

# Machine Learning Classification of Kidney and Lung Cancer Types

Vivek Jain, Weizhuang Zhou, Yifei Men  
Stanford University

Cancer type identification is often critical in disease management and extending life expectancy of patients. Conventional classification through pathological analysis is invasive and costly. With advancements in microarray tools, whole genome methylation profiling can now be performed quickly and cheaply. Since DNA methylation profiles are known to be different between normal and cancerous cells, they hold promise as a scalable avenue for cancer classification. In this project, we used machine learning algorithms to classify 4 different kidney and lung cancer types based on their methylation profiles, and show that our two best performing models have an accuracy exceeding 90%. We further demonstrate that this high level of prediction accuracy can be achieved with only 16 (transformed) features, a 30,000-fold reduction in feature space as compared to raw input from genome-wide methylation information.

## INTRODUCTION

Cancer cells exhibit a number of characteristics distinct from their healthy cellular counterparts due to misregulation of specific cellular pathways. These changes manifest in differential expression of genes. Cancer classification became arguably the first application of machine learning in medicine at the turn of the century, when both supervised and unsupervised learning were applied on gene expression data. These approaches achieved considerable success in discriminating between types of cancers<sup>1</sup> and in detection of cancer subtypes with clinical significance.<sup>2</sup> While some of these alterations in gene expression of cancer cells are due to changes in the DNA sequence, methylation changes of DNA molecules are increasingly acknowledged as key contributors.<sup>3</sup> The interest in methylation profiling has also been fueled by modern molecular biology technology that allows for high coverage analysis of DNA methylation sites<sup>4</sup> and profiling of cells at the level of the whole genome.

In this paper, we apply both unsupervised and supervised machine learning methods to whole-genome methylation data for lung and kidney cancers. We demonstrate that (i) methylation profiles can be used to build effective classifiers to discriminate lung and kidney cancer subtypes; and (ii) classification can be performed efficiently using

low-dimensional features from Principle Components Analysis (PCA).

By demonstrating that different cancer types have distinct methylation profiles, the findings of this paper are not only relevant as classification tools, but also set the foundation for targeted and specific epigenomic therapies for cancer. We also narrow down the list of methylation sites that show distinctly different profiles in different cancers, providing the potential for future work on the biological significance of the sites.

## METHODS

### Data

Methylation profiles for the following 4 cancer types: kidney renal clear cell carcinoma (KIRC, n=455), kidney renal papillary cell carcinoma (KIRP, n=156), lung adenocarcinoma (LUAD, n=441), and lung squamous cell carcinoma (LUSC, n=294) were obtained from The Cancer Genome Atlas (TCGA). For every sample, the degree of methylation at 485,577 positions across the whole genome was quantified numerically as beta values, assayed using Illumina HumanMethylation450 Beadchip.<sup>5</sup>

Imputation by mean values was used to fill up missing values in the dataset<sup>6</sup> and quantile normalization was performed to reduce batch-related variance/noise across the different samples, as per standard treatment.<sup>7</sup> A balanced data set was achieved by picking 150 samples randomly for each cancer type, for a total of 600 samples.

### K-Means

K-means was implemented in R using 2 and 4 centroids separately (Fig 2). K-means is a non-parametric, unsupervised machine learning algorithm which clusters given samples to  $k$  centroids by minimizing within-cluster sum of squares, where each  $x_j$  represents a 485,577-dimensional vector containing the methylation profile of a sample, and  $\mu_i$  is the mean of the points assigned to cluster  $S_i$ .

$$\arg \min_s \sum_{i=1}^k \sum_{x_j \in S} \|x_j - \mu_i\|^2$$

The algorithm was run till convergence or a maximum of 30 iterations.

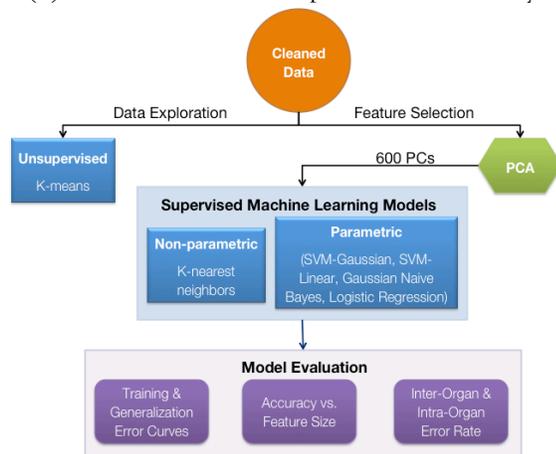


Fig 1. Outline of main processes used in study.

## Feature Selection

Principal Components Analysis (PCA) is a transformation that maps data to new dimensions (principal components, or PCs) that capture the greatest variance when the data is projected on it. PCA was performed on the data using Python's scikit-learn package,<sup>8</sup> yielding a total of 600 principal components.

## Supervised Learning

All of the following supervised learning algorithms were performed using lower-dimensional ( $n = 600$ ) features using scikit-learn packages. Run-time analysis was performed for varying number of features in 2-fold increments (Fig 3). Additionally, KNN and GNB were trained and tested using the full untransformed feature set.

## K Nearest Neighbors (KNN)

KNN classifies each test sample based on the majority label of the  $k$ -nearest neighbors, as determined from Euclidean distances to the test sample. We used  $K=5$  for our model.

## Gaussian Naive Bayes (GNB)

GNB builds a Bayesian probability model based on the frequency of observed features, assuming independence and Gaussian distribution. Laplace smoothing with a smoothing parameter = 1.0 was used.

## Support Vector Machine (SVM)

SVM works by finding points that can be used as "support vectors" to define the classification boundaries between the cancer types. 2 SVM models were generated, using linear and Gaussian kernels respectively. The Gaussian kernel is based on an infinite dimensional feature mapping. Both SVMs were trained with penalty term  $C = 1.0$ . Gaussian SVM was trained with kernel coefficient  $\gamma = \frac{1}{n_{\text{features}}}$ .

## Logistic Regression

Logistic regression models the probability of each of the 4 cancer subtypes as a linear sum of the features using a logistic function. The model was trained using L2-norm regularization with penalty  $C=1.0$ .

## Model Evaluation

For evaluation of generalization error, a random set of 15 samples were withheld from each cancer type, while the remaining data was used to train predictive models. Generalization and training errors were analyzed with different numbers of training examples in increments of 36, with equal training samples per class (Fig 5).

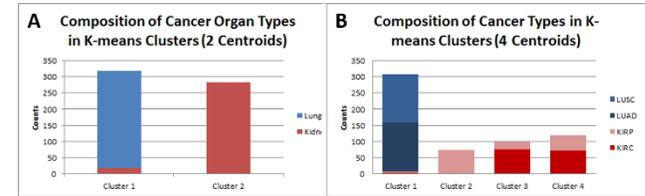
The effect of feature size on model performance was evaluated by varying the number of PCA features used to train different models. Features were selected from top-ranked principle components. The number of features used was incremented by 2-fold from 2 till 512 (Fig 7). 10-fold cross validation was used to assess performance.

## Selection of Significant Probes Set

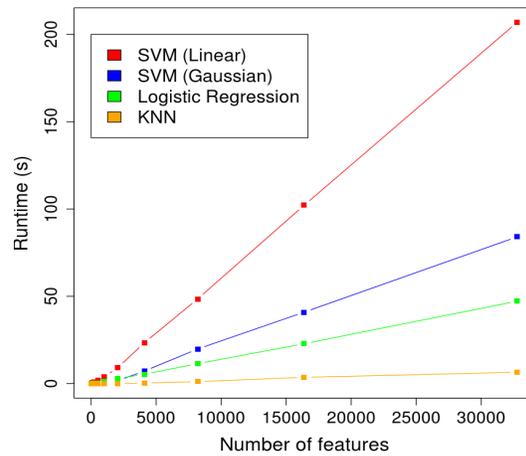
We recovered the contribution of each probe to eventual predictions. This was done by using a square matrix with 485,577 rows, where diagonal terms were maximum observed value for the corresponding probe (row) in the training set, off-diagonals were mean values of observations for given row from the training set. Matrix was transformed to lower dimensions using PCA loadings, then multiplied by logistic regression

coefficients. Probes whose contribution to the different cancer classes differed greatly (more than 3 standard deviations above the average range across all probes) were identified, and grouped by hierarchical clustering. These set of selected probes were then graphically depicted using a heatmap. (Fig 9)

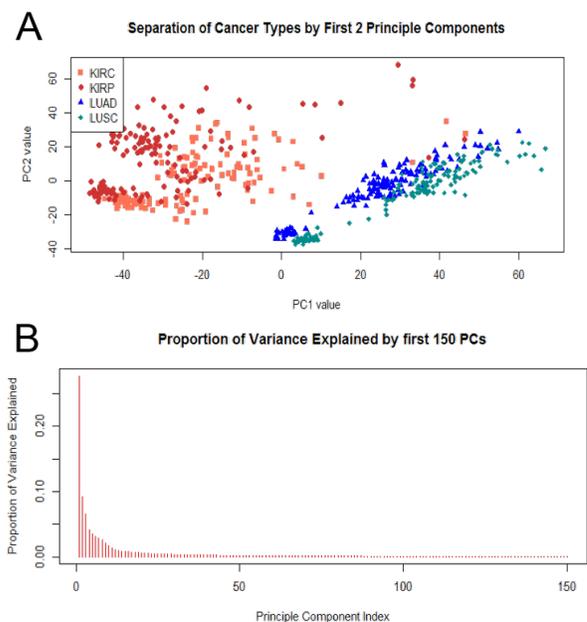
## RESULTS



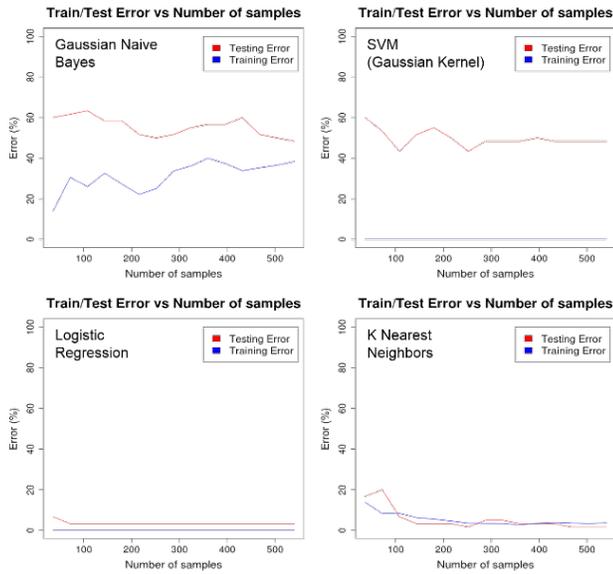
**Fig 2. Composition of cancer types in K-mean centroids.** (A) Cancer samples are well-clustered based on organ type with 2 centroids. (B) Cancer subtypes in the same organ system are not well separated using K-means.



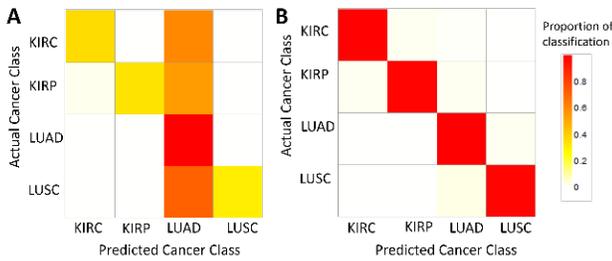
**Fig 3. Computational tractability of different machine learning models.** Runtimes based on models trained on 600 samples. Running times of all models appears to scale linearly with number of features.



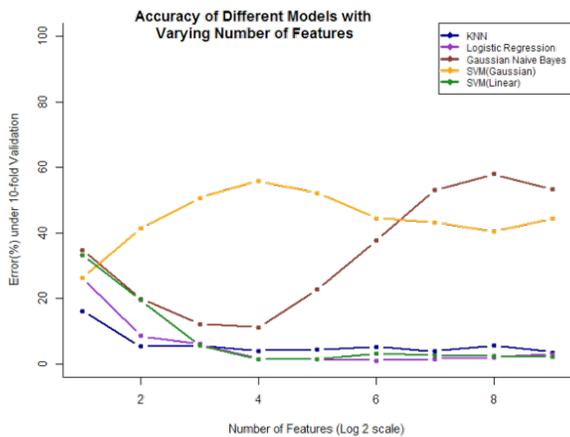
**Fig 4. Feature selection using PCA** (A) Cancers from different organs are well-separated spatially using the first 2 principle components. (B) **Scree plot.** The first few principle components explain a large proportion of variance explained in data.



**Fig 5. Training and Generalization (Testing) Error of Different Models with varying number of samples.** Logistic regression and Linear SVM (not shown) have similar profiles. Models were trained on 600 PCs and tested on a hold-out set.

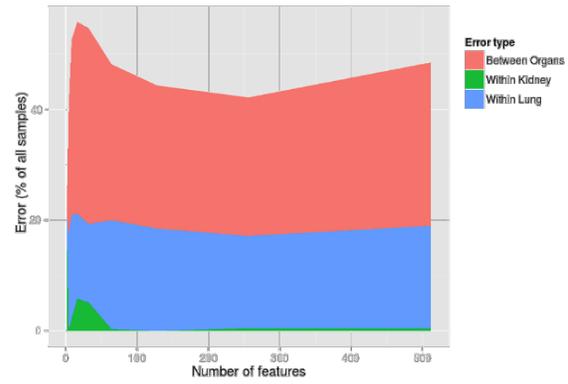


**Fig 6. Performance of SVM models.** Confusion matrix comparing accuracy of predictions using (A) SVM with Gaussian kernel and (B) SVM with linear kernel. Models were trained on 600 PCs and tested on a hold-out set.

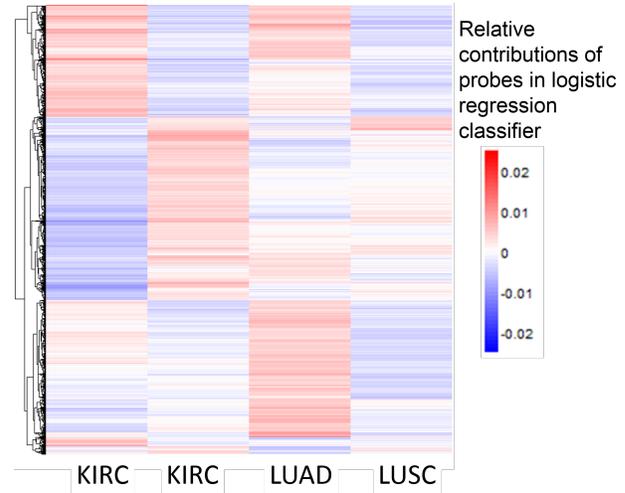


**Fig 7. Accuracy of models with varying number of features.**

Prediction error of 5 models against the number of PCA features. KNN, logistic regression and Linear SVM have similar observed trends.



**Fig 8. Dissection of Errors in SVM with Gaussian Kernel.** “Between Organs” refer to misclassifications between cancers from different organs, while “Within Kidney/Lung” refer to misclassifications between cancers of the respective organ. Beyond 80 features, error was dominated by “Between Organ” and “Within Lung” misclassifications. The breakdown of errors did not change significantly with change in number of features.



**Fig 9. Contribution of individual probes to logistic regression model.** 9843 probes (rows) with significantly different contributions to the logistic regression classification of the cancers subtypes are shown. A distinct profile is observed for each cancer type.

## DISCUSSION OF RESULTS

### Unsupervised Learning

We used K-means as an exploratory tool to determine whether there was an underlying nesting structure in the dataset. If the feature vectors of different cancer types were indeed different, K-means using 4 starting centroids should be able to achieve relatively good clustering according to organ systems and subtypes. We obtained encouraging results using 2 centroids, where the cancer types are well-segregated by organ types in the 2 clusters (Fig 2A), demonstrating that lung and kidney cancers have distinct methylation profiles. However, the cancer subtypes are not well segregated when 4 centroids were used (Fig 2B). The given clustering result suggests that the 2 lung cancer subtypes and 2 kidney cancer subtypes cannot be easily

distinguished using an unweighted linear combination of all 485,577 features.

## Feature Selection

To achieve better discrimination between the cancer subtypes, we applied supervised learning methodologies to the data set. Using the full feature set for all 600 samples, we were able to achieve an average of 97.3% accuracy using the Gaussian Naive Bayes model and a 94.7% accuracy using the KNN model based on 10-fold cross validation.

However, a simple benchmarking of the runtime of common machine learning algorithms on our dataset indicated that using the full 485,577-dimensional feature set was computationally expensive, and especially so for SVM (Fig 3). With improving technologies in whole-genome methylation profiling, we expect the number of methylation probes assayed to further increase in the near future. Using all probe values as training features will become increasingly challenging with time.

Given that many of the methylation sites are correlated with each other due to physical proximity in the genome and commonality of regulation pathways, the data lends itself naturally to dimensional reduction techniques, which will collapse covarying features. The reduced dimensions may reflect methylation hotspots or biological pathways that may be used for further research.

An even greater motivation to perform feature selection was based on the fact that the dimensionality of feature space was much greater than the number of training samples (600), which may result in overfitting during model construction.

We adopted principle component analysis (PCA) for feature selection as it offered an easy way to transform our data into lower dimensions ( $n=600$ ). PCA performed well for our dataset and the cancer samples were well-separated by organ types based on the first 2 PCs (Fig 4A), with some separation also observed between the subtypes. The first 30 PCs were also found to account for 77% of the variance in our data (Fig 4B).

## Supervised Learning

After obtaining a transformation of our data into a 600-dimensional subspace using the principle components, we applied several supervised machine learning algorithms to classify cancer subtypes using this new set of features.

Due to the short runtime and easy interpretation, we built a Gaussian Naive Bayes (GNB) classifier as our baseline supervised, parametric model. Despite GNB having good performance when trained on the full feature set consisting of 485,577 probes, the new model trained using the 600 PCs demonstrated high bias (Fig 5), with an unacceptable generalization (testing) error across all ranges of training size. This suggests that the PCs were not ideal features for training the GNB.

As an alternative baseline, we chose the non-parametric KNN classifier. The KNN model performed well on our new feature set, achieving a generalization error of 2%

when trained on all training samples (Fig 5), an improvement in accuracy as compared to the earlier KNN trained using the full feature set.

As a non-parametric algorithm however, KNN's runtime and storage size increases with the number of training samples. The model is also contingent on the assumption that new samples would be very similar to the known ones. Biological data is inherent noisy, and although we reduced variation via data pre-processing, the nature of KNN implies that it is sensitive to the differences between new test samples and our current data set, impairing its generalizability. Furthermore, KNN does not provide information on whether there are any methylation sites that are strongly correlated with the cancer types.

To gain better insight into the data, we applied linear discriminative models – logistic regression and SVM with a linear kernel. Both models gave similarly good performance, with zero training error and a low generalization error of <5% beyond a training size of 80 samples (Fig 5). Zero training error observed indicates that the given training data is linearly separable. Although minute, the decrease in generalization error with increasing training samples suggests that the models' performance can be further improved with more training data.

We also experimented with a non-linear SVM using a Gaussian kernel in an attempt to achieve even better classification accuracy. Unfortunately, this model performed poorly. Although the training error was zero, it had a high generalization error (>40%). Figure 6 shows a comparison of predictions generated by both SVM models; high intensity in the off-diagonal reflects high rates of misclassification in the Gaussian SVM model. In particular, the Gaussian SVM classified most samples wrongly as LUAD. This is likely due to over-fitting from high dimension feature space of the Gaussian kernel. While this can perhaps be resolved by more extensive parameter tweaking or further feature reduction (likely by exploring techniques other than PCA), we found this to be unnecessary due to the outstanding performance of our best models (logistic regression and linear SVM).

## Optimizing Prediction Model

The logistic regression and linear SVM models achieved good prediction accuracies on the test set, and the training curves do not suggest overfitting. However, in our analysis of PCA components, we noted that the lower-ranked PCs did not account for a large proportion of variance of the data. Intuitively, this means that the lower-ranked PCs should not have significant predictive value and could possibly be eliminated from the feature set to simplify models. To investigate this, we tested the accuracy of all previously described models with feature space restricted to the first  $2^n$  PCs, with  $n$  varying from 1 to 9. Models with good performance previously (KNN, linear SVM, logistic regression) achieved or approached their optimal performance with as few as 16 features (first 16 PCs). The

Gaussian Naive Bayes (GNB) model also achieved its best performance with 16 features, but errors increased when additional features were used. This suggests that the GNB model is overfitted when more than the 16 features are used.

In the previous section, we inferred that Gaussian SVM was possibly overfitted when trained on all 600 PCs, as evident from its high generalization error. Thus, we expected the model's generalization error to be lowered when fewer features are used. However, the model's performance remained consistently abysmal (Fig 7).

To understand the reasons for its poor performance, we examined the types of error made by the model. We grouped the errors as intra-organ ("Within Kidney/Lung") and inter-organ ("Between Organs") misclassifications. The breakdown of errors (Fig 8) is consistent with varying number of features. "Within Lung" and "Between Organs" misclassifications accounted for nearly all errors observed.

Since the performance of the Gaussian SVM did not improve with varying feature or sample size, it is likely that the set of features was unsuitable for this model. Alternatively, some form of regularization might be necessary to improve the performance of the model.

### Selection of Significant Probes Set

We found that the classifications made by linear SVM and logistic regression were in perfect agreement. We thus feel justified to focus only on the logistic regression model when selecting significant probes. We obtained a set of 9843 probes (Fig 9) that were found to be significant, using the procedure described in the methods section. We note that approximately three quarters of all 485,577 given probes were not predictive, with variation of contribution across the 4 classes for these probes being less than the average range for probes.

## CONCLUSION

In this paper, we show that logistic regression and linear SVM models can achieve >95% accuracy in classification of 4 different cancer types, using as few as 16 features obtained from top-ranking PCs. As compared to the full feature set of 485,577 probes, a 16 feature subset represents a 30,000 fold reduction in feature space, which drastically reduces computational overhead and complexity.

In particular, we note that both models were in full agreement on all predictions, which suggests that misclassifications by these two models are likely due to the inherent nature of the data, rather than models' bias or variance.

Future work would involve uncovering the biological significance of the identified methylation sites. We would also extend the classification model to include control (non-cancer) patients and more cancer types.

## REFERENCE

- [1] Golub, Todd R., Donna K. Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P. Mesirov, Hilary Coller et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *science* 286, no. 5439 (1999): 531-537.
- [2] Alizadeh, Ash A., Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature* 403, no. 6769 (2000): 503-511.
- [3] Laura, B. "Epigenomics: The new tool in studying complex diseases." *Nature Education* 1, no. 1 (2008).
- [4] Schumacher, Axel, Philipp Kapranov, Zachary Kaminsky, James Flanagan, Abbas Assadzadeh, Patrick Yau, Carl Virtanen et al. "Microarray-based DNA methylation profiling: technology and applications." *Nucleic acids research* 34, no. 2 (2006): 528-542.
- [5] Du, Pan, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren Kibbe, Lifang Hou, and Simon Lin. "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis." *BMC bioinformatics* 11, no. 1 (2010): 587.
- [6] Bolstad, Benjamin M., Rafael A. Irizarry, Magnus Åstrand, and Terence P. Speed. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." *Bioinformatics* 19, no. 2 (2003): 185-193.
- [7] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. "Missing value estimation methods for DNA microarrays." *Bioinformatics* 2001 Jun; 17(6):520-5.
- [8] Scikit-learn: Machine Learning in Python, Pedregosa et al., *JMLR* 12, pp. 2825-2830, 2011.
- [9] Model, Fabian, Peter Adorjan, Alexander Olek, and Christian Piepenbrock. "Feature selection for DNA methylation based cancer classification." *Bioinformatics* 17, no. suppl 1 (2001): S157-S164.