Abhijeet R Patil, Jongwha Chang, Ming-Ying Leung, and Sangjin Kim*

Analyzing high dimensional correlated data using feature ranking and classifiers

<https://doi.org/10.1515/cmb-2019-0008>

Received October 2, 2019; accepted December 2, 2019

Abstract: The Illumina Infinium HumanMethylation27 (Illumina 27K) BeadChip assay is a relatively recent high-throughput technology that allows over 27,000 CpGs to be assayed. The Illumina 27K methylation data is less commonly used in comparison to gene expression in bioinformatics. It provides a critical need to find the optimal feature ranking (FR) method for handling the high dimensional data. The optimal FR method on the classifier is not well known, and choosing the best performing FR method becomes more challenging in high dimensional data setting. Therefore, identifying the statistical methods which boost the inference is of crucial importance in this context. This paper describes the detailed performances of FR methods such as fisher score, information gain, chi-square, and minimum redundancy and maximum relevance on different classification methods such as Adaboost, Random Forest, Naive Bayes, and Support Vector Machines. Through simulation study and real data applications, we show that the fisher score as an FR method, when applied on all the classifiers, achieved best prediction accuracy with significantly small number of ranked features.

Keywords: Classification, DNA Methylation, High-dimensional Data, Information Theory, Feature Ranking

MSC 2010: 62P10, 62B10, 62G08, 62J99

1 Introduction

DNA methylation (DNAm) is an important epigenetic mechanism [1] controlling direct modification of DNA and abnormal DNA has been involved in the formation of diseases [2, 3]. DNAm is also involved in regulation of genes, genetic re-programming [4, 5], cell differentiation [6] and gene expression [7]. DNAm involves a process where the methyl group is added to the fifth carbon of a cytosine ring of DNA molecule, the presence of 5-methyl cytosine may change the activity around the DNA sequence. The methyl group can alter the transcription process of genes [8] and may lead to progression of tumor [9]. DNAm changes are linked with different types of diseases including neurological disorders, cardiovascular, and cancer [10, 11, 12]. Hence, studies on DNAm can also help in biomarker identification and disease classification [13, 14].

With the advent of high-throughput technology [15] large number of Illumina Infinium HumanMethylation27k (Illumina 27k) data sets have been amassed and made publicly available [16, 17]. Most of these data can be found in Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). The Infinium data sets have over 27000 CpGs to be assayed and are called CpG sites [18]. There are also Illumina Infinium HumanMethylation450k data sets being generated which have over 450k CpGs to be assayed which are ultra-high dimensional in nature.

Abhijeet R Patil: Computational Science, The University of Texas at El Paso, El Paso, TX 79968, USA, E-mail: arpatil@miners.utep.edu

Jongwha Chang: Department of Pharmacy Practice, The University of Texas at El Paso, El Paso, TX 79968, USA, E-mail: jchang@utep.edu

Ming-Ying Leung: Bioinformatics and Computational Science, University of Texas at El Paso, El Paso, TX 79968, USA, E-mail: mleung@utep.edu

***Corresponding Author: Sangjin Kim:** Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968, USA, E-mail: skim10@utep.edu

In this study, we focus on Illumina 27k that are high dimensional DNAm data. The results of the BeadChip arrays, the methylation status indicating the methylated and unmethylated DNA sequence measured using the β -values. For each of the specific CpG site, the β -value is calculated from the levels of the methylated and unmethylated alleles as the ratio of fluorescent signals [15, 19].

$$\beta = \frac{\max(M, 0)}{\max(U, 0) + \max(M, 0) + 100} \quad (1)$$

The β -value is between 0 (no methylation) and 1 (methylated), U is the fluorescent signal from an unmethylated allele, and M is that of a methylated allele.

There are relatively fewer studies conducted in the past for performing statistical analysis of the DNAm data sets generated from the Illumina 27k platform. Some of the existing studies for statistical comparison includes unsupervised clustering and normalization [20, 21], supervised feature ranking and classification methods [22, 23, 19, 24, 25]. feature ranking and classification have been extensively utilized in microarray gene expression studies [26, 27]. The key difference of DNAm data from gene expression is that the DNAm has continuous variables ranging between 0 and 1. There other difference is that there is a group structure with a gene. In the DNAm data, there are about 1-22 CpG sites per gene where the methylation levels of CpG sites within a gene are highly correlated.

In this article, we introduce the most widely utilized feature ranking method and classifier to the DNAm data. Feature ranking is a method employed to reduced the original dimension of the data to a lower dimension by removing the noise or unimportant features. The data in the field of data mining and bioinformatics [28, 29] such as microarray analysis [30, 31, 32], RNA-seq analysis [33], and mass spectra analysis appears to be high-dimensional in nature, where the number of features exceeds the number of samples. Dealing with high dimensional data is a challenging task and there is a need to reduce the dimension of data using the feature ranking techniques [34, 35, 36, 37, 38, 39, 40]. There are two traditional ways to deal with the high-dimensional data in biological application includes dimensionality reduction and feature ranking. The goal of the dimensionality reduction technique is to reduce the dimension of original feature space by transforming into a new feature space. Some of the most commonly used dimensionality reduction methods includes principal component analysis [41], and linear discriminant analysis [42]. The aim of feature selecting method is to select the significantly important features without modifying the original features. Several popular methods used for feature ranking in the literature includes information gain [43], fisher score [44], and minimum redundancy maximum relevance [45]. Both the dimensionality reduction and the feature ranking approaches are computationally efficient and are capable of improving the model performance. In this paper, we focus on the feature ranking methods, as they help in successfully preserving the original structure of the variables, whereas with the dimensionality reduction, the obtained new set of features are transformed with no significant meaning and cannot be used for further analysis. In the context of DNAm, when the feature ranking methods are employed the subset of most important CpG sites having an effect on the response variable are chosen and the original CpG sites are left unaltered. These important CpG sites can be used by biomedical researchers for further interpretability.

The overall purpose of this study is to compare the performances of popularly used feature ranking and classification methods on DNAm data and determine the performances utilizing the various performance metrics. To address these goals, we make use of a simulation study and two experimental Illumina 27k DNAm datasets which are obtained from GEO.

The remainder of the article is organized as follows. In section 2, the simulation and experimental data setup, various feature ranking and classification methods, and performance metrics are explained in detail. In section 3, we show the results of comparison FR and classifiers through simulation and experimental studies. We discuss our findings briefly in the section 4 and section 5 concludes the paper.

2 Materials and Methods

In this section, we describe the construction of the datasets; both synthetic and real data. This is followed by a description of the feature ranking and classification methods and then finally the definitions of the performance metrics applied.

2.1 Data Collection

We demonstrate by performing a simulation study and using the real datasets that the proposed procedure on DNA methylation can lead to better prediction than existing methods that ignore the association of these feature ranking methods.

DNAm data generated using the Illumina 27K array has large number of features representing the CpG sites and a relatively much smaller number of samples representing patients. This type of data is defined as high dimensional problem $p \gg n$ where p is the total number of CPG sites and n is the total number of samples.

DNAm data with p CpG sites and n samples can be represented by the following matrix:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix}_{n \times p} \quad (2)$$

X_i is a row vector that represents the total DNAm levels of sample i and X_{ij} is the methylation level of CpG site j of the sample i and all the values for X_i should lie between $[0,1]$.

Simulation data

In order to ascertain the results from the real datasets we include a synthetic dataset by simulation models which mimic the real dataset. In doing so we have to generate group correlated variables which lie between 0 and 1 inclusive. In our case, we have 475 genes in total consisting of 1-12 CpG sites such that the first group has 100 genes with one CpG site each, the second group has 75 genes and next subsequent groups have 30 genes with 3-12 CpG sites, respectively. Therefore we end up with a total of 2500 CpG sites representing 12 different groups. Now these CpG sites are formed using the inverse logit transformation of multivariate normal random variables. Thus, the methylation β values of the g -th gene for the i -th individual is calculated as;

$$X_{i,g} = \frac{\exp(t_{i,g})}{1 + \exp(t_{i,g})} \quad (3)$$

where, $t_{i,g} \sim \sqrt{p} N_{f_g}(\mu, \Sigma)$, p is a scale parameter and f_g is the size of the g -th gene, i.e., 12. For example, the 325th gene belongs to group 5 and $f_g = 5$ indicates the number of CpG sites for that particular gene. We set $\mu = (-0.1, \dots, -0.1)$ and $p = 2$ such that the methylation values have 0 as no methylation and 1 as completely methylated.

In order to the response variable, we assigned some of the variables as true important variables that is, those that are related to the cases (disease related sites). We did so by first selecting a gene from each of the 12 groups leading to 78 CpG sites each group having the corresponding number of CpG sites and are stored as CpG set. Mathematically, the true regression coefficient for the gene [19, 46] are as follows; CpG set is defined as follows,

$$\theta_{k,g} = (-1)^{g+1} \frac{\delta}{\sqrt{k}}, \text{ for all } k = 1, \dots, p_g; \quad g = 1, \dots, 12 \text{ and } \delta = 4. \quad (4)$$

Finally, the resulting response variable was simulated according to a Bernoulli distribution with the following model-based probabilities,

$$y_i \sim \text{Bernoulli}(p(x_i)), \quad p(x_i) = \frac{\exp(x_i^T \theta)}{1 + \exp(x_i^T \theta)} \quad (5)$$

where $x_i = (x_{i,1}^T, \dots, x_{i,475}^T)^T$ and $x_i = (\theta_{(1)}^T, \dots, \theta_{(475)}^T)^T$. The above procedure for generating the response variable was performed till we obtained the balanced class labels i.e., equal number of binary class labels '0' and '1'.

DNAm data is known for having strong correlation structures. The covariance matrix Σ within genes is defined by $\sum_{uv} = \rho^{|u-v|}$. The correlation of the models is AR(1). We simulated three different types of data based on different correlation structures. The simulation scenario S1 having low correlation $\rho = 0.2$, S2 with medium correlation $\rho = 0.5$ and S3 with high correlation $\rho = 0.8$. For each simulation scenario, samples were generated until we have 100 number of controls and cases respectively. Therefore, in each scenario, the total number of samples were 200 and the total number of CpG sites referred as features were 2500. For each scenario, We repeated the simulations 100 times.

Experimental data

For the real datasets, we downloaded two DNAm datasets from NCBI Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) which was generated from Illumina 27K. To compare the performances of the FR and Classifiers we obtained the data sets from Gene Expression Omnibus (GEO) having the accession numbers GSE26126 and GSE32393, designated as D1 and D2 respectively in the remaining part of the article. The D1 is a genome wide DNAm profiling of normal and tumor prostate samples, as well as cultured primary prostate cells overexpressing DNA Methyltransferases and EZH2. It involved the quantitative profiling of 95 primary prostate tumors and 98 healthy prostate tissue samples for their DNA methylation levels at 27578 CpGs representing 14,104 gene promoters by using the Illumina 27K platform. That is, it has 193 total samples in general and the data is balanced with almost equal number of cases and control.

The D2 is a epigenome analysis of breast tissue from women with and without cancer. The DNA methylation profiles across 27578 CpGs in breast tissues from women with and without breast cancer are obtained using Illumina 27K platform. Breast tissue samples were drawn from 114 breast cancer and 23 non-neoplastic breast tissues. That is, it has 137 total samples in general and the data is imbalanced with the cases being almost five times more than the control.

The real datasets were normalized using the Linear Models for Microarray and RNA-seq Data package (Limma). The "normalizeBetweenArrays" function in R is a quantile and cyclic loess normalization which was originally proposed by [47] for Affymetrix-style. It forces the entire empirical distribution of each column to be identical to achieve consistency between arrays. The pre-processing procedures were performed to handle these high-dimensional DNAm data. As first filtering step to overcome the redundant noisy CpG sites, the marginal maximum likelihood estimation (MMLE) was performed, the top 5000 CpG sites are ranked based on their level of significance. These ranked significant genes are further applied on the feature ranking methods.

2.2 Feature Ranking methods

Feature ranking (FR) methods are classified into supervised [35] and unsupervised [36] based on the presence of class labels. The simulation and the DNAm datasets used for this paper have labels hence qualifies under supervised FR. We used the filter based FR methods such as, Fisher Score [37], Information Gain [38], MRMR [39], and Chi-Square [40]. These methods measures the different characteristics of the data such as information, correlation structure, the distance and dependency between the variables and labels.

Fisher Score

Fisher Score (F-score) [37] is the most popular FR method used to select important features among all features which can be used for further classification. To achieve this, discriminative and statistical models are used. The general principle to select the features is that the calculated distance between the data points in the same class should be small and the distance between data points in the different class needs to be large. With this intuition, the fisher score for each feature f_i is calculated as,

$$F_i = \frac{\sum_{k=1}^K n_j (\mu_{ij} - \mu_i)^2}{\sum_{k=1}^K n_j \rho_{ij}^2} \quad (6)$$

where, F_i is the calculated score for the feature f_i , n_j is the number of instances in the class j , μ_i is the mean of the feature f_i , and μ_{ij} and ρ_{ij} are the mean and variance of the feature f_i in the class j , respectively.

Information Gain

Information Gain (IG) is used to measure the amount of information a feature gives us about the class [38]. Information Gain is one of the widely used FR methods because of its computational efficiency and its simple interpretation. The unrelated/ incorrect features gives no information and the features that perfectly divide gives the maximal information. This measure of purity is called the information and the measure of impurity is called entropy. The disadvantage of information gain is that it does not work well when there are redundant features.

The IG between the i -th feature X_i and the response labels Y are given as,

$$IG(X_i, Y) = H(X_i) - H(X_i|Y) \quad (7)$$

$$H(X_i) = \sum_i p(x_i) \log_2(p(x_i))$$

$$H(X_i|Y) = \sum_{y_j \in Y} p(y_j) \sum_{x_i \in X} p(x_i|y_j) \log_2(p(x_i|y_j))$$

where, $H(X)$ is the entropy of X_i , $H(X_i|Y)$ is the entropy of X_i given Y .

Minimum Redundancy Maximum Relevance

The Minimum Redundancy Maximum Relevance (MRMR) is a FR method. Features are selected by ranking them based on the maximal statistical dependency criterion [39]. The true important features are selected such that there is a minimum redundancy within the features and a maximum relevance to target variable. MRMR is also called a mutual information based method as it is used to determine both redundancy and relevance. The mutual information is defined as,

$$I(X, Y) = \iint \frac{P(X, Y) \log P(X, Y)}{P(X)P(Y)} \quad (8)$$

where X, Y are the vectors, $P(X, Y)$ represents joint probabilistic density, $P(X)$ and $P(Y)$ are marginal probabilities density.

The MRMR for discrete variables is given by,

$$\min W_I, \quad W_I = \frac{1}{|S|^2} \sum_{i,j \in S} I(i, j) \quad \text{and} \quad \max V_I, \quad V_I = \frac{1}{|S|^2} \sum_{i,s \in S} I(h, i) \quad (9)$$

The MRMR for continuous variables is given by,

$$\min W_C, \quad W_C = \frac{1}{|S|^2} \sum_{i,j \in S} |C(i, j)| \quad \text{and} \quad \max V_F, \quad V_F = \frac{1}{|S|^2} \sum_{i \in S} F(i, h) \quad (10)$$

where S is the set of features, $I(i, j)$ and $C(i, j)$ is mutual information and the correlation between features f_i and f_j respectively, h is the target class and $F(i, h)$ represents F-statistic.

Chi-square

Chi-square test (Chi2) is another popular FR method. It is a statistical test of independence which is used to determine the dependency of two variables [40]. Each sample in the data has several features and a response variable (or class label). The relationship between these features and response variable is observed and recorded using Chi2 statistics. The basic rule is that if there is a dependency between the response variable and the feature variable then it is considered as very important variable. Whereas, if the response variable is independent of the feature variable then such feature is ignored.

Two events X and Y are said to be independent of each other if:

$$P(XY) = P(X)P(Y) \text{ or } P(X|Y) = P(X) \text{ and } P(Y|X) = P(Y) \quad (11)$$

The two events in the context of feature selection corresponds to the occurrence of feature and the response variable. Based on the following equation, we can rank the terms.

$$\tilde{\chi}^2(D, f, y) = \sum_{e_f \in \{0,1\}} \sum_{e_y \in \{0,1\}} \frac{(N_{e_f e_y} - E_{e_f e_y})^2}{E_{e_f e_y}} \quad (12)$$

where e_f and e_y are the feature term and the response class respectively. N is the observed frequency in D and E is the expected frequency.

2.3 Classifiers

Many classification techniques are available in statistics and machine learning which can be applied to broad range of problems from business applications to biological research [48]. Depending on the type of data i.e. with or without labels, the algorithms are classified as supervised (such as classification and regression) and unsupervised (such as clustering). We focus on the supervised classification techniques as our data sets are labelled. The most commonly used advanced algorithms in the field of biomedical research such as Random Forests [49], Naive Bayes [50], Support vector machines [51], Logistic regression [48] and AdaBoost [52] are used here for comparing the results of FR method on different classifiers.

Random Forests

Decision trees are most popular among biomedical researchers as they are more decipherable when dealing with binary outcomes. A decision tree treats the given sample by going down the tree as a series of yes/no questions which eventually leads to the terminal node which carries predicted class.

Random forests (RF) [49] is combination of not one but multiple decision trees, where each one trains on different set of features (random sampling with replacement), the nodes are split in each tree taking limited number of features. The number of classifications equally correspond to the number of trees. The majority voting scheme is applied by aggregating all the trees and one final classification is obtained. The basic idea is to not rely on individual source but multiple. Through the concept of bagging, or averaging, the accuracy of unstable individual decision trees can be improved.

The interesting quality of random forests is it's out of bag technique. All the samples which were omitted of the bootstrap sample during training are considered as Out Of Bag (OOB) data, the error estimates are calculated by running the random forests through this OOB data. The average error estimate is taken by considering the error rate over all the trees and is called miss classification error estimate.

Naive Bayes

Naive Bayes (NB) is the most popular classification algorithm based on the probabilistic algorithms which take the advantage of probability theory and the Bayes theorem [50]. It is based on a simple rule that all features being classified are independent of each other and this turns out to be a major disadvantage because in real world situations the data always has some correlation.

Bayes theorem is used to find the probability of an event occurring given the probability of another event that has occurred already. It is stated mathematically as,

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (13)$$

where X and C represents the response variable and the feature vector.

Support Vector Machines

Support vector machines (SVM) are most commonly used among medical researchers in classification problems [51]. The main idea of SVM is to find a hyperplane that best divides the samples into two classes with the largest possible margin. The margin is referred as the distance between the hyperplane and the closest data points. In the linear condition, the goal of SVM is to find the hyperplane that maximizes the margin such that both the classes are well separable and classified correctly. But when we consider classification problems based on the real world datasets, the classes are never distinctly separated therefore the condition of classifying distinctly through a hyperplane becomes void.

To overcome this problem, SVM introduces two extensions: soft margins and kernel tricks. In the first approach, SVM tries to find the hyperplane that maximizes the margin but allows few misclassified dots. A threshold is chosen which acts as a penalty term such that the total distance between the dots and all misclassification dots remains smaller than this threshold. The latter approach is the more commonly used since the classes cannot be separated linearly in the original feature space, the kernel trick uses existing features, maps them into a higher dimensional space and applies transformation and creates new set of features. Those new set of features help SVM to find the non-linear decision boundaries. The dot product of input feature space can be computed as a kernel function

$$K(X^{(i)}, X^{(j)}) = \phi(X^{(i)})^T \phi(X^{(j)}) \quad (14)$$

Here the input data $X^{(i)}, X^{(j)}$ are mapped into feature space as $\phi(X^{(i)})^T, \phi(X^{(j)})$. There are different types of kernel which can be used: 'linear', 'polynomial', 'radial basis function', and 'sigmoid'. The most popular among the kernel functions are:

The k-degree polynomial given by,

$$K(X^{(i)}, X^{(j)}) = (1 + x^{(i)T} x^{(j)})^k \quad (15)$$

The radial basis kernel is given by,

$$K(X^{(i)}, X^{(j)}) = e^{\frac{-||x_i - x_j||^2}{c}} \quad (16)$$

and finally the sigmoid kernel which is defined as follows,

$$K(X^{(i)}, X^{(j)}) = \tanh(k_1 x^{(i)T} x^{(j)} + k_2) \quad (17)$$

Where k, k_1, k_2 and c are the required parameters which need to be specified. The boundaries which are non-linear on the original space are mapped as linear in a transformed high dimensional space through the use of kernel function and henceforth SVM gets its flexibility over other linear classifiers.

AdaBoost

AdaBoost (AB) also known as adaptive boosting is the popular boosting technique [52]. It is based on the idea of combining multiple “weak classifiers” or “inaccurate classifiers” into a single “strong classifier”. AB algorithm can be applied to any classification problem. Random subset of the training set is used to train each weak classifier, but the subsets are not chosen at complete random. AB algorithm assigns weight to each training item. A higher weight is assigned to the incorrectly classified item so that it appears with higher probability in the training subset of future classifier. Once each classifier is trained, based on the accuracy the weights are assigned. Thus, higher weight is assigned to more correct classifiers so that they will have higher probability in the final outcome. The rule of thumb for assigning weights to the classifier is that, a weight of zero is assigned to the classifier with fifty percent accuracy and a negative weight is assigned to the classifier with less than fifty percent accuracy.

2.4 Performance Metrics

Accuracy (Acc) is one of the widely used metrics for measuring the performance of classification models on the balanced data. It is calculated as the number of samples correctly classified to the total number of samples used in the classification model. Sensitivity (Sens) is used to test if the individual is correctly classified as ‘diseased’. Specificity (Spec) is used to test if the individual is correctly classified as ‘normal’. [53].

Accuracy, Sensitivity and Specificity are described in terms of true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP).

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{Sensitivity} &= \frac{TP}{TP + FN} \\ \text{Specificity} &= \frac{TN}{TN + FP} \end{aligned} \quad (18)$$

3 Results

We compare the performance of all FR and classifiers through extensive simulation studies with three different scenarios and to further support the results real DNAm dataset is used which are described in earlier section 2.

3.1 Simulation Data

The simulation studies are performed with three different scenarios: S1, S2 and S3 corresponding to the low, medium and high correlation structures are generated using the procedure explained in detail in simulation data section. Each simulated data was randomly partitioned into the training and testing dataset, where 70% of the samples were selected for the training dataset and the remaining 30% were selected for the testing dataset. The 70% of the training set was applied to the FR methods and the ranked features were given to the classification methods. For fair comparison, all of the used classification methods were evaluated for their classification performance metrics, averaged over 100 times. 2.1.

In this section, we show only the results for highly correlated simulation scenario S3. This scenario because of its high correlation structure matches with real DNAm data which are usually known for their high correlation structures. However, the results for low and medium correlation data (S2 and S3 respectively) along with the Figures and Tables depicting the accuracy, sensitivity and specificity are described in the appendix section.

Figure 1: Average accuracies over 100 iterations in simulation data S3

The average accuracies obtained for all the FR methods and classifiers on simulation scenario S3 are shown in Figure 1. The Table 1 shows the evaluation metrics accuracy, sensitivity and specificity along with the corresponding standard deviation. Here, we see that F-score outperforms all the FR methods by attaining the highest average prediction accuracies when applied on all the classifiers compared. Also, F-score attained the best prediction accuracy within top 30 features and is distinctly separated from other FR methods. The order of performance for the rest of FR methods within top 30 features with all classifiers is as follows: F-score, Chi2 and MRMR. These results shows that F-score method selects more number of true important variables given the range of ranked features and henceforth attains the best accuracies. F-score with Naive bayes classifier with the accuracy of more than 72% showed marginally better performance compared to F-score with Random forest and is the best performing combination among the remaining applications of FR methods and classifiers in the high correlation structure data. Also, we can note that the FR methods Chi2 and IG are having an overlapping performance between the 60 to 120 ranked features when applied on all the classifiers.

In the Table 1 we show the details of accuracy, sensitivity and specificity for the small, intermediate and large number of ranked features. i.e., at the 30, 90 and 200 ranked features. The order of accuracies for F-score method applied on Naive bayes and Random forest and AdaBoost classifier at the 90th ranked feature mark are 72.68%, 71.14% and 69.96% respectively. This shows that the optimal number of true important features were attained by the F-score within 90 ranked features and we see the drop in the performance of

all these classifiers at the 200th ranked feature mark. This can be justified because by adding more features to the model along with the less number of true variables more noise is being added. Henceforth the drop in performance is inevitable in the tree based methods such as random forests and AdaBoost. In contrast, we see that all the FR methods when applied on SVM showed an increase in accuracy with the increase in number of features, this is one of the important characteristics of support vector machines. Given a certain type of data, the support vector machines show an rise in performance when more number of features are added in the model. from the Table 1 we see that the SVM has the highest accuracies at the 200th ranked feature in comparison to the 30th and 90th mark among all the FR methods. Also, across all FR methods with SVM, the F-score with SVM method showed the best performance with an accuracy of 69.74%.

Although MRMR and IG are not showing the best performance, For analyzing the behavior of these FR methods on all the classifiers, we look at the corresponding accuracies and we can note that the MRMR and IG behaves similar to the SVM classifier with all FR methods. i.e., as the number of ranked features increases from 30 to 300, the accuracy of all the classifiers increases. This shows that both the MRMR and IG are not selecting more number of true important variables within the top 30 features. As the number of features are increased more number of true important variables are added along with which the accuracy increases. Henceforth, we can say that F-score which selects the large number of true important variables within the small range of ranked features is the best FR method compared to others.

3.2 Experimental Data

The experimental studies are performed with two DNAm datasets: D1 and D2 corresponding to the balanced and imbalanced samples. Both of the dataset D1 and D2 are obtained from GEO and the complete description of the data along with the data pre-processing steps is explained in section 2.1. Each DNAm dataset was randomly partitioned into the training and testing dataset, where 70% of the samples were selected for the training dataset and the remaining 30% were selected for the testing dataset. For fair comparison, all of the used classification methods were evaluated for their classification performance metrics, averaged over 100 times.

3.2.1 Experimental Data D1: GSE26126

The Figure 2 shows the average accuracies of all the classifiers with the ranked CpG sites using FR methods on real data D1. The detailed evaluation metrics is shown in Table 2. Here, we see that the F-score when applied on all the classifiers outperforms the other FR methods by attaining the highest average prediction accuracies within the top range of ranked CpG sites. These results are very similar to what we saw in the simulation data S3, which was a highly correlated data. from this matching behavior in terms of accuracy we can say that the experimental DNAm dataset D1 is also having very high correlation. For further illustrations of the correlation structure of the experimental data D1 we added the boxplot and histogram as shown in Figure 3, where we see that the median correlation for this data D2 is 0.76 which is closer to correlation structure of 0.8 from simulation scenario S3.

From the plots in Figure 2 we also see that there is an overlapping performance between the Chi2 and IG which was also seen in the simulation scenario S3. Th MRMR although it performed well on the tree based AdaBoost and random forest classifiers, it failed to show the consistency on naive bayes and support vector machines with the lowest accuracies among all FR methods.

In the Table 2, the evaluation metrics of all the FR with classifiers on the dataset D1 are shown. As in simulation studies, again we show the performance of all the FR methods with classifiers at the small, intermediate and large number of ranked CpG sites i.e., when p is 30, 90 and 200.

With the accuracy of 82.39% and standard deviation(sd) (5.59%), MRMR with SVM achieved the lowest performance among all the applications of FR method on classifiers. However, we also see that the F-score

Table 1: The averages of accuracy, sensitivity and specificity along with the standard deviation across 100 number of iterations for the simulation data S3

	#p	F-score			IG			Chi2			MRMR		
		30	90	200	30	90	200	30	90	200	30	90	200
AB	Acc	68.06 (sd) (6.37)	69.96 (6.70)	68.30 (6.95)	64.28 (6.68)	65.40 (6.94)	67.16 (6.67)	66.36 (7.93)	66.86 (7.87)	66.56 (7.36)	59.78 (7.69)	61.94 (7.80)	64.28 (7.73)
	Sens	69.08 (sd) (10.0)	70.65 (9.36)	69.34 (11.1)	64.31 (11.9)	65.45 (10.8)	68.58 (11.4)	65.70 (11.0)	66.76 (11.6)	67.63 (10.8)	59.03 (12.5)	62.46 (11.0)	64.47 (11.7)
	Spec	67.71 (sd) (9.52)	70.02 (9.68)	68.24 (10.2)	65.11 (10.3)	66.26 (10.5)	66.75 (9.07)	67.50 (11.0)	67.51 (10.0)	66.31 (10.5)	61.50 (9.81)	62.31 (11.4)	65.15 (10.6)
RF	Acc	68.74 (sd) (6.39)	71.14 (6.30)	70.22 (6.70)	64.84 (7.31)	67.98 (6.83)	68.02 (7.38)	67.56 (7.27)	69.10 (6.92)	69.06 (6.63)	61.12 (6.94)	63.52 (7.56)	65.36 (7.00)
	Sens	68.99 (sd) (10.6)	72.39 (10.6)	72.35 (12.2)	66.26 (11.2)	68.51 (11.6)	69.72 (13.4)	67.94 (10.9)	69.02 (11.3)	69.47 (13.0)	61.46 (11.8)	64.96 (12.1)	66.29 (13.7)
	Spec	69.42 (sd) (9.43)	71.15 (10.0)	70.05 (12.5)	64.50 (10.9)	68.72 (10.8)	68.42 (13.2)	68.00 (10.7)	70.23 (10.2)	70.43 (11.3)	62.19 (10.9)	63.61 (10.9)	66.06 (12.1)
NB	Acc	70.20 (sd) (6.63)	72.68 (6.40)	71.86 (6.18)	66.94 (7.20)	69.06 (6.61)	70.04 (6.49)	68.80 (6.40)	69.72 (6.96)	70.98 (6.51)	62.12 (7.19)	65.38 (7.08)	67.48 (6.12)
	Sens	70.82 (sd) (9.95)	73.55 (9.94)	71.63 (9.31)	67.85 (10.6)	69.14 (10.1)	70.48 (9.23)	69.04 (10.1)	69.64 (10.7)	70.65 (10.3)	61.70 (11.3)	66.13 (9.98)	67.95 (9.42)
	Spec	69.86 (sd) (9.25)	72.28 (8.62)	72.55 (9.43)	66.62 (9.95)	69.35 (9.45)	70.11 (9.50)	68.76 (10.3)	70.16 (10.0)	71.86 (9.22)	63.32 (10.2)	65.32 (10.0)	67.69 (8.91)
SVM	Acc	68.50 (sd) (6.57)	69.58 (6.64)	69.74 (7.35)	64.44 (7.33)	66.14 (7.10)	67.68 (6.38)	66.16 (7.22)	65.94 (8.42)	68.30 (6.70)	60.40 (7.43)	61.70 (7.56)	64.90 (7.10)
	Sens	67.95 (sd) (9.73)	69.20 (9.72)	69.92 (10.7)	64.30 (10.7)	65.48 (10.7)	66.21 (9.98)	67.23 (10.8)	65.38 (12.0)	68.21 (10.9)	59.96 (11.6)	62.30 (12.6)	64.72 (10.6)
	Spec	69.47 (sd) (10.2)	70.45 (10.5)	69.82 (9.69)	64.92 (11.2)	67.21 (9.84)	68.47 (10.3)	65.53 (11.4)	66.77 (11.0)	68.72 (8.74)	61.63 (11.3)	61.48 (10.4)	65.42 (9.58)

with SVM on the experimental data is seen performing better with the small number of ranked CpG sites with the accuracies of 88.58%, 86.05% and 83.55% respectively for the $p = 30, 90$ and 200 respectively.

Figure 2: Average accuracies over 100 iterations in experimental data D1

The F-score compared to other FR methods has remarkably good performance in selecting the true important variables within the top 30 ranked CpG sites. We also see that the F-score with Random Forests has an accuracy of 92.32% which is the highest considering all other applications of FR on classifiers. This shows that F-score attains the optimal number true important CpG sites required required for predicting the unknown samples within the small range of ranked CpG sites.

Table 2: The averages of accuracy, sensitivity and specificity along with the standard deviation across 100 number of iterations for the experimental data D1(GSE26126)

#p	F-score			IG			Chi2			MRMR		
	30	90	200	30	90	200	30	90	200	30	90	200
AB	Acc	89.65	90.01	89.89	88.22	88.67	89.51	88.22	88.65	89.37	89.03	89.25
	(sd)	(3.88)	(3.55)	(3.11)	(3.66)	(3.20)	(3.24)	(3.65)	(3.37)	(3.33)	(3.63)	(3.15)
	Sens	87.39	87.42	87.50	86.07	86.08	86.98	85.87	86.28	86.97	86.51	86.95
	(sd)	(6.27)	(6.23)	(5.51)	(6.29)	(6.32)	(6.04)	(6.45)	(6.28)	(6.06)	(6.80)	(6.08)
	Spec	91.78	92.46	92.15	90.27	91.20	91.90	90.48	90.97	91.73	91.51	91.47
	(sd)	(5.64)	(4.95)	(5.17)	(5.83)	(5.45)	(5.45)	(5.89)	(5.51)	(5.03)	(5.27)	(5.43)
RF	Acc	90.15	90.32	90.24	89.60	89.96	89.93	89.46	89.81	89.89	90.22	90.12
	(sd)	(3.47)	(3.12)	(3.14)	(3.19)	(3.24)	(3.17)	(3.10)	(3.21)	(3.24)	(3.26)	(3.13)
	Sens	87.81	87.87	87.57	87.10	87.55	87.44	87.20	87.47	87.49	87.69	87.91
	(sd)	(5.90)	(5.48)	(5.55)	(5.82)	(5.60)	(5.50)	(5.58)	(5.54)	(5.59)	(5.72)	(5.57)
	Spec	92.40	92.67	92.63	92.03	92.28	92.31	91.67	92.04	92.20	92.65	92.22
	(sd)	(5.10)	(4.88)	(4.89)	(5.13)	(5.11)	(5.24)	(5.18)	(5.07)	(5.09)	(5.15)	(5.18)
NB	Acc	89.22	89.06	89.01	89.00	89.03	89.13	88.94	89.12	89.13	88.15	88.88
	(sd)	(3.05)	(3.04)	(3.01)	(3.07)	(3.05)	(3.02)	(2.98)	(3.04)	(2.98)	(3.69)	(3.52)
	Sens	87.35	86.85	86.68	86.74	86.65	86.65	86.68	86.61	86.59	88.31	88.48
	(sd)	(5.53)	(5.27)	(5.07)	(5.39)	(5.12)	(5.12)	(5.12)	(5.35)	(5.11)	(5.90)	(5.43)
	Spec	90.94	94.79	91.19	94.78	91.24	91.47	91.18	91.45	91.52	87.98	88.60
	(sd)	(4.64)	(4.79)	(4.65)	(4.71)	(4.78)	(4.82)	(4.69)	(4.85)	(4.85)	(5.86)	(5.84)
SVM	Acc	88.58	86.05	83.55	87.13	84.18	83.32	87.43	84.74	83.36	86.58	82.39
	(sd)	(3.38)	(4.04)	(4.45)	(3.36)	(4.25)	(4.18)	(3.57)	(3.57)	(4.70)	(4.27)	(5.59)
	Sens	84.63	82.97	80.80	83.29	80.23	79.54	84.09	81.34	80.23	82.48	80.46
	(sd)	(6.22)	(6.42)	(8.51)	(6.97)	(7.83)	(8.03)	(6.53)	(7.38)	(7.83)	(7.98)	(8.10)
	Spec	92.43	89.12	86.36	90.90	88.11	87.02	90.65	88.08	86.23	90.47	84.33
	(sd)	(4.94)	(6.60)	(5.70)	(5.45)	(6.38)	(5.86)	(5.43)	(6.37)	(6.94)	(5.99)	(7.98)

Figure 3: The histogram and boxplot of pairwise correlation coefficients between the expression levels of 5000 filtered CpG sites. Two plots show that average pairwise correlation is 0.75 on the experimental data D1

3.2.2 Experimental Data D2: GSE32393

The average accuracies of all the classifiers with the ranked CpG sites using the FR methods on real data D2 is shown in Figure 4. The detailed results with the different evaluation measures based on the different number of ranked features can be seen in Table 3. Here, we see that the F-score method when applied on all the classifiers outperforms the other FR methods by attaining relatively better prediction accuracies within the top 30 ranked CpG sites. The results show that performances of all the classifiers are very close among all the FR methods. The correlation of the DNAm data D2 was evaluated to assess the performance of different FR methods on classification methods used in this study. In the Figure 5, we see that this dataset is seen to have a correlation of nearing to 0.5. The results of data D2 are found to be very similar to what we see in the low and medium correlation simulation scenarios S1 and S2, where the correlation were 0.5 and 0.8 respectively. The results for scenario S1 and S2 are explained in appendix section. The matching results in these scenarios and the experimental data D2 suggests that when there is a intermediate correlation in the DNAm data, the FR methods and classifiers tend to perform well. The F-score and MRMR methods on the naive bayes classifier are seen to have an overlapping performance with the F-score attaining the best score with the 200 ranked CpG sites by a very close margin. In the Table 3, the evaluation metrics of FR methods with different classifiers on the dataset D2 are shown.

The F-score compared to other FR methods has remarkably good performance in selecting the true important variables within the top 30 ranked CpG sites. We also see that the F-score with SVM has an accuracy of 96.57% when the ranked CpG sites were 30. This is the highest accuracy considering all other FR methods and classifiers the top 30 number of ranked CpG sites. This shows that F-score attains the optimal number true important CpG sites required required for predicting the unknown samples within the small range of ranked CpG sites.

The Figure 5 shows the correlation structure of the experimental data D2. We see that the data is having a intermediate correlation structure of near to 0.5.

Figure 4: Average accuracies over 100 iterations in experimental data D2

4 Discussion

High dimensional data is very complex to deal with. DNAm data with its high correlation structures adds more complexity. Many studies have been proposed to solve this problem. However, the best performing FR method on different classifiers to deal with this type of data is not known yet. For this purposed, In this paper, through our extensive simulation studies and experimental data we show the best performing FR with different classifiers to deal with the high-dimensional DNAm data. from the results, in both the simulation and experimental data we saw that among the FR methods definitely F-score had better performance than other FR methods by selecting the optimal number of true important CpG sites within the small range of ranked features. On the other hand, MRMR was having the least accuracies when applied on the classifiers compared here. The IG and Chi2 showed similar performance all along the ranked features.

Different type of classifier and FR methods tend to perform differently on certain types of data. It is partially biased to conclude one best performing combination in a high dimensional with high correlation setting such as DNAm data as it is extremely challenging problem to deal with. However, among the classifiers in our simulation data, Naive Bayes showed the better performance followed by Random Forest and AdaBoost respectively. But variation in the accuracies among all these were marginal. In the experimental data D1, tree-

Figure 5: The histogram and boxplot of pairwise correlation coefficients between the expression levels of 5000 filtered CpGs. Two plots show that average pairwise correlation is 0.42 on the experimental data D2

based ensemble methods Random Forests and AdaBoost showed marginally better performance compared to other classifiers. The fall in accuracies using SVM classifier in experimental data D1 explains how the inclusion of noise can affect performance and even lead to a high false positive rate. In the experimental D2, all the FR and classifier methods had accuracies ranging between 95-96%. This shows that all combination of the methods are performing best with closer margins.

However, among the FR methods we can easily decide the best performing method. The FR method as F-score showed the best performance on all the classifiers with small number of ranked important CpG sites. When dealing with high-dimensional data, FR approach is an important to reduce the number of features by eliminating the noise in the data and getting the truly significant features. It helps in reducing the computational time. The performances of all the experiments carried out in this research showed that having the small number of significant features can be sufficient to attain good performance.

5 Conclusion

The performances of the FR methods are seen similar in both the synthetic and real DNAm data. Among all the simulation scenarios and DNAm data sets, the F-score as a FR method when applied on all the classifiers achieved best prediction accuracy with significantly small number of ranked features. The second best performed FR method was IG followed by Chi2 which has least false positive rate especially when used with random forest classifier. That is, it has the lowest rate of predicting that a patient without a disease has a disease. The FR on all the classifiers used for comparison in this paper performed competitively with much less variation in the average accuracies. Also, the overall performances achieved by all the classifiers were closer when very few significant set of ranked features called CpG sites were considered. This showed that small set of significant features can classify the DNAm samples in more efficient manner not only by reducing the computational time but also in boosting the overall performance. As future research, we plan to focus on ultra-high dimensional DNAm data from Illumina 450K beadarrays which has over 450K CpG sites.

Table 3: The averages of accuracy, sensitivity and specificity along with the standard deviation across 100 number of iterations for the experimental data D2(GSE32393)

#p	F-score			IG			Chi2			MRMR		
	30	90	200	30	90	200	30	90	200	30	90	200
AB												
Acc	96.19	96.64	96.30	95.66	96.57	96.90	95.50	96.07	96.85	94.85	95.78	96.35
(sd)	(2.83)	(2.87)	(2.99)	(3.01)	(3.16)	(2.86)	(3.15)	(2.81)	(2.98)	(3.59)	(3.06)	(2.92)
Sens	98.34	98.48	98.66	97.92	98.51	98.75	97.73	98.01	98.68	98.04	98.36	98.64
(sd)	(2.42)	(2.46)	(2.40)	(2.75)	(2.83)	(2.33)	(3.02)	(2.82)	(2.42)	(2.88)	(2.66)	(2.38)
Spec	85.98	87.87	85.30	84.93	87.11	88.25	84.92	86.81	88.14	79.77	83.65	85.13
(sd)	(13.4)	(12.7)	(13.6)	(14.2)	(13.7)	(12.8)	(14.0)	(12.9)	(13.1)	(16.6)	(14.5)	(13.9)
RF												
Acc	96.21	96.14	96.33	95.83	96.71	96.50	95.80	96.69	96.97	94.95	95.52	96.07
(sd)	(2.87)	(2.70)	(2.73)	(3.03)	(2.92)	(2.92)	(3.06)	(2.61)	(2.76)	(3.47)	(3.21)	(2.91)
Sens	98.49	98.43	98.74	98.15	98.83	98.82	98.15	98.72	99.02	98.58	98.77	99.05
(sd)	(2.25)	(2.03)	(2.07)	(2.13)	(2.12)	(2.02)	(2.45)	(2.17)	(1.95)	(2.31)	(2.27)	(1.97)
Spec	85.62	85.41	85.06	85.26	86.84	85.83	85.21	87.05	87.30	77.85	80.35	81.95
(sd)	(15.1)	(15.3)	(15.1)	(14.9)	(14.8)	(15.1)	(14.7)	(14.6)	(14.5)	(17.2)	(15.5)	(14.8)
NB												
Acc	95.92	96.30	96.92	95.42	95.64	95.78	95.30	95.26	95.50	96.52	96.59	96.57
(sd)	(2.47)	(2.51)	(2.27)	(2.52)	(2.46)	(2.72)	(2.42)	(2.52)	(2.59)	(2.63)	(2.62)	(2.44)
Sens	95.59	96.10	96.78	95.13	95.41	95.82	94.97	95.09	95.45	98.01	98.05	98.06
(sd)	(2.92)	(3.00)	(2.76)	(2.95)	(3.00)	(3.04)	(2.93)	(3.00)	(3.07)	(2.38)	(2.15)	(2.08)
Spec	97.76	97.50	97.59	96.97	96.95	95.71	96.71	96.18	96.02	90.07	90.07	90.11
(sd)	(6.06)	(7.14)	(5.78)	(7.32)	(7.21)	(8.74)	(7.64)	(9.45)	(8.78)	(10.9)	(12.3)	(11.0)
SVM												
Acc	96.57	96.59	96.57	95.83	96.07	96.33	96.02	95.88	96.11	95.88	96.14	96.45
(sd)	(2.85)	(2.37)	(2.20)	(3.01)	(2.75)	(2.42)	(2.87)	(2.61)	(2.70)	(3.02)	(2.63)	(2.67)
Sens	98.23	98.19	98.34	97.72	97.83	98.06	98.09	97.72	98.02	98.26	98.06	98.45
(sd)	(2.41)	(2.12)	(2.14)	(2.57)	(2.32)	(2.20)	(2.34)	(2.39)	(2.32)	(2.17)	(2.19)	(2.06)
Spec	88.39	88.76	87.85	86.24	87.47	87.90	85.98	87.15	86.97	84.50	86.96	86.97
(sd)	(12.6)	(11.6)	(10.7)	(13.6)	(12.7)	(11.5)	(13.4)	(12.3)	(11.8)	(14.4)	(12.3)	(11.8)

A Appendix

The simulation scenarios S1 and S2 which corresponds to correlations of 0.2 and 0.5 respectively were carried out in this article. In this section, we show results of these scenarios. The Figure A.1 shows the averages of accuracy for low correlation simulation scenario: S1 and the Figure A.2 shows the performance of average

accuracies for intermediate correlation simulation scenario: S2. Each of the simulated data was randomly partitioned into training and testing data, 70% of the samples were applied on FR methods and the features were ranked based on their importance. The reduced number of top selected features were utilized for training the classification models. For fair comparison, all of the methods were evaluated with accuracy, sensitivity and specificity, averaged over 100 times and are shown in Tables A.1 and A.2.

Table A.1: The averages of accuracy, sensitivity and specificity along with the standard deviation across 100 number of iterations for the simulation data S1

	#p	F-score			IG			Chi2			MRMR		
		30	90	200	30	90	200	30	90	200	30	90	200
AB	Acc	55.70 (sd) (6.80)	54.36 (6.96)	54.80 (7.45)	54.88 (7.24)	53.80 (8.30)	54.42 (6.97)	53.80 (7.56)	56.16 (7.05)	55.64 (7.09)	52.94 (7.23)	53.46 (7.08)	54.48 (6.95)
	Sens	56.25 (sd) (9.32)	55.98 (10.7)	56.99 (12.4)	56.19 (11.5)	55.59 (11.5)	56.07 (11.6)	51.17 (10.4)	54.02 (10.8)	54.01 (12.9)	53.41 (10.5)	53.88 (10.7)	55.77 (12.3)
	Spec	55.94 (sd) (10.5)	53.87 (12.2)	53.97 (12.2)	54.73 (10.5)	53.27 (12.0)	54.33 (12.4)	57.11 (11.9)	59.23 (10.6)	58.47 (11.4)	53.30 (11.1)	54.00 (11.7)	54.26 (9.93)
RF	Acc	56.22 (sd) (7.27)	54.78 (6.94)	54.78 (7.19)	54.98 (7.29)	54.78 (7.19)	53.46 (6.43)	54.76 (7.41)	55.68 (7.65)	55.04 (7.27)	52.56 (7.72)	52.34 (7.32)	53.72 (7.26)
	Sens	57.81 (sd) (12.0)	56.78 (14.2)	57.87 (15.9)	56.32 (12.3)	56.06 (13.9)	54.91 (17.0)	52.43 (11.1)	52.88 (13.1)	52.76 (15.0)	53.38 (12.6)	52.94 (12.0)	55.29 (15.3)
	Spec	56.10 (sd) (11.6)	55.03 (12.9)	55.18 (17.2)	55.41 (12.5)	55.74 (13.3)	55.56 (17.7)	58.34 (11.7)	60.11 (12.5)	60.13 (15.8)	53.07 (12.3)	53.36 (12.7)	54.68 (15.6)
NB	Acc	56.14 (sd) (7.27)	56.32 (6.44)	57.20 (7.18)	56.00 (7.02)	55.56 (7.34)	54.92 (6.97)	55.12 (7.48)	56.88 (7.42)	56.80 (6.65)	52.96 (7.74)	54.20 (7.67)	55.10 (7.18)
	Sens	57.05 (sd) (11.3)	57.11 (10.8)	58.72 (11.9)	57.54 (10.8)	56.54 (11.1)	55.45 (10.8)	52.76 (11.3)	54.83 (11.2)	55.28 (11.6)	52.94 (12.3)	53.64 (10.4)	55.90 (10.8)
	Spec	55.90 (sd) (10.6)	56.53 (10.4)	57.17 (11.8)	55.45 (9.86)	55.61 (10.8)	55.43 (10.7)	58.11 (10.5)	59.69 (10.4)	59.38 (11.3)	53.60 (11.0)	55.38 (11.4)	54.85 (10.9)
SVM	Acc	55.66 (sd) (7.28)	54.66 (6.85)	56.80 (7.29)	55.88 (8.00)	54.14 (7.85)	54.00 (6.32)	55.08 (7.56)	55.20 (7.75)	57.02 (7.25)	52.88 (7.65)	53.02 (6.40)	54.74 (7.44)
	Sens	56.31 (sd) (11.14)	55.57 (10.1)	59.26 (11.0)	57.97 (11.4)	54.90 (12.0)	54.58 (10.2)	52.88 (10.6)	52.27 (10.9)	54.29 (11.0)	53.92 (11.2)	53.26 (10.3)	55.69 (10.2)
	Spec	55.70 (sd) (10.9)	54.19 (10.8)	55.58 (11.1)	54.58 (11.7)	54.32 (12.5)	54.24 (10.4)	58.07 (11.6)	58.64 (10.9)	59.07 (11.3)	52.48 (12.3)	54.18 (9.00)	54.16 (10.8)

Figure A.1: Average accuracies over 100 iterations in simulation data S1

Table A.2: The averages of accuracy, sensitivity and specificity along with the standard deviation across 100 number of iterations for the simulation data S2

#p	F-score			IG			Chi2			MRMR		
	30	90	200	30	90	200	30	90	200	30	90	200
AB	Acc	58.98	59.30	57.50	58.70	58.50	58.52	57.24	58.68	57.38	54.78	55.56
	(sd)	(7.12)	(6.76)	(6.56)	(7.67)	(7.87)	(6.90)	(7.55)	(7.43)	(6.39)	(6.64)	(7.23)
	Sens	58.95	59.35	58.58	57.06	57.16	59.68	57.66	59.01	59.19	54.81	56.22
	(sd)	(10.5)	(11.5)	(11.1)	(11.4)	(11.9)	(10.5)	(11.9)	(11.1)	(10.8)	(11.9)	(12.1)
	Spec	59.48	59.77	58.74	61.13	60.33	60.26	57.51	59.39	58.68	55.63	55.88
	(sd)	(10.5)	(9.53)	(10.8)	(11.3)	(10.5)	(11.1)	(9.94)	(11.1)	(10.3)	(9.79)	(11.2)
RF	Acc	60.02	61.54	61.34	58.58	58.78	59.38	58.12	59.32	59.72	55.28	55.32
	(sd)	(7.69)	(7.34)	(6.24)	(8.04)	(7.88)	(7.64)	(7.06)	(6.73)	(7.14)	(6.91)	(7.82)
	Sens	59.78	62.95	59.58	57.40	58.17	58.28	59.44	61.22	61.69	55.66	57.58
	(sd)	(11.1)	(11.8)	(14.9)	(12.4)	(13.9)	(15.9)	(11.3)	(12.6)	(15.2)	(12.9)	(14.3)
	Spec	61.36	61.74	61.82	60.83	61.22	63.43	58.31	59.40	60.85	56.31	55.49
	(sd)	(12.2)	(13.3)	(15.4)	(11.5)	(12.0)	(14.3)	(10.6)	(11.9)	(15.6)	(10.8)	(13.2)
NB	Acc	61.24	62.10	61.10	69.72	69.94	61.79	59.20	60.36	61.54	55.52	57.06
	(sd)	(7.65)	(7.45)	(7.56)	(7.38)	(7.46)	(6.84)	(6.89)	(7.82)	(6.68)	(7.01)	(7.55)
	Sens	60.68	62.37	61.22	58.45	59.09	60.26	59.67	61.06	62.23	55.24	57.65
	(sd)	(12.0)	(11.1)	(12.5)	(10.8)	(12.3)	(11.0)	(9.88)	(11.2)	(10.6)	(11.6)	(12.4)
	Spec	62.22	62.33	61.92	61.31	61.28	63.95	59.22	60.32	61.70	56.66	57.43
	(sd)	(10.7)	(11.9)	(11.4)	(10.1)	(10.5)	(10.3)	(9.73)	(11.0)	(10.7)	(9.81)	(10.7)
SVM	Acc	59.50	59.70	60.16	58.40	57.58	59.94	57.08	57.28	59.98	55.18	54.46
	(sd)	(7.64)	(7.09)	(7.57)	(7.75)	(7.27)	(7.16)	(7.83)	(7.19)	(6.69)	(7.74)	(7.05)
	Sens	59.07	60.69	60.31	57.54	56.40	59.29	57.09	57.16	60.71	55.12	54.00
	(sd)	(11.0)	(10.5)	(10.9)	(11.4)	(11.4)	(10.6)	(11.4)	(11.3)	(10.1)	(12.5)	(11.8)
	Spec	60.29	58.96	60.68	59.48	59.08	61.18	67.65	68.26	59.89	55.85	55.35
	(sd)	(12.0)	(11.0)	(11.3)	(11.2)	(10.6)	(10.6)	(10.4)	(10.2)	(9.96)	(10.4)	(11.2)

Figure A.2: Average accuracies over 100 iterations in simulation data S2

Acknowledgement: We would like to thank Dr. Jonathon E Mohl for reviewing and providing the feedback on the primary draft of the manuscript. We also thank Fredrick Ayivor for his helpful suggestions during the initial discussions. We finally thank the reviewers for their careful review and constructive comments on the manuscript.

References

- [1] Michael Weber, Ines Hellmann, Michael B. Stadler, Liliana Ramos, Svante Pääbo, Michael Rebhan, and Dirk Schübeler. Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome. *Nature Genetics*, 39:457–466, 2007.
- [2] Peter A. Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13:484–492, 2012.
- [3] Andrew P. Feinberg, Rolf Ohlsson, and Steven Henikoff. The epigenetic progenitor origin of human cancer. *Nature Reviews Genetics*, 7:21–33, 2006.
- [4] Konstantin Lepikhov, Julia Arand, Mark Wossidlo, and Joern Walter. Epigenetic reprogramming in mammalian development. 2012.
- [5] Thomas Mikeska and Alexander Dobrovic. Epigenetic basis of human cancer. 2017.

- [6] Robin David Holliday and James E. Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187 4173:226–32, 1975.
- [7] Mun kit Choy, Mehregan Movassagh, Hock-Guan Goh, Martin R. Bennett, Thomas A. Down, and Roger SY Foo. Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. In *BMC Genomics*, 2010.
- [8] Guiyou Liu, Fang Zhang, Yang Hu, Yongshuai Jiang, Zhongying Gong, Shoufeng Liu, Xiuju Chen, Qinghua Jiang, and Jun-wei Hao. Genetic variants and multiple sclerosis risk gene slc9a9 expression in distinct human brain regions. *Molecular Neurobiology*, 54:6820–6826, 2016.
- [9] Yuya Kobayashi, Devin M Absher, Zulfiqar G. Gulzar, Sarah R Young, Jesse K McKenney, Donna M. Peehl, James D Brooks, Richard M. Myers, and Gavin Sherlock. Dna methylation profiling reveals novel biomarkers and important roles for dna methyltransferases in prostate cancer. *Genome research*, 21 7:1017–27, 2011.
- [10] Rocío González Urdinguio, Jose Vicente Sanchez-Mut, and Manel Esteller. Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. *The Lancet Neurology*, 8:1056–1072, 2009.
- [11] Myungjin Kim, Tiffany I. Long, Kazuko Arakawa, Renwei Wang, Mimi C. Yu, and Peter W. Laird. Dna methylation as a biomarker for cardiovascular disease risk. In *PloS one*, 2010.
- [12] Manel Esteller. Aberrant dna methylation as a cancer-inducing mechanism. *Annual review of pharmacology and toxicology*, 45:629–56, 2005.
- [13] Stefanie Teschler, Julia Gotthardt, Gerhard Dammann, and Reinhard H. Dammann. Aberrant dna methylation of rdna and prima 1 in borderline personality disorder. 2019.
- [14] Aditi Mehta, Stephanie Dobersch, Reinhard H. Dammann, Saverio Bellusci, Olga N. Ilinskaya, Thomas Braun, and Guillermo Quiroga Barreto. Validation of tuba1a as appropriate internal control for normalization of gene expression analysis during mouse lung development. In *International journal of molecular sciences*, 2015.
- [15] Marina Bibikova, Zhenwu Lin, Lixin Zhou, Eugene Chudin, Eliza Wickham Garcia, Bonnie Wu, Dennis Doucet, Neal J. Thomas, Yunhua Wang, Ekkehard Vollmer, Torsten Goldmann, Carola Seifart, Wei Jiang, David L. Barker, Mark S. Chee, Joanna Floros, and Jian-Bing Fan. High-throughput dna methylation profiling using universal bead arrays. *Genome research*, 16 3:383–93, 2006.
- [16] Vardhman K. Rakyan, Thomas A. Down, Siarhei Maslau, Toby Andrew, Tsun-Po Yang, Huriya Beyan, Pamela Whittaker, Owen T. McCann, Sarah Finer, Ana M. Valdes, R David Leslie, Panagiotis Deloukas, and Timothy David Spector. Human aging-associated dna hypermethylation occurs preferentially at bivalent chromatin domains. *Genome research*, 20 4:434–9, 2010.
- [17] Andrew E. Teschendorff, Usha Menon, Aleksandra Gentry-Maharaj, Susan J Ramus, Daniel J. Weisenberger, Hui Shen, Michaela Campan, Houtan Noushmehr, Christopher G. Bell, Alexander Peter Maxwell, David A. Savage, Elisabeth Mueller-Holzner, Christian Marth, Gabrijela Kocjan, Simon A. Gayther, Allison Jones, Stephan Beck, Wolfgang Wagner, Peter W. Laird, Ian J. Jacobs, and Martin Widschwendter. Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome research*, 20 4:440–6, 2010.
- [18] Minoru Toyota, Nita Ahuja, Mutsumi Ohe-Toyota, James G. Herman, Stephen B. Baylin, and J P Issa. CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 96 15:8681–6, 1999.
- [19] Hokeun Sun and Shuang Wang. Penalized logistic regression for high-dimensional dna methylation data with case-control studies. *Bioinformatics*, 28 10:1368–75, 2012.
- [20] Devin C. Koestler, Carmen J. Marsit, Brock C. Christensen, Margaret R. Karagas, Raphael Bueno, David J. Sugarbaker, Karl T. Kelsey, and E. Andres Houseman. Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics*, 26 20:2578–85, 2010.
- [21] E. Andres Houseman, Brock C. Christensen, Margaret R. Karagas, Margaret R. Wrensch, Heather H. Nelson, Joseph Wiemels, Shichun Zheng, John K. Wiencke, Karl T. Kelsey, and Carmen J. Marsit. Copy number variation has little impact on bead-array-based measures of dna methylation. *Bioinformatics*, 25 16:1999–2005, 2009.
- [22] Fabian Model, Péter Adorján, Alexander Olek, and Christian Piepenbrock. Feature selection for dna methylation based cancer classification. *Bioinformatics*, 17 Suppl 1:S157–64, 2001.
- [23] Abeer A. Raweh, Mohammad Nassef, and Amr Badr. Feature selection and extraction framework for dna methylation in cancer. 2017.
- [24] Lili Milani, Anders Lundmark, Anna Kiialainen, Jessica Nordlund, Trond Flaegstad, Erik Forestier, Mats Heyman, Gudmundur K Jónmundsson, Jukka Kanerva, Kjeld Schmiegelow, Stefan Söderhäll, Mats G. Gustafsson, Gudmar Lönnérholm, and Ann-Christine Syvänen. Dna methylation for subtype classification and prediction of treatment outcome in patients with childhood acute lymphoblastic leukemia. *Blood*, 115 6:1214–25, 2010.
- [25] Fabrizio Celli, Fabio Cumbo, and Emanuel Weitschek. Classification of large dna methylation datasets for identifying cancer drivers. *Big Data Research*, 13:21–28, 2018.
- [26] Sangjin Kim and Jong-Min Kim. Two-stage classification with sis using a new filter ranking method in high throughput data. *Mathematics*, 7(6), 2019.
- [27] S Kim and S Halabi. High dimensional variable selection with error control. *BioMed Research International*, 2016:1–11, 2016.
- [28] Abhishek A. Singh. Gene selection using high dimensional gene expression data: An appraisal. 2016.

- [29] Shuangge Ma and Jian Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9 5:392–403, 2008.
- [30] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Reply to talloen et al.: Independent filtering is a generic approach that needs domain specific adaptation. *Proceedings of the National Academy of Sciences*, 107(46):E175–E175, 2010.
- [31] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 107 21:9546–51, 2010.
- [32] Jun Lu, Robnet T. Kerns, Shyamal D. Peddada, and Pierre R. Bushel. Principal component analysis-based filtering improves detection for affymetrix gene expression arrays. In *Nucleic acids research*, 2011.
- [33] Daniel Ramsköld, Eric T. Wang, Christopher B. Burge, and Rickard Sandberg. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. In *PLoS Computational Biology*, 2009.
- [34] Huan Liu and Hiroshi Motoda. Computational methods of feature selection (chapman & hall/crc data mining and knowledge discovery series). 2007.
- [35] Le Song, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation. In *ICML*, 2007.
- [36] Pabitra Mitra, C. A. Murthy, and Sankar K. Pal. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:301–312, 2002.
- [37] Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern classification, 2nd edition. 2000.
- [38] J. Ross Quinlan. C4.5: Programs for machine learning. 1993.
- [39] Hanchuan Peng, Fuhui Long, and Chris H. Q. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238, 2003.
- [40] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.
- [41] Shuangge Ma and Ying Dai. Principal component analysis based methods in bioinformatics studies. *Briefings in bioinformatics*, 12 6:714–22, 2011.
- [42] Yaqian Guo, Trevor J. Hastie, and Robert Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8 1:86–100, 2007.
- [43] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
- [44] Okeh Um and Ica Oyeka. Estimating the fisher’s scoring matrix formula from logistic model. 2013.
- [45] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, August 2005.
- [46] Caiyan Li and Hongzhe Li. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics*, 4 3:1498–1516, 2010.
- [47] Benjamin M. Bolstad, 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 2:185–93, 2003.
- [48] Trevor J. Hastie, Robert Tibshirani, and Jerome H. Friedman. The elements of statistical learning: Data mining, inference, and prediction, 2nd edition. In *Springer Series in Statistics*, 2005.
- [49] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
- [50] Nir Friedman, Dan Geiger, and Moisés Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.
- [51] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.
- [52] Yoav Freund. An adaptive version of the boost by majority algorithm. In *COLT*, 1999.
- [53] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In *Australian Conference on Artificial Intelligence*, 2006.