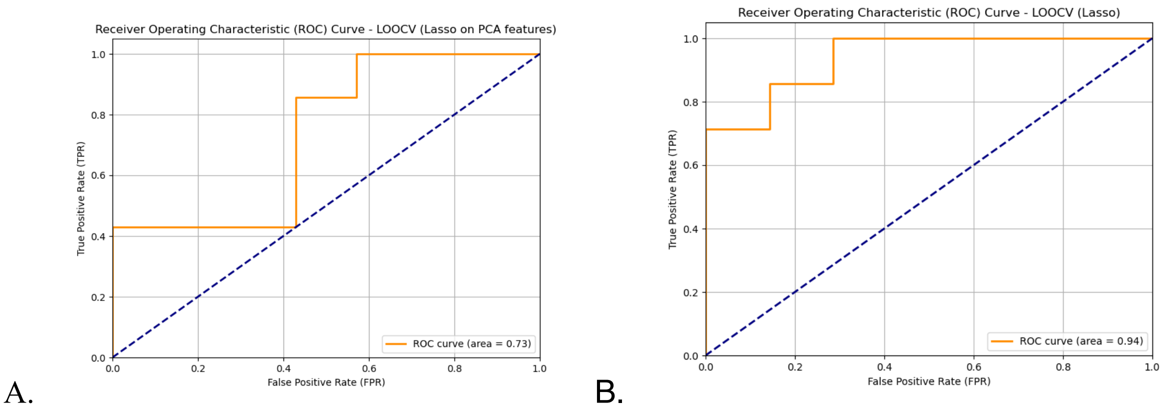
The models were judged on the accuracy using AUC-ROC and leave-one-out cross-validation (LOOCV). They were all run on Quartz in the same Conda environment with the same Slurm parameters. They will also be judged using metrics like CPU and RAM usage using Quartz's built-in job system. This will allow us to determine both the efficiency and accuracy of each approach. We hypothesize that the linear model will be more efficient than the non-linear one due to its lower complexity, but the non-linear model will perform better on features as SNPs often have complex, co-abundance effects on cancer rather than simple linear ones.

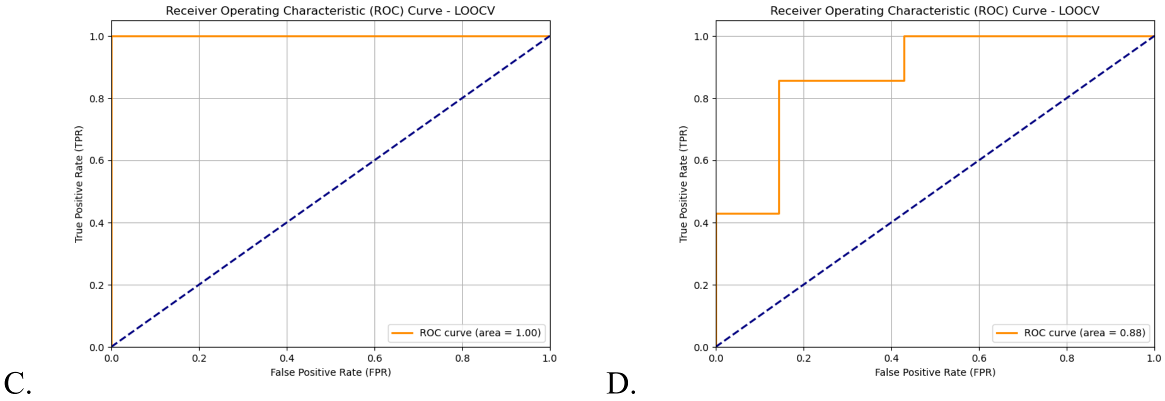
Results

The four tests were preformed: Lasso with PCA, Lasso with UMAP, Random forest with PCA, and Random forest with UMAP. For accuracy, leave-one-out cross-validation and ROC AUC were tested. The ROC curves are in figure 1.

Figure 1

ROC curves for all four tests





A). Lasso with PCA B). Lasso with UMAP C). Random forest with PCA D). Random forest with UMAP

A summary of all of the performance metrics can be found in Table 2.

Table 2

Performance Metrics Summary

Model	**ROC AUC**	**Accuracy**	**Time Spent (HH:MM:SS)**	**CPU Usage (%)**	**RAM usage (GB)**
Lasso PCA	0.73	0.7143	00:02:06	3.60%	2.52 GB
Lasso UMAP	0.94	0.9388	00:00:38	1.16%	1.87 GB
Random Forest PCA	1.00	1.0000	00:02:29	3.61%	1.71 GB
Random Forest UMAP	0.88	0.8776	00:00:44	1.83%	1.80 GB

Notes: Time spent, CPU usage, and RAM usage were calculated using the built in quartz job email statistics.

The rs IDS which make up each PCA component for each method are included in table 3.

Table 3

Most Prominent SNP IDS

Method	**PCA Object**	**SNP's included**
Lasso	1	rs381935: 0.0006 rs1476051255: -0.0006 rs1583758: -0.0006 rs1583787: -0.0006 rs1583782: -0.0006
Lasso	2	rs370884133: 0.0010 rs2602600: 0.0010 rs853379: 0.0010 rs369199122: 0.0010 rs853386: 0.0010
Lasso	3	rs181771137: 0.0014 rs9619551: 0.0014 rs373683622: 0.0014 rs1315361163: 0.0014 rs11792193: 0.0014
Lasso	4	rs1421833778: 0.0018 rs153445: 0.0018 rs1871062: 0.0018 rs753896795: 0.0018 rs1655446785: 0.0018
Lasso	5	rs191279075: 0.0015 rs4665882: 0.0015 rs374382085: 0.0015 rs1556240437: 0.0015 rs377357875: 0.0015
Random Forest	1	rs381935: 0.0006 rs1476051255 -0.0006 rs1583758 -0.0006 rs1583787 -0.0006 rs1583782: -0.0006

Random forest	2	rs370884133: 0.0010 rs2602600: 0.0010 rs853379: 0.0010 rs369199122: 0.0010 rs853386: 0.0010
Random forest	3	rs181771137: 0.0014 rs9619551: 0.0014 rs373683622: 0.0014 rs1315361163: 0.0014 rs11792193: 0.0014
Random forest	4	rs1421833778: 0.0018 rs153445: 0.0018 rs1871062: 0.0018 rs753896795: 0.0018 rs1655446785: 0.0018
Random forest	5	rs191279075: 0.0015 rs4665882: 0.0015 rs374382085: 0.0015 rs1556240437: 0.0015 rs377357875: 0.0015

Discussion

The most accurate method was Lasso with UMAP, which had an overall ROC AUC of 0.94, with Random Forest UMAP having a ROC AUC of 0.88, and Lasso PCA having the least ROC AUC with 0.73. With a perfect score of 1.00, it seems that Random Forest PCA overfitted even when the number of features was reduced from 5 to 2 due to the small sample size used. In terms of performance, both models were very close, using similar amounts of RAM and CPU and being within a minute of each other. Across the board, UMAP seemed to do better than its

PCA counterparts, taking less time, using less CPU, and using less RAM, with only the Random Forest UMAP using slightly more CPU than its PCA counterpart.

In terms of model vs model efficiency, it seems Random Forest took slightly longer, used slightly more CPU, and used less RAM than Lasso with both UMAP and PCA. This is in contrast with our hypothesis that the linear Lasso Method would consistently be more performant than Random Forest. It also seems that the maximum performance of Lasso was higher than Random Forest in this trial in contrast to our hypothesis. The true validity of these scores cannot be determined though, as the Random Forest PCA trial showed severe overfitting, so there is a chance these other high results are also suffering from a degree of overfitting. Further tests with higher sample sizes need to be done.

The consistent superiority of UMAP over PCA seems to be based on the biological context of SNPs. SNPs tend to affect disease in complex, non-linear ways, with many SNPs interacting in ways to increase risk rather than a directly linear relation of one SNP leading to more risk. As UMAP reduces data in a nonlinear way, it could be preserving more of these non-linear relationships, allowing for higher accuracy compared to PCA, which is purely looking at variance.

Conclusion

Machine learning has become an efficient approach in exploring and filtering expanding biomedical data and provides attractive means in disease risk prediction based on populational SNP makeup. The nonlinear interactions of SNPs and their associations on disease make nonlinear dimension reduction a powerful way to reduce the feature count and identify the most important SNPs. The result indicated that linear models offer slightly higher accuracy than nonlinear models and comparable performance across both models. Due to the small sample size, the models showed a tendency to overfit, especially the Random Forest PCA model, which had perfect accuracy even when LOOCV was used. This highlights the importance of doing further research with more samples to ensure that the in-sample error can more closely match the out-sample error with these models. We hope that the incorporation of advancing models and whole-genome sequencing can promote the accuracy of diagnosis and analysis through risk factors identification and interpretation. With complex associations between SNPs within diseases, it is common assumption that nonlinear models would outperform linear ones, but our research shows that linear models perform as well, if not better, than nonlinear ones, highlighting that simple models remain valuable approaches when considering modeling for complex biological questions.

References

American Cancer Society. (2024). *Breast cancer facts & figures 2024-2025*. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2024/breast-cancer-facts-and-figures-2024.pdf

Cuzick, J., Brentnall, A., & Dowsett, M. (2017). SNPs for breast cancer risk assessment. *Oncotarget*, *8*(59), 99211–99212. https://doi.org/10.18632/oncotarget.22278

Elgart, M., Lyons, G., Romero-Brufau, S., Kurniansyah, N., Brody, J. A., Guo, X., Lin, H. J., Raffield, L., Gao, Y., Chen, H., de Vries, P., Lloyd-Jones, D. M., Lange, L. A., Peloso, G. M., Fornage, M., Rotter, J. I., Rich, S. S., Morrison, A. C., Psaty, B. M., Levy, D., … Sofer, T. (2022). Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Communications biology*, *5*(1), 856. https://doi.org/10.1038/s42003-022-03812-z

Fang, L. T., Zhu, B., Zhao, Y., Chen, W., Yang, Z., Kerrigan, L., Langenbach, K., de Mars, M., Lu, C., Idler, K., Jacob, H., Zheng, Y., Ren, L., Yu, Y., Jaeger, E., Schroth, G. P., Abaan, O. D., Talsania, K., Lack, J., Shen, T. W., … Somatic Mutation Working Group of Sequencing Quality Control Phase II Consortium (2021). Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. *Nature biotechnology*, *39*(9), 1151–1160. https://doi.org/10.1038/s41587-021-00993-6

Lanchbury, J. S., & Pederson, H. J. (2023). An apparent quandary: adoption of polygenics and gene panels for personalised breast cancer risk stratification. *BJC reports*, *1*(1), 15. https://doi.org/10.1038/s44276-023-00014-w

Nozari, E., Bertolero, M. A., Stiso, J., Caciagli, L., Cornblath, E. J., He, X., Mahadevan, A. S., Pappas, G. J., & Bassett, D. S. (2024). Macroscopic resting-state brain dynamics are best described by linear models. *Nature biomedical engineering*, *8*(1), 68–84. https://doi.org/10.1038/s41551-023-01117-y

U.S. National Library of Medicine. (2025, March 25). *Single nucleotide polymorphisms (SNPs)*. MedlinePlus. https://medlineplus.gov/genetics/understanding/genomicresearch/snp/

Onay, V. Ü., Briollais, L., Knight, J. A., Shi, E., Wang, Y., Wells, S., Li, H., Rajendram, I., Andrulis, I. L., & Ozcelik, H. (2006). SNP-SNP interactions in breast cancer susceptibility. *BMC Cancer, 6*(114). https://doi.org/10.1186/1471-2407-6-114

Silva, P. P., Gaudillo, J. D., Vilela, J. A., Roxas-Villanueva, R. M. L., Tiangco, B. J., Domingo, M. R., & Albia, J. R. (2022). A machine learning-based SNP-set analysis approach for identifying disease-associated susceptibility loci. *Scientific reports*, *12*(1), 15817. https://doi.org/10.1038/s41598-022-19708-1