Biostatistical Methods in High-Dimensional Estimation and Prediction Problems

Chengzhou Wu

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation
https://digitalcommons.memphis.edu/etd/3618

This Dissertation is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.
Biostatistical Methods in High-Dimensional Estimation and Prediction Problems

by

Chengzhou Wu

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Major: Epidemiology and Biostatistics

The University of Memphis
August 2024
Acknowledgments

I am deeply grateful to everyone who has supported me throughout this journey. Foremost among them is my advisor, Dr. Xichen Mou. Your guidance and mentorship have been invaluable to me, shaping me into the researcher I am today. I truly appreciate your unwavering support, encouragement, and readiness to provide guidance whenever needed.

I extend my heartfelt appreciation to my collaborators, Dr. Hongmei Zhang, Dr. Yu (Joyce) Jiang, and Dr. Meredith A Ray. Working with you all has been both enriching and rewarding, and I am thankful for the knowledge and support you have generously shared.

To our group members, Dr. Rajesh Tirpaul Melaram, Dr. Liang Li, Lu Xie, Siyao Liu, Shengqiang Chen, Fen Yang, Samia Sultana, Naznin Alam, Mohammad Nahian Ferdous Abrar, Nahid Sultana, James O Adefisoye, Xin Jin, and Michael Arthur Ofori, I am grateful for the opportunity to work alongside such a talented group.

I would also like to express my gratitude to my committee member, Dr. Ching-Chi Yang, for your valuable insights and contributions to improving this dissertation. Your feedback has been instrumental in shaping its outcome.

Lastly, I would like to express my gratitude to my family, especially my wife, Dr. Wang, for their unwavering support and understanding throughout this journey. Their encouragement and patience have been instrumental in enabling me to pursue my goals and aspirations. I am truly grateful for their love and sacrifices.
Abstract

High-dimensional data, such as DNA methylation data, refers to datasets where the number of variables (features) exceeds the number of samples by a significant margin. In our research, our objective is to utilize advanced biostatistical methods to accurately estimate and predict outcomes from these complex datasets.

In the first project, we aim to identify Differentially Methylated Regions (DMRs) within the human genome using a novel biostatistical method. These genomic regions or specific positions exhibit distinct methylation patterns across various phenotypes. Despite existing methodologies like EWAS and dmrff, challenges such as low statistical power, high false positive rates, and complexities in confounder control persistently hinder progress in this field. To address these issues, our research focuses on developing an innovative approach using the Generalized Beta distribution, which effectively models DNA methylation data and accounts for correlation patterns through shared parameters. Inspired by the unique characteristics of DNA methylation, our method demonstrates significant power in identifying potential biomarkers through simulation studies and real-world data analyses.

In the second project, we aim to develop comprehensive prediction models for allergic diseases by integrating clinical variables with epigenetic risk factors identified through advanced feature selection methods. Asthma, characterized by varied clinical manifestations across different life stages, serves as our focal point. Our objective is to enhance predictive accuracy significantly through robust models that incorporate both clinical and epigenetic markers from DNAm profiles obtained at birth.
In the third project, we explore anxiety and depression prediction using machine learning algorithms, analyzing a very large dataset. Results demonstrate the superiority of random forest over alternative methods, evidenced by comparable accuracy metrics and superior area under the curve (AUC) scores. Feature importance analysis reveals crucial interpretable predictors including children’s demographics and parental mental health. The large number of testing participants makes the model and selected features robust for prediction. The study represents the largest dataset utilized to date for predicting children’s mental health, and its features are smaller in scale compared to previously reported methods. Overall, the study underscores the potential of some predictors in mental health prediction, offering algorithm insights for research and clinical applications.
To my family
# Table of Contents

Dedication ................................................................  v

List of Tables ................................................................. viii

List of Figures ................................................................. ix

Chapter 1 High-Dimensional Estimation with Generalized Beta Regression ........................................... 1
  1.1 Background .......................................................... 1
  1.2 Methods ............................................................ 5
    1.2.1 Segment CpG sites into blocks ............................ 5
    1.2.2 Model the blocks ............................................. 6
    1.2.3 Apply likelihood ratio test ................................ 8
  1.3 Results ................................................................. 9
    1.3.1 Simulations .................................................... 9
    1.3.2 Real data analysis ........................................... 13
    1.3.3 Consistency of the findings .............................. 14
    1.3.4 Biological relevant analysis ............................. 17
    1.3.5 Program complexity analysis ........................... 19
    1.3.6 R package implementation .............................. 20
  1.4 Materials ............................................................. 20
    1.4.1 Simulation method .......................................... 20
    1.4.2 The three real datasets ................................... 22
  1.5 Discussion .......................................................... 22
  1.6 Conclusion ........................................................ 25

Chapter 2 Integrating High-Dimensional Data Features for Predictive Modeling .......................................... 27
  2.1 Background .......................................................... 27
  2.2 Methods ............................................................. 29
    2.2.1 Feature selection ............................................ 30
Table of Contents

2.2.2 Prediction model ............................................ 32
2.2.3 Evaluation matric .......................................... 36
2.2.4 Clinical features in the real data .......................... 38

2.3 Results .......................................................... 38
  2.3.1 Feature selection comparison .............................. 38
  2.3.2 Training and testing in F1 generation ...................... 40
  2.3.3 Validation in the F2 generation ............................ 44

2.4 Discussion ..................................................... 46
2.5 Conclusion ...................................................... 48

Chapter 3 Understanding and Forecasting Childhood Depression and Anxiety .............................. 49
  3.1 Introduction .................................................... 49
  3.2 Aim ............................................................. 50
  3.3 Methods ........................................................ 50
    3.3.1 Participants ............................................... 50
    3.3.2 Features .................................................. 51
    3.3.3 Data analysis ............................................. 54
  3.4 Results ........................................................ 57
  3.5 Discussion and Conclusion ................................... 61

Bibliographic references ........................................ 63

Appendix A For Chapter 1 ........................................ 69
Appendix B For Chapter 2 ........................................ 104
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.1</td>
<td>Number of DMRs and DMPs</td>
<td>15</td>
</tr>
<tr>
<td>Table 1.2</td>
<td>Cross-validation of the identified CpGs in other datasets</td>
<td>17</td>
</tr>
<tr>
<td>Table 2.1</td>
<td>Clinical variables used for asthma prediction.</td>
<td>39</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Performance metrics of different algorithms at age 26.</td>
<td>44</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Percentage change for integrating epigenetic predictors into the model.</td>
<td>44</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>Performance metrics of different algorithms for the F2 dataset.</td>
<td>45</td>
</tr>
<tr>
<td>Table 2.5</td>
<td>Percentage change for integrating epigenetic predictors in validation.</td>
<td>45</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Clinical Variables Used for Anxiety Prediction.</td>
<td>53</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Clinical Variables Used for Depression Prediction.</td>
<td>55</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Performance metrics of different algorithms for the anxiety prediction.</td>
<td>57</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Performance metrics of different algorithms for depression prediction</td>
<td>58</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>Feature Importance for Anxiety Prediction.</td>
<td>60</td>
</tr>
<tr>
<td>Table 3.6</td>
<td>Feature Importance for Depression Prediction.</td>
<td>60</td>
</tr>
<tr>
<td>Table A.1</td>
<td>Running Complexity 1</td>
<td>92</td>
</tr>
<tr>
<td>Table A.2</td>
<td>Running Complexity 2</td>
<td>93</td>
</tr>
<tr>
<td>Table B.1</td>
<td>Performance metrics of different algorithms at age 10</td>
<td>105</td>
</tr>
<tr>
<td>Table B.2</td>
<td>Performance metrics of different algorithms at age 18</td>
<td>105</td>
</tr>
<tr>
<td>Table B.3</td>
<td>Performance metrics of different algorithms at age 26 (Persistent Asthma)</td>
<td>106</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Correlation heatmap for a segment of adjacent CpGs. The labels on the x and y axes are CpG names assigned by Illumina.</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>Power and false positive rates of gbdmr, dmrff, and EWAS across different block sizes.</td>
<td>11</td>
</tr>
<tr>
<td>1.3</td>
<td>Power over different signal strengths of gbdmr, dmrff, and EWAS across different block sizes.</td>
<td>12</td>
</tr>
<tr>
<td>1.4</td>
<td>Venn plot for identified CpGs across different approaches.</td>
<td>16</td>
</tr>
<tr>
<td>1.5</td>
<td>Extra CpGs genomic region distribution.</td>
<td>18</td>
</tr>
<tr>
<td>1.6</td>
<td>Power of dmrff when true $\gamma_{1b}$ ranges from 0 to 1.</td>
<td>24</td>
</tr>
<tr>
<td>2.1</td>
<td>Model workflow for allergic disease prediction.</td>
<td>30</td>
</tr>
<tr>
<td>2.2</td>
<td>Comparison of feature selection methods for individuals under the age of 10.</td>
<td>40</td>
</tr>
<tr>
<td>2.3</td>
<td>Comparison of epigenetic feature numbers for individuals under the age of 10.</td>
<td>41</td>
</tr>
<tr>
<td>2.4</td>
<td>Comparison of different algorithms for prediction asthma at age 10 and age 18.</td>
<td>42</td>
</tr>
<tr>
<td>2.5</td>
<td>Comparison of different algorithms for prediction asthma and persistent asthma at age 26.</td>
<td>43</td>
</tr>
<tr>
<td>3.1</td>
<td>Anxiety prediction algorithm comparison.</td>
<td>58</td>
</tr>
<tr>
<td>3.2</td>
<td>Depression prediction algorithm comparison.</td>
<td>59</td>
</tr>
<tr>
<td>B.1</td>
<td>Interpretable Feature Selection</td>
<td>104</td>
</tr>
</tbody>
</table>
Chapter 1
High-Dimensional Estimation with Generalized Beta Regression

1.1 Background

DNA methylation (DNAm), a biochemical procedure in which a methyl group is added to the cytosine, is a heritable epigenetic phenomenon. In mammals, DNAm mainly happens at the cytosine-phosphate-guanine (CpG) site. DNAm is critical in silencing retroviral elements, regulating tissue-specific gene expression, genomic imprinting, and X chromosome inactivation [1–3]. These processes influence the gene expression status and protein expressions, and they may be further involved in the development of diseases such as nervous disorders, cardiovascular diseases, and cancer [4–9].

DNAm is an important biomarker in the epigenetic study. CpG sites are methylated differently across phenotypes, including age, sex, and health conditions [10–12]. As DNAm plays a role in the biological pathways for various diseases, differentially methylated CpG sites help researchers identify risk factors of certain diseases and understand their underlying mechanisms. One classic strategy to study the association between CpG sites and phenotype is to regress each CpG’s DNAm levels on the phenotype through linear regression and screen for the most significant CpGs [13–15]. In this article, we refer to this approach as the traditional epigenome-wide association study (EWAS). However, phenotypes may be regulated jointly by CpGs in one region [16] or spatially close to each other. EWAS ignores such joint activities.
Thus, a method that has the ability to identify differentially methylated regions (DMRs) fits the underlying biological mechanisms and is expected to achieve better performance.

A variety of DMR detection methods have been developed. A widely used idea is to apply meta-analysis-based approaches to synthesize or summarize single CpG sites’ summary statistics from the traditional EWAS. For example, dmrff [17] divides the candidate regions based on the physical distances between CpG sites and applies the traditional EWAS on each CpG site in the same region. Then, it uses the inverse-variance weighted meta-analysis method to combine the EWAS outputs and searches for the most statistically significant sub-regions. Comb-p [18] starts from the EWAS p-values. It estimates spatial auto-correlation at different distance lags and adjusts p-values for adjacent CpG sites based on the Stouffer–Liptak–Kechris correction. The regions are determined based on the adjusted p-values, and the p-value of each region is re-calculated based on the Stouffer–Liptak–Kechris correction and auto-correlation function again. DMRcate [19] uses a Gaussian kernel smoother to adjust the p-values of the EWAS results for each CpG site. Then, the statistical significance is re-calculated based on the smoothed t-statistics, and the significant CpG sites within a specific distance are treated as DMRs. Similar to these methods, GlobalIP [20] also starts from the traditional EWAS outputs and uses the covariance matrix in the test statistic to account for the partial correlation among the CpG sites. In addition to the idea of summarizing EWAS results, seqlm [21] applies linear mixed models to segment CpG regions and fit the DNAm data simultaneously. [22] conducted a comprehensive comparison among the aforementioned methods, in which dmrff was shown to have the highest power in most settings. Following the findings of Lent et al., in this article, we set dmrff as the benchmark of DMR detection algorithms to
evaluate the proposed method.

Overall, both meta-analysis-based (e.g., dmrff) and model-based (e.g., seqlm) methods are restricted to the underlying framework of linear regression between DNA methylation (DNAm) level and phenotype. This framework assumes DNAm levels follow a normal distribution. However, the normality of DNAm data is questionable. Generated by Illumina’s 450k BeadChip or EPIC array, DNAm is calculated as \( \text{max}(|M|,0)/(|M| + |U| + c) \), denoted beta value with \( M \) and \( U \) being the intensities of methylated and unmethylated probes; \( c \) is a positive constant value to regularize the DNAm value when both \( M \) and \( U \) are too small [23]. Beta value is thus a ratio that ranges from 0 to 1, and it is natural to model beta values via beta distribution families. To meet the normality assumption, researchers often take logit transformation to convert a beta value to an M value. However, the normality of M value through such a transformation is problematic [24].

In this chapter, we propose an innovative model-based DMR detection approach, the generalized beta differentiated methylation region (gbdmr) detection method. Unlike conventional approaches that assume a normal distribution for the response variable, gbdmr employs Generalized Beta Regression that models multiple adjacent CpG sites jointly as ratios. To be specific, the gbdmr segments the candidate regions by physical coordinates and correlation patterns. It then uses the generalized beta distribution to straightforwardly model the DNAm of CpG sites in each region and calculates the corresponding p-values. The p-value for each region reflects statistical association significance between a phenotype and DNAm. Generalized beta distribution fits the definition of DNAm defined as beta values, and in addition, accounts for the
correlation structures among adjacent CpG sites. It has been used to model the
distribution of nominal family income and stock indexes in economics [25] and to
identify the intergenerational patterns of DNAm levels [26] in the epigenetic study.

To assess the performance of the proposed method, we conducted simulation studies
and real data analysis using gbdmr, dmrff, and EWAS. In simulation studies, gbdmr
achieved higher power than dmrff when the correlations between adjacent CpGs were
high; on the contrary, the power of dmrff is higher when the correlations are weaker.
We proved the performance decay of dmrff as the correlation increases theoretically,
and postulate this phenomenon may be universe among meta-analysis-based methods.
The gbdmr has been implemented in the R software. The R package is available
through GitHub: https://github.com/chengzhouwu/gbdmr.

The remaining article is structured as follows. The Method section includes the de-
tailed steps of gbdmr. The Results section presents the simulation results and real
data analysis. The Materials section provides technical details of simulation settings
and real data analysis. In the Discussion section, we explain the possible reasons for
the performance disparity between gbdmr, dmrff, and other DMR detection methods
under different correlations of adjacent CpG sites. The Conclusion section summarizes
the key findings and the implications. We include the theoretical proof, additional
simulation and data analysis results in the Supplementary Information.
1.2 Methods

The gbdmr consists of three key steps as illustrated in Section 1.2.1, 1.2.2, and 1.2.3 below.

1.2.1 Segment CpG sites into blocks

CpG sites are more likely to be correlated when their physical locations are close to each other [16, 27, 28]. We implement a strategy to identify highly correlated regions: CpG sites are ordered by the chromosome numbers and chromosome coordinates. In each chromosome, adjacent CpG sites’ Pearson correlation is calculated one by one. A chain of CpG sites forms a block if the correlation of each pair of neighboring CpG sites is higher than a certain threshold. We suggest a threshold of 0.5, i.e., consecutive CpG sites with neighboring correlation stronger than 0.5 form a block. If a CpG site does not have a sufficiently strong correlation with either of its neighbors, it forms a block by itself. For example, Figure 1.1 presents the correlation heatmap of a segment of adjacent CpGs in chromosome 1. We identify the sites and region based on the proposed strategy: The correlations of cg17177602 and cg08884932, cg08884932 and cg11225330 are larger than 0.5, which forms a block of three consecutive CpG sites. The other CpGs do not have a strong enough correlation with their neighbors. Thus, each of them forms a block consisting of a single CpG site.
1.2.2 Model the blocks

Suppose we divide $m$ consecutive CpG sites into $B$ blocks following the strategy of block segmenting in section 2.1. Denote $L_b$ the block size, i.e., number of CpG sites in the $b$th block, where $b = 1, \ldots, B$ and $\sum_{b=1}^{B} L_b = m$. When $L_b = 1$, the $b$th block is a single CpG site; when $L_b \geq 2$, the $b$th block is a region containing multiple CpG sites. We use generalized beta distribution to model the DNAm level(s) of CpG site(s) in each block.

Figure 1.1: Correlation heatmap for a segment of adjacent CpGs. The labels on the x and y axes are CpG names assigned by Illumina.
In the $b$th block, denote $Z_b = (Z_{1b}, \ldots, Z_{L_b b})$, DNAm levels of the $L_b$ CpG sites. Following [29], we define $Z_b$ follows a $L_b$-variate generalized beta distribution, denoted by $Gbeta(\alpha_b, \beta_b)$, if $Z_{lb} = P_{lb}/(P_{lb} + Q_b)$ for $l = 1, \ldots, L_b$, where $P_{lb} \sim Gamma(\alpha_{lb}, 1)$, $Q_b \sim Gamma(\beta_b, 1)$, $P_{lb}$’s and $Q_b$ are independent, and $\alpha_b = (\alpha_{1b}, \ldots, \alpha_{L_b b})$. Here $\alpha_{lb} > 0$ for $l = 1, \ldots, L_b$ and $\beta_b > 0$. The density function $f(Z_b|\alpha_b, \beta_b)$ can be expressed as

$$f(Z_b|\alpha_b, \beta_b) = \frac{\Gamma(\sum_{l=1}^{L_b} \alpha_{lb} + \beta_b) \prod_{l=1}^{L_b} \left\{ \left( \frac{Z_{lb}}{1-Z_{lb}} \right)^{\alpha_{lb} - 1} \left( \frac{1}{1-Z_{lb}} \right)^2 \right\}}{\Gamma(\beta_b) \prod_{l=1}^{L_b} \Gamma(\alpha_{lb}) \left\{ 1 + \sum_{l=1}^{L_b} \left( \frac{Z_{lb}}{1-Z_{lb}} \right) \right\} \sum_{l=1}^{L_b} \alpha_{lb} + \beta_b}$$ \hspace{1cm} (1.1)

Note that when $L_b = 1$, i.e., the block contains one single CpG site, the generalized beta distribution is trivialized to a univariate beta distribution. When $L_b \geq 2$, the marginal distribution of each CpG site also follows a univariate beta distribution. Thus, the generalized beta distribution has the ability to model each CpG with an ordinary beta distribution, which fits the definition of the beta value of a CpG site. Through the shared $Q_b$ in the definition, a generalized beta distribution incorporates correlation among CpG sites. This feature enables us to model correlation structures among adjacent CpG sites.

Now, we link the expected value of DNAm levels to the phenotype of interest through generalized beta regression. Denote $Z_{bi} = (Z_{i1b}, \ldots, Z_{iL_b b})$ the DNAm levels of the $b$th block for the $i$th sample, $i = 1, \ldots, n$. We assume $Z_{bi} \sim Gbeta(\alpha_{bi}, \beta_{bi})$, where $\alpha_{bi} = (\alpha_{i1b}, \ldots, \alpha_{iL_b b})$. Denote $X^i = (1, X_{i1}, \ldots, X_{ip})^\top$ a vector of independent variables for sample $i$. Here 1 is for the intercept; $X_{i1}$ is the phenotype of interest for the $i$th sample; $X_{i1}$ could be binary (e.g., sex) or continuous (e.g., age); $X_{i2}, \ldots, X_{ip}$ are the
\( p - 1 \) covariates or confounders that need to be adjusted for. We build a logit function to link the independent variables to the mean of \( Z^{i}_{lb} \) for \( l = 1, \ldots, L_b \) as below.

\[
\text{logit}\{E(Z^{i}_{lb})\} = \text{logit}\left(\frac{\alpha^{i}_{lb}}{\alpha^{i}_{lb} + \beta^{i}_{b}}\right)
\]

\[
= \log(\alpha^{i}_{lb}) - \log(\beta^{i}_{b})
\]

\[
= X^{\top} \gamma^{lb}, \quad (1.2)
\]

where \( \gamma^{lb} = (\gamma^{0lb}, \gamma^{1lb}, \ldots, \gamma^{plb})^\top \), in which \( \gamma^{0lb} \) denotes the CpG-specific intercept for the \( l \)th CpG in the \( b \)th block, and \( \gamma^{1lb}, \ldots, \gamma^{plb} \) denote the block-specific coefficients of independent variables for the \( b \)th block. Note that \( \gamma^{1lb} \) is of our interest indicating the effect of variable \( X_1 \) on DNAm in block \( b \).

1.2.3 Apply likelihood ratio test

Combining the density function (1.1) and Equation (1.2), we can obtain the likelihood function \( L(\theta_{b}) = \Pi_{i=1}^{n} f(Z^{i}_{lb} | \theta_{b}, X^{i}) \) for the \( b \)th block. Here \( n \) is the sample size, and we reparameterize the parameters \( (\alpha_{b}, \beta_{b}) \) to \( \theta_{b} = (\gamma^{lb}, \beta_{b}) \) based on Equation (1.2). To apply the likelihood ratio test, we calculate the difference of two log-likelihoods

\[
2 \ln\{\sup_{\theta_{b} \in \Theta} L(\theta_{b})\} - 2 \ln\{\sup_{\theta_{b} \in \Theta_{0}} L(\theta_{b})\},
\]

where \( \Theta \) denotes the whole parameter space for \( \theta_{b} \), and \( \Theta_{0} \) denotes the subspace constrained by the null hypothesis \( \gamma^{1lb} = 0 \). This difference follows a chi-square distribution with a degree of freedom 1. We apply the likelihood ratio test to obtain the p-value for each block. In the end, all p-values will be adjusted for multiple testing across the blocks. The statistically significant blocks are identified as DMRs.
1.3 Results

The performance of gbdmr was evaluated through both simulation studies and real data analyses. In a recent study that compared various DMR detection methods, including DMRcate, comb-p, seqlm, GlobalIP, and dmrff [22], it was found that dmrff had the highest power in most settings while maintaining a low false positive rate. Therefore, dmrff was chosen as the benchmark DMR detection method. Additionally, we include the traditional EWAS, which did not consider correlation structures, as a comparison approach to assess the benefit of DMR in detecting potentially informative CpGs. It is important to note that different methods for DMRs may not identify the exact same regions even if they are in close proximity. For instance, one method may detect CpG sites 1 to 3 as a DMR, while another method may detect CpG sites 2 to 6 as a DMR. Hence, using the number of DMRs alone is not an appropriate method to evaluate the performance of different detection methods. Instead, a more accurate approach is to consider the number of DMR CpGs, which refers to the total number of CpG sites within the positive DMRs identified by the method. Additionally, isolated single CpG sites that show differential methylation are referred to as differentially methylated positions (DMPs). We refer to these CpG sites as DMP CpGs in this paper.

1.3.1 Simulations

Simulations were used to evaluate how (1) the strength of correlations among adjacent CpG sites, (2) block size, and (3) effect size influence the performance of the three methods. We simulated DNAm data associated with a binary phenotype. Technical details are included in Section 1.4.1.
Figure 1.2 illustrates the power (sensitivity) and false positive rate (1 - specificity) across different correlation strengths of adjacent CpG sites, denoted by $\rho$, and block sizes. The power of all methods is calculated as the number of DMR CpGs divided by the total number of CpGs. When block size = 2, the power of traditional EWAS stays at a relatively stable level. GBDMR performs very similarly to EWAS when $\rho \leq$ the threshold 0.5. In this case, the correlation between the two adjacent CpG sites is not strong; we treat each as an independent CpG site and fit univariate beta regression on each CpG separately. We observe that the performance of univariate beta regression on single CpG sites is very close to that of EWAS which utilizes linear regressions. When $\rho > 0.5$, however, the power of GBDMR reaches almost 1, much higher than EWAS. It indicates that the GBDMR may perform better for strongly correlated CpG sites. In comparison with DMRF, we observe that DMRF has a higher power than GBDMR and EWAS when the correlation between adjacent CpG sites is low, while the power decreases as the correlation increases. We provide a theoretical explanation of this phenomenon in Section 1.5. As for specificity, all three methods keep the false positive rate at a low level. The patterns of the three methods for block size = 3 are similar to those of block size = 2. In Appendix 1.1 of Supplementary Materials, we visualized the frequency of block sizes in the three datasets of our real data analysis utilizing a histogram. We found more than 99.9% of block sizes range between 1 and 10, and over 99.8% of block sizes are less than 6. To reflect these block sizes in the real data and examine the robustness of GBDMR, we further applied a range of settings to compare the three methods, such as block sizes 4-10 and different correlation thresholds of $\rho$. The details of the extended simulations are in Appendices A.2 and A.3. In general, the simulation results are consistent with the findings presented in Figures 1.2.
Figure 1.2: Power and false positive rates of gbdmr, dmrff, and EWAS across different block sizes.

Figure 1.3 shows how power changes with signal strength when block sizes = 1, 2, 3, and 4. The $\rho$ is set to be 0.8 when the block size is larger than one. The signal strength in Figure 1.3 is defined by the difference of means in DNAm between the phenotype presence and absence groups divided by the standard deviation. The power is calculated as the number of positive DMP/DMR CpGs divided by the total number of CpGs when the block size equals/larger than 1. Note that when signal strength=0, i.e., the true distributions of the two groups are the same, the y-axis is a false positive
rate instead of power. When block size = 1, gbdmr works as well as EWAS, and both have slightly higher power than dmrff. For block size of 2, when signal strength = 0, i.e., the true distribution of the two phenotype groups are the same, all three methods identify no significant CpG sites, indicating a perfect specificity. When the signal strength lies between 0 and approximately 0.5, gbdmr has uniformly higher power than the other two methods. When the signal strength ≥ 0.5, the power of all three methods reaches one. Similar patterns are found when block size = 3 and
4. Various settings including larger block sizes are utilized to assess the performance of the three methods; these extended simulation results are presented in Appendix 2, in which gbdmr consistently exhibits greater statistical powers compared to both EWAS and dmrff.

In summary, gbdmr performs similarly to EWAS when the correlation between adjacent CpGs ≤ the threshold 0.5. Gbdmr is superior to the other two methods when CpG sites are highly correlated. On the contrary, dmrff shows a higher power when the correlation is weak, and the power decreases as the correlation grows stronger.

1.3.2 Real data analysis

We further applied the three methods to identify DMRs and DMPs in three DNA methylation datasets. One dataset is from a cohort study conducted on the Isle of Wight (IOW), United Kingdom[30], while the other two are from the Gene Expression Omnibus database repository (GSE) and focused on age-related DNA methylation profiles[31, 32]. We screened for DMRs and DMPs associated with various phenotypes, including age, sex, gestational age, and birth weight, and presented the number of identified DMRs and DMPs in Table 1.1. In total, ten analyses were conducted.

Table 1.1 shows that gbdmr outperformed dmrff in identifying DMR CpGs in most scenarios. Specifically, gbdmr detected more DMR CpGs than dmrff in eight out of ten analyses. Moreover, in eight out of ten analyses, the number of DMP CpGs identified by gbdmr was slightly higher than that by dmrff. In addition, both DMR detection methods outperformed EWAS in identifying a greater number of CpG sites (DMP CpGs + DMR CpGs).
Consistent with the simulation study in Figure 1.2, gbdmr is more sensitive than dmrff in identifying highly correlated DMRs in the real data analysis. Due to the block segmentation algorithm, gbdmr only identifies DMRs with neighboring DMR CpG’s correlation $\rho$ higher than the pre-specified threshold. In Table 1.1, compared to dmrff, gbdmr detects more DMR CpGs in most analyses, all of which have $\rho > 0.5$, while the correlations of adjacent DMR CpGs in dmrff vary in a wide range. In addition, we checked the overlapping DMR CpGs of both methods, and found gbdmr covered most of dmrff’s highly-correlated DMR CpGs. For instance, in Analysis 2, dmrff identified 91 regions with a block size equal to 2, and among them, 60 DMRs were highly correlated ($\rho > 0.5$), resulting in 120 CpGs. Gbdmr identified 104 of the 120 DMR CpGs identified by dmrff.

Furthermore, the DMRs identified in this study can be associated with biological and phenotype-related information using the EWAS Open Platform. This platform is a valuable resource that integrates knowledge from existing epigenome-wide association studies (EWAS) [33]. For instance, in Analysis 1, we were able to link the sex-related phenotype to 480 out of 900 DMRs identified by dmrff and 514 out of 1114 DMRs identified by gbdmr. This integration enables a deeper understanding of the potential biological significance of these DMRs in relation to specific traits or conditions.

1.3.3 Consistency of the findings

To ensure the robustness of our novel approach, we employed a dual strategy for validating its findings. Firstly, we examined the agreement of identified CpGs among various methods within the IOW dataset using Venn plots. Secondly, we conducted cross-validation of the positively identified CpGs across diverse datasets and for iden-
Table 1.1: Number of DMRs and DMPs

<table>
<thead>
<tr>
<th>Analysis index</th>
<th>dmrff</th>
<th>gbdmr</th>
<th>EWAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DMPs</td>
<td>DMRs</td>
<td>DMPs</td>
</tr>
<tr>
<td>1. IOW male vs female</td>
<td>2085</td>
<td>900 (2418)</td>
<td>1884</td>
</tr>
<tr>
<td>2. IOW male vs female</td>
<td>468</td>
<td>148 (391)</td>
<td>423</td>
</tr>
<tr>
<td>3. IOW male vs female</td>
<td>1050</td>
<td>418 (1163)</td>
<td>1092</td>
</tr>
<tr>
<td>4. IOW male vs female</td>
<td>4226</td>
<td>1986 (5314)</td>
<td>4300</td>
</tr>
<tr>
<td>5. IOW gestational age</td>
<td>908</td>
<td>333 (907)</td>
<td>996</td>
</tr>
<tr>
<td>6. IOW birth weight</td>
<td>210</td>
<td>78 (220)</td>
<td>219</td>
</tr>
<tr>
<td>7. GSE59065 male vs female</td>
<td>946</td>
<td>345 (1008)</td>
<td>1010</td>
</tr>
<tr>
<td>8. GSE59065 young vs old</td>
<td>20453</td>
<td>12376 (37651)</td>
<td>22814</td>
</tr>
<tr>
<td>9. GSE87571 male vs female</td>
<td>3381</td>
<td>1861 (5394)</td>
<td>3575</td>
</tr>
<tr>
<td>10. GSE87571 age</td>
<td>72070</td>
<td>28586 (87185)</td>
<td>101188</td>
</tr>
</tbody>
</table>

Using data in the IOW cohort and GSE datasets. DMPs: differential methylated CpG sites; DMRs: differential methylated regions that contain more than one CpG site. The number of DMR CpGs is included in the parentheses.

The adjacent correlation threshold used for gbdmr is 0.5.

IOW dataset age 26. IOW dataset age 18. IOW dataset age 10. IOW dataset age at birth. Gestational age is used to describe how far along the pregnancy is. It is measured in weeks.

Age-related profiling of DNA methylation in CD8+ T cells and age is recorded as a binary variable.

Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan Dataset.

tical phenotypes.

Figure 1.4 illustrates a Venn plot displaying sex-associated CpGs in Analysis 1 (Table 1.1). Remarkably, 3141 CpGs, including DMR CpGs (CpGs contained in DMRs) and DMP CpGs (single CpGs) are identified by all three methods. When using EWAS as the benchmark, 98.7% of CpGs from EWAS are identified by dmrff, while the overlapping between EWAS and gbdmr is 92.3%. The higher overlapping rate
between EWAS and dmrff can be attributed to dmrff’s reliance on EWAS summary results for subsequent analyses. In Appendix 3, DMPs identified by dmrff and gbdmr show strong consistency. Thus, the primary distinction between dmrff and gbdmr lies in their ability to identify DMRs. Further analysis shows gbdmr can identify a greater number of DMR CpGs, and more CpGs can be associated with the phenotype across different datasets.

![Venn plot for identified CpGs across different approaches.](image)

**Figure 1.4:** Venn plot for identified CpGs across different approaches.

To examine the consistency of findings across different datasets, we first applied dmrff, gbdmr and EWAS methods to identify CpGs associated with sex in IOW data at
age 26 (Analysis 1 in Table 1.1), then replicated these analyses in GSE59065 and GSE87571 and recorded the number of CpGs identified in the IOW cohort that are also detected in GSE59065 or GSE87571. We used the IOW data at age 26 to keep consistent with the age of adult samples in GSE59065 and GSE87571. The results are presented in Table 1.2. The gbdmr identified 767 CpGs (DMR plus DMP CpGs) in GSE59065 and 3158 CpGs in GSE87571 that overlap with those identified in IOW at age 26 years. Compared to dmrff, gbdmr identified less CpGs based on data in GSE59065 but more in GSE87571. When we exclude the DMP CpGs and focus only on the overlapping DMR CpGs, gbdmr identified more CpGs than dmrff in both datasets.

Table 1.2: Cross-validation of the identified CpGs in other datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Overlapping DMP + DMR CpGs</th>
<th>Overlapping DMR CpGs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EWAS</td>
<td>dmrff</td>
</tr>
<tr>
<td>GSE59065</td>
<td>549</td>
<td>797</td>
</tr>
<tr>
<td>GSE87571</td>
<td>2108</td>
<td>2731</td>
</tr>
</tbody>
</table>

Among the CpGs identified at age 26 in IOW (Analysis 1 of Table 1), the numbers of CpGs also identified in GSE59065 and GSE87571 across different approaches.

1.3.4 Biological relevant analysis

For the 2003 CpGs uniquely identified by gbdmr shown in Figure 1.4, we examined their biological functions through enrichment analyses. We first mapped the unique CpGs to their corresponding genes, and then conducted Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis. Finally, our analysis revealed the significance of two pathways: Tissue development
(adjusted p-value of 0.02) and Neuroactive ligand-receptor interaction (adjusted p-value of 0.03). With sex as a phenotypic variable of interest, the pathways based on mapped genes of CpGs uniquely identified using gbdmr underline sex differences and support the validity of these extra CpGs. For example, tissue development is the top pathway in which the corresponding genes are enriched and sex differences related to tissue development have been suggested across different studies[34–38]; A previous research report highlighted a significant sex-related difference in the Neuroactive ligand-receptor interaction pathway during Maternal Immune Activation[39].

Figure 1.5: Extra CpGs genomic region distribution.
DNAm microarray data typically spans various gene regions, encompassing TSS1500 (200–1500 bases upstream from the transcriptional start site, TSS), TSS200 (0–200 bases upstream from the TSS), 1st Exon, 5’UTR (5’ untranslated region), gene body, 3’UTR (3’ untranslated region), and IGR (intergenic region). We further examined the location of the extra CpGs from gbdmr with respect to genomic regions in Figure 1.5 and found that about one-third (29.6%) of the CpGs are in the region of TSS200 and TSS1500. The percentage increased to 42.0% with CpGs at 5’ UTR and 1st Exon also included. This indicates that many CpGs uniquely identified by gbdmr are informative in the regulation of gene expression.

1.3.5 Program complexity analysis

In Appendix 4, we included the computational complexity for gbdmr, EWAS, and dmrff across different sample sizes and numbers of CpGs. Specifically, Table 1 shows the increase of computing burden as sample size increases with the number of CpGs held constant, while Table 2 displays the changes of computing burden as the number of CpGs increases with sample size held constant. Our findings reveal that gbdmr scales effectively as the sample size and number of CpGs increase, with linear space complexity ($O(n)$) and linear time complexity ($O(n)$). In the comparison with EWAS and dmrff, gbdmr showed comparable memory-usage while a little bit longer running time. It is important to note that the difference in time efficiency between different methods primarily stems from the optimization processes. Both dmrff and EWAS use least square regression, a well-established technique with highly optimized R implementations. In contrast, gbdmr relies on the Nelder-Mead method, a general-purpose optimization algorithm focusing on numerical maximization of likelihood estimates [40, 41]. We integrated a parallel computing algorithm into gbdmr to ensure
its ability to efficiently handle genome-scale DNA methylation data with a reasonably large sample size. Coupled with its linear time complexity and the method’s capability to account for CpG dependence through generalized beta distribution, gbdmr emerges as a valuable tool in the realm of genomics.

1.3.6 R package implementation

The newly proposed method has been implemented in the R package, gbdmr. The package is available at GitHub: https://github.com/chengzhouwu/gbdmr with a detailed instruction file on how to find the clustered CpG sites, calculate the statistics and extract the CpG information. Gbdmr accommodates both categorical and continuous covariates of interest, and is applicable for various data sources, including 450k BeadChip, 850k BeadChip, or the EPIC array.

1.4 Materials

This section provides the specifics of our simulations and real data analyses.

1.4.1 Simulation method

To be consistent with the definition of beta values, we used beta distribution to simulate the DNAm levels of single CpG sites. In Figures 1.2 and 1.3, our phenotype is a binary variable representing the presence or absence of a certain phenotype. Given one CpG site, we generated 253/253 DNAm levels, representing samples whose trait is present/absent. We also explored scenarios where the presence/absence of DMRs is unbalanced by generating 422/84 DNAm levels. The results, as reported in Appendix 5, remain consistent with our findings. The total sample size 506 is the same as the
Isle of Wight dataset in the real data analysis. To control signal strength, DNAm levels of the two phenotype groups have the same standard deviation and different means by adjusting the shape and rate parameters of beta distributions. The signal strength is defined as the number of standard deviations between the mean of two phenotype groups.

To generate blocks of size $> 1$, we first simulated the DNAm levels of a single CpG site following beta distributions. Then, we generated a second vector such that it has a fixed correlation with the first CpG site with the same mean and standard deviation. We followed the same step to generate the third CpG site’s DNAm levels given the second, and so on. This procedure was used to simulate a chain of CpG sites with a given correlation between adjacent CpG sites.

When applying the three methods to the generated datasets, we used the simulated beta values for gbdmr, and used the transformed M values for dmrff and EWAS since these two approaches rely on linear regressions. For the power and false positive assessment in Figure 1.2, we generated CpGs’ values under different block sizes = 2 and 3 and different correlation strengths. For Figure 1.3, we generated DNAm following the same procedure as in Figure 1.2. When the block size is larger than 1, we set the correlation of CpG sites within each block to 0.8. For both simulations, 500 Monte Carlo replicates were generated for the purpose of power estimation and calculations of false positive rate and positive rate.

In addition to binary phenotypes, we extended our investigation to continuous phenotypes. Varied correlations between adjacent CpG sites and signal strengths were systematically employed to assess the performance of the three methods. Detailed
results of these simulations, including data generation settings, power, false positive rate, and effect size figures, are in Appendix 6. The simulation results are consistent with the findings for binary phenotypes.

1.4.2 The three real datasets

We analyzed three DNAm datasets. The first dataset was obtained from a birth cohort study conducted on the Isle of Wight in the United Kingdom. DNAm levels were measured in whole blood at different ages (birth, age 10, age 18, age 26) and preprocessed by background correlation, normalization, and batch effect removal. The final dataset contains 346,009 CpG sites with sample sizes ranging from 277 to 506. The other two datasets were obtained from the Gene Expression Omnibus (GEO) database repository. The first dataset (GSE59065) included 101 individuals (50 young and 51 old) and focused on age-related profiling of DNA methylation in CD8+ T cells. The second dataset (GSE87571) included 732 samples and investigated the continuous aging of the human DNA methylome throughout the lifespan.

For all the analyses, we used Bonferroni correction to adjust p-values. For dmrff, we divided regions using the default maximum distance of 500 bp. For gbdmr, we segmented blocks using a correlation threshold of 0.5.

1.5 Discussion

Different from the meta-analysis methods that summarize the EWAS results, the proposed method, gbdmr, is a model-based approach that fits the DNAm data by generalized beta distribution. In the simulation study, we demonstrated that the dmrff was less efficient when the correlation between adjacent CpG sites was strong.
This is counter-intuitive since DMR detection methods are expected to achieve a better performance in strong correlations. In Appendix 7, we show that the power of dmrff equals

\[
\text{Power} = P\left(Z > -\frac{\gamma_{1b}}{\sqrt{(1^\top \Omega^{-1} 1)^{-1}}} + z_{1-\alpha/2}\right) + P\left(Z < -\frac{\gamma_{1b}}{\sqrt{(1^\top \Omega^{-1} 1)^{-1}}} + z_{\alpha/2}\right),
\]

where \(\gamma_{1b}\) is the true effect of phenotype in Equation (1.2); \(1 = (1, \ldots, 1)^\top\);

\[
\Omega = \sigma_n^2 \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & 1 \end{bmatrix},
\]

an \(L_b\) by \(L_b\) matrix where \(L_b\) is the number of CpG sites in \(b\)th DMR;

\(\sigma_n = \sigma / \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}\); \(\sigma\) is the standard deviation of the DNAm levels of a CpG site; \(X_i\) is the phenotype of the \(i\)th sample; \(\bar{X} = \sum_{i=1}^n X_i / n\); \(\rho\) is the pairwise correlation among CpG sites in a DMR; \(Z\) is a random variable following standard normal distribution; \(z_{1-\alpha/2}\) and \(z_{\alpha/2}\) are \(1 - \alpha/2\) and \(\alpha/2\) quantile of the standard normal.

Figure 1.6 shows the power of dmrff versus the effect size \(\gamma_{1b}\) when \(\sigma_n = 1\), \(L_b = 2\), \(\rho = 0\), 0.5, and 0.9.

From Figure 1.6, we observe that the theoretical power of dmrff declines as \(\rho\) increases.
Here we provide an intuitive explanation of the theoretical results. The dmrff summarizes $p$ single CpG sites’ effects $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$ in the same region, where $\hat{\beta}$ is the EWAS effect estimates of $p$ CpG sites. Under a simplified condition that $\sigma$ is known, dmrff is equivalent to dividing the weighted average $A = (1\top \Omega^{-1} 1)^{-1} 1\top \Omega^{-1} \hat{\beta}$ by its standard error $\sqrt{(1\top \Omega^{-1} 1)^{-1}}$ as the test statistics. We examine two extreme cases: If the $p$ CpG sites are mutually independent, then $\rho = 0$ and $\Omega$ becomes a diagonal matrix. The dmrff is equivalent to a one-sample Z-test of sample size $p$ with i.i.d observations $\hat{\beta}_1, \ldots, \hat{\beta}_p$. On the other side, if $p$ CpGs are perfectly correlated, i.e., $\rho = 1$, 

**Figure 1.6**: Power of dmrff when true $\gamma_{lb}$ ranges from 0 to 1.
\(\hat{\beta}_1\) to \(\hat{\beta}_p\) will be exactly the same. In this case, there is only one effective observation, and the equivalent sample size is only one. This trend indicates that dmrff has a lower efficiency when the correlation grows stronger. Moreover, this explanation is not only restricted to dmrff, but may apply to a family of methods based on meta-analysis: Given a fixed number of CpG sites, a stronger inter-correlation means higher proportion of overlapping information among CpG sites, and thus fewer equivalent sample sizes can be used in summarizing the results. In contrast, gbdmr uses generalized beta distribution to directly model all CpG sites in a region and is not affected by the equivalent sample size shrinkage in meta-analysis approaches.

1.6 Conclusion

We proposed a novel model-based method for detecting DMRs, called gbdmr, which employs generalized beta regression to model correlated CpG sites. The package gbdmr, unlike traditional methods, does not necessitate the normality assumption on DNA methylation (DNAm) levels and is adept at considering the correlation structures among CpG sites. This approach exhibits a strong ability to identify informative CpG regions, especially in scenarios where there is a high degree of inter-correlation among these sites. The simulation studies show that gbdmr performed better than dmrff and traditional EWAS when the correlation between adjacent CpGs is high, while the dmrff achieves higher power when the correlation is weak. Both theoretical and heuristic explanations are provided for the performance decay of dmrff as correlation increases. Based on the real data analysis, gbdmr was able to identify a higher number of DMR CpGs compared to dmrff in most of the analyses. Further examination revealed that gbdmr was able to identify most of dmrff’s DMRs that
exhibit high correlations between adjacent CpGs. These findings are consistent with
the results obtained from the simulation study. In the future, a promising approach
would be to combine the strengths of dmrff and gbdmr to better adapt to different
situations.
Chapter 2

Integrating High-Dimensional Data Features for Predictive Modeling

2.1 Background

Asthma, a widespread chronic respiratory condition impacting individuals globally, poses a substantial health burden due to its diverse clinical presentations and profound effects on patients. According to the World Health Organization, asthma has an average global prevalence of 7%, with notable regional disparities and a rising incidence in numerous low to middle-income nations. This condition is typified by airway inflammation and hyperresponsiveness, resulting in symptoms such as wheezing, breathlessness, chest tightness, and coughing [42]. Childhood asthma and adult asthma exhibit distinct characteristics. Childhood asthma typically manifests with a higher prevalence among boys before puberty, often undergoes remission, and is associated with low mortality rates. Conversely, adult asthma shows a higher prevalence in females, tends to persist into adulthood, and carries a greater risk of mortality [43].

The prediction of asthma attacks has been a focus of research, with studies developing models based on various factors including biosignals such as Peak Expiratory Flow Rate (PEFR), Forced Expiratory Volume in one second (FEV1), and asthma symptoms, among others. Additionally, environmental conditions such as weather patterns and air pollution have also been integral components in the development of these predictive models. These predictive models aim to help asthmatic patients anticipate and prevent asthma attacks. Research has shown that combining patient
biosignals and environmental factors in predictive models is crucial for improved accuracy and sensitivity. Machine learning methods have been increasingly utilized to predict asthma exacerbations among asthmatic patients, showing promising results in terms of prediction performance[44–46]. However, challenges remain in terms of model generalizability and feature diversity, practicability, and the need for large datasets to enhance accuracy[44]. The incorporation of genomic risk scores to enhance childhood asthma prediction has been explored, albeit with reported improvements that may not reach statistical significance[47]. Despite this, there remains ample opportunity for further investigation in this area, particularly concerning the robustness of feature selection methodologies and comparative analyses of different algorithms. This suggests a promising avenue for future research endeavors aimed at refining predictive models for asthma.

Our study aims to advance the field of asthma disease prediction across various age demographics by incorporating both established clinical variables [48,49] and emerging epigenetic risk factors. To accomplish this, we did a comprehensive exploration of feature selection methods, leveraging real-world data to identify and prioritize the most informative epigenetic markers. By rigorously comparing these methods, we aim to pinpoint the most robust approach for identifying epigenetic risk factors.

Furthermore, our research delves into the application of diverse machine-learning algorithms to refine predictive models for asthma. Through comprehensive analysis, we endeavor to optimize the selection of algorithms that yield the best performance and reliable predictions across different age stages. This multifaceted approach holds promise for improving clinical decision-making and patient outcomes in asthma man-
agement.

The remaining for this chapter is structured as follows. The method section includes the method we used for building the model. The results section presents the feature selection comparison and real data analysis for different age stages. In the discussion section, we explain the performance of the method. The conclusion section summarizes the key findings and the implications.

2.2 Methods

We outline the comprehensive workflow for the method, encompassing various stages from data collection to model evaluation and validation. This workflow is illustrated in Figure 2.1. It contains the following steps: data collection, where information is gathered from various sources; data preprocessing, involving the selection of clinical variables devoid of missing values and DNA methylation transformation; feature selection employing different approaches to identify the most informative CpGs; data training utilizing diverse machine learning algorithms and employing leave-one-out cross-validation (LOOCV); evaluation of performance metrics; and validation of models using an independent dataset. Further discussion on these steps is provided in the subsequent sections of 2.2.
2.2.1 Feature selection

In machine learning and statistics, feature selection, also known as variable selection, plays a crucial role in choosing relevant factors for ML-based systems: it improves model performance by focusing on relevant features, reduces overfitting by simplifying the model, mitigates the curse of dimensionality, enhances interpretability by removing irrelevant features, boosts computational efficiency, and improves generalization to unseen data. By selecting the most informative features, feature selection streamlines model training improves accuracy, and facilitates understanding of the underlying relationships between predictors and the target variable. In this article, our objective is to identify informative CpGs that can improve prediction performance compared to models solely based on environmental and clinical variables.

Various feature selection techniques are commonly employed in constructing models for high-dimensional data, encompassing biomarker identification models, Lasso, meta-analysis, discriminative power analysis, and interpretable pipelines. In this paper, we undertook a benchmark study to compare different methods using real
datasets and identify the most suitable one.

A meta-analysis, which summarizes prior research, identifies the most significant correlations between DNAm and phenotypes, serving as a promising predictor pool [50]. In this study, we directly use the CpGs reported in the meta-analysis and also recorded in our real data to evaluate performance.

Lasso (Least Absolute Shrinkage and Selection Operator) is a widely used feature selection and regularization technique in machine learning [51]. It works by imposing a penalty on the absolute size of the coefficients, encouraging smaller coefficients and effectively performing feature selection by shrinking some coefficients to zero. Lasso minimizes the sum of squared errors between the observed target values and the predicted values, subject to the constraint that the sum of the absolute values of the coefficients is less than or equal to a specified constant (the regularization parameter, usually denoted as $\lambda$). This constraint is represented by the formula:

$$
\text{minimize} \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)
$$

Where $n$ is the number of observations, $p$ is the number of predictors (features), $y_i$ represents the observed value of the dependent variable for observation $i$, $x_{ij}$ represents the value of predictor $j$ for observation $i$, $\beta_0$ is the intercept term, representing the value of the dependent variable when all predictors are zero, $\beta_j$ is the coefficient of predictor $j$, $\lambda$ is the regularization parameter, controlling the strength of the penalty term. Additionally, Lasso typically selects features that are conditionally independent and has the ability to manage collinear variables.
Another classic feature selection method for DNAm data is to assess the degree of separation using Fisher’s criterion (Bishop, 1995) [52, 53]. To be specific, in the DNAm microarray dataset, it provides the discriminative power of the $k$th CpG as:

$$J(k) = \frac{(\mu_{k}^{\text{case}} - \mu_{k}^{\text{control}})^2}{\sigma_{k}^{\text{case}} + \sigma_{k}^{\text{control}}^2}$$

where $\mu_{k}^{\text{case/control}}$ is the mean and $\sigma_{k}^{\text{case/control}}$ is the standard deviation of all $x_{k}^i$ with $y_i = \text{case/control}$.

In this study, we also compared an interpretable pipeline for feature selection. Initially, we assessed the correlation between the outcome and phenotype, ranking all CpGs based on their correlation values and selecting the top 100 candidates. Subsequently, we evaluated the dispersion of each CpG and retained the top 50 CpGs. We then applied Recursive Feature Elimination (RFE) to identify the top 5 candidate CpGs. Given that RFE results may vary slightly across iterations, we performed 500 runs and selected CpGs based on their frequency of occurrence. The workflow of this method is detailed in Appendix B.1.

2.2.2 Prediction model

Logistic Regression

Unlike linear regression, which predicts continuous values, logistic regression models the probability of the outcome variable belonging to a particular class [54]. It achieves this by applying the link function to the linear combination of the input features and their associated coefficients. The logit link function transforms the output into a
range between $-\infty$ and $+\infty$. Logistic regression can be expressed as:

$$\log \left( \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} \right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$$

and further

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

where $P(Y = 1|X)$ is the probability of the positive class given the input features $X$, $\beta_0$ is the intercept term, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with the features $X_1, X_2, \ldots, X_n$, and $e$ is the base of the natural logarithm. In this project, we have two groups of features: clinical variables denoted by $X$ and epigenetic variables denoted by $E$. We aim to integrate the epigenetic risk factors and evaluate performance. Our final model is given by:

$$\log \left( \frac{P(Y = 1|X, E)}{1 - P(Y = 1|X, E)} \right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \beta_{n+1} E_1 + \ldots + \beta_m E_j$$

where $P(Y = 1|X, E)$ represents the probability of the positive class given the clinical variables $X$ and $E$, $\beta_0$ is the intercept term, $\beta_1, \ldots, \beta_n$ are the coefficients associated with the clinical variables $X_1, \ldots, X_n$, and $\beta_{n+1}, \ldots, \beta_m$ are the coefficients associated with the epigenetic variables $E_1, \ldots, E_j$.

Support Vector Machines

Support Vector Machines (SVM) are supervised learning models used for classifica-
tion and regression tasks [55, 56]. To be specific, given a dataset with \( m \) samples \((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\), where \( x_i \) represents the feature vector of sample \( i \) and \( y_i \) represents the corresponding class label (either \(-1\) or \(+1\) for binary classification), the objective of SVM is to find the hyperplane that best separates the data into two classes while maximizing the margin between the classes.

The decision boundary of the SVM is defined by the equation:

\[
w^T x + b = 0
\]

Where \( w \) is the weight vector (normal to the hyperplane) that determines the orientation of the decision boundary, \( b \) is the bias term (or intercept) that shifts the decision boundary away from the origin.

The decision function of the SVM for classifying a new sample \( x \) is given by:

\[
f(x) = \text{sign}(w^T x + b)
\]

Where: \( \text{sign}(\cdot) \) is the sign function, which outputs \(+1\) if its argument is positive, \(-1\) if its argument is negative, and \(0\) if its argument is zero.

Random Forest

Random Forest is an ensemble learning method used for classification and regression tasks[57]. It operates by constructing a multitude of decision trees during training and outputting the mode of the individual trees. Each tree in the forest is trained on
a random subset of the training data, and at each split, a random subset of features is considered.

Naive Bayes

Naive Bayes is a powerful probabilistic classifier widely used for classification tasks in machine learning [58]. It is based on Bayes’ theorem with the "naive" assumption of conditional independence among the features given the class label. Naive Bayes often performs well in practice, especially for text classification and other high-dimensional datasets.

The classification rule of Naive Bayes can be expressed as:

\[
\hat{y} = \arg \max_{y \in \mathcal{Y}} P(y|x) = \arg \max_{y \in \mathcal{Y}} P(y) \prod_{i=1}^{n} P(x_i|y)
\]

where \(\hat{y}\) is the predicted class label for input \(x\), \(\mathcal{Y}\) is the set of possible class labels, \(P(y)\) is the prior probability of class \(y\), \(P(x_i|y)\) is the conditional probability of feature \(x_i\) given class \(y\), and \(n\) is the number of features.

K-Nearest Neighbors

K-Nearest Neighbors (K-NN) is a non-parametric classification and regression algorithm [59]. It operates on the principle of similarity, where the prediction for a new data point is based on the majority vote or average of its \(k\) nearest neighbors in the feature space. K-NN is particularly suitable for classification tasks with non-linear decision boundaries and can be easily adapted for regression tasks as well.
The prediction of k-NN can be expressed as:

\[ \hat{y} = \text{majority vote}\{y_1, y_2, ..., y_k\} \]

where \( \hat{y} \) is the predicted output, and \( y_1, y_2, ..., y_k \) are the class labels of the k nearest neighbors of the input data point.

### 2.2.3 Evaluation matric

The project aims to enhance the performance of the prediction model via integrating epigenetic risk factors and clinical factors. We compared different algorithms’ performance via the metrics of Accuracy, Sensitivity (Se), Specificity (Sp), Positive Predictive Value (PPV), Negative Predictive Value (NPV), F1 score, Area Under The Curve (AUC) ROC (Receiver Operating Characteristics), Likelihood Ratio for a Positive Test Result (LR\(^+\)), and the Likelihood Ratio for a Negative Test Result (LR\(^-\)). The definitions of these metrics are:

- **Accuracy:**
  \[
  \text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Population}}
  \]

- **Sensitivity (Se):**
  \[
  \text{Se} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}
  \]
• Specificity (Sp):

$$Sp = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

• Positive Predictive Value (PPV):

$$PPV = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

• Negative Predictive Value (NPV):

$$NPV = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}}$$

• F1 Score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

• Likelihood Ratio for a Positive Test Result (LR+):

$$LR^+ = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$
• Likelihood Ratio for a Negative Test Result \((LR^-)\):

\[
LR^- = \frac{1 - \text{Sensitivity}}{\text{Specificity}}
\]

2.2.4 Clinical features in the real data

Diverse prediction models utilizing logistic regression have been documented for the IOWBC dataset. In Table 2.1, we present an overview of the clinical and environmental variables employed across various age stages in these models. Our next step involves integrating CpGs (CpG sites) into these existing models to assess the significance of incorporating epigenetic risk factors in disease prediction. This investigation aims to show the impact of including epigenetic risk factors in improving the performance and robustness of predictive models for allergic diseases.

2.3 Results

2.3.1 Feature selection comparison

We compared several different feature selection algorithms in the IOWBC dataset and used logistic regression as the standard algorithm. In Figure 2.2, we plot the Area Under the Curve (AUC) of ROC (Receiver Operating Characteristics) to compare the different methods within the same dataset. The x-axis represents the false positive rate (FPR), and the y-axis represents the sensitivity (Se). The ROC AUC summarizes a binary classification model’s ability to differentiate between classes with higher values indicating better performance in distinguishing positive and negative instances. Among those feature selection methods, Lasso exhibited the best performance. Con-
Table 2.1: Clinical variables used for asthma prediction.

<table>
<thead>
<tr>
<th>Model</th>
<th>Risk factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 10 Logistic regression</td>
<td>Nasal symptoms (at 1 y)</td>
</tr>
<tr>
<td></td>
<td>Recurrent chest infections (at 2 y)</td>
</tr>
<tr>
<td></td>
<td>Skin prick test response positivity (at 4 y)</td>
</tr>
<tr>
<td></td>
<td>Family history of asthma</td>
</tr>
<tr>
<td>Age 18 Logistic regression</td>
<td>Recurrent wheeze at age 4</td>
</tr>
<tr>
<td></td>
<td>Skin prick test at age 4</td>
</tr>
<tr>
<td>Age 26 Logistic regression</td>
<td>Recurrent wheeze at age 4</td>
</tr>
<tr>
<td></td>
<td>Skin prick test at age 4</td>
</tr>
<tr>
<td></td>
<td>Maternal rhinitis</td>
</tr>
<tr>
<td>PA age 26 Logistic regression</td>
<td>Recurrent wheeze at age 4</td>
</tr>
<tr>
<td></td>
<td>Skin prick test at age 4</td>
</tr>
<tr>
<td></td>
<td>Recurrent chest infection at age 2</td>
</tr>
</tbody>
</table>

Considering Lasso’s potential to control for collinear predictors, we ultimately selected Lasso as the feature selection method. Further discussions about other methods are provided in the discussion section. Also, we use the whole data as the training and testing dataset and select 3 CpG here.
2.3.2 Training and testing in F1 generation

We conducted a thorough analysis of the IOWBC dataset, concentrating on individuals from the F1 generation across three distinct age stages: 10, 18, and 26 years old, with asthma serving as the primary outcome measure. Also, persistent asthma was recorded for age 26 (having asthma at ages 10, 18, and 26).

Incorporating numerous variables (CpGs) may not necessarily enhance the model’s performance, particularly if the CpGs lack informativeness. Moreover, while incorpo-
rating numerous informative CpGs into the model may enhance its performance, it also runs the risk of overfitting issues. To prevent overfitting resulting from the inclusion of excessive CpGs, we optimized the number of CpGs to be added to the model, as illustrated in Figure 2.3. Our analysis reveals that integrating CpGs into the NB model can indeed enhance its performance. However, beyond a certain threshold, incorporating additional CpGs does not yield further improvements. Consequently, we decided to include only three CpGs in the model. This can both enhance the performance of the model and also avoid overfitting issue.

Figure 2.3: Comparison of epigenetic feature numbers for individuals under the age of 10.
In our pursuit of constructing a robust and versatile model and algorithm for asthma prediction, we systematically evaluated various algorithms across different age stages (10, 18, 26). As depicted in Figure 2.4, the Naive Bayes algorithm consistently outperformed the other algorithms at both the age 10 and age 18 stages. This trend persisted at age 26, as illustrated in Figure 2.5.

**Figure 2.4:** Comparison of different algorithms for prediction asthma at age 10 and age 18.
To comprehensively assess the model’s performance, we conducted a comparison based on various metrics including Accuracy, Sensitivity ($se$), Specificity ($sp$), Positive Predictive Value (PPV), Negative Predictive Value (NPV), F1-Score, ROC-AUC, Likelihood Ratio Positive ($LR^+$), and Likelihood Ratio Negative ($LR^-$). For instance, as illustrated in Table 2.2, which focuses on the performance at age 26, Naive Bayes demonstrates superior accuracy, F1-Score, AUC and $LR^+$. Other comparisons for various age stages are available in Appendix B.2, and the findings closely align with the results presented here.

We also assess the extent of model improvement achieved by integrating CpGs into the predictive model. Specifically, we calculate the percentage change for each metric, defined as (Full model - Clinical model) / Clinical model, where the full model incorporates both clinical variables and CpGs, while the clinical model considers only clinical variables. The results are summarized in Table 2.3. We observe from the table
that incorporating epigenetic factors into the model significantly enhances almost all evaluation metrics, particularly for age 10, age 26, and the age 26 persistent asthma model. Although the improvement in F1 and accuracy for age 18 is not as substantial, there is a considerable improvement in ROC. Therefore, we still recommend including CpGs in model building across different age stages as it proves to be beneficial. Given the robustness of the Naive Bayes model across different age stages, this method does help us predictor asthma in the future for the new participants.

Table 2.3: Percentage change for integrating epigenetic predictors into the model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Se</th>
<th>Sp</th>
<th>PPV</th>
<th>NPV</th>
<th>F1</th>
<th>AUC</th>
<th>LR +</th>
<th>LR -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 10</td>
<td>8.48</td>
<td>-27.09</td>
<td>24.19</td>
<td>17.94</td>
<td>-3.37</td>
<td>-6.72</td>
<td>8.95</td>
<td>42.34</td>
<td>18.36</td>
</tr>
<tr>
<td>Age 18</td>
<td>-1.24</td>
<td>2.79</td>
<td>-2.54</td>
<td>-4.40</td>
<td>0.49</td>
<td>-0.83</td>
<td>58.56</td>
<td>-7.01</td>
<td>-0.26</td>
</tr>
<tr>
<td>Age 26</td>
<td>2.57</td>
<td>8.83</td>
<td>1.55</td>
<td>7.51</td>
<td>1.55</td>
<td>7.97</td>
<td>20.79</td>
<td>15.08</td>
<td>-17.62</td>
</tr>
<tr>
<td>PA-Age 26</td>
<td>8.97</td>
<td>10.53</td>
<td>8.74</td>
<td>32.93</td>
<td>2.08</td>
<td>23.68</td>
<td>20.70</td>
<td>67.70</td>
<td>-34.31</td>
</tr>
</tbody>
</table>

2.3.3 Validation in the F2 generation

To validate our method proposed from the F1 generation dataset, we extended our analysis to the F2 generation dataset. In this validation phase, we focused on child-
Table 2.4: Performance metrics of different algorithms for the F2 dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Se</th>
<th>Sp</th>
<th>PPV</th>
<th>NPV</th>
<th>F1</th>
<th>AUC</th>
<th>LR +</th>
<th>LR -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.788</td>
<td>0.732</td>
<td>0.846</td>
<td>0.833</td>
<td>0.750</td>
<td>0.779</td>
<td>0.818</td>
<td>4.756</td>
<td>0.317</td>
</tr>
<tr>
<td>SVM</td>
<td>0.825</td>
<td>0.744</td>
<td>0.910</td>
<td>0.897</td>
<td>0.772</td>
<td>0.813</td>
<td>0.771</td>
<td>8.289</td>
<td>0.281</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.769</td>
<td>0.585</td>
<td>0.962</td>
<td>0.941</td>
<td>0.688</td>
<td>0.722</td>
<td>0.787</td>
<td>15.220</td>
<td>0.431</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.831</td>
<td>0.732</td>
<td>0.936</td>
<td>0.923</td>
<td>0.768</td>
<td>0.816</td>
<td>0.802</td>
<td>11.415</td>
<td>0.287</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.806</td>
<td>0.671</td>
<td>0.949</td>
<td>0.932</td>
<td>0.733</td>
<td>0.780</td>
<td>0.808</td>
<td>13.079</td>
<td>0.347</td>
</tr>
</tbody>
</table>

hood eczema as the primary outcome of interest. Employing the same approach, we executed the model-building process on this new dataset.

Table 2.5: Percentage change for integrating epigenetic predictors in validation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Se</th>
<th>Sp</th>
<th>PPV</th>
<th>NPV</th>
<th>F1</th>
<th>AUC</th>
<th>LR +</th>
<th>LR -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eczema</td>
<td>0.63</td>
<td>0.73</td>
<td>0.56</td>
<td>0.70</td>
<td>0.55</td>
<td>0.72</td>
<td>17.05</td>
<td>8.91</td>
<td>-2.48</td>
</tr>
</tbody>
</table>

As depicted in Table 2.4, Naive Bayes emerges as the top-performing algorithm, boasting the highest accuracy and F1 score, as well as the lowest LR+. Additionally, its ROC and LR+ values remain comparable. Our results further underscore the efficacy of integrating CpGs (epigenetic risk factors) into predictive models, resulting in significant enhancements in predictive performance. This emphasizes the pivotal role of epigenetic factors in comprehending and predicting allergic disease health outcomes. Moreover, these findings validate the importance of epigenetic markers in predictive modeling and emphasize the enduring effectiveness of Naive Bayes machine learning techniques in elucidating complex health dynamics across different generational cohorts.
2.4 Discussion

The results of this project provide valuable insights into the predictive modeling of allergic disease outcomes using epigenetic risk factors (CpGs) integrated with clinical variables.

Our analysis compared several machine learning algorithms across different age stages, including logistic regression, SVM, random forest, Naive Bayes, and K-NN. Among these algorithms, Naive Bayes consistently demonstrated superior performance in terms of AUC, F1 score, and other evaluation metrics. This indicates that naive Bayes might be especially suitable for modeling allergic disease outcomes, possibly owing to its capability to manage independent features and smaller sample sizes effectively.

Integrating epigenetic predictors (CpGs) into the predictive model led to significant improvements in predictive performance across various age stages. This enhancement underscores the importance of considering epigenetic factors in understanding and forecasting allergic disease health outcomes. Epigenetic markers provide valuable insights into the molecular mechanisms underlying allergic diseases, and incorporating them into predictive models can enhance the model’s ability to capture the complex interplay between genetic and environmental factors.

Our findings highlight the robustness of the predictive model across different age stages, with consistent improvements observed in predictive accuracy when incorporating epigenetic predictors. This robustness suggests that the predictive model may generalize well to diverse populations and age groups, making it a valuable tool for predicting allergic disease outcomes across generational cohorts.
Additionally, we conducted a simulation study to evaluate the robustness of the Lasso feature selection. Specifically, we randomly selected 3 independent CpGs out of 346K CpGs and defined the true model as:

\[
\log \left( \frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)} \right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}
\]

where \( i \) denotes the \( i \)th sample for \( i = 1, \ldots, n \). The coefficients are set as a vector: \( \{-7.0, -9.0, -8.5, 4.5\} \). Further,

\[
P(Y_i = 1|X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i})}}
\]

Then we simulate the \( Y_i \sim \text{binomial}(1, P(Y_i = 1|X_i)) \) for each participant. The Lasso selection method was used to select features among the 346K CpGs, and this process was replicated 500 times. The consistency rate of CpG selection was found to be 98.6%.

We also evaluated the nonlinear relationship via simulation using the model:

\[
\log \left( \frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)} \right) = \beta_0 + \beta_1 X_{1i}^2 + \beta_2 X_{2i} + \beta_3 X_{3i}
\]

The coefficients are set as a vector: \( \{-6.5, -2.0, -9.5, 6.5\} \). The simulation results showed a consistency rate of 99.0%.
2.5 Conclusion

Our study underscores the potential of integrating epigenetic risk factors with clinical variables to enhance the predictive modeling of allergic disease outcomes, particularly asthma. The proposed method for feature selection, along with the algorithm employed for disease prediction, demonstrates robustness across various age stages and datasets. This contributes significantly to the accurate prediction of allergic disease outcomes.

Moreover, the workflow we have developed is user-friendly and can be easily implemented in future studies to predict the risk scores of new participants for allergic diseases. This capability can facilitate early identification and intervention, thereby improving patient outcomes and advancing the field of allergic disease research.
Chapter 3
Understanding and Forecasting Childhood Depression and Anxiety

3.1 Introduction

The prediction of depression and anxiety, particularly in children, is a critical area of research aimed at understanding the risk factors associated with these mental health conditions. Some studies have utilized machine learning techniques to forecast the likelihood of depression and anxiety symptoms in schoolchildren. These conditions have a profound impact on children’s mental health growth, cognitive development, and overall well-being. Factors such as school violence, bullying, home violence, and family income have been identified as significant influencers of depression and anxiety scales in children [60].

Predicting depression and anxiety poses several challenges despite advancements in machine learning techniques. One major challenge is the identification of affected individuals and ensuring timely and appropriate treatment. Additionally, limitations in research studies, such as the low prevalence of anxiety and depression in study populations, small sample sizes, too many predictors, hinder accurate prediction models. These challenges underscore the complexity of predicting depression and anxiety accurately, emphasizing the need for further research to address these limitations and enhance predictive capabilities [61].
3.2 Aim

The application of machine learning (ML) in predicting mental health issues among children is relatively uncharted territory. Therefore, this study aimed to assess the efficacy of various ML techniques in predicting depression and anxiety by using a small number of interpretable predictors. Enhancing the robustness of the study, a comprehensive dataset with a larger sample size, broader outcomes, and recorded variables for each participant was utilized.

3.3 Methods

3.3.1 Participants

The National Survey of Children’s Health (NSCH) is a comprehensive survey that gathers data on various aspects of children’s lives, including physical and mental health, access to healthcare, family dynamics, neighborhood environments, and social contexts. Funded and directed by the Health Resources and Services Administration (HRSA) Maternal and Child Health Bureau (MCHB), the NSCH has undergone revisions over the years. The survey was conducted as a mail and web-based survey by the Census Bureau from 2016 to 2022, integrating content from both the NSCH and the National Survey of Children with Special Health Care Needs. In this study, we utilize data collected from the National Survey of Children’s Health spanning the years 2016 to 2019, encompassing a total sample size of 131,774 individuals. This comprehensive dataset provides a rich source of information, with a total of 775 variables recorded across various domains including but not limited to physical and emotional health, access to medical care, family dynamics, parental health, school experiences,
and neighborhood safety. By utilizing this extensive dataset, we aim to select the most informative features for predicting childhood depression and anxiety.

3.3.2 Features

For the prediction of anxiety, we initially identified 264 variables from the dataset. After removing variables with high rates of missing data (> 15%), we were left with 229 variables. To address issues of collinearity and overfitting during model building, we employed Lasso regularization for feature selection. Ultimately, this process resulted in a refined set of 62 variables, encompassing binary, ordinal, and continuous variables, which collectively contribute to the predictive accuracy of the model. Specifically, features included such as mental health care receipt, tourette syndrome prevalence, insurance coverage for mental health needs, child’s sex, autism occurrence, down syndrome presence, health affecting ability frequency, past-year stomach difficulties, adverse childhood experiences, severe headaches, adult mental health status, family frustration with service access, intellectual disability prevalence, parental perception of caregiving difficulty, bleeding gums occurrence, diabetes presence, child’s race, allergies, special health care needs, curiosity display, missed medical care, child’s age, parental nativity, shared decision-making partnership, cerebral palsy prevalence, emotional support from health providers, reduced hours due to health, breathing difficulties, functional difficulties, respiratory problems, household demographics, toothache occurrences, medical home presence, child’s developmental delay severity, absence of special health care needs, insurance benefits coverage, parental concern about child’s weight, young children count, care coordination adequacy, allowance to see providers, birth weight status, sealant application, current heart condition severity, effective care coordination, physical pain frequency, behavioral or conduct problems severity,
insurance coverage adequacy, parental aggravation from parenting, children’s overall health status, special services receipt, special education plan presence, satisfaction with communication among healthcare providers, speech problems prevalence, specialist visits, birth order, emotional support from religious leaders, communication satisfaction among healthcare providers, brain injury occurrence, chronic pain difficulty, and parental annoyance with child behavior. We ultimately selected the first thirty variables with the largest effect size to ensure the interpretability and computational efficiency of the model during building. Information about these variables is listed in Table 3.1.

Similarly, for the depression prediction, we have 49 features. To be specific, health insurance coverage for mental and behavioral needs, unmet mental health care needs, intellectual disability, adverse childhood experiences, reduced work hours due to health issues, sex, complexity of health care needs, age, frequency of health affecting ability, adult mental or emotional health status, parental perception of caregiving difficulty, curiosity display, specific types of special health care needs, frustration in accessing services, severity of child’s developmental delay, occurrence of frequent severe headaches, divorce or separation, shared decision-making partnership in families for health, parental concern about children’s weight, overall general health, frustration in service access, stomach difficulties, emergency room visits, adequacy of care co-ordination assistance, neighborhood vandalism, higher medical care needs or usage in children compared to others, speech problems, number of adverse childhood experiences, consistency in bedtime on weeknights, insurance benefits covering services, frequency of feeling bothered, occurrence of bleeding gums, children’s age categories, allowance to see healthcare providers, specialist visits, out-of-pocket healthcare pay-
Table 3.1: Clinical Variables Used for Anxiety Prediction.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ForgoneMH</td>
<td>Did not receive needed mental health care</td>
</tr>
<tr>
<td>Tourette</td>
<td>Children who have Tourette Syndrome</td>
</tr>
<tr>
<td>MENBEVCOV</td>
<td>Health Insurance - Cover Mental Behavioral Needs</td>
</tr>
<tr>
<td>SC_SEX</td>
<td>Sex</td>
</tr>
<tr>
<td>K2Q35A</td>
<td>Autism ASD</td>
</tr>
<tr>
<td>DownSynd</td>
<td>Children who have Down Syndrome</td>
</tr>
<tr>
<td>HCABILITY</td>
<td>Health Affected Ability - How Often</td>
</tr>
<tr>
<td>STOMACH</td>
<td>Difficulty Stomach</td>
</tr>
<tr>
<td>ACEmhealth</td>
<td>Adverse childhood experience</td>
</tr>
<tr>
<td>Headache</td>
<td>Children who have frequent severe headache</td>
</tr>
<tr>
<td>A1__MENTHEALTH</td>
<td>Parents mental or emotional health</td>
</tr>
<tr>
<td>Frustrated</td>
<td>Family frustrated in efforts to get services</td>
</tr>
<tr>
<td>IntDisab</td>
<td>Children who have an intellectual disability</td>
</tr>
<tr>
<td>DiffCare</td>
<td>Parent felt the child is much harder to care</td>
</tr>
<tr>
<td>Grumbled</td>
<td>Children who had bleeding gums</td>
</tr>
<tr>
<td>CSHCNComplex</td>
<td>CSHCN status and complexity of health care needs</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Children who have diabetes</td>
</tr>
<tr>
<td>SC__RACER</td>
<td>Race</td>
</tr>
<tr>
<td>ALLERGIES</td>
<td>Allergies</td>
</tr>
<tr>
<td>CSHCNtype</td>
<td>Children with specific types of special health care needs</td>
</tr>
<tr>
<td>K6Q71_R</td>
<td>Show interest and curiosity</td>
</tr>
<tr>
<td>ForgoneMed</td>
<td>Did not receive needed medical care</td>
</tr>
<tr>
<td>SC_AGE_YEARS</td>
<td>Age</td>
</tr>
<tr>
<td>HOUSE_GEN</td>
<td>parental nativity</td>
</tr>
<tr>
<td>ShareDec</td>
<td>Shared decision-making for their optimal health</td>
</tr>
<tr>
<td>Palsy</td>
<td>Children who have cerebral palsy</td>
</tr>
<tr>
<td>EmSProvider</td>
<td>Children whose parents received day-to-day emotional support</td>
</tr>
<tr>
<td>CUTHOURS</td>
<td>Cut hours because of health conditions</td>
</tr>
</tbody>
</table>
ments, household food situation, frequency of address changes, behavioral or conduct problems, problems with obtaining needed referrals, emotional support from healthcare providers, experience of living with mentally ill individuals, receipt of care from specialist doctors, satisfaction with communication among healthcare providers and educational institutions, family’s hopeful outlook in difficult times, hours spent providing home health care, and digestive issues occurrence. We also ultimately selected the first thirty variables with the largest effect size to ensure the interpretability and computational efficiency of the model during building. Information about these variables is listed in Table 3.2.

3.3.3 Data analysis

The data variables underwent cleaning, with all missing values removed prior to analysis. The dataset comprised 85,246 participants for anxiety and 65,430 participants for depression. Machine learning algorithms were then employed to predict children’s mental health depression and anxiety symptoms using R.

Various performance metrics were utilized to assess the predictive capabilities of the machine learning models for children’s depression and anxiety symptoms. These metrics included accuracy, specificity, precision, recall, and F1 score. The calculation equations for these performance measures are as follows:

Specificity = True Negative / (False Positive + True Negative)
Precision = True Positive / (True Positive + False Positive)
Recall = True Positive / True Positive + False Negative
F1 Score = (2 × Precision × Recall) / (Precision + Recall)
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MENBEVCOV</td>
<td>health insurance cover mental behavioral needs</td>
</tr>
<tr>
<td>ForgoneMH</td>
<td>Did not receive needed mental health care</td>
</tr>
<tr>
<td>IntDisab</td>
<td>Children who have intellectual disability</td>
</tr>
<tr>
<td>ACEmhealth</td>
<td>Adverse childhood experience</td>
</tr>
<tr>
<td>CUTHOURS</td>
<td>Cut hours because of health conditions</td>
</tr>
<tr>
<td>SC-SEX</td>
<td>Sex</td>
</tr>
<tr>
<td>CSHCNComplex</td>
<td>Complexity of health care needs.</td>
</tr>
<tr>
<td>SC-AGE-YEARS</td>
<td>Age</td>
</tr>
<tr>
<td>HCABILITY</td>
<td>Health affected ability - how often</td>
</tr>
<tr>
<td>A1M-ENTHEALTH</td>
<td>Parent mental or emotional health</td>
</tr>
<tr>
<td>DiffCare</td>
<td>Parent felt the child was much harder to care</td>
</tr>
<tr>
<td>K6Q71-R</td>
<td>Show interest and curiosity</td>
</tr>
<tr>
<td>CSHCNtype</td>
<td>Children with specific types of special health care needs</td>
</tr>
<tr>
<td>Frustrated</td>
<td>Family frustrated in efforts to get services</td>
</tr>
<tr>
<td>DevDelSev</td>
<td>Developmental delay</td>
</tr>
<tr>
<td>Headache</td>
<td>Severe headache</td>
</tr>
<tr>
<td>ACEdivorce</td>
<td>Parent or guardian divorced or separated</td>
</tr>
<tr>
<td>ShareDec</td>
<td>Families are partners in shared decision-making</td>
</tr>
<tr>
<td>WgtConcn</td>
<td>Parental concern about their children’s weight</td>
</tr>
<tr>
<td>K2Q01</td>
<td>General Health</td>
</tr>
<tr>
<td>C4Q04</td>
<td>Frustrated in efforts to get service</td>
</tr>
<tr>
<td>STOMACH</td>
<td>Difficulty stomach past 12 Months</td>
</tr>
<tr>
<td>HOSPITALER</td>
<td>Hospital emergency room visits</td>
</tr>
<tr>
<td>CareHelp</td>
<td>Got all needed extra help</td>
</tr>
<tr>
<td>K10Q23</td>
<td>Neighborhood - Vandalism</td>
</tr>
<tr>
<td>SC- K2Q13</td>
<td>Needs or Uses More Medical Care than Others</td>
</tr>
<tr>
<td>Speech</td>
<td>Children have speech problems</td>
</tr>
</tbody>
</table>
Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative)
3.4 Results

Comprehensive results for different machine learning algorithms are documented in Table 3.3 and Table 3.4. For anxiety prediction, a dataset comprising 32,972 participants was utilized to obtain the testing results, and 38,029 participants were used for depression prediction, representing the largest testing sample size reported to date. Upon examination of Table 3.3, it is evident that random forest demonstrates comparable results with SVM in terms of accuracy, F1-score, and LR, while outperforming in AUC and sensitivity. Based on these findings, random forest is selected as the optimal algorithm for anxiety prediction. Similar trends are observed in the depression prediction task, as evidenced by the results presented in Table 3.4.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Se</th>
<th>Sp</th>
<th>PPV</th>
<th>NPV</th>
<th>F1</th>
<th>AUC</th>
<th>LR+</th>
<th>LR-</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.86</td>
<td>0.75</td>
<td>0.87</td>
<td>0.44</td>
<td>0.96</td>
<td>0.55</td>
<td>0.82</td>
<td>5.98</td>
<td>0.28</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.83</td>
<td>0.80</td>
<td>0.83</td>
<td>0.39</td>
<td>0.97</td>
<td>0.52</td>
<td>0.89</td>
<td>4.84</td>
<td>0.24</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.34</td>
<td>0.97</td>
<td>0.47</td>
<td>0.86</td>
<td>3.89</td>
<td>0.25</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.83</td>
<td>0.78</td>
<td>0.83</td>
<td>0.38</td>
<td>0.97</td>
<td>0.51</td>
<td>0.87</td>
<td>4.68</td>
<td>0.26</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.38</td>
<td>0.97</td>
<td>0.52</td>
<td>0.87</td>
<td>4.67</td>
<td>0.22</td>
</tr>
</tbody>
</table>

**Table 3.3:** Performance metrics of different algorithms for the anxiety prediction.

The AUC curves depicted in Figures 3.1 and 3.2 further reinforce the superiority of the random forest algorithm over alternative methods. In both anxiety and depression prediction tasks, the AUC curve showcases the random forest’s superior performance compared to competing algorithms.
Table 3.4: Performance metrics of different algorithms for depression prediction.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Se</th>
<th>Sp</th>
<th>PPV</th>
<th>NPV</th>
<th>F1</th>
<th>A UC</th>
<th>LR +</th>
<th>LR -</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.92</td>
<td>0.75</td>
<td>0.93</td>
<td>0.39</td>
<td>0.98</td>
<td>0.52</td>
<td>0.88</td>
<td>10.79</td>
<td>0.27</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.86</td>
<td>0.88</td>
<td>0.85</td>
<td>0.27</td>
<td>0.99</td>
<td>0.41</td>
<td>0.93</td>
<td>5.99</td>
<td>0.14</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.80</td>
<td>0.88</td>
<td>0.79</td>
<td>0.20</td>
<td>0.99</td>
<td>0.33</td>
<td>0.90</td>
<td>4.36</td>
<td>0.14</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.26</td>
<td>0.99</td>
<td>0.40</td>
<td>0.89</td>
<td>5.89</td>
<td>0.17</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.87</td>
<td>0.88</td>
<td>0.87</td>
<td>0.29</td>
<td>0.99</td>
<td>0.43</td>
<td>0.92</td>
<td>6.64</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Figure 3.1: Anxiety prediction algorithm comparison.
We visualized the feature importance using the random forest prediction algorithm, and the top ten features are delineated in Table 3.5 and 3.6. Notably, prominent features include children’s sex, physical health status, insurance status, and parental mental health. In the context of depression prediction, factors encompassing child experiences, healthcare needs, adverse childhood experiences, and parental experiences. Several features overlap between anxiety and depression prediction models, contributing to enhanced model generalization across mental health domains.

Figure 3.2: Depression prediction algorithm comparaison.
Table 3.5: Feature Importance for Anxiety Prediction.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Important Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health insurance - cover mental behavioral needs</td>
<td>1918.73</td>
</tr>
<tr>
<td>Age</td>
<td>1117.61</td>
</tr>
<tr>
<td>Complexity of health care needs.</td>
<td>780.29</td>
</tr>
<tr>
<td>Health Affected Ability - How Often</td>
<td>657.72</td>
</tr>
<tr>
<td>Parent Mental or Emotional Health</td>
<td>506.99</td>
</tr>
<tr>
<td>Parent felt child is much harder to care</td>
<td>495.84</td>
</tr>
<tr>
<td>Children with specific types of special health care needs</td>
<td>469.92</td>
</tr>
<tr>
<td>Children whose parents received day-to-day emotional support</td>
<td>394.58</td>
</tr>
<tr>
<td>Children with one or more functional difficulties</td>
<td>341.86</td>
</tr>
</tbody>
</table>

Table 3.6: Feature Importance for Depression Prediction.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Important Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health Insurance Cover Mental Behavioral Needs</td>
<td>879.03</td>
</tr>
<tr>
<td>Age</td>
<td>692.25</td>
</tr>
<tr>
<td>Complexity of health care needs.</td>
<td>327.16</td>
</tr>
<tr>
<td>Number of Adverse Childhood Experiences</td>
<td>261.65</td>
</tr>
<tr>
<td>Parent Mental or Emotional Health</td>
<td>228.76</td>
</tr>
<tr>
<td>Health Affected Ability How Often</td>
<td>217.04</td>
</tr>
<tr>
<td>Parent felt child is much harder to care</td>
<td>203.56</td>
</tr>
<tr>
<td>Go to bed at the same time on weeknights</td>
<td>192.40</td>
</tr>
<tr>
<td>General Health</td>
<td>174.71</td>
</tr>
<tr>
<td>Show Interest and Curiosity</td>
<td>166.78</td>
</tr>
</tbody>
</table>
3.5 Discussion and Conclusion

Based on the comprehensive analysis conducted on different machine learning algorithms for anxiety and depression prediction, we observed significant insights into the predictive capabilities of these models. Utilizing the largest dataset reported till now for anxiety and depression prediction, we assessed the performance metrics across various algorithms. Random forest exhibited comparable results with Support Vector Machine (SVM) in terms of accuracy, F1-score, and LR$^+$, while surpassing in AUC and sensitivity for anxiety prediction. Similarly, for depression prediction, random forest outperformed other algorithms in AUC, indicating its robustness in distinguishing between positive and negative cases. However, for depression prediction, the logistic regression also showed a good candidate algorithm for the outcome prediction. The superior performance of random forest was further corroborated by visualizing the AUC curves, which consistently demonstrated its effectiveness over alternative methods across both anxiety and depression prediction tasks.

Furthermore, the examination of feature importance revealed crucial insights into the key determinants influencing anxiety and depression outcomes. Prominent features such as children’s sex, physical health status, insurance status, and parental mental health emerged as significant predictors for both anxiety and depression. This underscores the multifaceted nature of mental health outcomes and the importance of considering various socio-demographic and health-related factors in predictive modeling. Moreover, the overlap observed in several features between anxiety and depression prediction models highlights the connection of these mental health domains and underscores the potential for developing generalized predictive models that can
effectively capture the underlying complexities of mental health outcomes.

In conclusion, our findings suggest that random forest emerges as the optimal algorithm for anxiety and depression prediction, offering valuable insights into the determinants of mental health outcomes and paving the way for more accurate and generalized predictive models in mental health research and clinical practice.


43. Trivedi, M. & Denton, E. Asthma in children and adults—what are the differences and what can they tell us about asthma? *Frontiers in pediatrics* 7, 256 (2019).


47. Kothalawala, D. M. *et al.* Integration of genomic risk scores to improve the prediction of childhood asthma diagnosis. *Journal of personalized medicine* 12, 75 (2022).


Appendix A

For Chapter 1

Appendix A offers an extensive examination of various aspects related to Chapter 1. It encompasses additional simulations, a thorough analysis of real-world data, insights into computational complexity, and a comprehensive theoretical power analysis specifically tailored for evaluating the dmrff method.

Appendix 1

To comprehensively compare different methods’ power and false positive rate across different block sizes, we systematically explored various settings, encompassing adjustments in block size, mean values, and standard deviations for each group.

Appendix 1.1

To enhance the realism of our simulation, we present the distribution of block sizes within the real data (IOW, GSE59065, GSE87571) for both gbdmr and dmrff. In gbdmr, blocks are determined based on correlation. A chain of neighboring CpGs forms a block if the correlation of each neighboring pair is larger than the correlation threshold. In the histogram, we set the threshold to be 0.5. In dmrff, a block is formed based on the distance. If the distance between two neighboring CpGs is less than 500 bp, then they belong to the same block. As illustrated in Appendix 1.1, the majority of candidate block sizes are clustered within the range of one to ten, with larger regions being infrequent and characterized by very low frequencies. Specifically,
we found more than 99.9% of block sizes range between 1 and 10, and over 99.8% of block sizes are less than or equal to 6 across the three datasets.
Appendix  1.2

In this appendix, we assess the power and false positive rate of gbdmr, dmrff, and EWAS across additional settings. Specifically, we simulated 253/253 DNAm to represent the trait present/absent groups following gamma distribution. The average DNAm of the two groups are set to be M1 and M2. We examined a spectrum of M1 & M2 and correlation thresholds for $\rho$. All simulations are repeated 500 times, and the average power/false positive rates are presented in the figures.

When the average DNAm is 0.3:
Correlation of adjacent CpG sites (block size = 2)

Power

False positive rate

Correlation of adjacent CpG sites (block size = 3)

Power

False positive rate

Correlation of adjacent CpG sites (block size = 4)

Power

False positive rate

M1 0.3 M2 0.29 sd 0.03 gbdmr cutoff 0.7
When the average DNAm is 0.5:
When the average DNAm is 0.7:
Correlation of adjacent CpG sites (block size = 2)

Power

False positive rate

Correlation of adjacent CpG sites (block size = 3)

Power

False positive rate

Correlation of adjacent CpG sites (block size = 4)

Power

False positive rate
Appendix 1.3

In addition to simulations in 1.2, we expanded our analysis to incorporate larger block sizes ranging from 5 to 10. This broader scope allowed us to comprehensively evaluate the system’s performance by assessing both power and false positive rates. These evaluations were conducted under average DNAm at 0.3, 0.5, and 0.7. All simulations are repeated 500 times, and the average power/false positive rates are presented in the figures.

When the average DNAm is 0.3:
When the average DNAm is 0.5:
M1 0.5 M2 0.49, gbdmr cutoff 0.5
Correlation of adjacent CpG sites (block size = 8)

Power
False positive rate

Correlation of adjacent CpG sites (block size = 9)

Power
False positive rate

Correlation of adjacent CpG sites (block size = 10)

Power
False positive rate

81
When the average DNAm is 0.7:
Appendix 2

In Appendix 2, we examined the relationship between signal strength and the power of gbdmr, dmrff, and EWAS. Same as Section 1, we simulated 253/253 DNAm to represent the trait present/absent groups following gamma distribution. The signal strength is defined as the mean difference between trait present/absent groups divided by the standard deviation. We checked the performance across different combinations of the block size(1-10), mean, standard deviation of DNAm, gbdmr correlation threshold, and the true correlations between adjacent CpG sites when block size > 1. All simulations are repeated 500 times, and the average power/false positive rates are presented in the figures.

When the average DNAm starts from 0.3:
Start mean 0.3, cor 0.8, gbdmr cutoff 0.3, sd 0.3

Signal strength (block size = 1)

Signal strength (block size = 2)

Signal strength (block size = 3)

Signal strength (block size = 4)

Start mean 0.3, cor 0.8, gbdmr cutoff 0.5, sd 0.3

Signal strength (block size = 1)

Signal strength (block size = 2)

Signal strength (block size = 3)

Signal strength (block size = 4)
When the average DNAm starts from 0.5:

![Graphs showing power distribution for different block sizes and signal strengths.](image-url)
Start mean 0.5, gbdmr cutoff 0.5

signal strength (block size = 9)

Power
dmrff
EWAS
gbdmr

signal strength (block size = 10)

Power
When the average DNAm starts from 0.7:
Appendix 3

Figure 4 of the main paper presents the Venn plot of CpG sites (DMP CpGs + DMR CpGs) identified by EWAS, dmrff, and gdmr. Appendix 3 presents the Venn plot of DMP CpGs only.
Appendix 4

In Appendix 4, we recorded the execution time and memory usage of various methods for different sample sizes and CpG numbers. The gbdmr method was implemented using Intel Skylake Gold 6148 Processors (core frequency: 2.40 GHz) with the processing distributed across 8 parallel tasks. Table 1 lists the results based on 346k CpGs with sample sizes ranging from 50 to 500, and Table 2 is for a fixed sample size of 200 but varying numbers of CpG sites ranging from 100k and 350k.

Table A.1: Running Complexity 1

<table>
<thead>
<tr>
<th>Sample size</th>
<th>dmrff</th>
<th>EWAS</th>
<th>gbdmr</th>
<th>dmrff</th>
<th>EWAS</th>
<th>gbdmr</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>1</td>
<td>18</td>
<td>22</td>
<td>687</td>
<td>463</td>
<td>1380</td>
</tr>
<tr>
<td>100</td>
<td>1.2</td>
<td>20</td>
<td>26</td>
<td>970</td>
<td>883</td>
<td>1660</td>
</tr>
<tr>
<td>150</td>
<td>1.2</td>
<td>22</td>
<td>30</td>
<td>1250</td>
<td>1302</td>
<td>1930</td>
</tr>
<tr>
<td>200</td>
<td>1.5</td>
<td>25</td>
<td>34</td>
<td>1527</td>
<td>1723</td>
<td>2210</td>
</tr>
<tr>
<td>250</td>
<td>1.5</td>
<td>27</td>
<td>38</td>
<td>1810</td>
<td>2143</td>
<td>2490</td>
</tr>
<tr>
<td>300</td>
<td>1.6</td>
<td>27</td>
<td>42</td>
<td>2090</td>
<td>2562</td>
<td>2760</td>
</tr>
<tr>
<td>400</td>
<td>1.6</td>
<td>28</td>
<td>51</td>
<td>2647</td>
<td>3402</td>
<td>3320</td>
</tr>
<tr>
<td>500</td>
<td>1.7</td>
<td>29</td>
<td>60</td>
<td>3207</td>
<td>4242</td>
<td>3870</td>
</tr>
</tbody>
</table>

Fix the CpG numbers as 346k; sample size ranges from 50-500. A parallel computation algorithm has been implemented in gbdmr, and the data processing is carried out with eight parallel tasks.
Table A.2: Running Complexity 2

<table>
<thead>
<tr>
<th>CpG numbers (K)</th>
<th>Time (Min)</th>
<th>Memory (Mb)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dmrff</td>
<td>EWAS</td>
<td>gbdmr</td>
</tr>
<tr>
<td>100</td>
<td>0.6</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>150</td>
<td>0.7</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>200</td>
<td>0.9</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>250</td>
<td>0.9</td>
<td>16</td>
<td>24</td>
</tr>
<tr>
<td>300</td>
<td>0.9</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>350</td>
<td>1</td>
<td>19</td>
<td>35</td>
</tr>
</tbody>
</table>

Fix the sample size as 200; CpG number ranges from 100-350K. A parallel computation algorithm has been implemented in gbdmr, and the data processing is carried out with eight parallel tasks.

Appendix 5

In Appendix 5, we revisited the key simulations from Appendices A and B to assess the three methods under an unbalanced design. To this end, we altered the simulation parameters, setting the DNAm at 422/84 instead of the equal 253/253 distribution for a binary exposure such as disease.

Power and false positive rate by correlation strengths of adjacent CpGs:
Power by signal strength:

Unbalanced, start mean 0.7, cor 0.5, gbdmr cutoff 0.3, sd 0.3

Unbalanced, start mean 0.7, cor 0.8, gbdmr cutoff 0.3, sd 0.3
Appendix 6

In Appendix 6, we repeated the key simulations from Appendices A and B using a continuous phenotype instead of a binary phenotype. Specifically, we first simulated the DNAm following gamma distribution such that the mean equals $\mu$ and standard deviation equals $\sigma$. The $\sigma$ is fixed, and the $\mu$’s varied across an evenly spaced grid ranging from 0.69 to 0.70. For each $\mu$ level, we generated the corresponding continuous phenotype using a normal distribution with the same mean ($\mu$) and a standard deviation of 0.002. To generate blocks of size $> 1$, we first simulated the DNAm of a single CpG site following beta distributions. Then, we generated a second vector such that it has a fixed correlation with the first CpG site with the same mean and standard deviation. We followed the same step to generate the third CpG site’s DNAm given the second, and so on. This procedure was used to simulate a chain of CpG sites with a given correlation between adjacent CpG sites. Our analysis includes the assessment of power and false positive rate at different correlation strengths of adjacent CpGs (as in Appendix 1.2) and power at different signal strengths (as in Appendix 2). We examined the block sizes varying from 1 to 10, and all simulations are repeated 500 times.

Power and false positive rate by correlation strengths of adjacent CpGs:
Correlation of adjacent CpG sites (block size = 8)

Power

False positive rate

Correlation of adjacent CpG sites (block size = 9)

Power

False positive rate

Correlation of adjacent CpG sites (block size = 10)

Power

False positive rate

M1 0.7 M2 0.69, gbdmr cutoff 0.5

Correlation of adjacent CpG sites (block size = 8)

False positive rate

Correlation of adjacent CpG sites (block size = 9)

False positive rate

Correlation of adjacent CpG sites (block size = 10)

False positive rate
Power by signal strength:

Start mean 0.69, gbdmr cutoff 0.5
Start mean 0.69, gbdmr cutoff 0.5

Power
dmrff
EWAS
gbdmr

signal strength (block size = 9)

signal strength (block size = 10)
Appendix 7

Technical details in the power of dmrff:

Let

\[ Y_{ij} = \mu_j + X_i \beta + \epsilon_{ij} \]

denote the DNA methylation (DNAm) levels of \( j \)th CpG site of the \( i \)th sample, where \( i = 1, \ldots, n; j = 1, \ldots, L_b \) (assuming the \( b \)th block of CpG sites consist of \( L_b \) CpG sites). We assume the phenotype \( X_i \)'s have the same effect \( \beta \) on \( L_b \) CpG sites in the differentiated DNAm region (DMR), and \( \epsilon_{ij} \) and \( \epsilon_{ij'} (j \neq j') \) are correlated, with 

\[
(\epsilon_{i1}, \ldots, \epsilon_{iL_b})^\top \sim MVN(0, \Sigma), \]

where \( 0 = (0, \ldots, 0)^\top \), and

\[
\Sigma = \sigma^2 \begin{bmatrix}
1 & \rho & \ldots & \rho \\
\rho & 1 & \ldots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \rho & 1
\end{bmatrix}.
\]

The estimate of \( \beta \) using the \( j \)th CpG site's DNA methylation, \( \hat{\beta}_j = e_2 (X^\top X)^{-1} X^\top Y_j \), where \( e_2 = (0, 1)^\top \),

\[
X = \begin{bmatrix}
1 & X_1 \\
\vdots \\
1 & X_n
\end{bmatrix}.
\]
and \( Y_j = (Y_{1j}, \ldots, Y_{nj})^\top \). Following these notations, \( \hat{\beta}_j \) is an ordinary linear regression estimator with \( E(\hat{\beta}_j) = \beta \). The covariance between \( \hat{\beta}_j \) and \( \hat{\beta}_{j'} \),

\[
\text{cov}(\hat{\beta}_j, \hat{\beta}_{j'}|X) = \begin{cases} 
\sigma_n^2, & \text{if } j = j' \\
\rho \sigma_n^2, & \text{if } j \neq j', 
\end{cases}
\]

where \( \sigma_n^2 = \sigma^2 e_2^\top (X^\top X)^{-1} e_2 = \sigma^2 / \sum_{i=1}^n (X_i - \bar{X})^2 \) and \( \bar{X} = \sum_{i=1}^n X_i/n \). Thus, we have the distribution of \( \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_{1b})^\top \) follows a multivariate normal distribution with mean \( \beta = (\beta, \ldots, \beta)^\top \) and covariance matrix

\[
\Omega = \sigma_n^2 \begin{bmatrix}
1 & \rho & \ldots & \rho \\
\rho & 1 & \ldots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \rho & 1
\end{bmatrix}.
\]

Denote by \( A = (1^\top \Omega^{-1} 1)^{-1} 1^\top \Omega^{-1} \hat{\beta} \). Then \( E(A|X) = (1^\top \Omega^{-1} 1)^{-1} 1^\top \Omega^{-1} \beta = \beta \) and \( \text{var}(A|X) = (1^\top \Omega^{-1} 1)^{-1} 1^\top \Omega^{-1} \Omega^{-1} 1(1^\top \Omega^{-1} 1)^{-1} = (1^\top \Omega^{-1} 1)^{-1} \). Note that the \( \sigma \) in
Ω, the true standard deviation of the error term, is not known. Dmff estimates Ω by

\[ \hat{\Omega} = \hat{\sigma} \hat{\sigma}^\top \odot \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}, \]

where \( \hat{\sigma} = (\hat{\sigma}_1, \ldots, \hat{\sigma}_{L_b})^\top \), \( \hat{\sigma}_j \) is the standard error of \( \hat{\beta}_j \), and \( \odot \) is the element-wise multiplication of matrix. For simplicity, we assume \( \sigma_n \) is known, and thus the estimator of dmff is expressed as \( A/\sqrt{\text{var}(A|X)} \). To derive its power, note that

\[ \frac{A}{\sqrt{\text{var}(A|X)}} = \frac{A - \beta}{\sqrt{\text{var}(A|X)}} + \frac{\beta}{\sqrt{\text{var}(A|X)}}. \]

Thus, the power of dmff is

\[ P\left( \frac{A - \beta}{\sqrt{\text{var}(A|X)}} > Z_{1-\alpha/2} - \frac{\beta}{\sqrt{\text{var}(A|X)}} \right) + P\left( \frac{A - \beta}{\sqrt{\text{var}(A|X)}} < Z_{\alpha/2} - \frac{\beta}{\sqrt{\text{var}(A|X)}} \right) \]
Appendix B offers an extensive examination of various aspects related to Chapter 2.

Appendix 1

Below is the interpretable feature selection method proposed in the chapter 2:

Figure B.1: Interpretable Feature Selection

Appendix 2

In this section, we present the comparison results of algorithms for age 10 and 18, as well as for persistent asthma at age 26, based on the incorporation of clinical vari-
ables and three CPGs. Table B.1 illustrates that Naive Bayes demonstrates superior performance across various comprehensive evaluation metrics, including Accuracy, ROC-AUC, and LR⁺, while also achieving comparable results in terms of F1 score and LR⁻.

Table B.1: Performance metrics of different algorithms at age 10.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Se</th>
<th>Sp</th>
<th>PPV</th>
<th>NPV</th>
<th>F1</th>
<th>AUC</th>
<th>LR⁺</th>
<th>LR⁻</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.631</td>
<td>0.642</td>
<td>0.625</td>
<td>0.444</td>
<td>0.789</td>
<td>0.525</td>
<td>0.670</td>
<td>1.713</td>
<td>0.573</td>
</tr>
<tr>
<td>SVM</td>
<td>0.669</td>
<td>0.532</td>
<td>0.733</td>
<td>0.482</td>
<td>0.771</td>
<td>0.505</td>
<td>0.676</td>
<td>1.993</td>
<td>0.639</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.638</td>
<td>0.601</td>
<td>0.655</td>
<td>0.448</td>
<td>0.779</td>
<td>0.514</td>
<td>0.652</td>
<td>1.742</td>
<td>0.609</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.713</td>
<td>0.462</td>
<td>0.830</td>
<td>0.559</td>
<td>0.768</td>
<td>0.506</td>
<td>0.710</td>
<td>2.723</td>
<td>0.648</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.695</td>
<td>0.486</td>
<td>0.792</td>
<td>0.522</td>
<td>0.768</td>
<td>0.503</td>
<td>0.689</td>
<td>2.339</td>
<td>0.649</td>
</tr>
</tbody>
</table>

For the age 18 group, Table B.2 showcases a consistent trend, with Naive Bayes emerging as the most robust algorithm across various evaluation metrics. This finding underscores the reliability and effectiveness of Naive Bayes in predictive modeling for these specific age demographics.

Table B.2: Performance metrics of different algorithms at age 18.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Se</th>
<th>Sp</th>
<th>PPV</th>
<th>NPV</th>
<th>F1</th>
<th>AUC</th>
<th>LR⁺</th>
<th>LR⁻</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.647</td>
<td>0.620</td>
<td>0.659</td>
<td>0.437</td>
<td>0.802</td>
<td>0.513</td>
<td>0.694</td>
<td>1.817</td>
<td>0.577</td>
</tr>
<tr>
<td>SVM</td>
<td>0.677</td>
<td>0.508</td>
<td>0.749</td>
<td>0.464</td>
<td>0.781</td>
<td>0.485</td>
<td>0.664</td>
<td>2.029</td>
<td>0.656</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.649</td>
<td>0.525</td>
<td>0.702</td>
<td>0.429</td>
<td>0.776</td>
<td>0.472</td>
<td>0.657</td>
<td>1.760</td>
<td>0.677</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.704</td>
<td>0.514</td>
<td>0.785</td>
<td>0.505</td>
<td>0.791</td>
<td>0.510</td>
<td>0.705</td>
<td>2.393</td>
<td>0.619</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.694</td>
<td>0.402</td>
<td>0.819</td>
<td>0.486</td>
<td>0.762</td>
<td>0.440</td>
<td>0.643</td>
<td>2.218</td>
<td>0.730</td>
</tr>
</tbody>
</table>

Persistent asthma data for age 26 is also included in our dataset, where a case is defined as a participant having asthma at all three age points: 10, 18, and 26 years.
old. Table B.3 presents the comparison results, revealing that Naive Bayes remains the optimal method for model building in this context. This underscores the robustness and consistency of Naive Bayes as a predictive modeling technique, even when considering persistent asthma across multiple age stages.

Table B.3: Performance metrics of different algorithms at age 26 (Persistent Asthma).

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Se</th>
<th>Sp</th>
<th>PPV</th>
<th>NPV</th>
<th>F1</th>
<th>AUC</th>
<th>LR +</th>
<th>LR -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.841</td>
<td>0.615</td>
<td>0.891</td>
<td>0.552</td>
<td>0.914</td>
<td>0.582</td>
<td>0.803</td>
<td>5.633</td>
<td>0.432</td>
</tr>
<tr>
<td>SVM</td>
<td>0.766</td>
<td>0.808</td>
<td>0.756</td>
<td>0.420</td>
<td>0.947</td>
<td>0.553</td>
<td>0.794</td>
<td>3.314</td>
<td>0.254</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.752</td>
<td>0.731</td>
<td>0.756</td>
<td>0.396</td>
<td>0.928</td>
<td>0.514</td>
<td>0.805</td>
<td>2.999</td>
<td>0.356</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.855</td>
<td>0.808</td>
<td>0.866</td>
<td>0.568</td>
<td>0.954</td>
<td>0.667</td>
<td>0.820</td>
<td>6.007</td>
<td>0.222</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.855</td>
<td>0.500</td>
<td>0.933</td>
<td>0.619</td>
<td>0.895</td>
<td>0.553</td>
<td>0.742</td>
<td>7.438</td>
<td>0.536</td>
</tr>
</tbody>
</table>