COSC681 - AI Project

Classifying smoking exposure with epigentic-trained machine learning

Lachlan Jones

2025

Supervisors:
Prof. Greg Jones (University of Otago)
Prof. Richard Green (University of Canterbury)

Department of Computer Science
University of Canterbury

Abstract

Contents

List of Figures

List of Tables

1 Introduction

1.1 Tobacco Related Health Issues

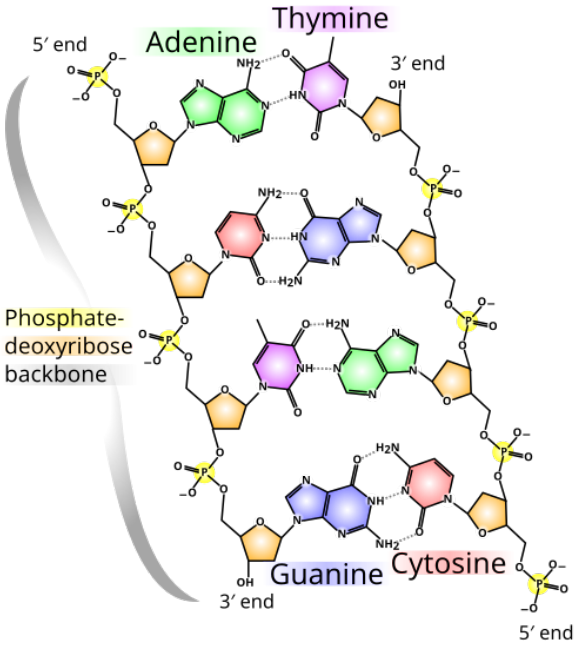
The harms associated with tobacco use are well recognised. Tobacco kills up to half its users who do not quit and more than 8 million people per year, including an estimated 1.3 million non-smokers due to second hand smoke [1]. Smoking causes cancer, heart and lung disease, stroke, type 2 diabetes, and harmful reproductive effects [2]. There is a growing body of evidence suggesting a causal relationship between smoking and mental health issues [3]. Clearly, such negative impacts on patient health due to tobacco use are undesirable, just as they are avoidable. For these reason, tobacco usage is of great concern to health professionals. The World Health Organization asserts that surveillance is key for addressing the tobacco epidemic, as tracking tobacco usage indicates how to shape policy [1].

1.2 Self-Reported Smoking Status

Current surveillance relies on self-reported smoking data. That is, a patient's smoking history is recorded by them personally recalling and reporting. It is a convenient and cost-effective way of collecting smoking statistics. There are two main types of smoking data used to measure tobacco exposure: smoking status and smoking pack-years. Smoking status is label based on the history and habits of tobacco use. Indiviudals are binned into never smokers, ex-smokers and current smokers. Smoking pack-years is a calculated score that tries to quantify tobacco use. It is calculated as the number of packs of cigarettes smoked per day multiplied by years of smoking [4]. For example, one pack-year is one pack per day for one year, or half a pack per day for two years. Therefore, smoking pack-years quantifies both the degree of exposure and duration of exposure equally.

Self-reported smoking data has several limitations. Relying on individuals recounting information can introduce bias. Self-reported smoking data is prone to inaccuracy due to stigma, recall bias and a lack of information on second-hand exposure [5, 6]. That is, the social pressure to deny partaking in stigmatised behaviours, forgetting details and information, and not being aware of sources of second-hand exposure can all influence the results of self-reported smoking data. A method of using objective evidence to determine smoking history could overcome these issues. On the other hand, the inaccuracy of self-reported smoking data can differ between population groups. For example, studies suggest that teens are more likely to provide false responses in smoking surveys [5]. Moreover, tobacco consumption differs between social groups, with smoking more prevalent in low-education and low-socio-economic groups [7].

To this end, developing diagnostic tests to collect smoking data that do not share the biases of self-reported methods are of interest for improving the monitoring of health. One such approach is the use of epigenetic biomarkers.



Chemical structure of DNA [8]

1.3 Epigenetics

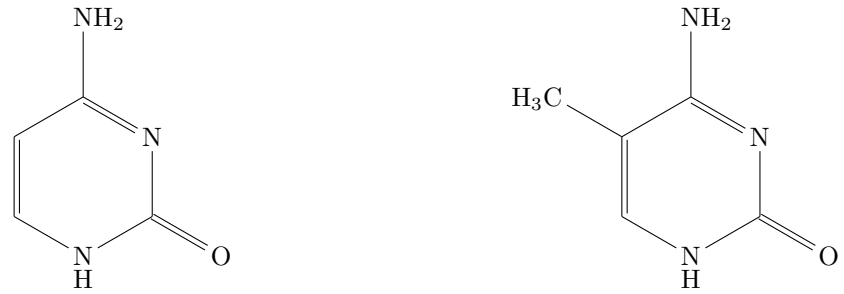
Epi- is a Greek prefix meaning upon or on. Therefore, epigenetics is the study of factors on top of or upon genetics. Specifically, it is the study of how environmental factors and behaviours affect, modify and regulate your genetics and their expression, without changing the DNA itself. We consider one type of epigenetic modification: DNA methylation.

1.3.1 DNA Methylation

DNA is a sequence of one of four nucleotide bases: adenine (A), cytosine (C), guanine (G) and thymine (T), linked together by a deoxyribose sugar and phosphate backbone. It is this sequence that provides genetic instructions. Like a human reading a book, strings of these bases are converted into information that tells cells how to function, called DNA transcription. Continuing the book analogy, a sentence of such instructions is called a gene, found in a chapter called a chromosome.

DNA methylation involves the addition of a methyl group (CH_3) to the 5-carbon position of cytosine nucleotides. This cytosine modification makes it harder for transcription to occur, which can modulate, or even completely silence, gene expression. This is relevant when a guanine is directly followed by a cytosine in the DNA sequence. Because of the phosphate connecting these bases, such a region is called a CpG site. CpG sites are the main form of DNA methylation that occurs in mammalian cells.

A biological sample used to assess DNA methylation will contain many thousands of copies of DNA. When measuring a CpG site for DNA methylation, the reported value is the percentage of DNA copies



(a) Cytosine (b) 5-Methylcytosine

Modification of Cytosine into 5-Methylcytosine caused by DNA methylation

from the sample which are methylated. That is, a value between 0 (all DNA copies un-methylated) and 1 (all DNA copies methylated).

While the genetic sequence of DNA is stable, methylation is not. It is a dynamic state that depends on factors such as behaviours and environmental exposure [9]. Exposure to a wide range of environmental factors including air pollutants [10], diet [11], physical activity and even psychosocial stress [12] have been shown to be associated with specific changes in DNA methylation. Additionally, sufficent lack of exposure to such factors can reverse such changes. As previously mentioned, DNA methylation affects the expression of genes. Methylation at a CpG site can silence the expression of the gene that site is located in, where more methylation at a site leads to stronger silencing. Moreover, DNA methylation is not random. There is strong correlation between methylation of specific sites with specific factors [13]. With these two points in mind, this means that DNA methylation of CpG sites can be used as a biomarker indicative of the factors that caused it, while also describing changes in cellular function. Therefore, DNA methylation is a biomarker not only useful for reporting on environmental exposures, but also predicting future health outcomes or risks. Examples of this include prediction of cardiovascular diseases [14], neurological diseases [15], type 2 diabetes [16], pace of aging [17], and cancer [18]. Furthermore, DNA methylation is not self-reported, and therefore overcomes the biases associated with self-reported data.

Altogether, this motivates the use of DNA methylation data to develop methods for collecting smoking history of indiviudals. Work using such data is referred to as an epigenome-wide association study (EWAS).

1.3.2 DNAm Platforms

The human genome contains $\sim$28 million CpG sites. This is often a computationally infeasible domain for a dataset, due to massive dataset sizes, processing requirements, and noise contained in the signal. Instead, most EWAS use a biologically relevant and informative subset of CpG sites. There are two commonly chosen platforms used to achieve this: Illumina 450k [19] and Illumina EPIC [20]. 450k was the first array developed by Illumina, consisting of 485,577 CpG sites chosen for their quality and

usefulness. The EPIC array was developed as a successor, increasing to 865,859 CpG sites. However, only around $> 90\%$ of the sites were retained from 450k. This is something that needs to be considered when developing scores or screening tests from an EWAS if compatability with multiple datasets or cohorts is of concern (see 2.4).

Additionally, these arrays are compatible with DNA from multiple different cell types. Some examples include whole blood, purified cells and fresh-frozen tissue [21]. Moreover, the methylation signal can differ across different cell types [22], which means selection of cell type can be an import factor when designing an EWAS. Typically, whole blood is used, specifically white blood cells, as red blood cells in humans do not carry a copy of the DNA.

1.4 Machine Learning in Epigenetics

Broadly, machine learning algorithms are split into one of two tasks: regression or classification. The most significant distinction between these two tasks is the choice of supervised learning labels. Regression algorithms are trained against continuous, numeric scores, while classification algorithms are trained against discrete class labels. The choice of label in turn determines the output produced by the algorithm. Machine learning has already seen use in many areas of clinical epigenetics. We begin with a review of some developed methods, addressing both classification and regression tasks.

1.4.1 Applications

1. Malta et al. [23] proposed a method for assessing oncogenic dedifferentiation (cells becoming cancerous). This approach seeks to model a "stemness index" which indicates how similar a cell is to stem cell - a trait found in cancerous cells. Of relevance is the developed epigenetic approach using one-class logistic regression. The training features consisted of 219 hypermethylated CpG sites associated with stem cells. Training data only consisted of a single, positive, class: stem cells. The resulting model can then be fed non-stem cells to compare how similar they are to stem cells, i.e. cancerous cells.

2. Adorján et al. [24] proposed a method for using DNA methylation to classify cancer tissues. CpG sites were ranked using a two sample t-test, and then fed into a support vector machine. Models were evaluated using the average of 50 runs of 8-fold cross-validation. The top two CpG sites could classify leukaemia from healthy cells with 84% accuracy, while the top 60 sites achieve 94% accuracy.

3. Dogan et al. [25] proposed a method for integrated genetic and epigenetic classification of coronary heart disease. The training dataset consisted of 1,545 indiviudals. An approach combining undersampling and ensemble learning [26] was used to address class imbalance, creating 8 training

sub-datasets. Point biserial correlation and Pearson correlation were used for feature selection, resulting in 107,799 CpG sites for training. These features were ranked using ROC AUC. Random Forest classifiers were then trained on the 8 training sub-datasets, with majority voting used for ensembling. Hyperparameters were tuned using 10-fold cross-validation. The final model used 4 CpG cites, two genetic variables, age and sex. This achieved an accuracy, sensitivity and specificity of 78%, 0.75 and 0.80, respectively.

1.5 Smoking Algorithms

In the context of smoking, the two most significant machine learning epigenetic scores use Elastic Net regression.

1.5.1 Elastic Net Regression

Elastic Net [27] is a regularised form of linear regression that includes two additional penalty terms. Given n examples, p features with inputs $x \in \mathbb{R}^{n\times(p+1)}$ and corresponding ground-truth $y \in \mathbb{R}^n$, we find coefficients $\beta \in \mathbb{R}^{p+1}$ that produces an output:

$$\hat{y} = x\beta \in \mathbb{R}^{p+1}$$

and minimises the function:

$$\mathcal{L}(y, \hat{y}, \beta) = ||y - \hat{y}||_2^2 + \alpha\lambda||\beta||_1 + \alpha\frac{1-\lambda}{2}||\beta||_2^2$$

where

$$||\beta||_1 = \sum_{i=1}^{p} |\beta_i|$$

and:

$$||\beta||_2 = \sqrt{\sum_{i=1}^{p} \beta_i^2}$$

We can see this is simply mean-squared error, with the L_1-norm and L_2-norm included as penalisation terms on β. The two hyperparameters α and λ control the strength of regularisation and ratio between L_1-norm and L_2-norms, respectively. Elastic Net is a combination of two other modifications to linear regression: lasso [28] and ridge [29] regression. The two mathematical regularisation terms have interpretable effects on the convergence of β. Gradient updates for $||\beta||_1$ are uniform for all non-zero values of β_i, which typically results in most β_i set to zero, with some having large values. This is referred to as a sparse solution. Gradient updates for $||\beta||_2^2$ are large for large β_i, and small for small β_i, which typically results in all β_i having similar values. This is referred to as shrinkage. Together, this results

in Elastic Net models promoting the grouping effect, where strongly related features are all included or excluded together. This grouping effect is particularly useful when $p > n$, as it provides better feature selection than lasso [27]. Because of this, Elastic Net sees wide use in epigenetics, with biological age clocks being one area of particular use [30].

1.5.2 DNAmPACKYRS

Lu et al. [31] proposed DNAmPACKYRS as a DNA methylation based score for calculating smoking pack-years. This score was originally developed as surrogate biomarker for use in the DNAm GrimAge and DNAm GrimAge v2 epigenetic clocks [31, 32]. DNAm GrimAge is a regression model for estimating mortality risk. The DNAm GrimAge score is calculated using covariates sex, age, 7 surrogate biomarkers of plasma proteins, and of relevance, the surrogate biomarker for smoking pack-years: DNAmPACKYRS. Elastic Net regression was used to train both DNAm GrimAge and the surrogate biomarkers of plasma proteins and pack-years. Training data consisted of 1731 individuals from the Framingham Heart Study dataset [33]. Indiviudals from this dataset had a mean age of 66 years. 54% of individuals were female, leaving 45% as male. The intersection of sites available on Illumina 450k and Illumina EPIC were chosen as the available CpG sites for training. This was to ensure compatability and future-proofing with new datasets. A total of 450,161 CpG sites were available. 10-fold cross-validation was used for hyperparameter tuning the regularisation strength of the Elastic Net model. This resulted in the DNAmPACKYRS score which used 172 CpG sites.

As surrogate biomarker for lifespan, DNAmPACKYRS performs better than self-reported pack-years. Firstly, DNAmPACKYRS can be used to predict lifespan in never-smokers [31], whereas self-reported pack-years cannot (all self-reported values would be 0). Additionally, DNAmPACKYRS is a more significant predictor of lifespan than self-reported pack-years. Across 4 out of 5 datasets DNAmPACKYRS had a smaller Cox regression p-values when compared to self-reported pack-years [31].

There are some limitations with the DNAmPACKYRS score. Since DNAmPACKYRS is only evaluated as a surrogate biomarker in the GrimAge paper, there is no reporting on its performance of predicting smoking status. However, this can be done independently (see 3.2). When performed, we can see that while DNAmPACKYRS achieves good predictive performance seperating never smokers from current smokers (ROC AUC = 0.991) and ex-smokers from current smokers (ROC AUC = 0.915), the score is not optimised for seperating never smokers from ex-smokers (ROC AUC = 0.798). Additionally, the choice of self-reported pack-years as a regression label introduces a source of inaccuracy, discussed below (see 1.5.3: mCigarette).

1.5.3 mCigarette

Chybowska et al. [34] proposed mCigarette as a DNA methylation based score for calculating pack-years. This score was part of the very recent (2025) study comparing brain- and blood-based DNA methylation associated with smoking. This worked aimed to overcome the limitations of self-reported smoking data, as well as the modest performance of seperating never smokers from ex-smokers of previous studies. Several techniques were employed, including high resolution approaches involving $\sim$20 million CpG sites. The most relevant method is the developed smoking biomarker: mCigarette. Similar to DNAmPACKYRS, this is a methylation based score for smoking pack-years trained with Elastic Net Regression. Training data consisted of 17,865 indiviudals from the Generation Scotland dataset [35], which was developed on the Illumina EPIC array. Individuals from this dataset had a mean age of 47.6 years. 59.1% of the individuals were female. CpG sites were filtered based on statisticcal association with tobacco use at a false discovery rate (FDR) < 0.05, resulting in 18,760 CpG sites per indiviudal. This filtering was ran in a dataset seperate to the training dataset, using the Illumina 450k array. Note that this limits training features those CpG sites found on both the 450k and EPIC arrays. Elastic Net regression was then used to train the mCigarette score. Hyperparameters were tuned using 10-fold cross-valdiation, which set $\lambda = 0.012577$ and resulted in a model using 1255 CpG sites.

As a result, mCigarette achieves incremental performance on previous scores. In an independent validation cohort, mCigarette achieves ROC AUCs of 0.98, 0.90 and 0.85 for current vs never, current vs ex- and never vs ex- smokers, respectively.

There are limitations associated with the mCigarette score. The choice of smoking pack-years may be inaccurate as a training marker, due to the potential to label two individuals with siginificantly different smoking exposure as the same. As an example, consider two ex-smokers both of 70 years of age. Individual A gave up smoking 1 year ago, but smoked half a pack of cigarettes per day for 40 years prior, which equates to 20 pack-years. Individual B smoked 2 packs of cigarettes per day from ages 15 to 25 but has quit since, which also equates to 20 pack-years. Both individuals have the same smoking pack-years score, but different lengths of cessation (1 year vs 45 years). This is problematic, as risk of chronic disease (e.g. lung cancer and cardiovascular disease) would be expected to be much higher in individual A than individual B, as duration of smoking carries greater risk than intensity of smoking [36]. Moreover, we would expect different smoking-related methylation signals for the two individuals. Thus, individuals with less intense but longer duration of smoking exposure may have underestimated smoking exposure by such a model. Therefore, smoking pack-years is an inaccurate choice of training label, as it does not capture the importance of length of smoking cessation. Note that this limitation is also associated with the DNAmPACKYRS score.

1.6 Aim of This Work

As mentioned above, there are limitations associated with using smoking pack-years as a training label, largely due to not acounting for length of smoking cessation. This could be addressed by correcting smoking pack-year values for length of smoking cessation, however, both self-reported smoking pack-years and self-reported length of smoking cessation have concerns for bias associated with self-reported data. Instead, self-reported smoking status (current, ex- or never smoker) is deemed to be a more reliable and complete measure of smoking exposure.

Therefore, we reframe modeling smoking history as a classification task rather than a regression task (2.1). This changes aims to overcome the described limitations of using smoking pack-years as a training label to achieve better class seperation between never smokers and ex-smokers. We first use individuals' self-reported smoking status to identify differentially methylated CpG sites as a feature selection step in our training dataset (2.4.1). These sites are then used to train a multi-class logistic regression model (2.5), which is validated in a hold-out test dataset (3.1) and independent validation cohort (3.2.1). Additionally, the independent validation cohort is used to compare our model with the two existing gold standard models: DNAmPACKYRS and mCigarette (3.2.2).

2 Method

2.1 Algorithm

We begin by transforming Elastic Net from a regression problem into a classification problem. In the binary classification case, this is a straightforward replacement of the mean-square error term in the loss function with a binary-cross entropy error term, alongside transforming the linear prediction into a probability via the sigmoid function. However, in the multiclass classification case the modification also affects the regularisation terms.

Given n examples, p features, K classes with inputs $x \in \mathbb{R}^{n\times(p+1)}$ and corresponding ground-truth (as one-hot encoded vectors) $y \in \mathbb{R}^{n\times K}$, we find coefficients $\beta \in \mathbb{R}^{K\times(p+1)}$ that produces logits:

$$z = x\beta^{\mathsf{T}} \in \mathbb{R}^{n\times K}$$

We transform this into probabilities with the softmax function:

$$\text{softmax}(t_1, ..., t_k) = \begin{bmatrix} \frac{\exp(t_1)}{\sum_{j=1}^{K}\exp(t_j)} \\ \vdots \\ \frac{\exp(t_k)}{\sum_{j=1}^{K}\exp(t_j)} \end{bmatrix}$$

The loss function then becomes:

$$\mathcal{L}(y, z, \beta) = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log(\text{softmax}(z_i)) + \alpha \sum_{k=1}^{K} (\lambda ||\beta_k||_1 + (1-\lambda)||\beta_k||_2^2)$$

where the L_1-norm and L_2-norm are as before.

Colloquially, this model can be thought of as training K indiviudal logistic regression sub-classifiers, where each sub-classifier is tasked only with predicting one of the K classes. The softmax function ensures that for each example, the K predicted probabilities will all add to 1. Regularisation terms work as before, but are now per-sub-classifier.

Implementation of multi-class Elastic Net logistic regression is provided by scikit-learn [37], via the `linear_model.LogisticRegression` class.

2.2 Datasets

This work uses two datasets, a discovery cohort and a valdiation cohort.

2.2.1 Cohort 1: Discovery

Cohort 1 contains DNA methylation data for 943 individuals. Across these individuals, 100% are male, and there is a mean age of 72 years. Data was collected using the Illumina 450k array, which measured the methylation of 449,521 CpG sites using the white blood cells found in whole blood samples. To ensure model compatability with independent datasets, available CpG sites were restricted to those found on both Illumina 450k and EPIC. This reduced available CpG sites by $\sim$90%. Note that this practice is consistent with methods seen in previous work (see 1.5.2 and 1.5.3). Alongside epigenetic data, age, sex, self-reported pack-years and self-reported smoking-status were also recorded. Of the 943 individuals, 235 are never smokers, 599 are ex-smokers and 109 are current smokers. 90% of this cohort was used as the model's training dataset, while the remaining 10% was reserved as a test set to evaluate model performance. This was done in a stratified manner, which preserved the class balance from the original cohort.

2.2.2 Cohort 2: Validation

Cohort 2 contains DNA methylation data for 984 individuals. Across these individuals 67% are male, and there is a mean age of 68 years. Data was collected using the Illumina EPIC array, which measured 865.859 CpG sites using the white blood cells found in whole blood samples. Because the model was trained on a dataset using the 450k array, the available sites in this evaluation dataset were restricted to those found on both Illumina 450k and EPIC. This reduced available CpG sites by $\sim$50%. Alongside epigenetic data, age, sex, self-reported pack-years and self-reported smoking-status were also recorded.

Table 1: Datasets Comparisons

Measure	Cohort 1 - Train	Cohort 1 - Test	Cohort 2
Class Balance			
Num. Individuals	848	95	984
Never Smokers	211 (25 %)	24 (25 %)	403 (41 %)
Ex-Smokers	539 (63 %)	60 (63 %)	476 (48 %)
Current Smokers	98 (12 %)	11 (12 %)	105 (11 %)
Cohort Distribution			
Percentage Male	100 %	100 %	67 %
Mean Age	72	71.5	68
Num. CpG Sites			
Raw Illumina Array	485,577	485,577	865,859
Post-intersection	449,521	449,521	449,521
Percentage Retained	93 %	93 %	52 %

Of the 984 individuals, 403 are never smokers, 476 are ex-smokers and 105 are current smokers. In addition to independently validating the trained model's performance, this dataset is used to compare the trained model to the externally-derived, existing gold standard models.

2.3 Hardware and Software

- Apple M2
- Memory (and limitations)
  - relevant to aspects of training and preprocessing (see Kruskal-Wallis)
- Python version and package versions (sklearn, numpy, scipy)
- bench mark training times and memory use

All pre-processing, training and the production of results was ran on an M2 MacBook Air. This device had an Apple M2 processor with 8 CPU cores and 8 GPU cores, 16GB of unified memory, and the macOS Sonoma version 14.2 operating system.

2.4 Pre-Processing

- Limiting both datasets to the intersection of 450k and EPIC
- Formatting to csv/transposing for ML
-

- came in format for EWAS, not machine learning
- alignment of headers
- structure wrong to read off a full set of features for a given individual
- required transposing, memory heavy operation
-

2.4.1 Feature Selection

- Kruskal-Wallis feature selection [38]
  - identifying "differentially methylated" cpg sites

2.5 Training

Cross-validation and grid search were used to select model hyperparameters. In particular, the training dataset was split into 10 stratified folds for use in cross-valdiation. A grid search was used to evaluate the different combinations of regularisation strength (α) and

[37]

- Cross Validation
  - 10 folds
  - stratified folds (same proportion of classes in each fold)
  - coarse to fine approach
  - f1-macro scoring (instead of neg log loss)
- choice of optimiser/solver
- choice of loss function
- class weight balancing for gradient updates

3 Results

1. Kruskal-Wallis: $\sim$480k sites $\rightarrow$ 3122 sites
2. Training: 3122 sites $\rightarrow$ $\rightarrow$ 2381 sites
3. Performance on Cohort 1 test set
4. Comparison to existing models

(a) comparisons of sites used in all 3 models

(b) comparisons ROC performance

3.1 Cohort 1: Training Results and Model Performance

3.2 Cohort 2: External Validation Cohort

3.2.1 Model Performance

3.2.2 Comparison with Prior Results

4 Discussion

Choices made in development

- class weight balancing
- feature pruning for cohorts, ensuring compatability
- train/test set choices in computer science vs biomedical science
- direction of discovery/validation cohorts, sex being a cofounder
- k-fold cross validation, fold size, using stratified folds
  - instead of validation set, keeps training sizes larger
- coarse-to-fine cross-validation strategy
- choice of scoring metric (f1-macro)
- number of features used (model didn't converge at 10,000 features, did at 100 and 1000)
- compute power (training times) wasn't an issue, but memory was (preprocessing)

Comparisons to other models in validation cohort

- obviously ROC curves, etc.
- how many features selected during training
-

Limitations of my model vs other models

- trained only on males
- much smaller dataset size (n)
- comparison of distributions of sex, age, etc in training cohorts (potential confounders)
  - this is likely just a limitation of the dataset and the number of participants

5 Footnotes

5.1 Ethics Statement

All participants gave written informed consent, and the study was approved by the national ethics committee. Data was anonymised and only age, biological sex and self-reported smoking values were extracted for comparison with matching whole blood DNA methylation values.

5.2 Acknowledgements

References

[1] World Health Organization. *Tobacco*. Accessed: 2024-11-04. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/tobacco.

[2] U.S. Department of Health and Human Services. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Accessed: 2024-04-24. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK179276/pdf/Bookshelf_NBK179276.pdf.

[3] Gemma M. J. Taylor and Marcus R. Munafò. "Does smoking cause poor mental health?" In: *The Lancet Psychiatry* 6.1 (2019). Accessed: 2024-11-04, pp. 2–3. DOI: 10.1016/S2215-0366(18)30459-0. URL: https://doi.org/10.1016/S2215-0366(18)30459-0.

[4] *Smoking Pack Years Calculator*. URL: https://www.smokingpackyears.com/ (visited on 05/19/2025).

[5] Myung Bae Park et al. "The Correlation of Different Cotinine Levels With Questionnaire Results: A Comparative Study for Different Measurement Methods of the Adolescent Smoking Rate in Korea". In: *Asia-Pacific Journal of Public Health* 27.5 (July 2015). Epub 2015 Jan 1, pp. 542–550. DOI: 10.1177/1010539514565447.

[6] Sarah Connor Gorber et al. “The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status”. In: *Nicotine and Tobacco Research* 11.1 (2009), pp. 12–24.

[7] Centers for Disease Control and Prevention. “Disparities in Adult Cigarette Smoking — United States, 2018”. In: *Preventing Chronic Disease* 16 (2019). Accessed: 2024-11-04, p. 180553. DOI: 10.5888/pcd16.180553. URL: https://www.cdc.gov/pcd/issues/2019/18_0553.htm.

[8] Madeleine Price Ball. *DNA Chemical Structure*. CC0 (Public Domain Dedication). 2010. URL: https://commons.wikimedia.org/wiki/File:DNA_chemical_structure.svg.

[9] Maxim VC Greenberg and Deborah Bourc’his. “The diverse roles of DNA methylation in mammalian development and disease”. In: *Nature reviews Molecular cell biology* 20.10 (2019), pp. 590–607.

[10] Elizabeth M Martin and Rebecca C Fry. “Environmental influences on the epigenome: exposure-associated DNA methylation in human populations”. In: *Annual review of public health* 39.1 (2018), pp. 309–333.

[11] Jiantao Ma et al. “Whole blood DNA methylation signatures of diet are associated with cardiovascular disease risk factors and all-cause mortality”. In: *Circulation: Genomic and Precision Medicine* 13.4 (2020), e002766.

[12] Lauren A Opsasnick et al. “Epigenome-wide association study of long-term psychosocial stress in older adults”. In: *Epigenetics* 19.1 (2024), p. 2323907.

[13] EWAS Atlas Consortium. *EWAS Atlas: A curated knowledgebase of epigenome-wide association studies*. 2024. URL: https://ngdc.cncb.ac.cn/ewas/atlas (visited on 05/27/2025).

[14] Vicky A Cameron et al. “DNA methylation patterns at birth predict health outcomes in young adults born very low birthweight”. In: *Clinical epigenetics* 15.1 (2023), p. 47.

[15] Samareh Younesian et al. “The DNA Methylation in Neurological Diseases”. In: *Cells* 11.21 (2022). ISSN: 2073-4409. DOI: 10.3390/cells11213439. URL: https://www.mdpi.com/2073-4409/11/21/3439.

[16] Yipeng Cheng et al. “Development and validation of DNA methylation scores in two European cohorts augment 10-year risk prediction of type 2 diabetes”. In: *Nature Aging* 3.4 (2023), pp. 450–458.

[17] Daniel W Belsky et al. “DunedinPACE, a DNA methylation biomarker of the pace of aging”. In: *eLife* 11 (Jan. 2022). Ed. by Joris Deelen et al., e73420. ISSN: 2050-084X. DOI: 10.7554/eLife.73420. URL: https://doi.org/10.7554/eLife.73420.

[18] Huiyan Luo et al. “Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer”. In: *Science translational medicine* 12.524 (2020), eaax7533.

[19] Inc. Illumina. *Infinium HumanMethylation450 BeadChip Datasheet*. Tech. rep. Illumina, 2012. URL: https://www.illumina.com/documents/products/datasheets/datasheet_humanmethylation450.pdf.

[20] Inc. Illumina. *Infinium MethylationEPIC BeadChip Datasheet*. Tech. rep. Illumina, 2015. URL: https://support.illumina.com/content/dam/illumina-support/documents/downloads/productfiles/methylationepic/infinium-methylation-epic-ds-1070-2015-008.pdf.

[21] Basharat Bhat and Gregory T Jones. “Data Analysis of DNA MethylationEpigenome-Wide Association Studies (EWAS): A Guide to the Principles of Best Practice”. In: *Chromatin: Methods and Protocols* (2022), pp. 23–45.

[22] Yen-Tsung Huang et al. “Epigenome-wide profiling of DNA methylation in paired samples of adipose tissue and blood”. In: *Epigenetics* 11.3 (2016), pp. 227–236.

[23] Tathiane M Malta et al. “Machine learning identifies stemness features associated with oncogenic dedifferentiation”. In: *Cell* 173.2 (2018), pp. 338–354.

[24] Péter Adorján et al. “Tumour class prediction and discovery by microarray-based DNA methylation analysis”. In: *Nucleic acids research* 30.5 (2002), e21–e21.

[25] Meeshanthini V Dogan et al. “Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham Heart Study”. In: *PloS one* 13.1 (2018), e0190549.

[26] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. “Exploratory undersampling for class-imbalance learning”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2008), pp. 539–550.

[27] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.

[28] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[29] Arthur E Hoerl and Robert W Kennard. “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1 (1970), pp. 55–67.

[30] Andrew E Teschendorff and Steve Horvath. “Epigenetic ageing clocks: statistical methods and emerging computational challenges”. In: *Nature Reviews Genetics* (2025), pp. 1–19.

[31] Ake T Lu et al. “DNA methylation GrimAge strongly predicts lifespan and healthspan”. In: *Aging (albany NY)* 11.2 (2019), p. 303.

[32] Ake T Lu et al. "DNA methylation GrimAge version 2". In: *Aging (Albany NY)* 14.23 (2022), p. 9484.

[33] Thomas R. Dawber, Gilcin F. Meadors, and Felix E. Moore. "Epidemiological Approaches to Heart Disease: The Framingham Study". In: *American Journal of Public Health and the Nations Health* 41.3 (1951). PMID: 14819398, pp. 279–286. DOI: 10.2105/AJPH.41.3.279. eprint: https://doi.org/10.2105/AJPH.41.3.279. URL: https://doi.org/10.2105/AJPH.41.3.279.

[34] Aleksandra D Chybowska et al. "A blood-and brain-based EWAS of smoking". In: *Nature Communications* 16.1 (2025), p. 3210.

[35] Blair H Smith et al. "Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability". In: *BMC medical genetics* 7 (2006), pp. 1–9.

[36] Alexandra L Potter et al. "Pack-year smoking history: An inadequate and biased measure to determine lung cancer screening eligibility". In: *Journal of Clinical Oncology* 42.17 (2024), pp. 2026–2037.

[37] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[38] William H. Kruskal and Wilson Allen Wallis. "Use of Ranks in One-Criterion Variance Analysis". In: *Journal of the American Statistical Association* 47 (1952), pp. 583–621. URL: https://api.semanticscholar.org/CorpusID:51902974.