Early Stage Cancer Detector: Identifying Future Lymphoma Using Epigenomics Data

Category: Life Sciences

Ayush Agarwal (ayush), Sai Anurag Modalavalasa (anuragms), Sarah Egler (segler)

Stanford
CS229, Fall 2018

Objective

DNA methylation is an epigenetic process affecting gene expression which has been linked to cancer. We use this biomarker to **classify future Lymphoma,** a group of cancers beginning in white blood cells of the immune system. The Optimizing Metric for the classification model is **F1 Score**.

Data

DNA methylation

- 566 blood samples from two cohorts (m)
- 444,000 genomic probes (n)

y	sex	A_23_P100001	A_23_P100011	A_23_P100022	A_23_P100056
1	1	6.402494	5.799493	3.447526	5.439588
1	2	6.786831	3.320382	4.830281	4.955008

Immune cell fractions

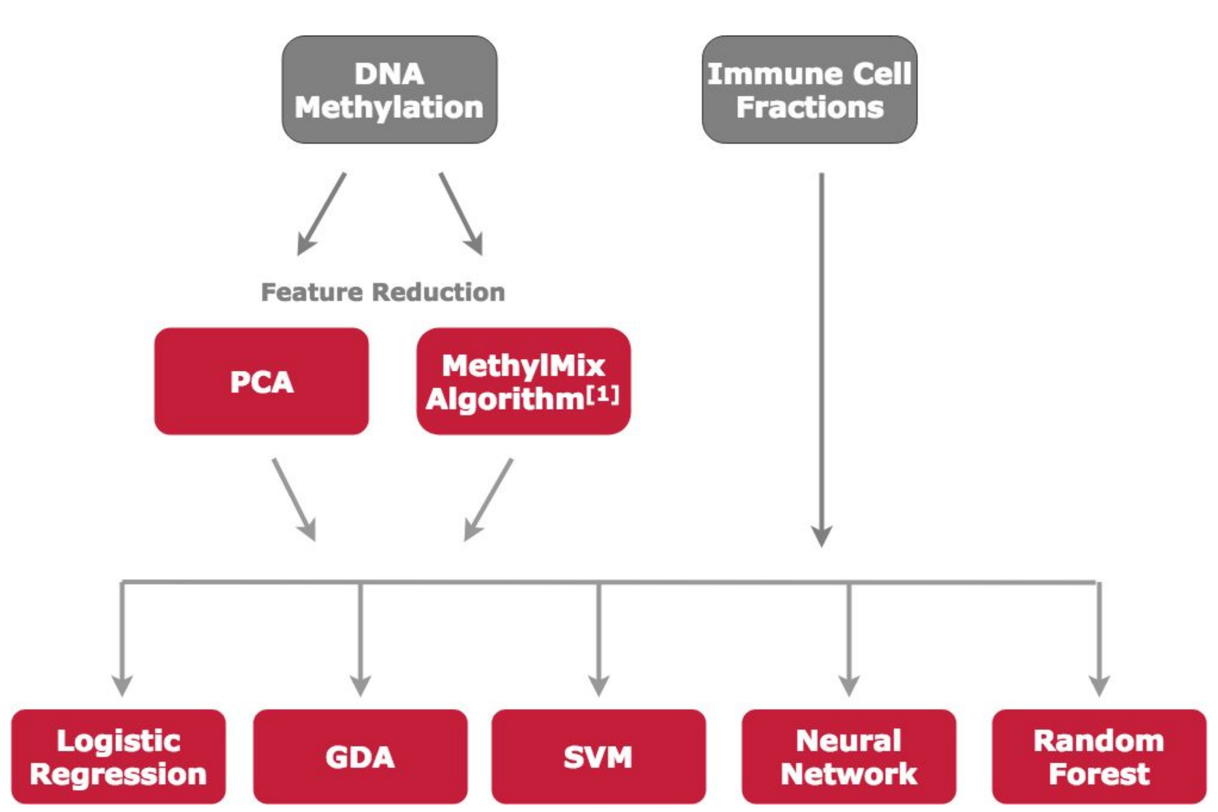
- 196 blood samples (m)
- 23 fractional components of blood (n)

Y	B.cells.naive	B.cells.memory	Plasma.cells	T.cells.CD8	T.cells.CD4.naive
0	0.0	0.030000	0.027737	0.012134	0.090172
0	0.0	0.036672	0.025804	0.001736	0.000000

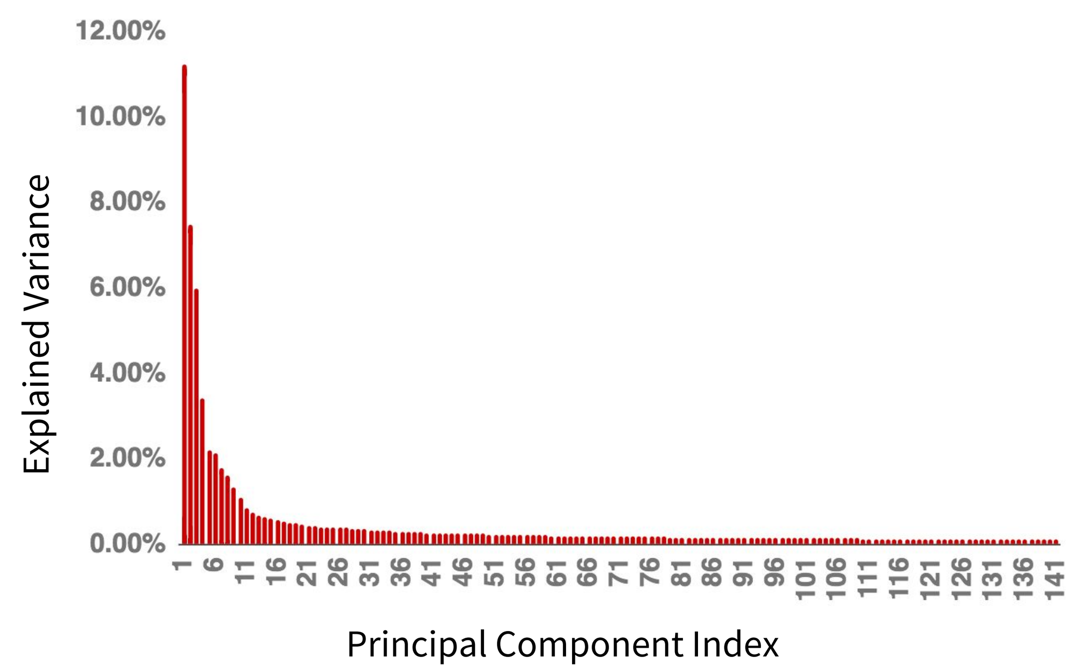
Challenges : Small data set, biological noise

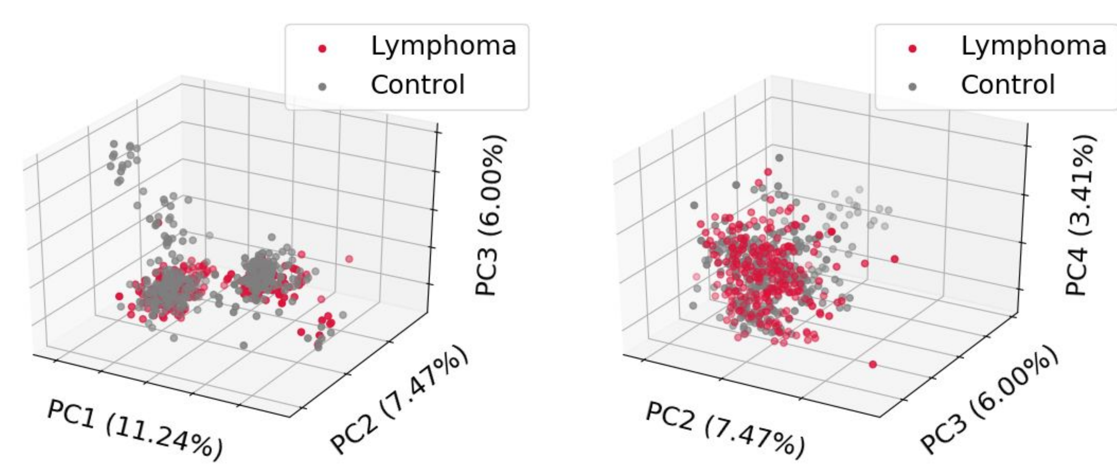
Methods

DNA methylation data is noisy with correlation across gene probes. **Feature selection** and **normalization** techniques are useful precursors to supervised learning techniques.



Feature Selection Techniques





Classification Models

Logistic Regression

$\phi(z) = log(1 + e^{(-z)})$

Techniques: L2 Regularization, Ensembling, k-Fold Cross Validation

Support Vector Machines

$\phi(z) = \max(1 - z, 0)$

Techniques: Polynomial, Gaussian RBF Kernel, Regularization

Gaussian Discriminant Analysis : MLE
Techniques : Box-Cox Transforms

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln y_i & \text{if } \lambda = 0, \end{cases}$$

Neural Networks

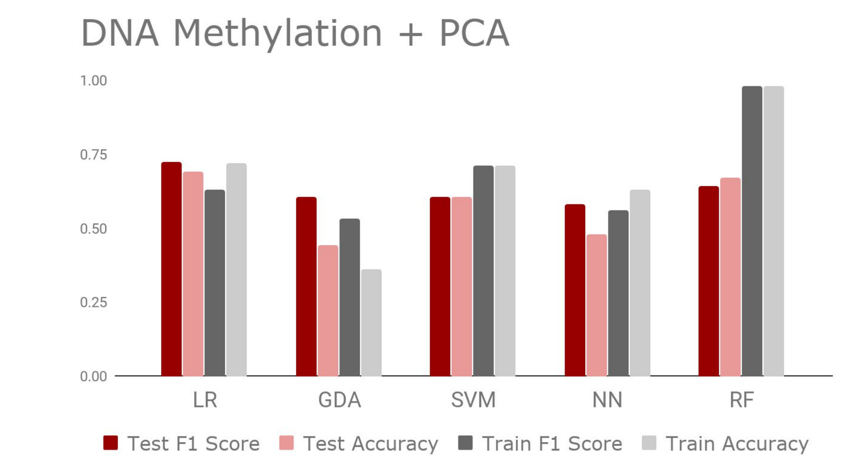
$J(y, \hat{y}) = -(Ylog(\hat{y}) + (1-y)\ log(1-\hat{y}))$

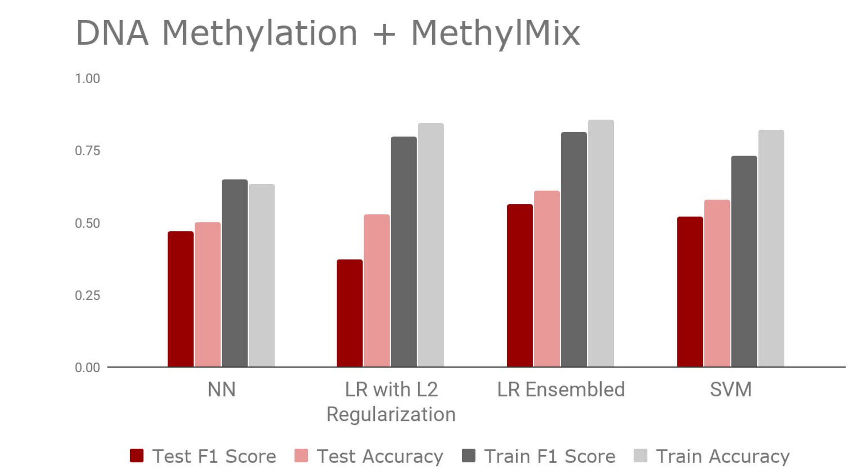
Techniques: ReLU, Sigmoid activations, L2 Regularization, Drop Out, Early Stopping, Learning Rate Decay, Adam Optimizer, Hyperparameter Tuning, Threshold Variation

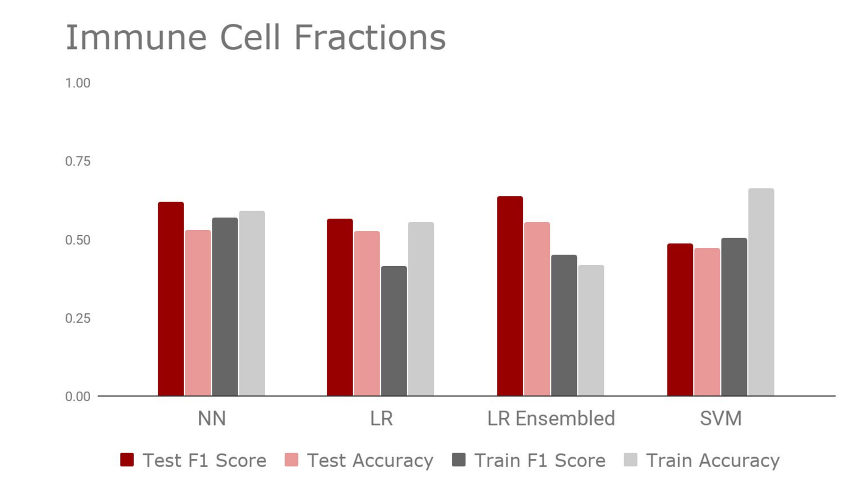
Random Forest : Gini Loss $\sum_c \hat{p}_c(1 - \hat{p}_c)$

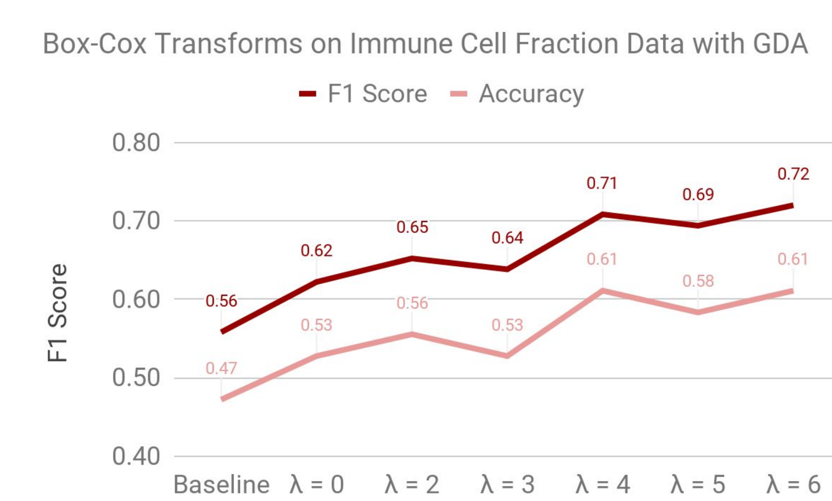
Techniques : Boosting, Hyperparameter Tuning

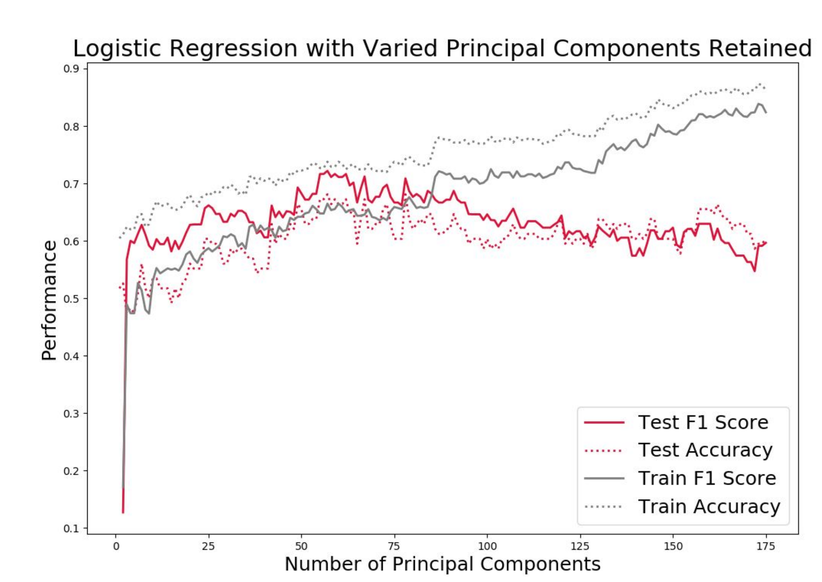
Results











Discussion

Best Lymphoma Predictor
Dataset : DNA Methylation
Features : First 59 Principal Components
Model : Logistic Regression
Test F1 Score : 72%, Test Accuracy : 69%

PCA outperforms MethylMix algorithm

The MethylMix Algorithm is used to identify disease related hyper- and hypo-methylated genes[1]. However, all models performed better on DNA Methylation + PCA dataset as compared to DNA Methylation + MethylMix dataset. PCA may be a better feature reduction technique in the context of lymphoma detection.

Immune Cells Fractions: Transform induces normality

GDA works well if the data is Gaussian. GDA's performance improved with when x was transformed using Box-Cox transforms. Normalization of the data is useful given the biological noise.

Bias Variance Trade-Offs

1. Logistic Regression: Ensembling reduced variance
2. Neural Networks, Random Forests: High Variance
3. GDA: High Bias, best model for immune cell fractions dataset (small dataset) with power transform

Future

- Model and dataset ensembling
- Pair with microRNA expression data
- Map feature importance to genes
- Softmax classification of lymphoma subtypes

References

[1] P.-L. Cedoz, M. Prunello, K. Brennan, and O. Gevaert, "MethylMix 2.0: an R package for identifying DNA methylation genes," Bioinformatics, vol. 34, no. 17, pp. 3044–3046, 2018.

Acknowledgements

Thanks to Dr. Almudena Espin Perez in the department of Biomedical Informatics for the data and mentorship.