Deep IV with Whole-Omics Data to Identify Novel Carcinogenesis-Mediating Pathways

Jack Andraka
jandraka@stanford.edu

Billy Ferguson
billyf@stanford.edu

Charlie Walker
cwalker4@stanford.edu

Problem

Breast cancer accounts for over 25% of cancer diagnoses and 15% of cancerous deaths in women [1]. Despite extensive research with methods such as genome-wide association studies (GWAS), only 84% of breast cancer heritability can be explained by currently known genes, most likely due to GWAS's inability to detect additive, small gene signals (the dominant hypothesis of phenotype formation) [2]. Additionally the GWAS study design is plagued by the confoundedness between the transcriptome and expressed phenotypes, despite sophisticated statistical techniques to deal with complicated interactions [3]. A new method is needed for ascertaining the causative relationship between transcriptome profiles and carcinogenesis.

Dataset

To assess the ability of the Deep IV model to uncover carcinogenesis-mediating pathways using genomic mutations as the instrument, transcriptome profiles as the treatment variable, and cancerous (1) vs healthy (0) as the outcome variable, we employed whole-genome and whole-transcriptome data from the Genotype-Tissue Expression (GTEx) project and The Cancer Genome Atlas (TCGA). The GTEx dataset includes 11,688 samples that are non-cancerous with matched whole-genome and whole-transcriptome data. The TCGA dataset includes 978 breast cancer samples with matched whole-genome and whole-transcriptome data. When the two datasets are interjoined there are 13,980 genes and 53,196 transcripts measured. To make network training computationally tractable, we reduced the transcript space to 2,344 transcripts of genes that are hypothesized to have a role in carcinogenesis.

Objective Function

We employed the Adam Optimizer with the default parameters [4]. We utilized the following integral loss function over training data D of size T = |D|.

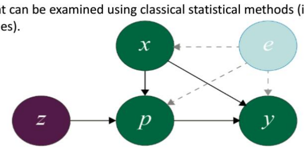
$$\mathcal{L}(D;\theta) = |D|^{-1}\sum_t \left(y_t - \int h_\theta(p, x_t)\, d\hat{F}_\phi(p|x_t, z_t)\right)^2.$$

We use the following unbiased estimate of the loss function to train our model, replacing the integral with a sum over samples from the fitted treatment distribution function.

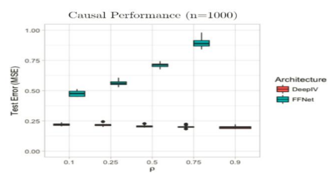
$$\mathcal{L}(D;\theta) \approx |D|^{-1}\sum_t \left(y_t - \frac{1}{B}\sum_{\dot{p}\sim\hat{F}_\phi(p|x_t,z_t)} h_\theta(\dot{p}, x_t)\right)^2 := \hat{\mathcal{L}}(D;\theta).$$

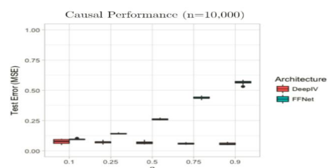
Network Architecture

In this project, we applied the method described in "Deep IV: A Flexible Approach for Counterfactual Prediction. To do this, we utilize a two-stage Deep Neural Network (DNN), where the first DNN, the policy network, takes input z (gene mutations; exogenous) and creates predictions of p (transcript abundance; treatment variable) [5]. These predictions of p are then fed into the second DNN, the response network, to predict y (phenotype, cancerous or health; outcome variable), thus removing the endogeneity between p and y, enabling the determination of causal relationships between the transcriptome and carcinogenesis. To interpret these causal relationships, the response network is used as a data-generating process to create a simulated dataset containing the causal relationship uncovered by the Deep IV network, that can be examined using classical statistical methods (i.e. Trees).



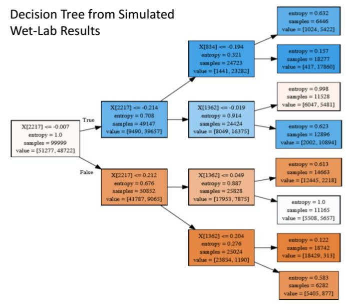
Using simulated data with an underlying causal relationship, we demonstrated that the Deep IV model's performance is unaffected by increasing confoundedness, while the performance of a standard Feed-Forward Network does poorly at recovering the true counterfactual function.





Results

We performed a grid search, selecting for number of layers, number of nodes, and the hyperparameters for learning rate, L2 regularization, and dropout rate. Below we show the training and validation for both of our stages with the preferred model specification highlighted in green. To optimize the full DeepIV model, we select the best model specification for the first stage, and then use the predicted outcomes from this optimal first stage to train the optimal second stage [5]. Ultimately, using this two stage DeepIV framework, we ended with a .5% classification error on the test set. Our second stage actually models the causal relationship between gene expression and cancer. Thus, we can use the trained second stage to simulate outcomes for varying gene expressions as if we had performed controlled scientific experiments in the lab to overexpress and underexpress certain genes. Using simulated outcomes form the second stage, we constructed a decision tree to interpret the causal relationship of complicated interactions between gene expression levels.



1st Stage

Model Architecture (layers/nodes;LR; L2; Dropout Rate; activation/output)	Training MSE	Validation MSE
([100,100,100], 1.00E-3, 0, 0, tanh/linear)	0.974	0.976
([200,100,50], 1.00E-3, 0, 0.4, tanh/linear)	0.806	1.085
([200,200,200], 1.00E-3, 0, 0.5, tanh/linear)	0.983	0.975
([200,200,200], 1.00E-3, 0, 0, tanh/linear)	0.974	0.976
([200,200,200], 1.00E-3, 1.00E-4, 0, tanh/linear)	1.002	0.993

2nd Stage

Model Architecture (layers/nodes; LR; L2; Dropout Rate; activation/output)	Training Cross-Entropy	Validation Cross-Entropy
([50, 50], 1.00E-5, 0, 0, sigmoid/sigmoid)	0.019	0.019
([100, 50], 3.00E-5, 1.00E-4, 0, tanh/sigmoid)	0.036	0.038
([100, 50], 3.00E-5, 0, 0.1, tanh/sigmoid)	0.013	0.014
([100, 50], 1.00E-5, 0, 0, tanh/sigmoid)	0.014	0.017
([100, 50, 10], 1.00E-05, 0, 0, sigmoid/sigmoid)	0.044	0.046

Future Work

Although we discovered a number of transcripts that are influential in mediating carcinogenesis in breast tissue as well as quantifying their causative effect on carcinogenesis, we must validate these results with wet-lab experiments. This would consist of manual up and downregulation of the identified genes to reduce the abundance of the corresponding transcripts followed by monitoring for carcinogenesis behavior (i.e. measuring mutant p53 levels). Additionally, this study notably suffers from batch effect due to all health samples coming from the GTEx dataset and all cancerous samples coming from the TCGA dataset. In order to validate our model further, we have partnered with Dr. Assimes of the Stanford School of Medicine to work with NIH data that does not exhibit batch effects as well as increase the sample size, utilize the complete transcriptome space, and try more extensive hyperparameter tuning.

References

1.) Torre, L. A., et al. "Global Cancer Incidence and Mortality Rates and Trends--An Update." *Cancer Epidemiology Biomarkers & Prevention*, vol. 25, no. 1, 2015, pp. 16–27., doi:10.1158/1055-9965.epi-15-0578.

2.) Skol, Andrew D., et al. "The Genetics of Breast Cancer Risk in the Post-Genome Era: Thoughts on Study Design to Move Past BRCA and towards Clinical Relevance."*Breast Cancer Research*, vol. 18, no. 1, 2016, doi:10.1186/s13058-016-0759-4.

3.) Campos, Gustavo De Los, et al. "Predicting Genetic Predisposition in Humans: the Promise of Whole-Genome Markers." *Nature Reviews Genetics*, vol. 11, no. 12, 2010, pp. 880–886., doi:10.1038/nrg2898.

4.) Kingma, Diederik, et al. "Adam: A Method for Stochastic Optimization." arXiv:1412.6980. 2017.

5.) Jason Hartford et al. "Deep IV: A Flexible Approach for Counterfactual Prediction." In: *International Conference on Machine Learning*. 2017, pp. 1414-1423.