Early Stage Cancer Detector: Identifying Future Lymphoma Using Epigenomics Data

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

***Abstract*— DNA Methylation is an epigenetic process affecting gene expression which has been linked to cancer [4]. A combination of supervised and unsupervised machine learning techniques have been implemented on epigenomics datasets to build a classification model that can predict whether a person will develop lymphoma (a group of cancers beginning in white blood cells of immune system) in the future. An F1 score of 72% and accuracy of 69% have been obtained on test dataset using a combination of PCA (with the projections of dataset on the first 59 principal components) and logistic regression.**

I. INTRODUCTION

Genome-wide methylation was first associated with future lymphoma by Georgiadis et al. in 2017, who found that epigenetic changes are already present in blood samples 2.1 to 15.9 years prior to diagnosis [4]. As a result, if such epigenetic pattern changes can be observed in blood samples, cancer can be detected years prior to diagnosis, increasing the likelihood that patients will receive better medical aid. This possibility of impacting the lives of those likely to develop lymphoma was a natural motivator in tackling such a big problem. The goal of this study was to build machine learning models to predict future Lymphoma. We used two main biomarkers that have been linked to the pathogenesis of cancer: DNA methylation and fractional components of immune cells.

DNA methylation is an epigenetic process whereby methyl groups are added to DNA molecules without changing the DNA sequence itself, typically acting to suppress gene expression. Measures of fractional components of immune cells are derived from gene expression. Together, these biomarkers provide insight into the differential expression of genes and the pathogenesis of lymphoma and were used as parallel inputs to our problem [4].

As on date, it is not possible for medical experts to look at the data and predict the likelihood of a person having lymphoma in future. As a result, the Bayes limit is currently unknown for this problem.

The input to the algorithm includes 1) the DNA methylation levels (floats) across different genomic probes and 2) fractional components of immune cells (floats representing the fraction of each component). We use unsupervised feature reduction along with several supervised learning techniques (logistic regression, SVMs, GDA, neural networks, and random forests) to output binary predictions of future lymphoma.

II. RELATED WORK

Georgiadis et al. tried to perform pathway analysis to identify the relevant genes and underlying biology pertaining to lymphoma[4]. They also perform several supervised and unsupervised learning techniques to assess classification accuracy. As in our case, they implemented multiple classification models in order to find the best, including SVMs with both gaussian and linear kernels as well as random forests. While we use GDA as a generative model, they use the Naive Bayes classifier.

Feature selection is one of the main objectives in genomic data used for disease classification. In fact, the number of genes needed for discriminant analysis in disease classification is likely much lower than 50 [2][7]. The challenge of feature selection in similar genomics data has given rise to many novel approaches. The MethylMix algorithm is a relatively recent approach which identifies disease-specific hyper- and hypo-methylated genes using a beta mixture model [3][5]. The novelty of MethylMix lies in the metric of differential, as opposed to absolute levels of methylation in cancer. Many approaches to feature selection seem to have in common that they take into consideration biological relevance. For example, Georgiadis et al. implemented PCA as a feature reduction technique in conjunction with the identification of biologically relevant genes found via a separate model. A biologically driven approach may be extremely powerful if the assumptions of the model fit.

III. OBJECTIVE

Since recall is an important parameter to understand the effectiveness of model, based on discussions with our mentor in the Department of Bioinformatics, we selected F1 score as the optimization metric for the classification model. Obtaining an accuracy of 50% is a satisfying baseline metric for the model.

IV. DATASET AND FEATURES

A. Dataset

The data used in this study was obtained from Stanford Department of Biomedical Informatics. There are three datasets: 1) DNA methylation, 2) Immune Cell Fractions, 3) MethylMix + DNA Methylation. The DNA methylation data set contains 566 pre-diagnostic blood samples (m) with 444,000 features (n). The features are methylation levels across 444,000 genomic probes given as M-Values, which are logarithmic ratios between methylated and unmethylated signals. The Immune Cell Fractions dataset contains 196 pre-diagnostic blood samples (m) with 23 features (n). The features are fractional values of various components of these immune cells. The MethylMix + DNA Methylation dataset is a subset of blood samples and features from the original

DNA methylation dataset. There are 196 blood samples (m) with 101 features. The features have been selected from the original DNA methylation levels using the MethylMix algorithm, which identifies probes with disease related hyper- and hypomethylated states. The selected features represent differential levels of DNA methylation at these probes. The set of 566 examples spans two cohorts with 234 total cases of future lymphoma, while the sets of 196 examples include 76 cases of future lymphoma. Each of the three datasets were split into train/validation/test sets of approximately 70/10/20.

B. Feature Selection Techniques

The small number of blood samples, large number of features in the DNA methylation dataset, and inherent biological noise present a set of challenges common across genomic applications of machine learning. Moreover, DNA methylation levels are correlated across gene probes. It is therefore desirable to capture the essence of the DNA methylation dataset in a smaller feature space prior to applying supervised learning techniques. Two feature reduction techniques are used in parallel: MethylMix (data provided) and PCA. We then applied the following supervised learning models to the three datasets: logistic regression, GDA, SVMs, neural networks, and random forests.

Application of the MethylMix algorithm on the DNA methylation dataset reduced the number of features from nearly half a million to 101. The objective of the MethylMix algorithm is identification of disease specific hyper/hypo methylated genes [3][5]. However, it is not certain whether this is the best feature selection technique for lymphoma prediction. Therefore, we also implemented PCA on the original DNA methylation dataset to reduce the number of features and biological noise associated with the data, retaining 70% variance. We experimented with higher variance retention but found 70% satisfactory given that fewer genes are likely necessary in this problem [2][7]. Most of the meaningful information corresponding to the original matrix can be captured using the first several principal components.

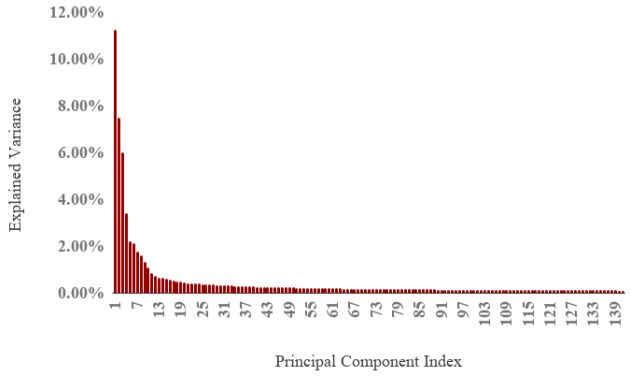

Fig. 1. Explained Variance vs Principal Components Index Plot (70% Variance obtained by using 176 Principal Components, out of which 141 had more than 0.1% variance

As shown in figure 1, most of the information can be extracted using a few principal components. To visualize the data, plots were made by taking projects of the data on first two principal components. It was found that drawing a decision boundary was not possible using the projections of the DNA methylation data on first two principal components. Separability increases when three principal components are used instead (Figure 2A). It is highly likely that separability increases as the number of principal components goes up. However, owing to the limited number of examples, using extremely high number of principal components will force us to operate in null space. This will result in overfitting and variance related problems. Thus, the number of principal components to be used has been treated as a hyperparameter while tuning the models.

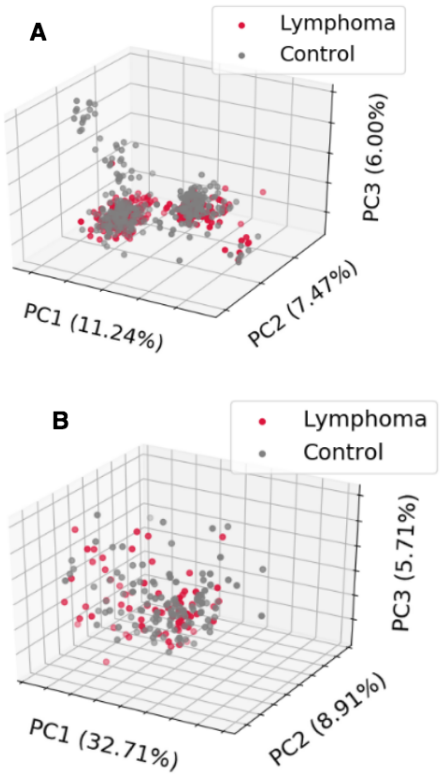

Fig. 2. **A)** The first three principal components with PCA applied to the DNA methylation dataset, **B)** The first three principal components with PCA applied after MethylMix

Given the success of PCA as compared to MethylMix, we decided to see if the combination of the two would yield better results, to visualize the data better and understand if the number of parameters can be reduced. We applied PCA to the dataset output from the MethylMix algorithm, retaining 95% variance with 49 principal components. However, drawing a decision boundary is not possible using the first two components of MethylMix data. Further, separability associated with MethylMix data for the first 3 principal components was lower as compared to separability associated with DNA methylation data (Figure 2B). Because the number of features has already been reduced by the application of MethylMix algorithm, further parameter reduction might

result in loss of relevant information. We also had far fewer examples for the MethylMix methylation dataset. Hence, to avoid bias, we decided to use the entire column space corresponding to MethylMix data while applying algorithms.

V. Classification Models and Methods

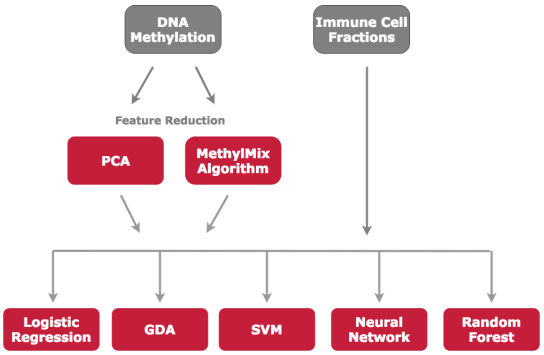

Fig. 3. Overview of methods used.

A. Logistic Regression

Logistic regression model works well as a baseline. It is easy to implement and usually gives good insights. Since we began with limited information about the distribution, we felt the application of a logistic regression model was a good way to get started quickly and iterate upon. This decision was further supported by the performance of logistic regression in conjunction with parameter reduction in cancer detection problems with similar data [2][6][7]. Given the relatively small size of our dataset, we used Newton's method with the logistic loss function (Equation 1).

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

L2 regularization and ensembling techniques were used to address the relatively small size of the dataset and overfitting when training in a high dimensional feature space. Ensembling techniques used include traditional bagging as well as a more novel and linear approach in which we simply average the model weights for k-fold training samples.

B. Gaussian Discriminant Analysis

Gaussian Discriminant Analysis is a generative learning model that models the probability of the data given the labels, as opposed to modeling the probability of the labels given the data. GDA thus models attributes of the biomarkers for the disease versus healthy state and uses this model to predict future lymphoma given an unlabeled sample. GDA models usually perform better if the distribution is Gaussian in nature and if the number of examples is low. Since the number of examples is low in our case and since we did not know anything about the distribution, we decided to apply GDA. Our mentor has suggested the data likely has a bimodal distribution, and that it is likely that non-gaussian statistical models will perform better on any DNA methylation data [8]. Hence, we decided to apply power transform techniques to induce normality [1]. Box-Cox Transforms (Equation 2) have improved the performance of model using the maximum likelihood estimation for lambda.

$$x_i^\lambda = \begin{cases} \dfrac{x_i^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ log(x_i) & if \lambda = 0 \end{cases} \tag{2}$$

The model's parameters have been learned using Maximum Likelihood Estimation (MLE). Using the MLE parameters, predictions are made on the test set to test the effectiveness of the model.

C. Support Vector Machines

Visualization of data plots using principal components (Figure 2) gave us an impression that the data can be separated better using kernels to map the first several principal components into higher dimensions. The intuition was further supported by the fact that SVMs were used by Georgiadis et al. when they tried to build classification models for this problem[4]. Building on the intuition we developed after visualization of data plots using principal components, we implemented SVMs using polynomial, Gaussian RBF, and linear kernels. Hinge loss function was used (shown in equation 3).

$$\phi(z) = max((1 - z), 0) \tag{3}$$

Gaussian and high dimensional polynomial kernels faced overfitting problems. We then tried a linear kernel, as in Georgiadis et al., which performed best. The performance was improved after tuning regularization and gamma margin parameters using validation data. After learning the parameters, the model was tested on test set.

D. Neural Networks

Since the logistic regression model gave good results, we felt neural networks might extract deeper information and give even better results. The number of parameters corresponding to neural networks are higher than logistic regression model and the number of samples available was limited, so we decided to compensate by using fewer principal components. Sigmoid activation layer was used for the final layer; ReLU and sigma activation functions were tried for hidden layers. Weighted binary cross entropy loss function, shown in equation 4, was implemented with Adam Optimizer (mini batch gradient descent) using TensorFlow and Keras frameworks[10][11]. A weighted binary cross entropy loss function was used to tackle the slight data imbalance problem. Since recall is a very important parameter for the model, application of weighted binary cross entropy loss is justified; it penalizes the model if an actual true is predicted as false. The threshold was selected empirically based on the performance use the F1 metric on the training set.

$$J(y, \hat{y}) = -(Wylog(\hat{y}) + (1 - y)log(1 - (\hat{y}))) \tag{4}$$

Increasing the number of layers improved the training accuracy but ran into over fitting problems; the model architecture was tuned to reduce variance. Further, to combat the overfitting problem, different regularization techniques such as L2 regularization (for kernels), early stopping, learning rate decay, and drop out were used. The hyperparameters tuned include the weight in weighted cross entropy loss function, threshold, learning rate, learning rate decay, number of epochs, activation function for hidden layers.

E. Random Forests

Decision trees are another useful model for non-linear decision boundaries. However, decision trees are high variance models prone to overfitting, as can be imagined by a tree where there is a distinct leaf for each training example. Random forests are useful techniques for bagging decision trees by training a bootstrapped sample on each tree and averaging these models. We applied random forest models with Gini Loss function, shown in equation 5, which similar to the cross-entropy loss function for decision trees can help to maximize the information gain from one level of the tree to the next.

Similar to our neural network model, given the high variance nature of decision trees and the limited number of samples available, we trained the model using fewer principal components as compared to other models.

$$L_{gini} = \sum_{c} (\hat{p}_c) * (1 - \hat{p}_c) \tag{5}$$

The hyper parameters tuned include minimum leaf size and maximum features considered in order to reduce overfitting that was quite apparent. Application of AdaBoost improved the performance of the model by weighting misclassified examples during training in attempts of creating a stronger learner out of a set of weaker learners.

VI. RESULTS, INFERENCES AND DISCUSSION

A. Best Predictor Combination

The best combination of model and dataset for lymphoma detection was logistic regression with L2 regularization on the DNA Methylation + PCA dataset with 59 principal components (55% variance), achieving an F1 score of 72% with 69% accuracy (Figure 5). With an increasing number of principal components, logistic regression overfits the data even with L2 regularization (Figure 6).

B. PCA outperforms MethylMix algorithm

All the classification models performed better on the DNA methylation dataset where feature selection has been done using PCA as compared to the dataset where feature selection has been done using MethylMix algorithm (Figures 4A and 4B). PCA may capture the information better than the MethylMix algorithm in the context of lymphoma detection. Applying PCA to the reduced MethylMix dataset did not improve the best model, falling short with a 61% F1 Score and 50% accuracy in comparison. While MethylMix might not be a good fit, the poor performance could also be a

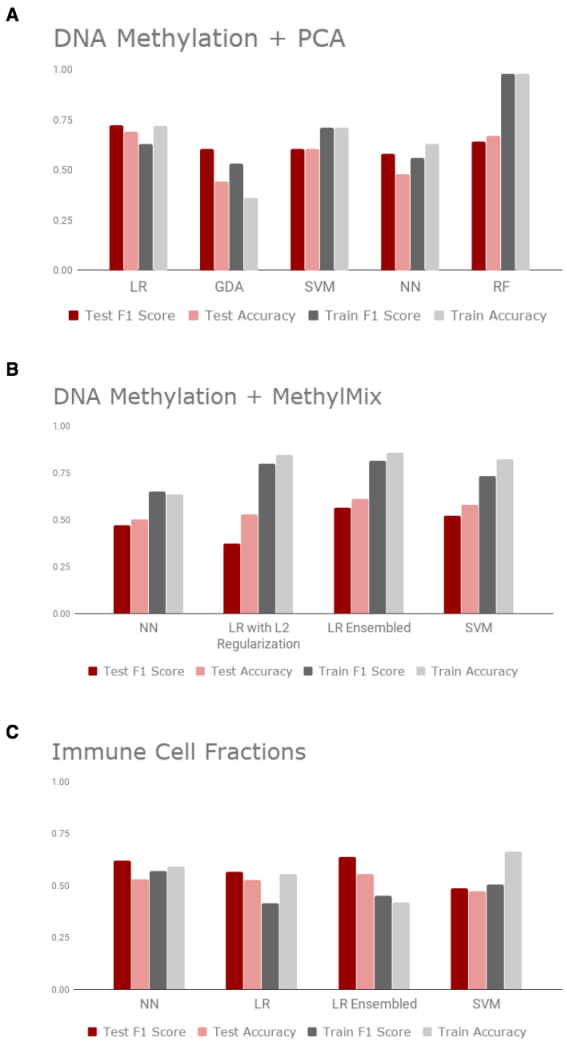

Fig. 4. Best performance of each model on **A)** the DNA Methylation data with PCA, **B)** The DNA Methylation data with MethylMix, **C)** The Immune Cell Fractions data

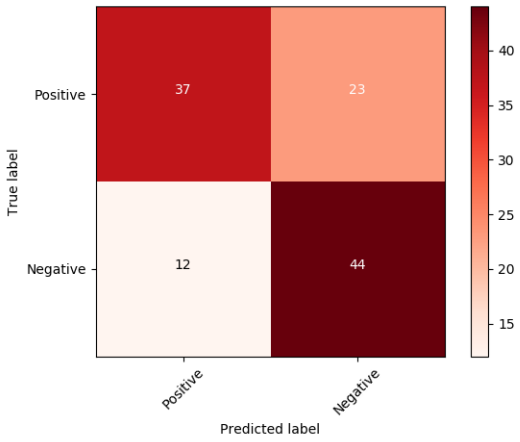

Fig. 5. Confusion matrix of logistic regression on the DNA methylation + PCA dataset using the first 59 principal components.

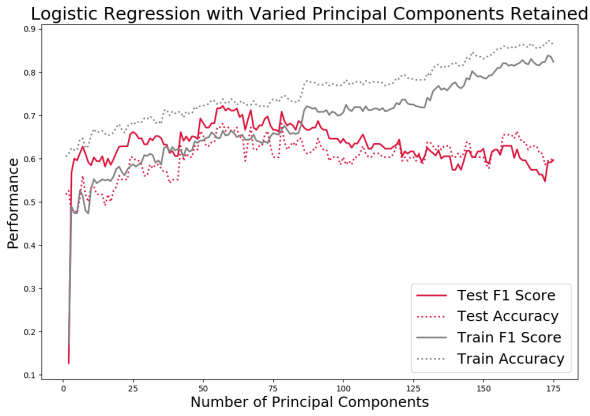

Fig. 6. Logistic regression with L2 regularization $\lambda = 0.01$ run on the DNA methylation + PCA dataset varying the number of the first principal components used. Overfitting occurs with increasing principal components.

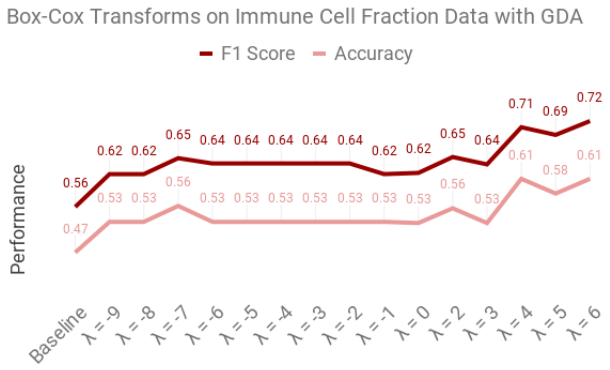

Fig. 7. Peformance of GDA on the Immune Cell Fractions dataset with Box-Cox Transformations with varied values of λ compared to the baseline GDA on this dataset with no transformation. The MLE estimate of lambda is -9.

result of the smaller size of the MethylMix reduced dataset (196 examples) compared to the PCA reduced dataset (566 examples).

C. Immune Cells Fractions: Transform induces normality

GDA works better if the input data is gaussian in nature. GDAs performance was improved when the input data corresponding to Immune Cells Fractions database was transformed using Box-Cox transforms (Figure 7). Thus, it can be inferred that Immune Cells Fractions dataset is inherently not Gaussian and that transforms, most likely, induce normality.

D. Bias Variance Analysis

Overfitting was a common problem that most models faced in the methylation datasets. Ensembling techniques reduced the variance pertaining to the logistic regression model. The variance problem was bigger in models involving Neural networks and Random Forests; several regularization techniques were implemented to reduce this variance (at the cost of reduced accuracy). Regularized Logistic regression, most likely, finds the right balance in bias variance trade-off and hence performed the best for this data set. A larger number of samples and optimized feature selection technique may help to overcome this variance in the future.

VII. CONCLUSION/FUTURE WORK

Logistic regression model gave the best results after the number of features has been reduced using PCA techniques. While neural networks could capture the information better and perform better on training datasets, they ran into overfitting problems and several regularization techniques have been implemented to address this problem. Similarly, random forests performed nearly as well as logistic regression models, but faced the overfitting problem. Logistic regression model likely performed better than GDA on the methylation dataset, because of the fact that methylation data does not follow a Gaussian distribution if it is not transformed. However, GDA performed almost as well logistic regression on the immune cells dataset, which is smaller and may have an underlying Gaussian distribution which can be revealed by de-noising and power transformations.

Future work should explore application of model and dataset ensembling techniques as they might reduce the variance and help in obtaining better results. Application of similar models and ensembling techniques on an available microRNA expression dataset can also, probably, help to attain better results and obtain new insights as this is a related biomarker. When the number of samples in the dataset grows over a period of time, classification of lymphoma subtypes using a softmax algorithm will be an interesting problem to tackle. Ultimately, we would like to map the principal components that were important predictors back to the corresponding genes for biological experts to understand the underpinnings of this disease.

ACKNOWLEDGMENT

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

CONTRIBUTIONS

Name	Contribution
Ayush	Data Pre-Processing, SVM
Anurag	Data Pre-processing, Structuring the Machine Learning Project, Metrics Code, PCA, Neural Networks, Analysis of Project, Figures, Poster, Report
Sarah	Data Pre-Processing, Structuring the Machine Learning Project, PCA, Logistic Regression, GDA, Random Forests, Analysis of Project, Figures, Poster, Report

CODE

Github Link: https://github.com/sarahegler/CS229-LymphomaDetection

REFERENCES

[1] G.E.P. Box and D.R. Cox, An Analysis of Transformations, Journal of the Royal Statistical Society B, 26, 211-252 (1964).

[2] J. Liao and K.-V. Chin, Logistic regression for disease classification using microarray data: model selection in a large p and small n case, Bioinformatics, vol. 23, no. 15, pp. 19451951, 2007.

[3] O. Gevaert, R. Tibshirani, and S. K. Plevritis, Pancancer analysis of DNA methylation-driven genes using MethylMix, Genome Biology, vol. 16, no. 1, p. 17, 2015.

[4] P. Georgiadis, I. Liampa, D. G. Hebels, J. Krauskopf, A. Chatziioannou, I. Valavanis, T. M. D. Kok, J. C. Kleinjans, I. A. Bergdahl, B. Melin, F. Spaeth, D. Palli, R. Vermeulen, J. Vlaanderen, M. Chadeau-Hyam, P. Vineis, and S. A. Kyrtopoulos, Evolving DNA methylation and gene expression markers of B-cell chronic lymphocytic leukemia are present in pre-diagnostic blood samples more than 10 years prior to diagnosis, BMC Genomics, vol. 18, no. 1, 2017.

[5] 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. 30443046, 2018.

[6] X. Zhou, K.-Y. Liu, and S. T. Wong, Cancer classification and prediction using logistic regression with Bayesian gene selection, Journal of Biomedical Informatics, vol. 37, no. 4, pp. 249259, 2004.

[7] W. Li and Y. Yang, How Many Genes are Needed for a Discriminant Microarray Data Analysis, Methods of Microarray Data Analysis, pp. 137149, 2002.

[8] Z. Ma, A. Teschendorff, H. Yu, J. Taghia, and J. Guo, Comparisons of Non-Gaussian Statistical Models in DNA Methylation Analysis, International Journal of Molecular Sciences, vol. 15, no. 6, pp. 1083510854, 2014.

[9] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[10] Martn Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Man, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Vigas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[11] Chollet, François and others. Keras. 2015.