COSC681 - AI Project

Classifying lifetime smoking exposure with DNA methylation trained machine learning

Lachlan Jones

2025

University of Canterbury
Department of Computer Science

Abstract

Smoking is the leading cause of preventable death. Current surveillance relies on self-reported smoking data, which is prone to inaccuracy and bias. As an alternative, DNA methylation can be used as a biomarker for lifetime smoking exposure, and is not limited by the same biases as self-reported data. Prior research used self-report pack-years as a training label, which we deem to be an unreliable and biased score. We propose Elastic Net Smoking-Status (ENSS), a DNA methylation-based machine learning model which classifies lifetime smoking exposure. As an alternative to self-reported pack-years, we used self-reported smoking-status as a training label. Individuals' self-reported smoking status was used to identify differentially methylated CpG sites via the Kruskal-Wallis test by ranks. These sites were then used to train a multi-class logistic regression model, ENSS, which used Elastic Net regularisation. ENSS was evaluated in a hold-out test dataset as well as an independent test cohort. Additionally, the independent test cohort was used to compare ENSS with two existing gold standard models: DNAmPACKYRS and mCigarette. We evaluated ENSS' ability to separate classes of smoking-status (never, ex, and current) pairwise, with AUCs of 0.977, 0.941, 0.909 on a hold-out test dataset, and AUCs of 0.971, 0.859, 0.686 on an independent test cohort. While ENSS did not outperform prior research, we indentified several limitations of the model that could be addressed to improve performance and mitigate overfitting.

Contents

List of Figures

List of Tables

List of Equations

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 is the leading cause of preventable death in both the United States [2] and New Zealand [3]. Smoking causes cancer, heart disease, 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 [4]. Such negative impacts on patient health due to tobacco use are undesirable, as well as are avoidable. For these reasons, 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 record of a patient's smoking history by personal recalling and reporting. This 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. Individuals are binned into never-smokers, ex-smokers and current-smokers. Alternatively, 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 [5]. 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 [6, 7]. Specifically, the social pressure to deny partaking in stigmatised behaviours, memory lapses and not being aware of sources of second-hand exposure can all influence the accuracy 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 [6]. Moreover, tobacco consumption differs between social groups, with smoking more prevalent in low-education and low-socio-economic groups [8].

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

Figure 1: Chemical structure of DNA [9]

of epigenetic biomarkers.

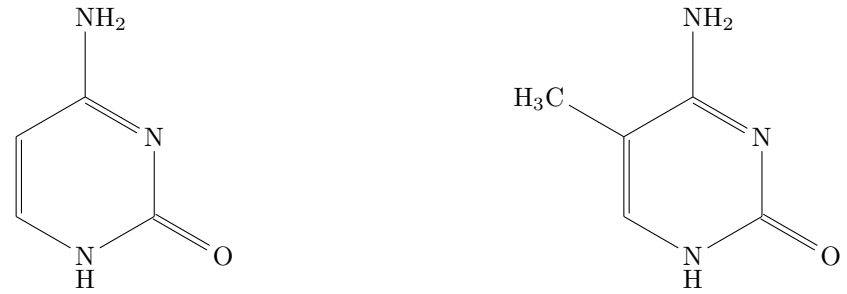
1.3 Epigenetics

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 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 (see Diagram 1). 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 (DNAm) involves the addition of a methyl group (CH_3) to the 5-carbon position of cytosine nucleotides (see Diagram 2). This chemical modification to the cytosine makes it harder for transcription to occur, which can modulate, or even completely silence, gene expression. This is the most relevant and reproducible 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



(a) Cytosine (b) 5-Methylcytosine

Figure 2: Modification of cytosine into 5-methylcytosine caused by DNA methylation

main form of DNAm that occurs in mammalian cells.

A biological sample used to assess DNAm will contain many thousands of copies of DNA. When measuring a CpG site for DNAm, the reported value is the percentage of DNA copies from the sample which are methylated. That is, a value between 0 (all DNA copies unmethylated) 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 [10]. Exposure to a wide range of environmental factors including air pollutants [11], diet [12], physical activity and even psychosocial stress [13] have been shown to be associated with specific changes in DNAm. Moreover, sufficient lack of exposure to such factors can reverse such changes. As previously mentioned, DNAm 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, DNAm is not random. There is strong correlation between methylation of specific sites with specific factors [14]. In the context of this work, smoking is strongly associated the methylation of cg05575921, found in the aryl hydrocarbon receptor repressor (AHRR) gene [15]. With these 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, DNAm is a biomarker not only useful for reporting on environmental exposures, but also predicting future health outcomes and risks. Examples of this include prediction of cardiovascular diseases [16], neurological diseases [17], type 2 diabetes [18], pace of ageing [19], and cancer [20]. Furthermore, DNAm is not self-reported, and therefore overcomes the biases associated with self-reported data.

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

1.3.2 DNAm arrays

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, referred to as an array. There are two commonly chosen DNAm arrays used: Illumina 450k [21] and Illumina EPIC [22]. 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 the array size 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 compatibility with multiple datasets or cohorts is of concern (see 2.4). These arrays are compatible with DNA from multiple different cell types. Some examples include whole blood, purified cells and fresh-frozen tissue [23]. Moreover, the methylation signal can differ across different cell types [24]. This 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) for its convenience [25].

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

Malta et al. [26] 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 hyper-methylated 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.

Adorján et al. [27] proposed a method for using DNAm data 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.

Dogan et al. [28] proposed a method for integrated genetic and epigenetic classification of coronary heart disease. The training dataset consisted of 1,545 individuals. An approach combining undersampling and ensemble learning [29] 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 sites, two genetic variables, age and sex. This achieved an accuracy, sensitivity and specificity of 78%, 0.75 and 0.80, respectively.

Adeoye et al. [30] proposed a method for using saliva samples to detect oral cancer. This study aimed to propose an alternative to tissue biopsy, which can be invasive and prone to false-negatives. Training data consisted of saliva samples of 33 randomly selected patients with oral cancer. Eight machine learning models were compared (RBF kernel, SVMs, AdaBoost, kNNs, random forest, decision trees, ExtraTrees, and gradient boosting machines) alongside three feature selection techniques (ANOVA, MRMR and LASSO). The SVM was the best performing model, with a recall, specificity and precision of 0.94, 0.93 and 0.94, respectively, while LASSO was the most robust feature selection technique across all machine learning models.

Cheng et al. [18] proposed a method for predicting the risk of developing type 2 diabetes within the next 10 years. Training data consisted of 374 cases and 9461 controls, for a total of 9835 individuals. Several penalised regression models were trained, including Cox proportional-hazards with LASSO regularisation, random survival forests, and survival bayesian additive regression trees. Evaluation compared models using only traditional risk factors against models that included DNAm factors. All models showed incremental gains in performance with the addition of DNAm factors, with an increase in ROC-AUC values. The best performing model was the Cox proportional-hazards with LASSO regularisation, which reported an ROC-AUC of 0.872.

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 [31] is a regularised form of linear regression that includes two additional penalty terms. Given n examples, p features with data $x \in \mathbb{R}^{n \times (p+1)}$ 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:

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

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 regularisation strength and ratio between L_1-norm and L_2-norms, respectively. Elastic Net is a combination of two other forms of regularised linear regression: lasso [32] and ridge [33] 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 [31]. Because of this, Elastic Net sees wide use in epigenetics, with biological age clocks being one area of particular use [34].

1.5.2 DNAmPACKYRS

Lu et al. [35] proposed DNAmPACKYRS as a DNAm based score for calculating smoking pack-years. This score was originally developed as surrogate biomarker for use in the DNAm GrimAge [35] and DNAm GrimAge v2 epigenetic clocks [36]. DNAm GrimAge is a regression model for estimating mortality risk. The DNAm GrimAge score is calculated using the covariates sex, age, 7 surrogate biomarkers of plasma proteins, and of relevance, a 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 (FHS) dataset [37]. Individuals 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 compatibility 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 [35], 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 [35]. 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 3.5). When performed, we can see that DNAmPACKYRS achieves ROC-AUCs of 0.991, 0.915 and 0.798 for separating never- vs current-, ex- vs current- and never- vs ex-smokers, respectively.

There are some limitations with the DNAmPACKYRS score. While the score achieves excellent performance separating current-smokers from the other smokers, performance separating never-smokers and ex-smokers is modest. Additionally, the choice of self-reported pack-years as a regression label introduces a source of inaccuracy, discussed below (see 1.5.3).

1.5.3 mCigarette

Chybowska et al. [38] proposed mCigarette as a DNAm based score for calculating pack-years. This score was part of the very recent (2025) study comparing brain- and blood-based DNAm associated with smoking. This worked aimed to overcome the limitations of self-reported smoking data, as well as the modest performance of separating never-smokers from ex-smokers of previous studies. Several techniques were employed, including high resolution approaches involving ~ 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 individuals from the Generation Scotland (GS) dataset [39], 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 statistical association with tobacco use at a false discovery rate (FDR) < 0.05, resulting in 18,760 CpG sites per individuals. This filtering was run in a dataset separate 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-validation, which

set $\alpha = 0.012577$, while λ was fixed at 0.5. This resulted in a model using 1255 CpG sites. As a result, mCigarette achieved improved performance compared to previous scores. In an independent validation cohort, mCigarette achieved 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 label, due to the potential to identified two individuals with significantly different smoking exposures as similar. 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 [40, 41, 42, 43]. 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 and biased choice of training label, as it assigns equal importance to smoking duration and smoking intensity [44]. As mentioned above, this limitation is also associated with the DNAmPACKYRS score, which also used smoking pack-years for training labels.

1.6 Aim of this work

This work aims to use DNA methylation-based machine learning to classify lifetime smoking exposure, using self-reported smoking status as a training label.

As discussed, there are limitations associated with using smoking pack-years as a training label, namely, not accounting for the length of smoking cessation. This could be addressed by adjusting smoking pack-year values for the length of smoking cessation. However, this would require the use of additional self-reported data, which could also introduce further bias. As an alternative to smoking pack-years, self-reported smoking status (current-, ex- or never-smoker) is a more consistent measure of lifetime smoking exposure. In contrast to prior research which uses smoking pack-years, we propose Elastic Net Smoking-Status (ENSS) as a method which uses smoking-status as a training label for supervised machine learning. Since smoking-status is a discrete variable (rather than continuous like smoking pack-years), this required reframing the modelling of smoking history as a classification task rather than a regression task. While current methods for predicting smoking history perform well separating current-smokers from other smokers, performance separating never-smokers from ex-smokers is modest. Because of the

described limitations of smoking pack-years, the change to smoking-status in our proposed method aims to achieve better class separation between never-smokers and ex-smokers.

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 achieved by a straightforward modification to the loss function, replacing the mean-square error term with a binary-cross entropy error term, alongside transforming the linear prediction into probabilities via the sigmoid function. However, in the multi-class classification case the modification is more complex, and also affects the regularisation terms. We modify the algorithm seen in 1.5.1 as follows:

Given n examples, p features, K classes with data $x \in \mathbb{R}^{n\times(p+1)}$ and corresponding ground-truth $y \in \mathbb{R}^{n\times K}$ (as one-hot encoded vectors), 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 a row of logits into probabilities with the softmax function:

$$\hat{y_i} = \text{softmax}(z_i) \in \mathbb{R}^{n\times K}$$

where:

$$\text{softmax}(t_1, \ldots, t_k) = \left[\frac{\exp(t_1)}{\sum_{j=1}^{K}\exp(t_j)} \cdots \frac{\exp(t_k)}{\sum_{j=1}^{K}\exp(t_j)}\right]$$

The loss function then becomes:

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

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

The output of this model $\hat{y}$ is matrix of K probability predictions, one for each class, for n different examples. For each example, the softmax function ensures the K predicted probabilities from these logits will sum to 1. To make a prediction for the i^{th} example, the model chooses the class with the largest probability: $\text{argmax}(\hat{y_i})$. Alternatively, this model can be thought of as training an ensemble of K individual logistic regression predictors. Each of K rows of β is a class-specific linear predictor that produces logits for predicting a single class. Regularisation terms work as before, but are now regularising coefficients per-class-specific linear predictor, rather than regularising all coefficients in the model together.

2.2 Datasets

This work used two datasets, a discovery cohort and an independent test cohort.

2.2.1 Cohort 1: Discovery

Cohort 1 contained DNAm data for 943 individuals. Across these individuals, 100% were male, and there was 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 compatibility with independent datasets, available CpG sites were restricted to those found on both the Illumina 450k and EPIC arrays. This reduced available CpG sites by $\sim$10%. Note that this practice was consistent with the methods seen in prior research (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 were never-smokers, 599 were ex-smokers and 109 were current-smokers. 90% of this cohort was used as the model's training dataset, while the remaining 10% was reserved as a hold-out test dataset to evaluate model performance. This split was done in a stratified manner, which preserved the class balance from the original cohort.

2.2.2 Cohort 2: Evaluation

Cohort 2 contained DNAm data for 984 individuals. Across these individuals 67% were male, and there was 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. Of the 984 individuals, 403 were never-smokers, 476 were ex-smokers and 105 were current-smokers. In addition to independently testing ENSS' performance, this dataset was used to compare ENSS to the externally-derived, existing gold standard models.

2.2.3 Choice of discovery and evaluation cohorts

This approach of using two independent datasets to train and evaluate a model is not conventional in machine learning. Instead, all available data is split in training, validation and testing sets. However, independent testing data is common practice in clinical research. Not only does this evaluate overfitting, it also evaluates generalisability across populations and ascertainment bias, both relevant challenges when developing a clinical test. Because of the multidisciplinary nature of this project, a combination of both approaches was used. This then required deciding which cohort would provide training data, and which would be used for evaluation. Cohort 1 was selected for training as it contained only males. Given

Table 1: Training and testing dataset 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 %

that sex is a likely confounder of lifetime smoking exposure (males tend to smoke more than females [45, 46]), Cohort 1 was deemed to be a less-confounded dataset.

Summaries of the datasets used in this study can be seen in Table 1.

2.3 Hardware and software

Pre-processing, training and the production of results were run 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. Code implementations were written in Python (version 3.9.10), using the scikit-learn (version 1.6.0) [47], SciPy (version 1.13.1) [48] and NumPy (version 1.26.4) [49] packages.

2.4 Pre-processing

Both datasets required preparation before training. Raw Illumina array text files came in a format suitable for an EWAS, but not machine learning. Preparation steps included conversion into comma separated value format, merging and formatting of headers, transposing whole files, and splitting into separate labels and data files. Because of the large file size ($\sim$3.5 GB per dataset), not all operations could be run in-memory. To address this, the pre-processing pipeline made multiple writes to the disk, requiring $\sim$10 GB of storage per dataset (including the initial dataset file). These memory limitations were most noticeable when transposing whole EWAS files, requiring a package to process files in chunks [50].

Available CpG sites for training were filtered in two different ways. Firstly, as previously mentioned, both datasets were filtered to the intersection of CpG sites available on both the 450k and EPIC arrays, to ensure compatibility. Secondly, training data required pre-training feature selection. High dimensional data is a well observed technical challenge for machine learning [51]. From a technical perspective, early experimentation showed model training not converging with an input space of the order of 10,000 features, despite supervised feature selection being built into the model (see Equation 2). However, an input space of the order of 1000 or 100 features converged adequately. The concept of sample complexity (the number of examples needed for a model to achieve sufficient accuracy) is well recognised [52], suggesting additional feature selection, alongside the regularisation built into the model, was required for the given sample size. From a practical perspective, a test which uses fewer features is more cost-effective and interpretable, hence more likely to be translatable to clinical applications. Therefore, developing a model which achieves sufficient accuracy with the fewest features possible is desirable. To implement this feature selection, we used the Kruskal-Wallis test by ranks.

2.4.1 Feature selection: Kruskal-Wallis test

The Kruskal-Wallis test by ranks [53] is a non-parametric statistical test that identifies if there are statistically significant differences between the distributions of observations for two or more samples. In the context of this work, the test was run per CpG site, where the samples are classes of smoking status (current-, ex-, never-smokers) and observations are DNAm values for that site. Applying this test to the training dataset identifies the CpG sites which are the most differentially methylated across all three classes of smoking status. More commonly used statistical tests are the one-way ANOVA [54] or Mann-Whitney U [55] tests. One-way ANOVA is a parametric test for two or more samples, while Mann-Whitney is a non-parametric test for exactly two samples. A parametric test assumes the samples are drawn from populations which are normally distributed, and there is no evidence to make this assumption for DNAm data. A goal of this work was to improve class separation between all three classes, motivating the use of a test on two or more samples, rather than running a test on exactly two samples pair-wise three times. For these reasons, these tests were rejected in favour of the Kruskal-Wallis test. In fact, Kruskal-Wallis is the non-parametric equivalent of one-way ANOVA, and extension to more than two samples of Mann-Whitney, making this a natural choice.

To run the test, first all observations are ranked (sorted in ascending order, with correction for ties). The test statistic is then given by:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{C} \frac{R_i^2}{n_i} - 3(N+1) \tag{3}$$

where C is the number of samples, n_i is the number of observations in the i^{th} sample, $N = \sum n_i$ is

the number of observations in all samples, and R_i is the sum of the ranks in the i^{th} sample. The null-hypothesis is the distributions of all samples are the same. A large value of H is indicative of rejecting the null-hypothesis. Moreover, approximating the distribution of the H statistic as a chi-squared distribution allows the calculation of p-values. However, because the test was run hundreds of thousands of times, there is an increased chance at least one test was deemed significant (i.e. a false positive) due to random chance. To control the false discovery rate (FDR), we used Benjamini-Hochberg correction [56] to turn p-values into false discovery corrected q-values. To achieve feature selection, we retained all CpG sites with a q-value < 0.05. Implementations of the Kruskal-Wallis test and Benjamini-Hochberg correction are provided by SciPy, via the `scipy.stats.kruskal` and `scipy.stats.false_discovery_control` functions.

2.5 Training

The 3122 sites identified by the Kruskal-Wallis test were then utilised in machine learning training. Using the 848 individuals from the cohort 1 training set, DNAm values from these sites were used to train a multi-class Elastic Net logistic regression model. The implementation of this model was provided by scikit-learn, via the `linear_model.LogisticRegression` class. This model used the Elastic Net penalty and SAGA solver, as this is the only solver compatible with Elastic Net regularisation. However, this does make the model stochastic. Additionally, the model used balanced class weights. Because of the class imbalance in the training dataset (see Table 1) the model was performing poorly when classifying never-smokers and current-smokers. By using balanced class weights, gradient updates are scaled inversely-proportional to the frequency of the given class. While this fixed the misclassification issues, it also can increase the chance of overfitting if the class balance present in the training sample is not representative of the true population.

Two of the hyperparameters required tuning: regularisation strength (α) and L_1 ratio (λ). Values for these were determined using a combination of k-fold cross-validation and grid search. In particular, the training dataset was split into 10 stratified folds for use in cross-validation. A choice of $k = 10$ was used to ensure each run in the cross-validation loop contained sufficient training data, because of the relatively low training dataset size. Folds were chosen to be stratified to because of the class imbalance. Cross-validation was used instead of a typical validation hold-out set in order to preserve the size of the training dataset. A grid search was used to evaluate the different combinations of hyperparameters. F_1-macro was selected as the scoring function to evaluate each fit, as each class is equally represented in the score, in spite of class imbalance. Outputs of cross-validation included the mean and standard deviation of F_1-macro scores across the 10 folds per fit. These were then used to calculate 95% confidence intervals (see Figure 3). Because the majority of hyperparameter choices had overlapping 95% confidence intervals, they were deemed to not be statistically different. Therefore, hyperparameters

Figure 3: F_1-macro scores for each fit during cross-validation. Error bars indicate 95% confidence intervals. Because of the overlapping confidence intervals, the choices of hyperparameters were not deemed to be statistically different.

were selected to increase regularisation strength and minimise the number of features used in the final model. This resulted in a choice of $\alpha = 0.2$ and $\lambda = 0.85$, which were used to retrain the Elastic Net model on the entire training dataset. Henceforth, we refer to this trained model as Elastic Net Smoking-Status (ENSS). Implementations of cross-validation and grid search were provided by scikit-learn, via the `sklearn.model_selection.StratifiedKFold` and `sklearn.model_selection.GridSearchCV` classes.

Additionally, a linear regression Elastic Net model was trained to benchmark against, using the same algorithm seen in DNAmPACKYRS and mCigarette (see Equation 1). This model used the same choice of hyperparameters as the logistic regression model, and used self-reported pack-years as a training label. The implementation of this model was provided by scikit-learn, via the `linear_model.ElasticNet` class.

3 Results

3.1 Site reduction from feature selection and regularisation

As discussed in the method (2.4.1), the Kruskal-Wallis test was used to select CpG sites as a feature selection step. The test was run independently on the 485,577 CpG sites common to the 450k and EPIC arrays, using the observations (DNAm values) from cohort 1. After controlling for FDR, 3122 CpG sites remained with a q-value < 0.05.

From the 3122 sites identified by the Kruskal-Wallis test, the Elastic Net logistic regression model selected 2381 different sites used in predicting smoking status. These 2381 sites consisted of three sets of CpG sites, where each set contained the sites used to predict the probabilities for never-, ex- and current-smokers, respectively. The intersections of these sets can be seen in Figure 4. While 2381 total sites were used for the entire model, only 520 were common to predicting all 3 classes of smoking status. While each class-specific linear predictor regularised to a similar number of CpG sites (1653 for never, 1627 for ex and 1514 for current), there were 1373 sites used in the prediction of exactly two of the three classes, and 488 sites used in the prediction of exactly one class. The limited overlap between the features used in each class-specific predictor was unsurprising, as the L_1-norm is applied per class (see Equation 2) rather than across all parameters used in the model. However, this property is not desirable, as it led to a larger number of CpG sites used in the entire model.

Figure 4: Intersection of CpG sites identified per class-specific predictor. All three class-specific linear predictors regularised to a similar number of features used during training.

3.2 Evaluation metrics

The potential utility of a smoking model lies in its ability to differentiate pairs of classes (i.e. never- vs current-, ex- vs current- and never- vs ex-). Moreover, the multi-class model can be thought of as an ensemble of three classifiers that each predict the probability of a different class (see 2.1). Therefore, the model's ability to separate classes was evaluated using receiver operating characteristic (ROC) curves, using a one vs one (OvO) strategy to compare pairs of classes [57]. Two types of ROC curves were evaluated. First, ROC curves of the probabilities for each of the class-specific predictors were evaluated. These show a single set of coefficients ability to separate two classes. Subsequently, macro-averaging [57] was used to aggregate the outputs of both relevant class-specific predictors, which generated a single ROC curve used for evaluating OvO class separation of the entire model. The red points on ROC curves indicate the optimal threshold for separation, calculated using Youden's index [58]. Area under the curve (AUC) was used as an aggregated score for how well the model separated the two classes, reflecting both sensitivity (true positive rate) and specificity (true negative rate). The rubric used throughout this work for evaluating AUCs can be seen in Table 2.

Table 2: AUC evaluation rubric

AUC range	**Discrimination ability**	**Usefulness**
0.9 to 1.0	Excellent	Near perfect separation of the two groups, with 1.0 being perfect discrimination
0.8 to 0.9	Good	Generally useful at separating the two groups, but there is room for improvement
0.7 to 0.8	Acceptable/Modest	Has some value, but ability to separate the two groups is limited
0.6 to 0.7	Poor	Limited value in separating the two groups
0.5	No discrimination	No better performance than random guessing

3.3 Test set model performance (Cohort 1)

ENSS achieved excellent predictive performance on the Cohort 1 hold-out test set. The classification performance on this dataset can be seen in Figure 5. Out of the 95 individuals in this dataset, only 15 were misclassified by the model. Per class, 92 % of never-smokers were correctly classified, while 82 % of both ex-smokers and current-smokers were correctly classified.

The model's ability to separate classes was also evaluated. Macro-averaged OvO ROC curves for the entire model can be seen in Figure 6. These were generated by aggregating OvO ROC curves of the class-specific predictors, seen in Figure 7. The model achieved macro-averaged OvO ROC-AUCs of: 0.977 for separating never-smokers from current-smokers, 0.941 for separating ex-smokers from current-smokers, and 0.909 for separating never-smokers from ex-smokers. Overall, we can see the model achieved excellent performance separating each of the pairs of classes.

Finally, distributions of probabilities per class, per class-specific predictor, can be seen in Figure 8. Here we can see each class-specific predictor generally predicted its positive class with high probability, and the two negative classes with low probability. This separation was distinct for the predictor of current-smokers and never-smokers, with little to no overlap between box plots. However, separation is less distinct for ex-smokers, as there is some overlap between the box plots of ex-smokers and the box plots of the other classes. This is perhaps unsurprising, given that ex-smokers are the intermediate class, and a biological state between current- and never-smokers.

(a) Counts

(b) Proportions

Figure 5: Confusion matrices (Cohort 1 - test set). The first confusion matrix displays classification and misclassification counts, while the second confusion matrix normalises these into proportions. The model correctly classified each class with high accuracy in this cohort.

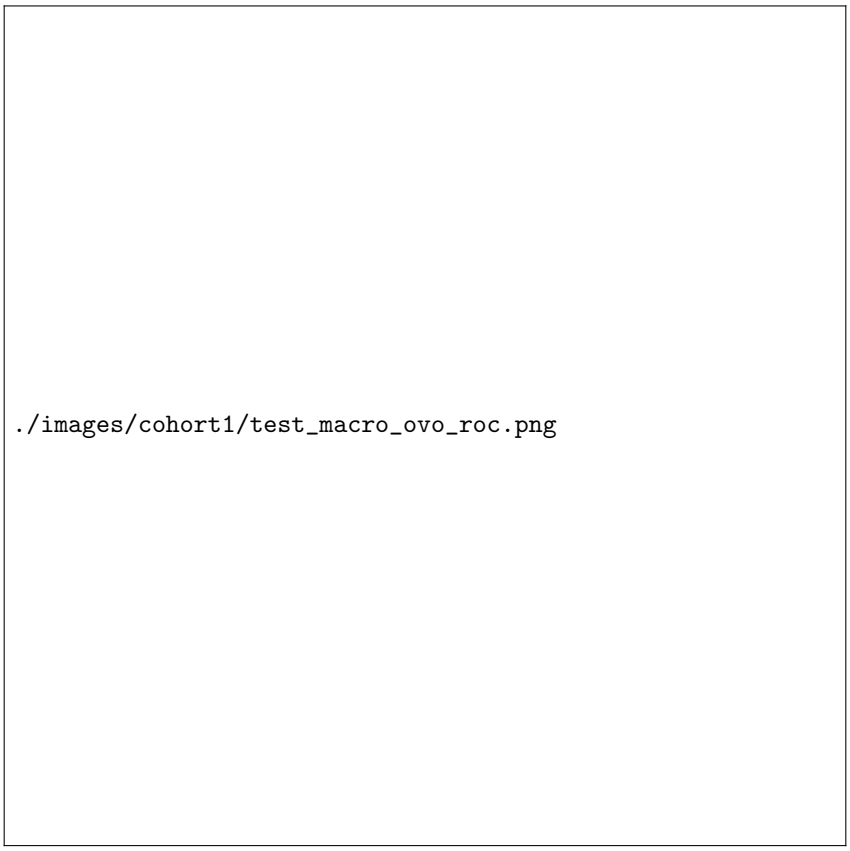


Figure 6: Macro-averaged ROC curves (Cohort 1 - test set). Each curve aggregates the two relevant class-specific predictors into a single ROC curve, giving a single curve which evaluates the entire model's performance separating two classes. The model's separation of all three classes was excellent in this cohort.

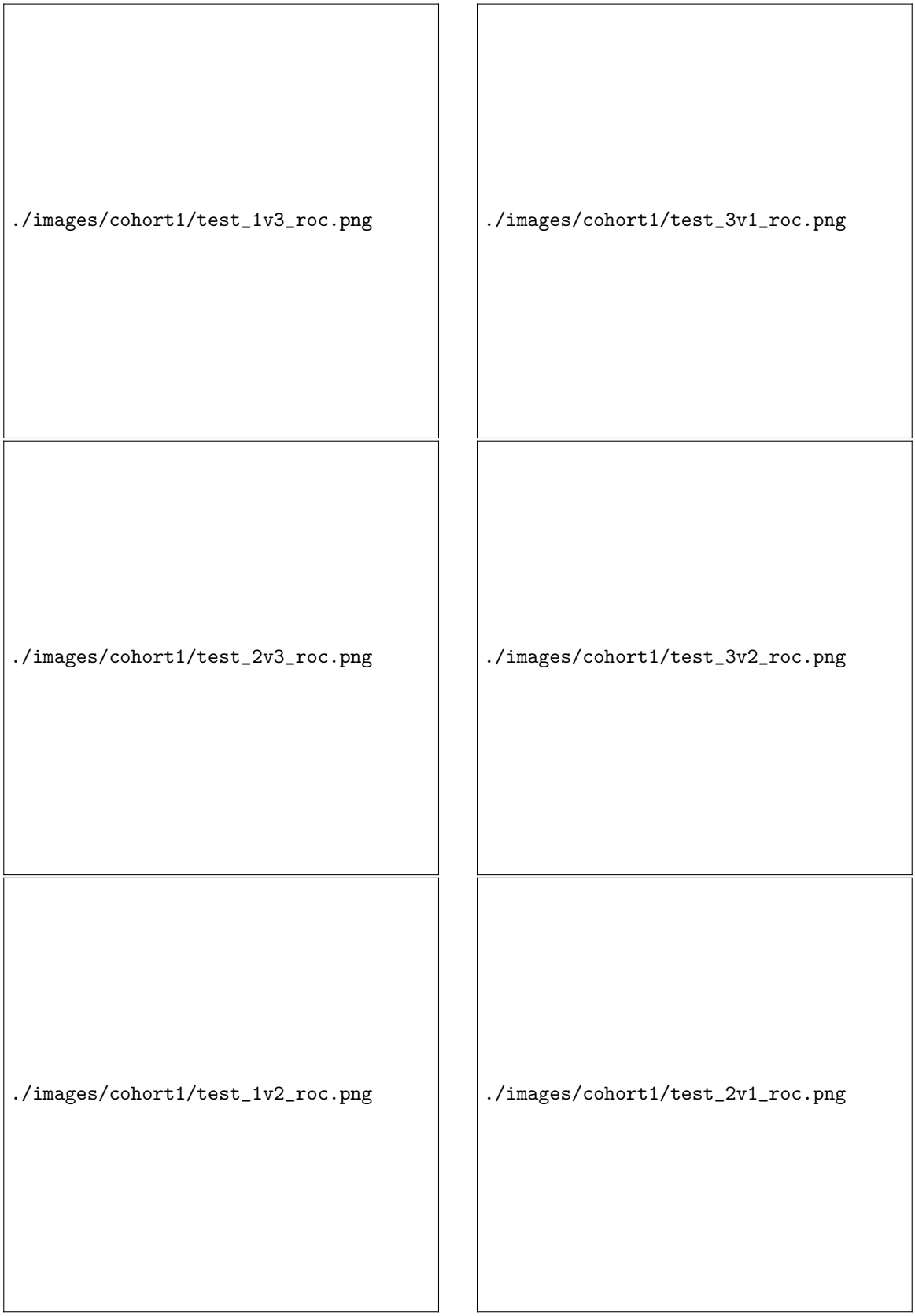


Figure 7: OvO class-specific predictor ROC curves (Cohort 1 - test set). Each curve shows a single class-specific predictor's ability to separate two classes, using the probability outputs for the given predictor. The predictor used for each curve corresponds to the positive class. Each class-specific predictor achieved excellent performance separating each pair of classes in this cohort.

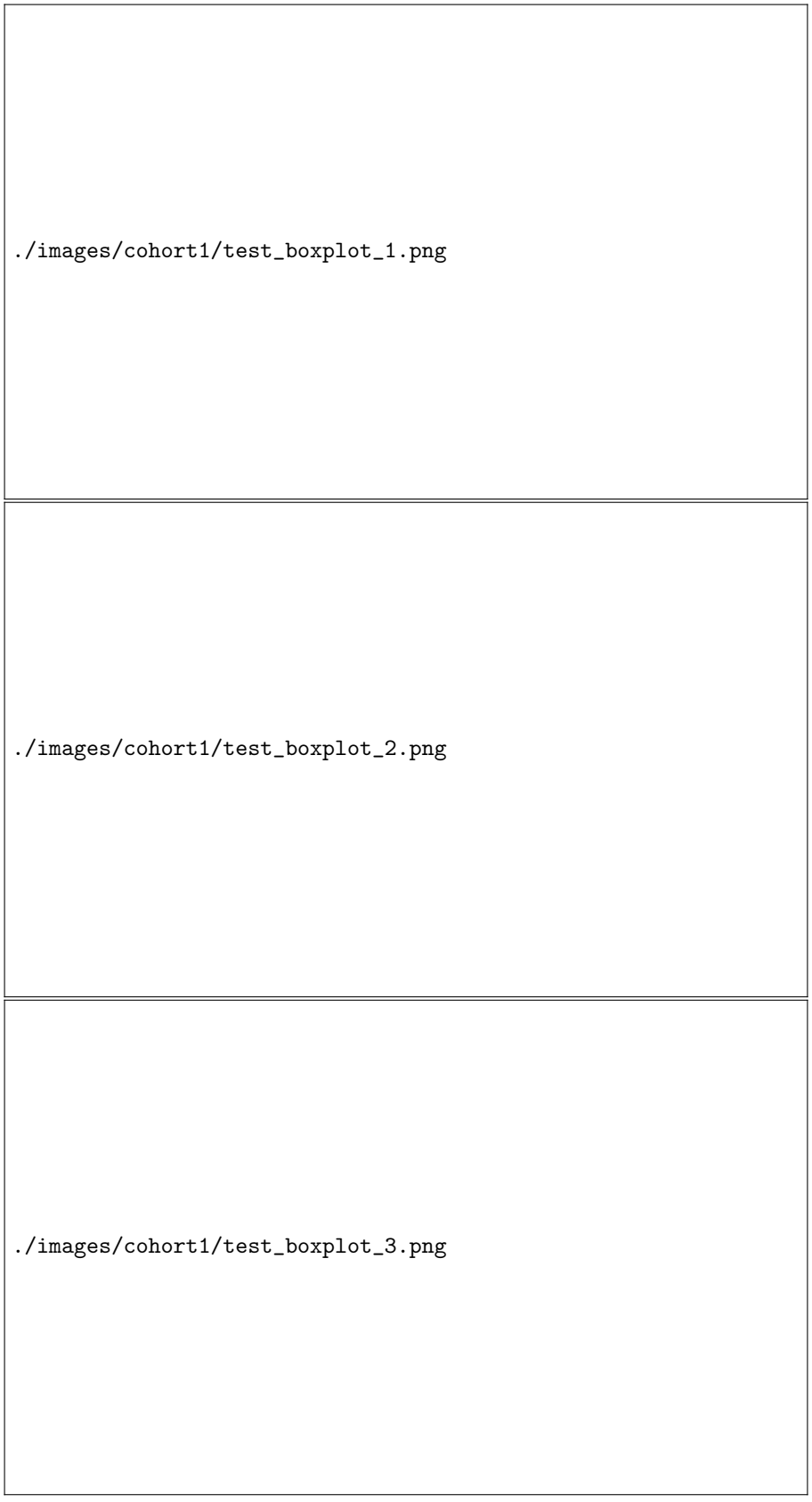


Figure 8: Class-specific probability box plots (Cohort 1 - test set). Each box plot shows the range of probability outputs for a given class-specific predictor, separated by class. In this cohort, each class-specific predictor assigned high probabilities to its positive class, and low probabilities to the other classes.

3.4 Independent cohort model performance (Cohort 2)

ENSS achieved modest predictive performance on the Cohort 2 independent test dataset. The classification performance on this dataset can be seen in Figure 9. Out of the 984 individuals in this dataset, 271 were misclassified by the model. Per class, 73 % of current-smokers and 78 % of ex-smokers were correctly classified, while only 0.36 % of never-smokers were correctly classified, with 0.62 % of never-smokers being confused for ex-smokers. Compared to the performance in Cohort 1, there was a clear decrease in predictive accuracy in Cohort 2.

The model's ability to separate classes was also evaluated. Macro-averaged OvO ROC curves for the entire model can be seen in Figure 10. These were generated by aggregating OvO ROC curves of the class-specific predictors, seen in Figure 11. The model achieved macro-averaged OvO ROC-AUCs of: 0.971 for separating never-smokers from current-smokers, 0.859 for separating ex-smokers from current-smokers, and 0.686 for separating never-smokers from ex-smokers. Comparing the performance in Cohort 1 to Cohort 2, we can see the model maintained the ability to discriminate never-smokers from current-smokers, but the ability to separate ex-smokers from current-smokers was only good, and the ability to separate never-smokers from ex-smokers was poor.

Finally, distributions of probabilities per class, per class-specific predictor, can be seen in Figure 12. As seen in the ROC curves, the model was able to separate current-smokers from the other two classes, predicting high probabilities in the current-smoker class-specific predictor, and low probabilities in the other two predictors. However, we can see between the never-smoker and ex-smoker class-specific predictor box plots there was significant overlap, leading to misclassification and confusion between these two classes.

(a) Counts

(b) Proportions

Figure 9: Confusion matrices (Cohort 2). The model classified ex- and never-smokers with moderate accuracy in this cohort, but misclassified most never-smokers as ex-smokers.

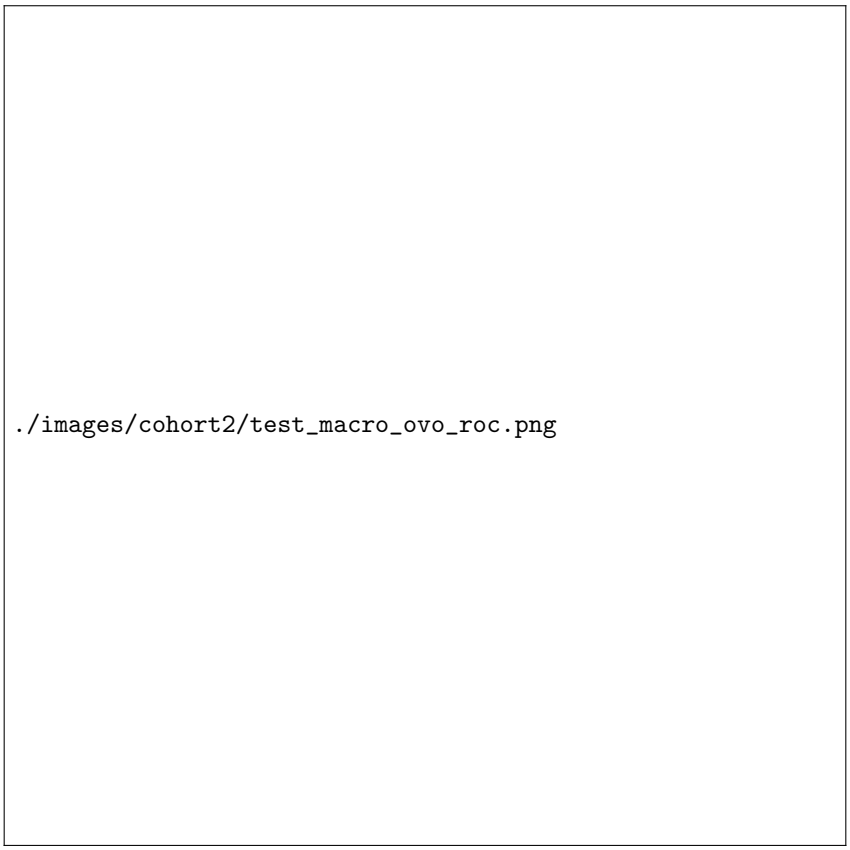


Figure 10: Macro-averaged ROC curves (Cohort 2). Each curve aggregates the two relevant class-specific predictors into a single ROC curve, giving a single curve which evaluates the entire model's performance separating two classes. The model's separation of never- and current-smokers was excellent in this cohort, while separation of ex- and current-smokers was good, and separation of never- and ex-smokers was poor.

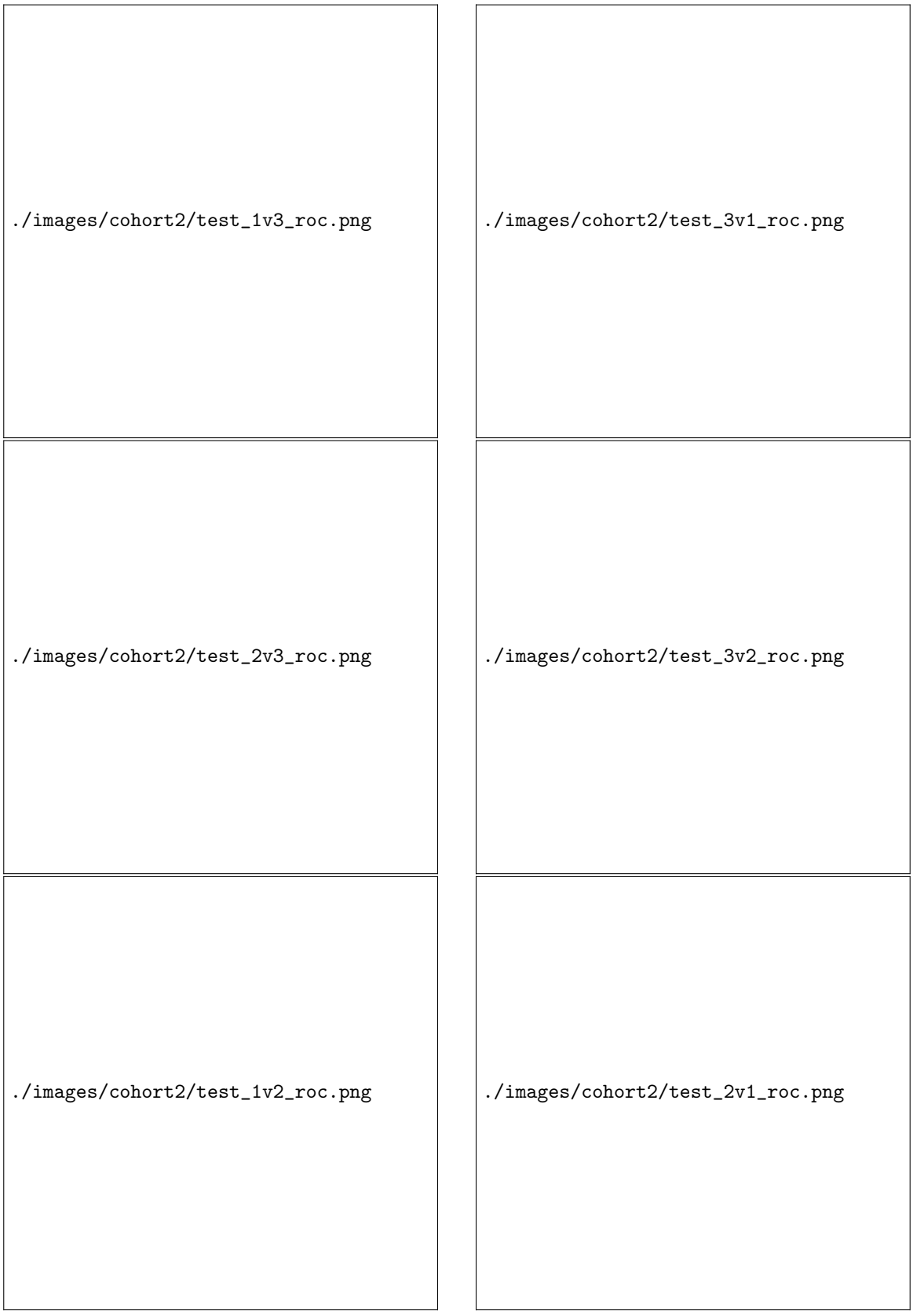


Figure 11: OvO class-specific predictor ROC curves (Cohort 2). Each curve shows a single class-specific predictor's ability to separate two classes, using the probability outputs for the given predictor. The predictor used for each curve corresponds to the positive class. Separation of each pair of classes varied in this cohort, with class-specific predictors achieving excellent performance for never vs current, excellent to good performance for ex vs current, and poor to acceptable performance for never vs ex.

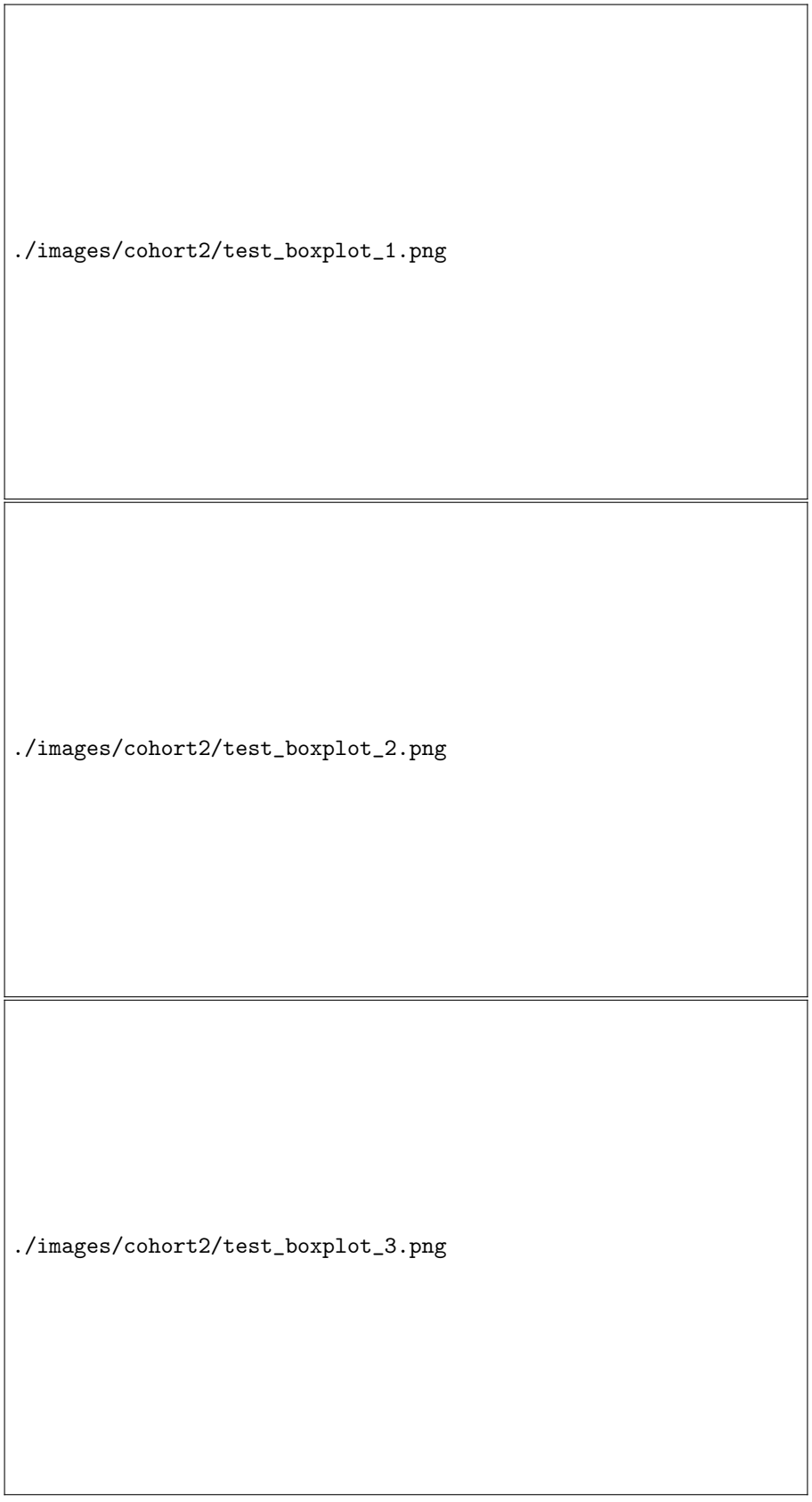


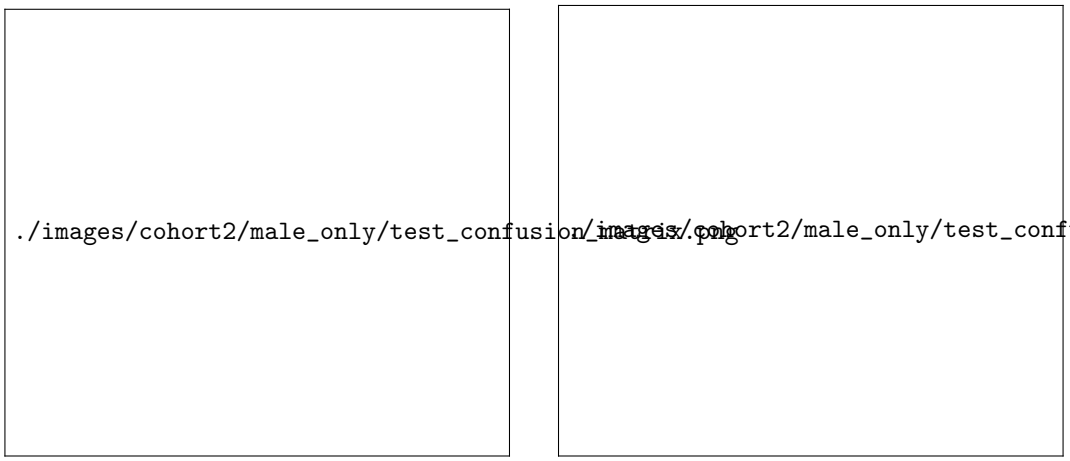
Figure 12: Class-specific probability box plots (Cohort 2). Each box plot shows the range of probability outputs for a given class-specific predictor, separated by class. In this cohort, probabilities were correctly assigned by the current-smoker class-specific predictor, but both the never- and ex-smoker class-specific predictors had significant overlap in the distributions of their probabilities.

3.4.1 Model performance separated by sex

ENSS was trained on Cohort 1, which only included male individuals. However, independent evaluation used Cohort 2, which contains both males and females. This raises the concern of sex being a potential confounding variable. This is especially relevant for smoking, as tobacco use is much higher in males than in females [45, 46]. In order to assess this, evaluation was re-ran in Cohort 2 while separating for sex.

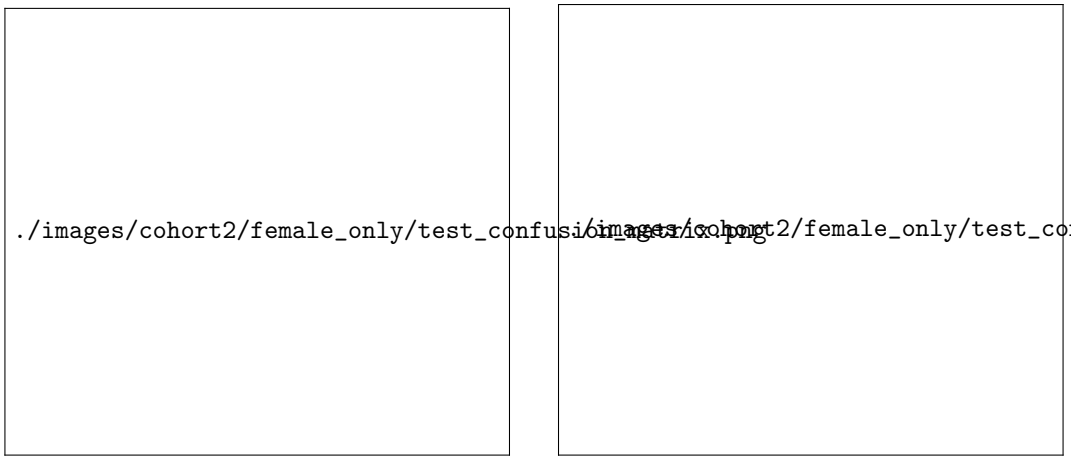
Classification performance on Cohort 2 males can be seen in Figure 13, while classification performance on Cohort 2 females can be seen in Figure 14. Performance for classifying ex-smokers and current-smokers appeared to be consistent, with 78 % of male ex-smokers and 77 % of female ex-smokers correctly classified, and 73 % of male current-smokers and 74 % of female current-smokers correctly classified. However, classification performance for never-smokers varied between males and females. 73 % of female never-smokers were misclassified as ex-smokers, with 23 % correctly classified, while 54 % of male never-smokers were misclassified as ex-smokers, with 45 % correctly classified. While classification of never-smokers was more accurate in males than females, performance was still poor in males, suggesting sex was only a partial confounding variable, with other factors leading to poor performance.

As before, the model's ability to separate classes was also evaluated. Macro-averaged OvO ROC curves for Cohort 2 males and Cohort 2 females can be seen in Figure 15. AUCs for these curves were similar across sexes. ENSS achieved AUCs of: 0.980 on males and 0.964 on females for separating never-smokers from current-smokers, 0.879 on males and 0.840 on females for separating ex-smokers from current-smokers, and 0.772 on males and 0.637 on females for separating never-smokers from ex-smokers. Separation performance appeared to be consistent with the trend seen in classification, with the largest decrease in performance between sexes for never-smokers. As before, separating the cohort by sex only led to a minor increase in performance, suggesting sex was only a partial confounding variable.



(a) Counts (b) Proportions

Figure 13: Confusion matrices (Cohort 2, male only). Despite restricting to the only sex present in the training dataset, never-smokers are still largely misclassified.



(a) Counts (b) Proportions

Figure 14: Confusion matrices (Cohort 2, female only). The model performs worse classifying female never-smokers than male never-smokers.

(a) Male only (b) Female only

Figure 15: Macro-averaged ROC curves (Cohort 2, separated by sex). The model's ability to separate classes of smoking-status is marginally better in males than females.

3.5 Comparison of ENSS to prior research (Cohort 2)

3.5.1 Comparison of CpG sites used in all 3 models

The intersections of the different sites used in DNAmPACKYRS and mCigarette with the never-, ex- and current-smoker class-specific predictors can be seen in Figure 16. There was limited overlap between the CpG sites used in all three models. Only 11 sites were common to all three models. Per class-specific predictor, 11 sites were common to the never-smokers predictor, DNAmPACKYRS and mCigarette, 10 sites were common to the ex-smokers predictor, DNAmPACKYRS and mCigarette, and 10 sites were common to the current-smokers predictor, DNAmPACKYRS and mCigarette. However, only 14 CpG sites were common between DNAmPACKYRS and mCigarette, with 158 and 1241 CpG sites unique to those models, respectively. While there does not appear to be significant overlap between the features used in ENSS and either DNAmPACKYRS and mCigarette, neither was there significant overlap between the features used in DNAmPACKYRS and mCigarette, despite all three models being produced from datasets using the same 449,521 CpG sites. Therefore, the intersection of (or lack thereof) ENSS with prior models appears to be consistent. This could be indicative of redundancy within the methylation signal, where different models can each select several different CpG sites to achieve the same effect.

The effects of regularisation on the model were not consistent with prior research. Comparisons of feature reduction steps can be seen in Table 3. Regularisation in the training of DNAmPACKYRS selected 172 features out of 450,161, while regularisation in the training of mCigarette selected 1255 features out of 187,600. However, the three class-specific predictors of ENSS each selected 1653, 1627 and 1514 features out of 3122, respectively. While the hyperparameter choices for DNAmPACKYRS are not publicly available, mCigarette used a choice of $\alpha = 0.012577$ and $\lambda = 0.5$, whereas ENSS was trained using $\alpha = 0.2$ and $\lambda = 0.85$. Therefore, ENSS was trained using stronger regularisation than mCigarette yet proportionally selected more features, suggesting the model overfitted the training data.

An Elastic Net linear regression model was trained as a benchmark to evaluate this hypothesis. Regularisation in the training of this linear benchmark selected 57 features out of 3122, which is in line with the regularisation seen in DNAmPACKYRS and mCigarette.

Figure 16: Intersections of CpG sites with prior models. While the intersections of the class-specific predictors of ENSS with both DNAmPACKYRS and mCigarette was limited, so was the intersection between DNAmPACKYRS and mCigarette.

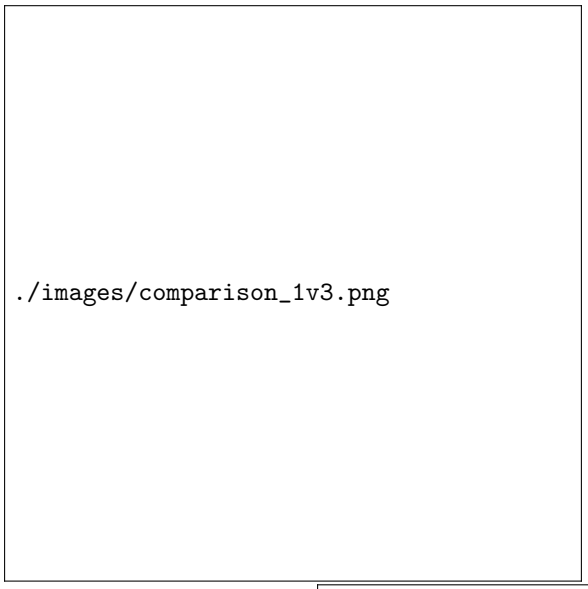
Table 3: Comparison of feature reduction steps

Reduction step	**DNAmPY**	**mCigarette**	**ENSS**			**Linear**
			Never	Ex	Current	
Starting CpG sites	450,161	449,521	449,521	449,521	449,521	449,521
Post statistical test	N/A	187,600	3122	3122	3122	3122
Post ML regularisation	172	1255	1653	1627	1514	57

3.5.2 Evaluating performance with ROC comparisons

To avoid discovery bias associated with the training cohort, model comparisons were only evaluated using Cohort 2. OvO ROC curves comparing Elastic Net Smoking-Status, DNAmPACKYRS and mCigarette can be seen in Figure 17. All three models achieved excellent performance separating never-smokers from current-smokers. Namely, ENSS achieved an ROC-AUC of 0.971, DNAmPACKYRS achieved an ROC-AUC of 0.991 and mCigarette achieved an ROC-AUC of 0.994. Both DNAmPACKYRS and mCigarette achieved excellent performance separating ex-smokers from current-smokers, with ROC-AUCs of 0.915 and 0.923, respectively, while Elastic Net Smoking-Status achieved good performance, with an ROC-AUC of 0.859. For separating never-smokers from ex-smokers, mCigarette achieved good performance with an ROC-AUC of 0.838, DNAmPACKYRS achieved acceptable performance with an ROC-AUC of 0.798, while ENSS achieved poor performance with an ROC-AUC of 0.686. A summary of these AUCs can be seen in Table 4.

Comparing ENSS to the models from prior research shows worse performance in the independent cohort, indicative of overfitting. In order to diagnose this, a linear model (using the same algorithm as DNAmPACKYRS and mCigarette, see Equation 1) was trained to benchmark the utility of the training dataset. OvO ROC curves of this model can be seen in Figure 17. The linear benchmark achieved ROC-AUCs of: 0.990 for separating never-smokers from current-smokers, 0.918 for separating ex-smokers from current-smokers, and 0.773 for separating never-smokers from ex-smokers. This performance was similar to that of DNAmPACKYRS and mCigarette, and outperformed ENSS (see Table 4). Comparing linear to logistic models, both internally and externally derived linear models outperformed the internally derived logistic model. Moreover, the linear benchmark was not tuned or optimised, and used the same hyperparameters as the logistic model. This result suggests that the poor regularisation of ENSS was due to the model itself, rather than the quality of the training dataset.



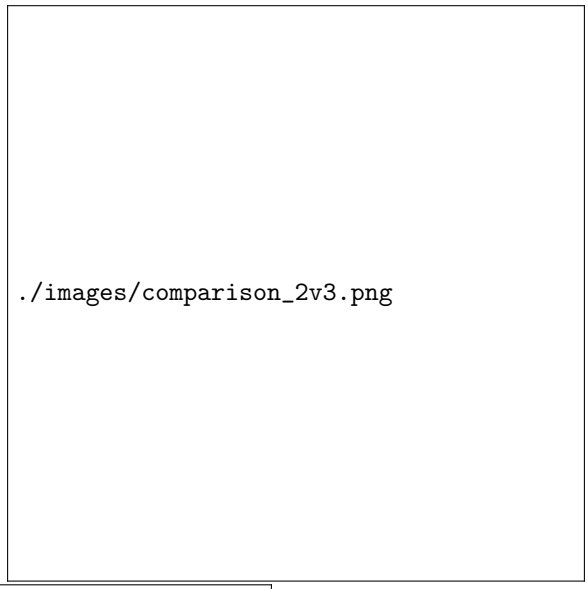


Figure 17: Comparison of ROC curves with scores from prior research. Both internally and externally derived linear models outperformed the internally derived logistic model (ENSS), with greater AUCs across all classes.

Table 4: Summary of model ROC-AUCs in Cohort 2 independent test set

Model	**AUC**		
	Never vs Current	Ex vs Current	Never vs Ex
DNAmPACKYRS	0.911	0.915	0.798
mCigarette	0.994	0.923	0.838
ENSS	0.971	0.859	0.686
Linear benchmark	0.990	0.918	0.773

4 Discussion

While ENSS achieved excellent discrimination in the hold-out testing dataset (ROC-AUCs: 0.977, 0.941, 0.909), performance ranged from excellent to poor in the independent testing dataset (ROC-AUCs: 0.971, 0.859, 0.686). When compared the current gold-standard smoking models DNAmPACKYRS (ROC-AUCs: 0.991, 0.915, 0.798) and mCigarette (ROC-AUCs: 0.994, 0.923, 0.838), we can see ENSS achieved similar performance separating never-smokers from current-smokers, but worse performance separating ex-smokers from current-smokers and never-smokers from ex-smokers. All three of these models were developed using a similar selection of CpG sites (intersection of the 450k and EPIC arrays), and took similar approaches with feature selection, cross validation and training. Therefore, the most likely cause of the discrepancies seen in performance are the differences between the training datasets used, namely the distribution of sex and sample size. While ENSS was produced from a dataset only containing males, the DNAmPACKYRS and mCigarette scores were not, and had both males and females present in their training data (see Table 5). However, separating males from females in the independent testing dataset resulted in only an incremental gain in performance, suggesting sex was not a substantial confounding variable. In terms of sample size, the training dataset for ENSS only contained 848 individuals, compared to 1731 (FHS) for DNAmPACKYRS and 17,865 (GS) for mCigarette (see Table 5). Moreover, ENSS was significantly more complex than DNAmPACKYRS or mCigarette, using more CpG sites (see Figure 16 and Table 3). Producing a simpler model (Elastic Net linear regression) on the same training dataset resulted in performance comparable to DNAmPACKYRS and mCigarette (ROC-AUCs: 0.990, 0.918, 0.773). This suggests that the additional complexity of the logistic regression model is promoting overfitting, which is consistent with machine learning theory [59]. An increased training sample size can reduce overfitting [59], which could explain the increased performance of DNAmPACKYRS and mCigarette in the independent dataset. Similarly, this could potentially resolve overfitting issues of ENSS. The other conventional approach to overfitting is increasing regularisation strength [60]. However, it

Table 5: Training dataset comparisons with prior studies

Measure	**Cohort 1 - Train**	**FHS**	**GS**
	Class balance		
Num. individuals	848	1731	17,865
Never-Smokers	211 (25 %)	710 (41 %)	3277 (18 %)
Ex-Smokers	539 (63 %)	883 (51 %)	9414 (53 %)
Current-Smokers	98 (12 %)	138 (8 %)	5174 (29 %)
	Cohort distribution		
Percentage male	100 %	46 %	59.1 %
Mean age (years)	72	66	47.6

should be noted that this did not work in practice as ENSS was trained with stronger regularisation than mCigarette, yet proportionally selected more features to be used in the final model. One explanation for overfitting could be due to the fact that the sample complexity for logistic regression grows with the number of irrelevant features [61]. This suggests that no amount of hyperparameter tuning would have improved model performance in the independent test dataset, without sufficient sample size in the training dataset.

4.1 Limitations of this work

There were several limitations associated with this work. Firstly, the labels for supervised learning used self-reported data. Self-reported data is prone to inaccuracy (see 1.2), and therefore, potentially introduces bias into a trained model. In training, an incorrectly labelled example could lead to gradient updates which are not related to the true relationship between training data and model output. In evaluation, an incorrectly labelled example could lead to a correct model prediction being identified as erroneous, skewing the perception of the model's accuracy. This limitation is not unique to this work, and would be expected to be present for the DNAmPACKYRS and mCigarette scores as well.

As discussed, the trained logistic regression model (ENSS) overfits the training data. While traditional machine learning evaluation showed excellent performance separating all classes, independent evaluation representative of practical applications showed ENSS overfitting, with class separation ranging from excellent to poor. However, a linear regression model (linear benchmark) trained on the same dataset outperformed the logistic approach. We hypothesised that this was largely due to differences in model complexity, and more training data would be required when using a logistic approach. However, even with this hypothesis verified, this means the required sample size limits the utility of our proposed method. Large sample sizes are not always available, especially for epigenetic data. Therefore, even with overfitting addressed, the method proposed in this work maybe be unsuitable for some applications.

4.2 Future Directions

Recent work has shown promise integrating principal component analysis (PCA) into Elastic Net model training [62]. Using only a single additional step, PC-based Elastic Net showed improvements in reliability and performance, due to reducing the effect of noise present in the dataset. Moreover, using Elastic Net models trained on principal components (rather than CpG sites) has benefits for interpretability and regularisation. While this method has shown improvements for DNAm-based aging clocks, little work has been done to evaluate the benefits for DNAm-based smoking scores. Moreover, improvements associated with regularisation and noise could help reduce the overfitting problem present with the method proposed in this work.

This worked aimed to improve class separation by reframing the problem of modelling lifetime smoking exposure as a classification task rather than a regression task. To achieve this, we adapted the existing gold-standard techniques, which use linear regression, into logistic regression algorithms. However, there are other choices of classification algorithms to explore. Decision trees are one such example that show promise. Decision trees are relatively simple models, and therefore could perform well with smaller training datasets, addressing a limitation of ENSS. Moreover, decision tree training requires iteratively selecting features in training, which could leverage the existing Kruskal-Wallis feature selection used in this method. Additionally, ensemble methods such as random forests or gradient boosting (e.g. XGBoost) provide options for increased accuracy, however, at the cost of interpretability.

5 Footnotes

5.1 Acknowledgements

I would like to thank Professor Richard Green from the Department of Computer Science at the University of Canterbury and Professor Greg Jones from the Department of Surgical Sciences at the University of Otago for supervising this project. Without your support, guidance and resources this project would not have been possible.

5.2 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.

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] Janine Nip et al. *Smoking prevalence and trends: Important findings from the 2023–24 New Zealand Health Survey*. Public Health Expert Briefing, Public Health Communication Centre (PHCC). Ac-

cessed: 2025-06-24. Nov. 2024. URL: https://www.phcc.org.nz/briefing/smoking-prevalence-and-trends-important-findings-202324-new-zealand-health-survey.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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).

[15] Lindsay M Reynolds et al. "DNA methylation of the aryl hydrocarbon receptor repressor associations with cigarette smoking and subclinical atherosclerosis". In: *Circulation: cardiovascular genetics* 8.5 (2015), pp. 707–716.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] E Andres Houseman et al. "DNA methylation in whole blood: uses and challenges". In: *Current environmental health reports* 2 (2015), pp. 145–154.

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

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

[28] 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.

[29] 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.

[30] John Adeoye et al. "Machine learning-based genome-wide salivary DNA methylation analysis for identification of noninvasive biomarkers in oral cancer diagnosis". In: *Cancers* 14.19 (2022), p. 4935.

[31] 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.

[32] 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.

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

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

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

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

[37] 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.

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

[39] 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.

[40] T Remen et al. "Risk of lung cancer in relation to various metrics of smoking history: a case-control study in Montreal". In: *BMC cancer* 18 (2018), pp. 1–12.

[41] W Dana Flanders et al. "Lung cancer mortality in relation to age, duration of smoking, and daily cigarette consumption: results from Cancer Prevention Study II". In: *Cancer research* 63.19 (2003), pp. 6556–6562.

[42] Jay H Lubin and Neil E Caporaso. "Cigarette smoking and lung cancer: modeling total exposure and intensity". In: *Cancer Epidemiology Biomarkers & Prevention* 15.3 (2006), pp. 517–523.

[43] Richard Doll and Richard Peto. "Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers." In: *Journal of Epidemiology & Community Health* 32.4 (1978), pp. 303–313.

[44] 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.

[45] World Health Organization. *Global report on trends in prevalence of tobacco use 2000-2025, fourth edition*. Accessed: 2025-06-22. Geneva, 2021. URL: https://www.who.int/publications/i/item/9789240039322.

[46] Stephen T Higgins et al. "A literature review on prevalence of gender differences and intersections with other vulnerabilities to tobacco use in the United States, 2004–2014". In: *Preventive medicine* 80 (2015), pp. 89–100.

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

[48] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[49] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[50] Julian Lehrer. *transposecsv: Transpose large CSV files in Python*. urlhttps://pypi.org/project/transposecsv/. Version 0.0.5. Accessed: 2025-03-24. 2022.

[51] Jie Cai et al. "Feature selection in machine learning: A new perspective". In: *Neurocomputing* 300 (2018), pp. 70–79.

[52] Avrim L Blum and Pat Langley. "Selection of relevant features and examples in machine learning". In: *Artificial intelligence* 97.1-2 (1997), pp. 245–271.

[53] 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.

[54] Ronald A Fisher. "On the" probable error" of a coefficient of correlation deduced from a small sample". In: *Metron* 1 (1921), pp. 3–32.

[55] Henry B Mann and Donald R Whitney. "On a test of whether one of two random variables is stochastically larger than the other". In: *The annals of mathematical statistics* (1947), pp. 50–60.

[56] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[57] Scikit-learn developers. *Multiclass Receiver Operating Characteristic (ROC)*. https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html. Accessed: 2025-06-02.

[58] William J Youden. "Index for rating diagnostic tests". In: *Cancer* 3.1 (1950), pp. 32–35.

[59] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[60] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[61] Andrew Y Ng. "Feature selection, L_1 vs. L_2 regularization, and rotational invariance". In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 78.

[62] Albert T Higgins-Chen et al. "A computational solution for bolstering reliability of epigenetic clocks: implications for clinical trials and longitudinal tracking". In: *Nature aging* 2.7 (2022), pp. 644–661.