A scalable deep learning framework for breast cancer prediction using DNA methylation data.

Nijhum Paul[1], Rick Jansen[1], Rahul Gomes[2], Nichole He[2], Aaron Huber[2]
[1] North Dakota State University, Public Health, [2] University of Wisconsin-Eau Claire, Computer Science

BACKGROUND

- DNA methylation is a key epigenetic modification that can modulate gene expression to influence the cell functionality.
- This process often affects tumor suppressor genes and oncogenes leading to cancer.
- DNA methylation can be measured by high-throughput sequencing technology that is able to read methylation markers (CpGs) across the majority of the human genome.
- Patterns in CpG markers can be used to improve cancer prediction accuracy.
- Feature engineering and deep learning[1] methods have shown promise in creating successful prediction models of cancer using methylation data.

DATASET

- We obtained methylation values from 1188 normal and tumor samples of the Breast Invasive Carcinoma project using Illumina 27K and 450K platforms from GDC Data Portal.
- Total number of CpG markers is 487177.
- Sample size is 1188.

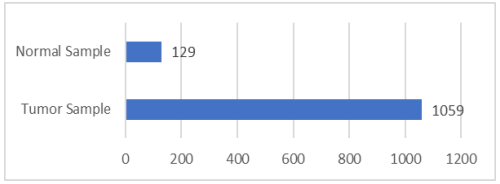

Figure 1. Distribution of tumor and normal samples.

METHODS

Handling null value

- As machine learning algorithms don't work well with 'no data', we removed CpG markers with more than 20% NaN values since statistical tests could only validate up to 30% of markers with no data imputed[2].
- We then applied mean imputation on CpG markers with less than 20% NaN values separately on the cancer and non-cancer samples based on their mean values to preserve class-based variation.
- We arrived at 23378 significant CpG markers for the 1188 samples.

Handling class imbalance

- The class imbalance of tumor and normal samples is very high (89.1% and 10.9% respectively) that can cause biased prediction and misleading accuracy .
- We used Synthetic Minority Oversampling Technique (SMOTE) for oversampling the minority class giving us a 1:1 ratio of tumor and normal samples.

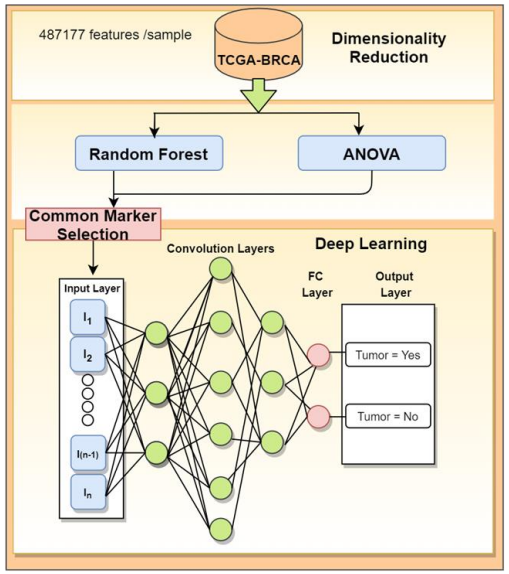

Figure 2. Workflow for the final data processing procedure used.

Feature selection

- We trained **ANOVA F-test** model on 23378 features.
- Markers with a p-value > 0.05 were removed from the total features, resulting in 7284 remaining.
- The 7284 features are applied to the **random forest** model and the number of remaining markers is 884.

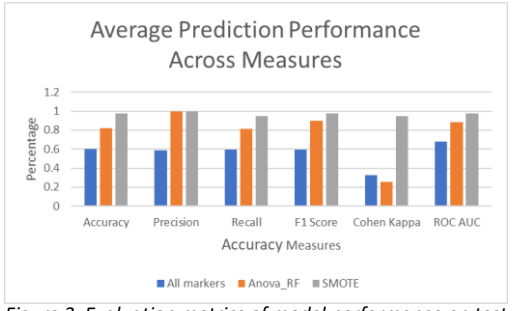

Figure 3. Evaluation metrics of model performance on test dataset based on if feature selection processes and oversampling were performed or not.

Deep learning

- Deep learning model as shown in Figure 2 with three hidden layers was then applied on three separate datasets to ensure the effectiveness of preprocessing steps, which are
  - Complete number of 23378 features.
  - Reduced features after applying ANOVA F-test and random forest feature selection models.
  - Reduced features after applying SMOTE, ANOVA F-test and random forest feature selection
- The model was trained for 30 epochs with 70-30 train-test split.
- Results are displayed in Figure 3

Gene Set Enrichment Analysis (GSEA)

- GSEA was performed on the genes associated with the reduced set of 884 Breast Cancer CpG markers.
- GSEA organizes groups of the significant genes into those processes, components, functions and pathways most related (high –log10 FDR) to the development of breast cancer.
- One of the most significant genes in cellular component associated with breast cancer is cytoplasmic membrane-bounded vesicle (Figure 4).

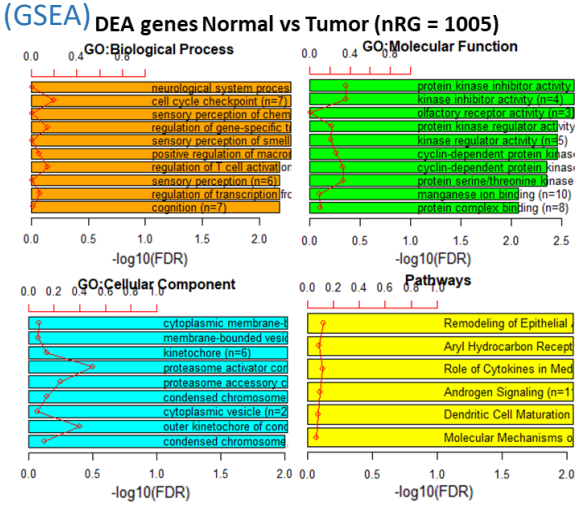

Figure 4. GSEA on gene set of reduced markers.

Discussion

- We observed that the model performance was significantly better using selected features and more so with the balanced dataset produced by SMOTE (Figure 3).
- We can conclude that integrating feature engineering, oversampling, with deep learning model provides better performance in predicting breast cancer using methylation data.

References

1. B. Liu, Y. Liu, X. Pan, M. Li, S. Yang, and S. C. Li, "DNA Methylation Markers for Pan-Cancer Prediction by Deep Learning", 2019 Oct.
2. P. Di Lena, C. Sala, A. Prodi, and C. Nardini, "Missing value estimation methods for DNA methylation data," Bioinformatics, vol. 35, no. 19, pp. 3786–3793, Oct. 2019.

Acknowledgements

- Funding support NIH grant (P20GM109024) and NSF MRI Award #2019077.
- The computational resources of the study were provided by the Blugold Center for High-Performance Computing under NSF grant CNS-1920220.