Netrias, LLC
3033 Wilson Boulevard
Suite 700
Arlington, VA 22201
703.647.4100

Summary

Deep architectures (such as Deep Neural Networks (DNNs)) are well-equipped to handle high dimensional, sparse, noisy data with nonlinear relationships [1]. We stood up a Deep Learning (DL) framework to rapidly test different DL models for cancer type classification, and our results show that this DNN framework can play a significant role in understanding relationships between genes that can discriminate between cancers. We will adapt and apply this framework to more data modalities (genomic, transcriptomic, and proteomic) on different cancer types to identify and test effective strategies for dimensionality reduction coupled with DL to select the best model. **This is an unsolved problem** [1]. As we engineer the framework to identify a model that performs well in discriminating between diseases, we will use the model itself to reverse engineer relationships between the input data.

Goals and Impact

With the increased availability of multi-omics data for various types of cancer and the advancement of analytical tools, particularly DL frameworks to analyze large sets of heterogeneous data, we set out to benchmark the application of these tools to discriminate between cancer types. Specifically, our goal was to apply state-of-the-art DL techniques to understand how high-dimensional, heterogenous data can bias a model, and how DNN topology affects classification. A thorough understanding of the mapping between these ‘omics datasets and model performance will inform how to effectively use these models to identify appropriate biomarkers and enable Precision Medicine.

Data

The source of data was the Cancer Cell Line Encyclopedia (CCLE), a non-uniform, open dataset containing DNA copy number, single-nucleotide polymorphism (SNPs), indels (Affymetrix SNP6.0 arrays) and messenger ribonucleic acid (mRNA) expression (Affymetrix U133+2 arrays) for 18,889 genes across 1,036 cancer cell lines. There were no controls in this dataset. Figure 1 below shows counts of cancer types across the 1,036 lines.

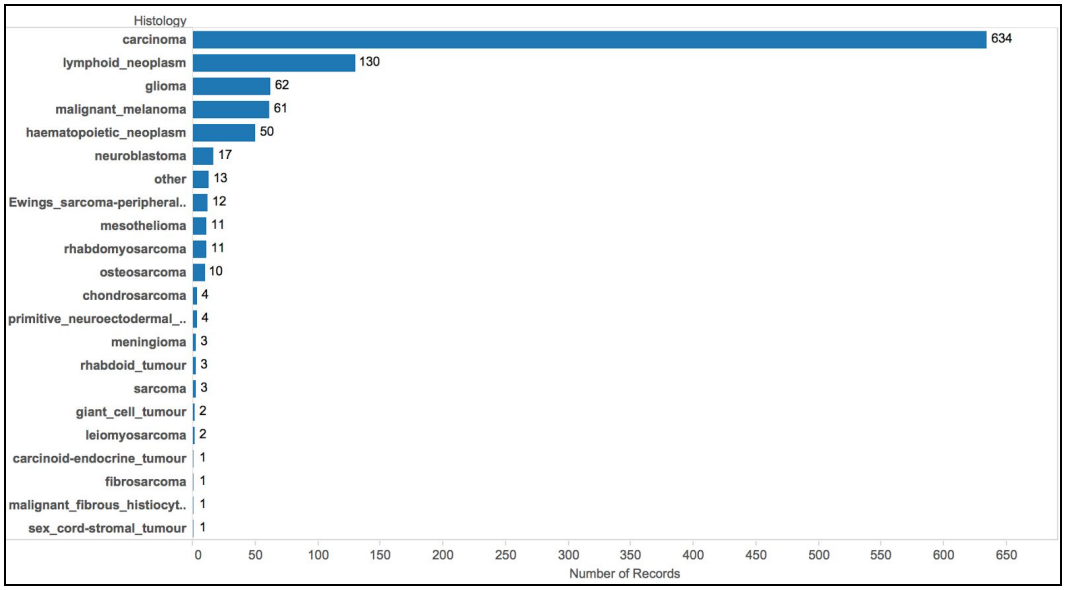

Figure 1: Number of cancer types by record in CCLE.

We focused our analysis on cancer classification at the transcription level because its characteristics were best suited for our effort as compared to cancer subtype. The data is stored as Gene Cluster Text (*.gct), files. These files are binary representations of a large matrix whose rows are genes and columns are cancer cell lines. We focused on cancer types with 50 or more samples to ensure adequate training examples. This resulted in the desired classification of five cancer types: glioma, carcinoma, malignant melanoma, lymphoid neoplasm, and hematopoietic neoplasm.

Analysis and Results

We built a DL framework to ingest sets of multi-omics data using Google TensorFlow's DNN Classifier. We used the framework to test models of three layers of 15, 20, and 15 hidden units per layer with rectified linear unit (ReLu) activation functions for high compression of input variables, as well as a network with three hidden layers of 1000, 2000, and 1000 neurons per layer. The motivation for selecting these frameworks was to test the performance of the model at two different scales. From cancer types with at least 50 samples, we created dynamic test and training sets, randomly sampling from 80% of the data to train and leaving aside 20% to test. Test and training sets were re-computed for each model run. All models were run for at least 6000 epochs with 10-fold cross validation. The classifier was configured to predict five classes of data (our five cancer types) for four different scenarios: 1) Use of all 18,889 transcripts in data, 2) Use of random selection of 100 genes, 3) Use of another random selection of 100 genes with no overlap to (2), 4) Use of 38 genes that were used as drug targets for drugs in the CCLE database.

Our first test was to determine how a different set of samples can bias the model.

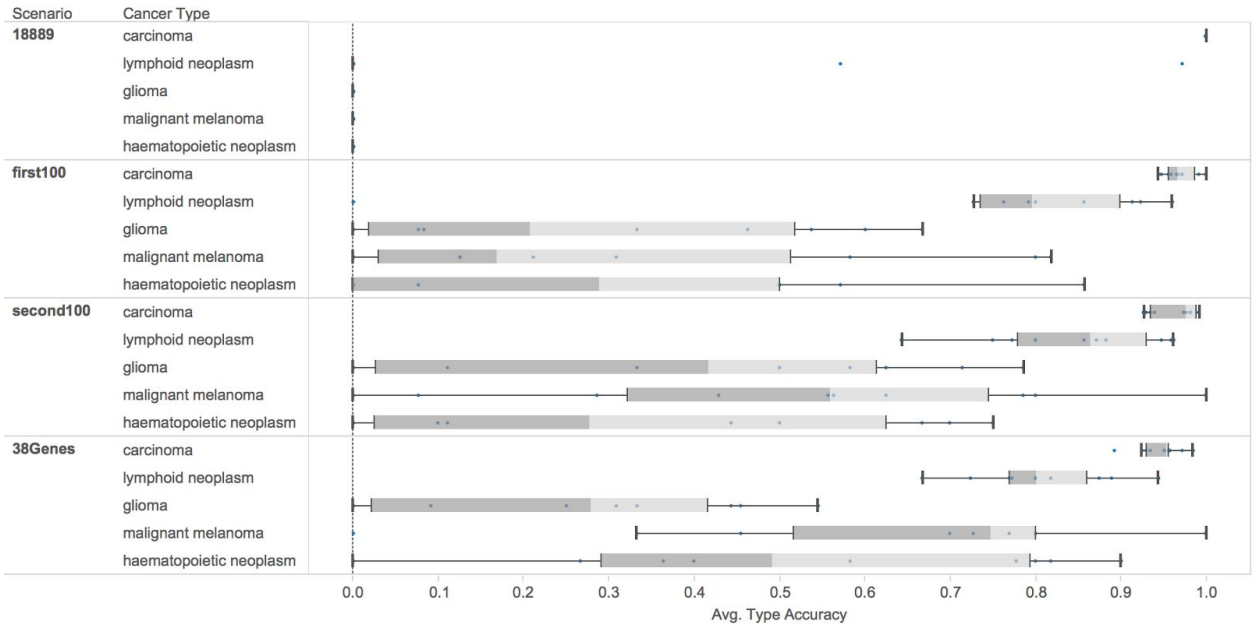

Figure 2: Average classification accuracy for unbalanced datasets using 10-fold cross validation on a Deep Neural Network classifier. Bias toward cancer types with more data and higher dimensions (input genes).

The results show clear indications of overfitting towards cancer types with more data in CCLE, notably carcinoma and lymphoid neoplasm, especially with higher dimensions (input genes). Dimensionality reduction (subselection of input genes) significantly improves the potential for the model to classify the cancer types with smaller training samples (the median accuracy for all cancer types in the case of 18,889 genes is zero, while it grows for the other categories - and has the least overall variance for 38 genes). This implies that in clinical settings that inherently suffer from the problem of imbalanced sets of data, effective dimensionality reduction techniques must be applied for models with high-compression ratios to discriminate between disease types.

We then took the same model and balanced the input datasets to classify between carcinoma and non-carcinoma cancers with the latter three experiments so that we could have a balanced dataset between carcinoma and non-carcinoma samples.

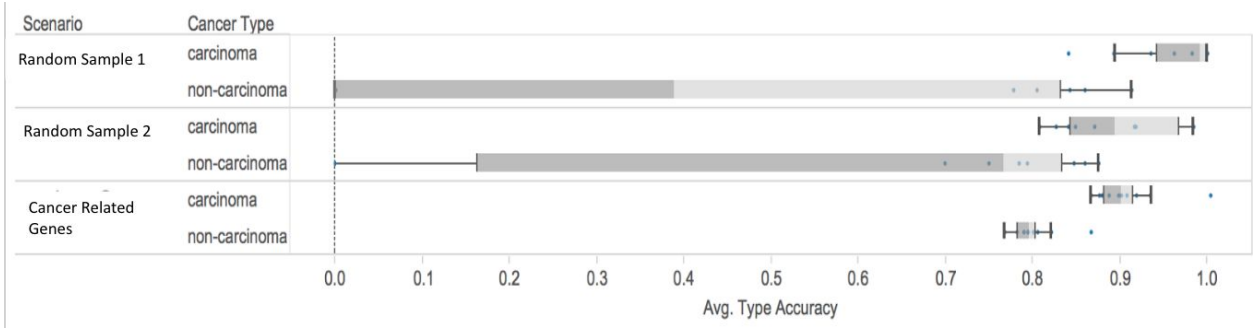

Figure 3: Average classification accuracy and variance from 10-fold cross validation on a Deep Neural Network classifier for Carcinoma/Non-carcinoma classification balanced dataset. Random sampling of transcripts (Random

Samples 1 and 2) and incorporation of prior knowledge through feature selection of cancer related genes are used as input.
Testing/training methodology remained the same - selecting 80/20% within each cancer class. In this test, we saw a significant increase in accuracy as well as a substantial decrease in the variability of the performance of the model. This seems to imply that balancing the dataset and an effective incorporation of prior knowledge can significantly improve the model.

Future Work

Our analysis shows that Deep Neural Networks (DNNs) can play a significant role in understanding relationships between genes that can discriminate between cancers. These genes and their relationships have the potential to act as biomarkers for cancer. The sensitivity in choosing the right data set with the right topology remains an open area of research. Chen, et. al. [2] conducted tests on all permutations of 3000, 6000, and 9000 neurons for a three-layer DNN to achieve accuracies of ~70% with very low variance (0.08). Fakoor, et. al. [3] compared the use of Principal Component Analysis (PCA), sparse, and stacked autoencoders to first reduce the dimensionality of gene expression data and then perform classification. All but one of their 13 datasets were of a single cancer type doing binary classification. Thus, identifying the requirements of selecting the appropriate model (layers, activation function, learning rates) and dimensionality reduction techniques (sparse/stacked autoencoders, PCA, non-negative matrix factorization [NMF]) is **an unsolved problem**. It is our goal to not only develop the appropriate requirements for data modality (genomic, transcriptomic, and proteomic data), but once a model is found to perform well in discriminating between diseases, using the model itself to reverse engineer relationships between the input data. Future work will benefit greatly from larger sets of data with stable controls.

References

[1] Mamoshina, P., Vieira, A., Putin, E., & Zhavoronkov, A. (2016). Applications of Deep Learning in Biomedicine. *Molecular Pharmaceutics*, *13*(5), 1445–1454. http://doi.org/10.1021/acs.molpharmaceut.5b00982

[2] Chen, Y., Li, Y., Narayan, R., Subramanian, A., Xie, X.; *Gene expression inference with deep learning*. Bioinformatics 2016; 32 (12): 1832-1839. doi: 10.1093/bioinformatics/btw074

[3] Fakoor, R., Ladhak, F., Nazi, A., Huber, M. *Using deep learning to enhance cancer diagnosis and classification*. Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR:
W&CP volume 28.