UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Simo Pähk

Prediction of Cell Counts from DNA Methylation

Bachelor's Thesis (9 ECTS)

Supervisor: Ahto Salumets, MSc

Tartu 2022

Prediction of Cell Counts from DNA Methylation

Abstract:
DNA methylation is an epigenetic factor that modulates gene expression. The close relationship between gene expression and cell differentiation serves as a basis for methylation-based cell mixture deconvolution — a method for determining the proportions of constituent cell types in a biological sample. Previous work has demonstrated its usefulness in predicting lymphocyte subtypes in blood samples, but has neglected T_{EMRA}, a type of senescent lymphocyte associated with aging and autoimmune diseases. This thesis sets out to explore the feasibility of estimating the proportions of T cells in various stages of differentiation, including T_{EMRA}, from methylation sequencing data using machine learning. The results show that while prediction accuracy is lower for T_{EMRA} subtypes than for general subtypes such as T cells, it is nonetheless a viable approach for this task, especially since DNA sequencing is cheaper and more scalable than traditional laboratory methods for blood sample analysis.

Keywords:
Methylation, cell mixture deconvolution, TEMRA, machine learning, regression

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Rakkude arvukuse mudeldamine DNA metülatsiooni põhjal

Lühikokkuvõte:
DNA metülatsioon on geenide avaldumist mõjutav epigeneetiline mehhanism. Otsene seos geenieksрessiooni ja raku diferentseerumise vahel on aluseks metülatsioonipõhisele rakutüübi dekonvolutsioonile, mis on meetod rakutüüpide proportsioonide määramiseks bioloogilises proovis. Varasemad uurimused on seda meetodit edukalt rakendanud lümfotsüütide alamtüüpide arvuliseks hindamiseks vereproovis, kuid on suuresti kõrvale jätnud T_{EMRA}, mis on eluea lõpufaasis olev lümfotsüüdi tüüp, mida seostatakse vananemise ja autoimmuunhaigustega. Käesoleva töö eesmärk on välja selgitatada, kas ja kui edukalt on võimalik sekveneeritud metülatsiooniandmeid kasutada T rakkude arvukuse hindamiseks nende erinevates elufaasides, sh T_{EMRA} jaoks, kasutades masinõpet. Töö tulemusel selgub, et kuigi T_{EMRA} alamtüüpide arvukuse hindamine ei ole nii täpne kui T rakkude üldiste alamtüüpide puhul, on töös käsitletud metoodika siiski tulemuslik, seda eriti, kuna DNA sekveneerimine on odavam ja skaleeruvam alternatiiv tavapärastele laboroorsetele meetoditele vereproovide analüüsimiseks.

Võtmesõnad:
Metülatsioon, rakutüübi dekonvolutsioon, TEMRA, masinõpe, regressioon

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Contents

1 Introduction

The trillions of cells we are made up of contain identical DNA, yet differ widely in shape and function. These differences are an example of epigenetic phenomena — heritable traits that involve no alterations in DNA sequences. DNA methylation, the most studied such modification, is an epigenetic marker that regulates gene expression, for example by altering the spatial structure of DNA [84].

The aim of this thesis is to use methylation profiles obtained from whole blood samples to predict the proportions of different immune cells. This is related to a method called cell type deconvolution, pioneered by Houseman et al. [45], which is based on the identification of differentially methylated regions in DNA. The assumption is that if a region is uniquely (un)methylated for some cell type, the region's average methylation obtained from a biological sample should reflect its proportion in it. In this thesis, methylation states of individual sites, instead of whole regions, will be used.

Using data from 187 individuals, separate regression models will be created for predicting 9 different cell proportions, but focusing on a subtype of T lymphocytes called T_{EMRA}. The reason behind this is that there is a growing interest in T cell aging due to their significant impact on overall immune responses, and T_{EMRA} cells are associated with aging of the immune system, albeit with somewhat controversial results [13, 62]. As of now, T_{EMRA} has been largely neglected from deconvolution analyses, and as it is promising to be a cost-efficient alternative to standard lab procedures, this thesis seeks to explore its feasibility.

The thesis is structured as follows. Chapter 2 provides the relevant biological background on genetics and the immune system. DNA methylation is covered in depth, and the effects of aging on the immune system is explored in association to changes in lymphocyte function and composition, including those of T_{EMRA}. This chapter also discusses related work and the novelty of this thesis in relation to it. Chapter 3 is devoted to machine learning and covers the different regression methods used for building predictive models, along with metrics for evaluating their performance. Chapter 4 details the data, its processing, and its usage in model selection. The results are presented in chapter 5, together with discussion and comparison to previous work.

2 Background and related work

To properly understand the origin and function of T_{EMRA}, and how DNA methylation relates to these cells, a primer in genetics and the immune system is needed. The following sections focus precisely on that.

2.1 Genetics

The following overview is based on the book by Strachan and Read [83, 85]. DNA is the carrier of heritable information in almost all living organisms. It is a sequence of nucleotides — molecules made up of a deoxyribose sugar, a phosphate group and a nitrogenous base. The four nucleotides found in DNA, defined by their respective nitrogenous base, are adenine (A), thymine (T), guanine (G) and cytosine (C). Nucleotides are attached to each other by the 3' and 5' carbons of the deoxyribose sugar via a phosphodiester bond, giving DNA molecule a direction. All biological mechanisms read these sequences in a 5' to 3' direction.

The most stable form of DNA is the double-stranded helix, in which two complementary DNA strands are bound together by hydrogen bonds. Complementary here means two things: (i) that the two chains run in opposite directions; and (ii) hydrogen bonds only form between A — T and C — G (Figure 1).

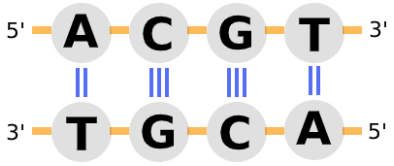

Figure 1. Structure of double-stranded DNA. Both strands, supported by the phosphate backbone (orange lines), are bound to each other by hydrogen bonds (blue lines).

In cell nuclei, DNA is wound around histones, forming nucleosomes. Nucleosomes are the building blocks of chromosomes, whose function is to compactisize the otherwise long DNA chains. This packaging has a direct effect on gene expression by controlling the availability of DNA to various molecular mechanisms.

A gene is a nucleotide sequence that serves as a template for making a functionally important RNA molecule. Genes are expressed when RNA polymerase *transcribes* a complementary RNA molecule from the DNA template, replacing thymine (T) with uracil (U). Some transcripts serve functions of their own, such as regulating gene expression, while others — coding RNA[1] — are used as templates in protein synthesis. *Translation* of RNA to a protein is based on the genetic code, which maps triples of nucleotides

[1] Also known as messenger RNA or mRNA.

to a chain of amino acids. Only some parts of a gene — exons — contribute to the growing peptide chain, the rest having either a regulatory role (e.g. the promoter) or being discarded (e.g. introns). This flow of genetic information from DNA to RNA to protein is the central dogma of molecular biology (Figure 2).

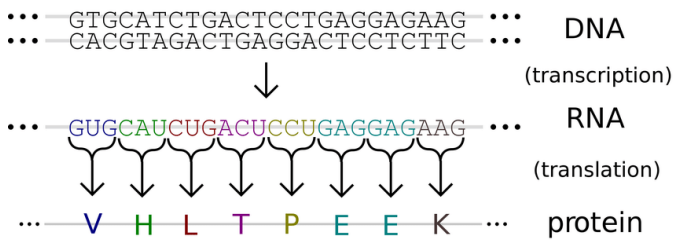

Figure 2. The central dogma of molecular biology establishes the unidirectional relationship between DNA, RNA and proteins [36].

The size of the human genome is estimated at 3.1Gbp[2] (giga base pairs), with only ~1% of it belonging to the ~20000 protein-coding sequences [46]. These sequences are flanked by non-coding repetitive and transposable elements, which make up the overwhelming majority of DNA [46].

2.2 Epigenetics

Epigenetics is the study of heritable information that occurs without accompanying changes in the genomic DNA sequence. Epigenetic marks, represented by covalent chemical modifications to the DNA or histones, affect gene expression by altering the structure of nucleosomes, or by modifying the binding of certain transcription factors [84].

2.2.1 DNA methylation

The most studied epigenetic marking is DNA methylation. In mammalian genomes, the primary targets of methylation are cytosines in the symmetrical CpG[3] context [84]. Cytosine is methylated by the covalent addition of a methyl (CH_3) group to the fifth carbon of its nucleobase ring, resulting in 5-methylcytosine (5mC). Initial methylation patterns are set up in early development, but are subject to changes throughout life and are maintained between subsequent cell divisions by enzymes called DNA methyltransferases [84].

[2]Of the 23+1 unique chromosomes — humans have 46 chromosomes, and thus about twice the amount of base pairs in their DNA.

[3]CpG denotes adjacent C and G nucleotides connected via a phosphodiester bond, which is symmetric with its complement on the opposite strand.

Methylation is not widespread in the human genome. The frequency of CpG dinucleotides, where methylation usually occurs, is less than 1% — much less than the statistically expected frequency of 4% [46]. A likely reason for this is that 5mC readily converts into thymine by spontaneous deamination[4], resulting in a loss of CpGs over the course of evolution [84]. An estimated 60-80% of CpGs in the human genome are methylated, and reside primarily in regions with low CpG density [81]. This contrasts to CpG islands (CGI) — CpG-rich regions about 1000 base pairs in length — which house less than 10% of all CpG sites and are mostly unmethylated [81].

2.2.2 Functions of methylation

The effect of methylation depends on the location of CpGs in the genome (Figure 3). About 60% of genes have CGIs at their promoters, where lack of methylation is associated with nucleosome depletion [47]. These regions are not bound by histones and are often occupied by transcription factors, thus enabling gene expression [52]. Conversely, methylation in promoter CGIs is associated with long-term suppression of gene expression, e.g. in X-chromosome inactivation [92] and gene imprinting [6], both of which are responsible for parent-specific gene expression.

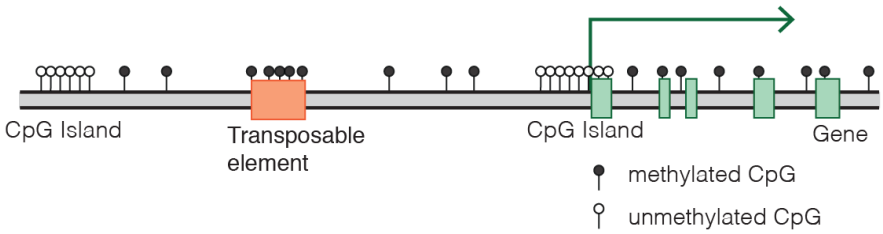

Figure 3. Typical mammalian DNA methylation landscape [19]. Unmethylated CpG islands enhance transcription, which is why they're often found at the promoters of actively expressed genes. The opposite applies to harmful sequences, such as transposable elements, for which methylation ensures their silencing. CpGs in gene bodies, while rare, are often methylated and therefore prone to spotaneous deamination, i.e. point mutations.

CpG sites in gene bodies, although not widespread, are extensively methylated [47]. Contrarily to promoters, methylation in gene bodies is usually associated with higher gene expression [47]. Gene bodies may also contain CGIs, but methylation there is tissue-specific and its function has not been clearly established [47]. Due to spontaneous deamination of 5mC, CpG-rich gene bodies are prone to mutations. An estimated 30% of point mutations that cause genetic diseases originate from methylated CpG sites, and are a major cause of cancer [78].

[4]The removal of an amine (NH_2) group from the nucleobase ring.

Methylation also occurs in non-coding regions of the genome, where it serves to repress the expression of potentially harmful genetic elements [96]. This includes transposable elements (TE) — parasitic sequences capable of self-replication [96] — which make up nearly 45% [46] of mammalian genomes. Methylation ensures that these sequences are stably silenced across cell divisions [96].

2.2.3 Environmental factors

Because methylation patterns are herited during mitosis, DNA methylation acts as a persistent interface between the genome and the environment. For example, phenotypic differences between monozygotic twins can be explained by differences in methylation patterns acquired during lifetime [31]. A study on mice [65] revealed that differences in maternal care during infancy affected methylation patterns in brain tissue, which persisted into adulthood. Similar relationship has been established for prenatal exposure to famine [41] and chemicals [42] in humans.

Aging has been described as an epigenetic phenomenon, characterized by changes in methylation patterns associated with genes involved in basic cellular functions such as DNA repair [3]. This theory is further supported by methylation-based age prediction models [44], whose estimates not only correlate with chonological age, but which have been proven to be powerful estimators of biological age. Acceleration of epigenetic (biological) age compared to chronological age is a strong indicator of disease [55], and has been found to be a predictor of all-cause mortality in later life [60].

2.3 Immune system

The human immune system is a complex network of cells, tissues and organs. It consists of innate mechanisms and adaptive mechanisms, which complement each other to provide defence against a wide array of pathogens.

2.3.1 Immune cells

White blood cells (WBCs), also known as leukocytes, constitute the principal part of the immune system. Leukocytes originate from common stem cells in the bone marrow, and use cell-surface receptors for recognizing molecular patterns inherent to specific pathogens. [15]

Lymphocytes are leukocytes which are responsible for aspects of the adaptive immunity. The genes of their receptors undergo somatic recombination, a process of DNA fragment shuffling, which yields a pool of lymphocytes with very diverse receptors. The main types of lymphocytes are T and B cells, and the molecules their receptors recognize are called antigens. T cells, which mature in the thymus, include helper T cells, which coordinate immune responses via signaling molecules — cytokines and

lymphokines — and cytotoxic T cells, which directly neutralize pathogens. Among B cells are plasma cells, which produce antigen-specific antibodies for tagging and neutralizing pathogens. [15]

Innate immune cells perform diverse functions from blood clotting (thrombocytes) to microbe ingestion (macrophages), and are widely present in the tissues lining the entrances to our bodies (mast cells). Their receptors recognize pathogen-associated molecular patterns, which, unlike antigens, are common to a wider range of pathogens. This makes innate immune cells less specific, but a good first line of defence. [15]

Leukocytes can be identified by the various proteins they express on cell surface, which define their *phenotype profile*. More than 500 such markers have been identified and assigned a cluster of differentiation (CD) number [38]. Table 1 provides an excerpt of some markers relevant to this thesis.

Table 1. Some surface markers and their associated T lymhpocyte subtypes [38].

Marker	**Identifies**
CD3d,e,g	All T cells
CD45RA	Naive T cells
CD4	Helper T cells
CD8a,b	Cytotoxic T cells

Surface markers are used in flow cytometry for sorting cells by their type [64]. Cytometry can also be used for determining the proportions of cell types in a biological sample — a common diagnostic procedure, since changes in lymphocyte composition may be a sign of disease or other pathological processes [64]. For example, CD4 lymphocytes are monitored in HIV patients for determining when to start antriretroviral therapy [4]. However, cytometry might not be suitable for large-scale use due to its cost, which has created an incentive to find alternatives, among them the DNA methylation based method explored in this thesis.

2.3.2 Immune response

The immune response is initiated by innate immune cells upon recognizing a foreign pathogen. Toll-like receptors on mast cells, for example, react to viral nucleic acids and bacterial products, while macrophages recognize microbal proteins expressed on the surface of infected host cells. Once these leukocytes are activated, they induce inflammation and secrete cytokines to stimulate other immune cells to flood the site of infection. [66]

The innate immune system interacts with the adaptive immune system via antigen presentation. Phagocytic cells, which process microbial material into antigen, travel to lymph nodes, where they present the antigen to naive T and B cells. The immense

diversity of lymphocyte receptors ensures the activation of at least some T and B cells with receptors matching the antigen. [66]

Once activated, naive lymphocytes undergo clonal expansion — a process of cell division and differentiation into effector and memory types (Figure 4) [1]. Effector CD8 T cells "show a major cytotoxic activity against cells infected with intracellular microbes[5] and against tumor cells," [15] while effector CD4 cells respond well to extracellular bacteria and direct the inflammatory response [15]. The contraction that follows the acute phase reduces the clonally expanded set into a stable population of memory T cells, which become dormant in the body and form the basis of immunological memory [1].

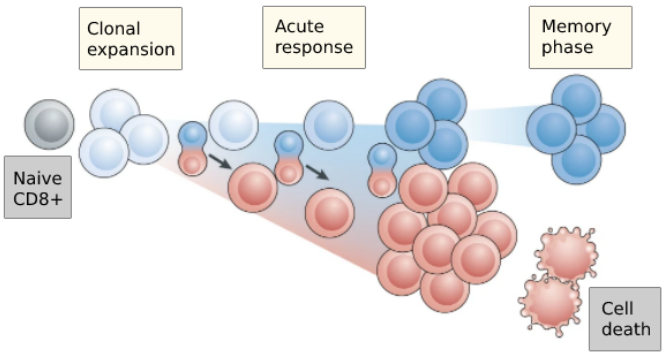

Figure 4. The composite model of differentiation for CD8 T lymphocytes, which includes aspects from several competing differentiation models (image adapted from source) [1]. After the initial expansion phase, cells might become biased to end up as either effector (red) or memory (blue) cells, but retain their flexibility to both fates. The acute phase is followed by contraction, during which most effector cells reach terminal differentiation and die off, leaving behind a smaller population of long-term memory cells.

Memory T cells include central memory (CM) cells, which reside in the lymphoid organs and are thought to be effective against chronic infections, and effector memory (EM) cells, which are more active in the peripheral tissues [61]. Compared to naive types, memory T cells have a lower threshold of activation, and provide a faster and more robust response during infection recurrence [61]. Their long-term nature is also the reason behind the effectiveness of vaccines [21].

2.3.3 Effects of aging

Aging is accompanied by a decline in immune competence, termed immunosenescence, which is characterized by an increased susceptibility to infections, autoimmune disorders, and chronic inflammatory diseases [34]. Poor vaccine efficacy leads to complications

[5] Viruses, for example.

in common viral infections, such as influenza, and is a frequent cause of death in elderly individuals [73]. Immunosenescence affects the immune system as a whole, but is especially pronounced in the adaptive immunity. The mechanisms underlying this deterioration include involution of the thymus — already underway at the first year of life — which inhibits the maturation of naive T cells [69]. Aging is also associated with a decrease in haematopoietic tissue in the bone marrow, which affects the production of lymphocyte progenitors [74].

Imunosenesence not only affects lymphocyte production but also their function. Changes in T cell populations are very apparent, as the overall diversity of T cells decreases and becomes increasingly skewed towards previously encountered antigens, notably those implicated in chronic infections, e.g. cytomegalovirus [48]. Immune cells are also affected by the common hallmarks of aging, such as telomere attrition, genomic instability and irregularities in protein synthesis [59].

Infections, especially chronic, cause large amounts of memory T cells to accumulate in the body [48]. While most of these are reactivated in case of a recurrent infection [16], this is not the case with a subtype called effector memory T cells re-expressing CD45RA (EMRA). T_{EMRA} have reduced capacity to activate in response to cytokine stimulation, are associated with cell death, and exhibit a low proliferation rate, making them largely dysfunctional in comparison to other memory cells [35]. Their accumulation has been identified in transplant recipients, being a cause of inflammation [88] and late graft dysfunction [95]. While chronological age strongly correlates with the amount of T_{EMRA} cells [34], research has been inconclusive regarding their role in immunosenescence. High levels of T_{EMRA} have been implicated in reduced immune competence [13], but also correlate with high life expectancy [62].

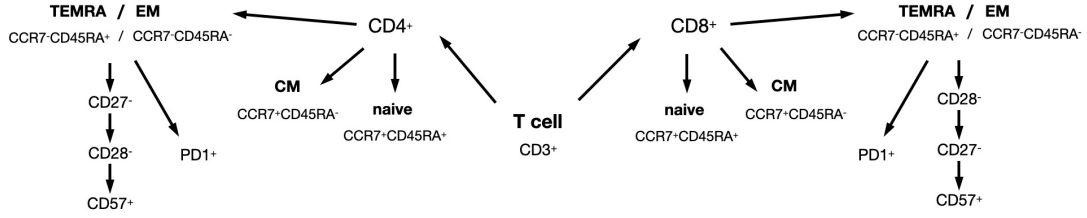

Figure 5. The currently accepted phenotyping model for distinguishing subtypes of T cells in various stages of differentiation [33, 34, 51, 53]. Starting from $CD3^+$, which identifies all T cells, each subtype follows from its parent by the addition (or removal) of some surface marker(s). For example, naive cytotoxic T cells have the phenotype profile $CD3^+$ $CD8^+$ $CCR7^+$ $CD45RA^+$, which differ from cytotoxic T_{EMRA} by the expression of CCR7, a marker associated with lymph node homing [33].

Figure 5 shows the differences in surface markers between naive and memory T

cells, including T_{EMRA}. Further subtypes of T_{EMRA} can be identified by the (lack of) expression of markers associated with senescence and exhaustion. Specifically, the flow cytometry data provided to the author of the thesis includes the proportions of T_{EMRA} in various stages of senescence ($CD28^-$, $CD28^-$ $CD27^-$ and $CD28^-$ $CD28^-$ $CD57^+$) and exhaustion ($PD1^+$). While the loss of CD27 and CD28 "can be considered to be indicative of impaired telomere function in T cells and denotes progression towards replicative senescence" [51], truly senescent cells also express CD57, which inhibits proliferation [51]. PD1, an important immunoregulator, is responsible for limiting excessive T cell activation, but is a sign of dysfunction in case of prolonged expression and can be considered a marker of cell exhaustion [53].

2.4 Related work

Recent years have witnessed a surge in the use of methylome data in bioinformatics. It has been successfully employed in clinical research for cancer classification [80], disease prognosis [91] and drug development [54]. Relevant to this thesis are the various age prediction models based on DNA methylation data, such as those by Hong et al. [43] and Horvath [44]. The former identified just 7 CpG sites in DNA extracted from saliva, based on which a predictive model was created that exhibited 0.89 correlation between predicted and actual age, while the latter, based on 353 CpGs, yielded accurate estimates for samples from many different tissue types. These models have proved to be useful not only in forensics, but have provided insight into processes underlying aging [93]. A recent study even used methylation-based biological age estimation for developing a protocol for age reversal [23].

Work relating specifically to this thesis has been attempted before, and is known in literature as cell-type deconvolution (CTD). The idea of using differentially methylated regions for CTD was first investigated by Baron et al. [5], and was put into practical use by Houseman et al. [45]. While different approaches have emerged, such as the unsupervised CTD method by Zou et al. [97], the method of Houseman et al. remains in wide use to this day. CTD specifically for immune cells has been done before, but mostly focusing on the more well-known lymphocytes, such as $CD8^+$ and granulocytes [20, 63]. A notable exception to this is the work by Bergstedt et al. [8], who built regularized regression models for predicting proportions of a wide range of immune cells, including T_{EMRA}. The results of their work will be referred to later in the analysis of the results of this thesis.

3 Machine learning

The focus of this thesis is to build and evaluate a number of regression models for estimating cell type proportions from DNA methylation. In machine learning, this is a task of supervised learning — learning a model from labelled data. The following sections detail some of these models and their evaluation strategies.

3.1 Linear models

A linear regression model relating the output (target) variable y to the input (feature) vector $\mathbf{x} = (x_1, ..., x_k)$ has the form

$$f_\beta : y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon,$$

where the coefficients β_i are the parameters of the model. The term ϵ is a random variable called the noise term, which accounts for errors between the data and the model and is assumed to have a mean of zero. The term linear only applies to the coefficients — a model with the term $\beta_1 x_1^2$ is also considered linear. [57]

Fitting a linear model to a set of examples $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ amounts to finding the coefficients that produce the least error. A common approach is the (ordinary) least squares method, in which model parameters are estimated by minimizing the sum of squared residuals:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - f_\beta(\mathbf{x}_i))^2.$$

When ϵ is assumed to be normally distributed, a closed-form solution exists for $\hat{\beta}$, which amounts to a fairly simple algorithm based on matrix operations. [57]

3.1.1 Regularization

The advantage of linear models is that they express the relationship between inputs and outputs in a straight-forward way. However, for small datasets (small n) or a lot of features (large k), they are prone to overfitting [57]. When this happens, the model picks up the noise in training data and might not generalize on new, yet unseen samples.

Regularization is a method for reducing overfitting by applying additional constraints to the cost function [26]. Cost functions with a regularization term have the form

$$\sum_{i=1}^{n} (y_i - f_\beta(\mathbf{x}_i))^2 + \gamma R(\beta),$$

where R is a regularization function and γ controls its strength [57]. Penalizing the parameters serves to reduce the complexity of the model and increases its generalization ability [57].

A common regularization approach is the ridge regression, for which $R(\beta)$ is the squared L_2 norm of the parameter vector:

$$\sum_{i=1}^{k} \beta_i^2.$$

Similarly to ordinary least squares, ridge-regularized least squares has a closed-form solution and is therefore computationally inexpensive. In another method called lasso[6] regression, regularization is based on the L_1 norm instead:

$$\sum_{i=1}^{k} |\beta_i|.$$

While seemingly similar, both methods yield different results with increasing values of γ. In ridge regression, the parameter values $\beta_1, ..., \beta_k$ are pushed towards zero, while lasso favours sparse solutions in which some β_i are set to exactly 0. In essence, lasso can act as a feature selector by selectively excluding input variables from the model. [57]

3.2 Tree models

Tree models are one of the most versatile models in machine learning. They are best understood in the context of decision trees, which partition the input space into non-overlapping regions by a set of rules (Figure 6). These rules are applied to an input in the internal nodes of the tree, which collectively decide the path that is taken to reach a leaf node. Each leaf node is associated with a value that is then taken as the input's corresponding prediction. [29]

Decision trees are grown (learned) by recursively splitting existing nodes into new subtrees. The objective is to end up with leaves that are maximally pure, i.e. containing training samples with very similar outputs. As each leaf is later associated with a single prediction value, purity ensures that this generalization has the smallest error. For regression trees, a good impurity[7] measure is variance:

$$Imp(Y) = \frac{1}{|Y|} \sum_{y \in Y} (y - \bar{y})^2,$$

where Y is the set of numeric outputs of the subtree, and $\bar{y}$ is the arithmetic mean of Y. When considering between multiple sets of splits, the one with the lowest weighted average impurity should be chosen. Therefore, the cost function has the form

[6] An acronym of Least Absolute Shrinkage and Selection Operator.

[7] Unlike purity, impurity is directly translatable into a cost function.

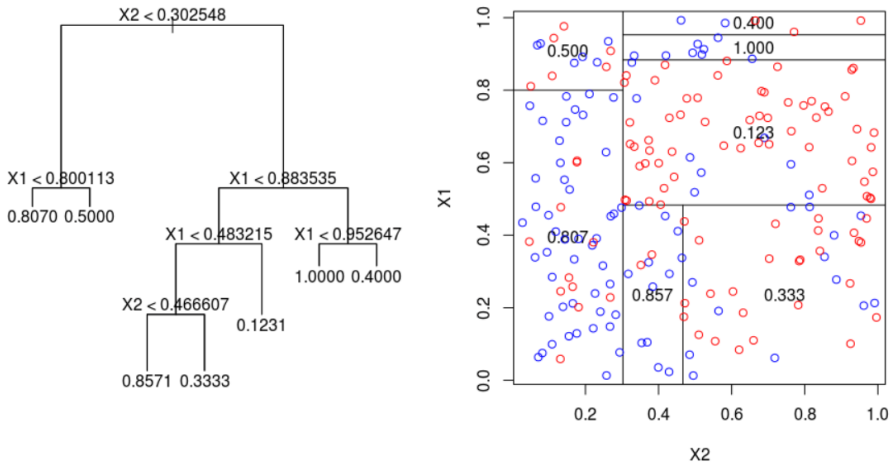

Figure 6. An example of a decision tree model for regression [76]. Internal nodes (left) denote decision boundaries based on the values of features X_1 and X_2, which recursively split the input space into non-overlapping regions (right).

$$Imp(\{Y_1, ..., Y_l\}) = \sum_{i=1}^{l} \frac{|Y_j|}{|Y|} Imp(Y_i),$$

where $Y = Y_1 \cup ... \cup Y_l$. [29]

A popular tree learning algorithm for continuous data is CART (classification and regression tree), also used by scikit-learn's various tree models [71]. The splits created by CART are binary and univariate, resulting in decision rules of the form $x_i \leqslant C$ [58]. The variable x_i and the corresponding cutoff value C are both optimized to yield the biggest decrease in impurity [58]. Since each node is associated with a single variable, the decrease in impurity can also be used to deduce feature importances, which is exploited by feature selection algorithms such as Boruta [50].

Just like simple linear models, decision trees can easily overfit the training data. For tree models, a common approach to reducing this dependency is pruning. Pre-pruning refers to conditions placed on the depth of the tree (*max-depth*), or the amount of training examples each leaf must represent (*min-samples-leaf*) [12]. Post-pruning is more intricate and involves a separate pruning dataset, which is used to remove subtrees where splitting provides no decrease in impurity [12]. In scikit-learn, pre-pruning hyperparameters[8] can be provided to the model prior to training [71].

[8]Hyperparameters are provided to the model prior to training, while parameters are the values that the model fits the training data to [75].

3.2.1 Random forest

A single decision tree might not be accurate enough to warrant its use in practice. An interesting result from the 18th century, called the Condorcet Jury Theorem, provides the basis to the idea of averaging the predictions of many independent classifiers to increase the overall prediction accuracy. According to the theorem, when individual voters have a probability larger than 0.5 of making the correct decision, the prediction accuracy when voted by the individuals as a group can only increase. This principle underlies a machine learning approach called ensemble learning. [79]

Random forest is an ensemble method based on decision trees. To ensure the diversity of individual classifiers, each decision tree is learned on a subset of the underlying training data. In a commonly used method called bagging, short for bootstrap aggregating, this subset is obtained by random sampling with replacement. The probability of a particular data point not being included in the bootstrap sample of size n is $(1 - 1/n)^n$, which has a limit of $1/e$ for $n \to \infty$. Therefore, bootstrap samples leave out about a third of the data points, providing a simple means of introducing diversity into the ensemble. [28]

3.3 Feature selection

Reducing the amount of features prior to modeling has many benefits. According to Guyon and Elisseeff [40], it facilitates the interpretation of data, reduces the training time, and enables the discovery of features relevant to the problem. Furthermore, it helps to reduce overfitting, especially if the number of features compared to the number of samples is large. Nilsson et al. [68] divide feature selection algorithms into two categories: (i) those that attempt to find the minimal feature set optimal for prediction (the *minimal-optimal* problem); and (ii) those that are set to identify all features relevant to the target variable (the *all-relevant* problem).

The method of choice of this thesis is the Boruta feature selection algorithm. According to Kursa and Rudnicki [50], Boruta is a wrapper around the random forest algorithm, and solves the all-relevant problem in feature selection. Although univariate random forest models are capable of ranking features by their decrease in impurity, these values shouldn't be used for this task as-is because they do not account for random variation [50]. Boruta solves this problem by the introduction of shadow variables — shuffled copies of feature vectors — whose importance measures are compared to those of the actual features. In an iterative algorithm, a new tree is built in each step with a new set of shadow features. Features less important than the best shadow feature are discarded in each loop cycle, which runs until only important features remain.

3.4 Model evaluation

In the presence of countless different machine learning methods, how to pick the one most suitable for the problem at hand? Practical machine learning is based on machine learning experiments, in which different models, possibly with different sets of hyperparameters, are trained on the same example data [27]. Models can be then compared to each other quantitatively by various evaluation metrics, e.g. on prediction accuracy or convergence speed.

To avoid bias, evaluation is best performed on data different from the one used in training. This can be achieved by splitting data into two sets — train and test. After a model is fit onto the train set, it is evaluated on the test set by comparing the model's predictions to the actual values [27]. Various metrics exists that aggregate these comparisons into a single value.

For regression models, these metrics are variations of the prediction error, defined for each sample i as $e_i = y_i - \hat{y}_i$ (where y_i is the actual value and $\hat{y}_i$ is the prediction). The most widely used are the mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE) [11]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|,$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2} = \sqrt{MSE}.$$

There are conflicting opinions on which one is a more appropriate measure [14]. It is known that for any distribution, using the expected median for prediction minimizes MAE, while using the expected mean minimizes (R)MSE — thus for asymmetric distributions, MAE might introduce bias as it does not properly account for skew [49]. Absolute values are also mathematically harder to work with than squares, which makes

(R)MSE more attractive for theoretical tasks. On the other hand, MAE expresses the average error and is therefore easily interpretable, in addition to being more robust to outliers [49].

MSE and RMSE yield the same results when used for model evaluation, but MSE might be preferred in this case since it requires one less operation. For other purposes, RMSE is a better choice due to having the same dimension as the underlying data, which facilitates interpretation [11].

Models can also be evaluated visually by plotting predictions against actual values. Accuracy can then be assessed by inspecting the spread of points $(\hat{y}_i, y_i)$ around the line $y = \hat{y}$, or quantitatively with (sample) Pearson correlation coefficient [9]:

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}}.$$

If the model's predictions are accurate, these points are spread evenly around the diagonal and exhibit high correlation, i.e. have r close to 1.

3.4.1 Cross-validation

Setting aside a single test set might not allow for accurate evaluation, since it is not representative of data in the wild. Even more, when data is scarce, the evaluation metrics will likely depend on how the data was divided into train and test sets. To overcome this, model evaluation can be performed with cross-validation (CV). [27]

A common cross-validation scheme is k-fold CV, in which the original data is, usually randomly, split into k equal-sized non-overlapping samples [27]. In each iteration, one of the folds becomes a test (validation) set, while the rest are used for training. This scheme ensures that each sample in the original data gets to be a test sample exactly once.

Evaluation metrics (e.g. MSE) are calculated on the test set of each split and averaged over all iterations. This aggregate score does not measure the performance of any single model but that of the model selection process — if the CV results are satisfactory, the same model can be retrained on all data for further use in practice [75].

In nested cross-validation, each training set itself becomes the basis of cross-validation (Figure 7) [75]. The inner CV loop can then be used for hyperparameter optimization, i.e. choosing the best hyperparameters to use for evaluation in the outer loop.

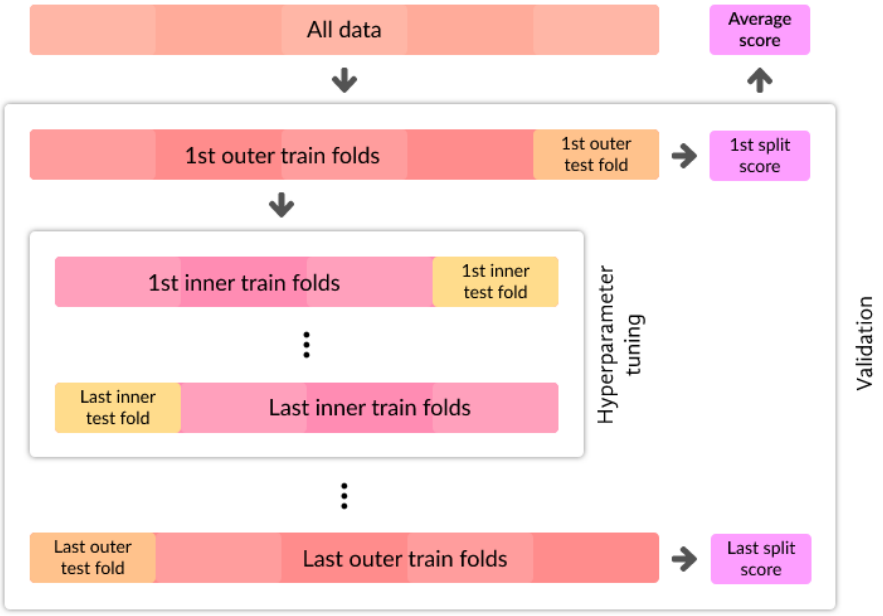

Figure 7. Nested cross-validation. The inner splits are used for determining the optimal hyperparameters to use for training a model on the outer split's training set, which is then evaluated against the corresponding test set.

4 Methods

This chapter covers the steps from processing raw data to evaluating different regression models for predicting lymphocyte counts. Most scripts, intermediate files, generated images and reports that the following sections are based on are available in a public GitHub repository[9].

4.1 Data overview

The work at hand is based on data collected from 183 subjects, recruited mostly from Tartu University Hospital, aged 5–96 but focusing on the elderly (>65 years). Subjects are identified by codes prefixed with IGA, IGN and EDU, dividing them into three subgroups — IGA are patients from the Internal Medicine Clinic, IGN from the Dermatology Clinic, and EDU are non-hospitalized volunteers. Overview of the subjects' age and sex can be seen in Figure 8.

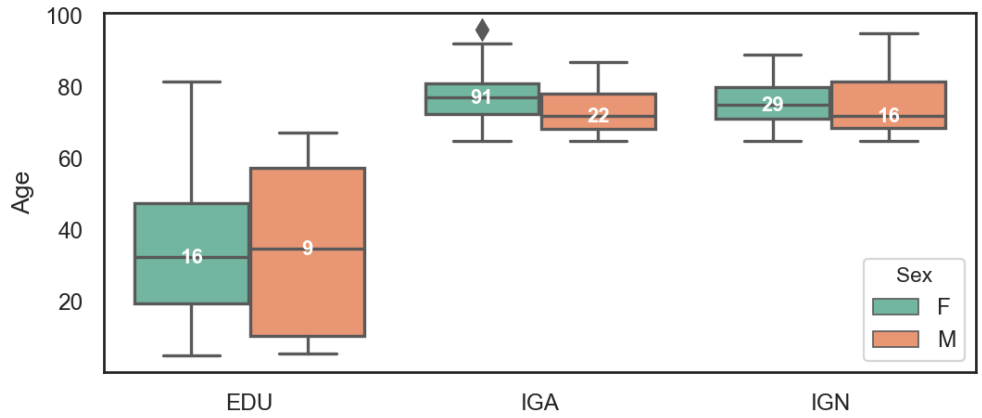

Figure 8. Boxplots of age across the three subject groups, color-coded by sex, with numbers denoting the size of each subgroup. IGA and IGN consist solely of elderly individuals (>65). EDU, on the other hand, is age-wise more diverse (5–81), but represents only 14% of the data. Women make up the majority of subjects at 74%.

Methylation profiles were obtained from bisulfite-sequenced DNA extracted from the subjects' blood samples. Prior to sequencing, DNA was PCR-amplified with primers targeting methylation sites of interest, which were determined by collaborators based on their previous studies. Lymphocyte subtype composition of the blood samples was measured with FACS[10] and expressed as proportions among other subtypes, however

[9] `https://github.com/riimeik/bsc-thesis`.

[10] Short for Fluorescence Activated Cell Sorting, a flow cytometry technology [90].

only the proportions among leukocytes were considered. IGA and IGN had more FACS data available, thus some models are based on data from only these two subject groups.

4.2 Bisulfite sequencing data

The exact nucleotide sequences in DNA can be determined with DNA sequencing. The first step in this process is PCR-amplification (Figure 9) of raw DNA with pre-selected primers, which creates millions of copies of targeted genomic regions [82]. These copies — amplicons — are then ligated with short nucleotide sequences called adapters, necessary for their attachment onto the sequencer's substrate [2]. Sequencers read these amplicons for a specified length and store the results in a FASTQ file (Figure 10), which includes not only the sequences but also a quality estimate for each base recall [18].

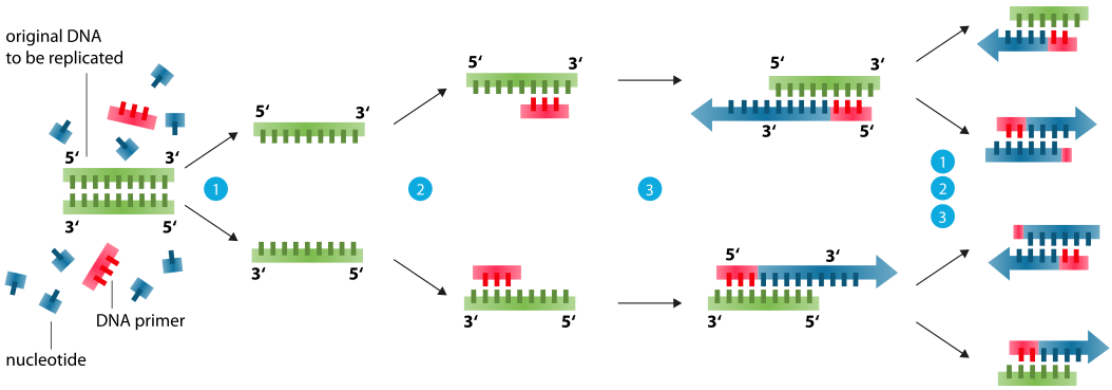

Figure 9. Polymerase chain reaction (PCR) [72]. The book by Strachan and Read [82] explains the steps as follows. A mixture of DNA, primers, nucleotides and DNA polymerases is heated to about 94 °C, which breaks hydrogen bonds between the two strands (1). After cooling, primers attach to strands at complementary sequences, creating replication initiation sites (2). The mixture is then heated to ~72 °C to allow DNA polymerases to synthesize new complementary strands (3). In subsequent cycles, newly synthesized strands themselves become templates of replication. After 30 or so cycles, the regions selected by primers will be the predominant product in the solution.

```
1  @M01338:111:000000000-B53M3:1:1101:12590:1999 1:N:0:1
2  AGGAAGGTTTATGTGTTGGAGGAAGTATGTTTGAAGAGTAGTAGGTTTTATAGAGTTTGTTTTTTAATATT
3  +
4  3>33AFAFFFFFGGGGGGGEFFEHHHFHHHHHHFHHFGFHHGGHHGHHGHHHHHHFGHHGHHHGGHHHHHH
```

Figure 10. An excerpt from FASTQ file. First line is the amplicon's identifier, second its nucleotide sequence, and fourth the estimated Phred quality of each base recall (encoded in ASCII with an offset of 33) [18].

Sequencing does not distinguish between methylated and unmethylated cytosines, and thus cannot be directly used for analyzing DNA methylation. A workaround for this involves the prior treatment of DNA with sodium bisulfite ($NaHSO_3$), which converts unmethylated cytosines to uracils with no effect on methylated cytosines (Figure 11) [32]. Uracil, a nucleotide not found in DNA, is then converted to thymine in the subsequent PCR amplification step. After sequencing the converted DNA and comparing it to a reference genome, methylation states of individual sites can be inferred from C — T (or G — A) mismatches [10].

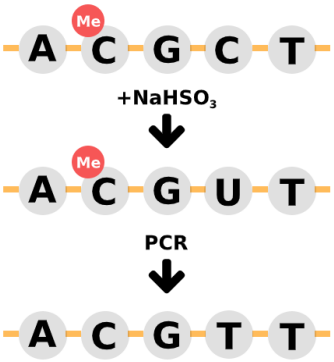

Figure 11. Preparation of DNA in bisulfite sequencing. Unmethylated cytosine reacts with bisulfite to form uracil, which is then corrected to thymine during PCR.

The above-mentioned bisulfite sequencing protocol was applied to the subjects' blood samples and sequenced with a paired-end sequencer. In paired-end sequencing, each amplicon is read in both directions, which later helps to resolve ambiguities when aligning the reads onto a reference genome [2]. These reads are stored in separate FASTQ files, which in this case totalled 520 from 260 sequencing experiments. The discrepancy between the number of subjects (183) and the number of sequencing experiments (260) comes from the fact that several subjects' DNA was sequenced multiple times.

4.3 Trimming and quality control

Individual quality control reports per FASTQ file were generated with FastQC [24] and aggregated into a summary report with MultiQC [22]. The initial summary report indicated widespread adapter contamination which, however, is expected of amplicons with insert[11] size less than the read length of the sequencer (250 in this case). Trim Galore [89] was chosen for data cleansing and consisted of the following procedures: (i) trimming of low quality (<20 Phred score) tails; (ii) adapter trimming from both ends; and (iii) removal of short (<20bp) reads.

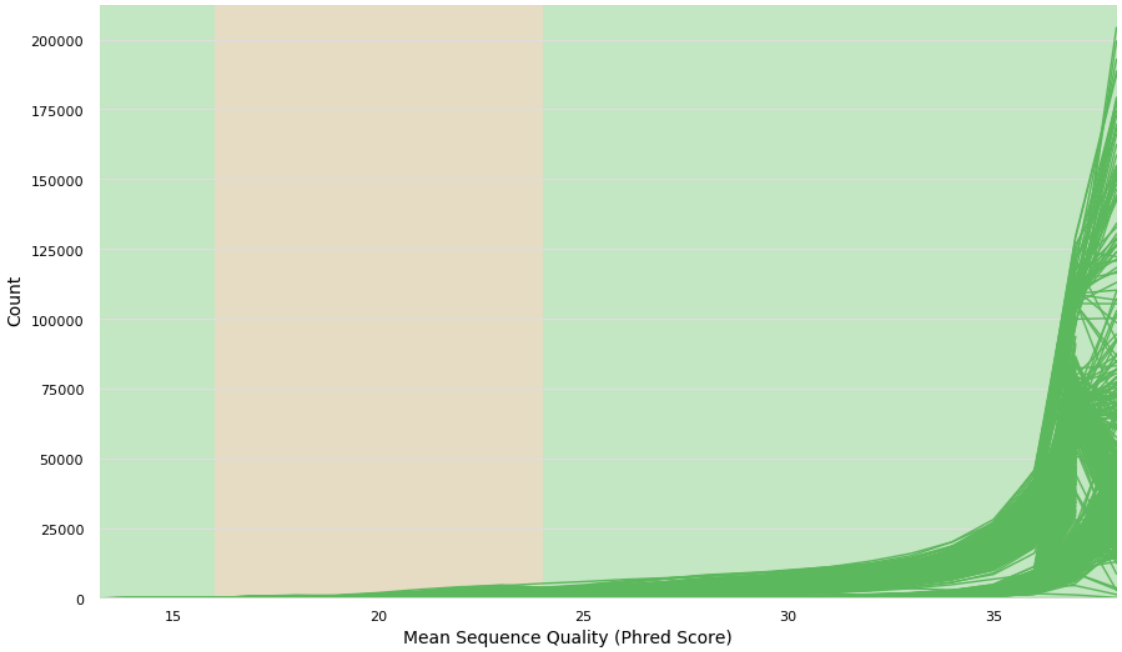

Figure 12. Distribution of the reads' average Phred quality scores after trimming with Trim Galore. Each green line, obtained by graphing the amount of reads (vertical axis) corresponding to each quality score bucket (horizontal axis) and connecting them, represents one FASTQ file.

A MultiQC report generated after adapter trimming indicated no adapter contamination and a sufficiently high quality for most reads (Figure 12). In addition, the success of bisulfite treatment of DNA could be confirmed by observing a high T to C ratio (Figure 13). The reasoning is that since genomic cytosines are largely unmethylated (~97%, knowing that the frequency of C-s is ~21%, the frequency of CpG-s is ~1%, and methylation among CpG-s is ~70% [46]), it is expected that most of them undergo conversion to thymine via bisulfite treatment.

[11] Insert is the sequence that remains when adapters are removed from both ends of an amplicon.

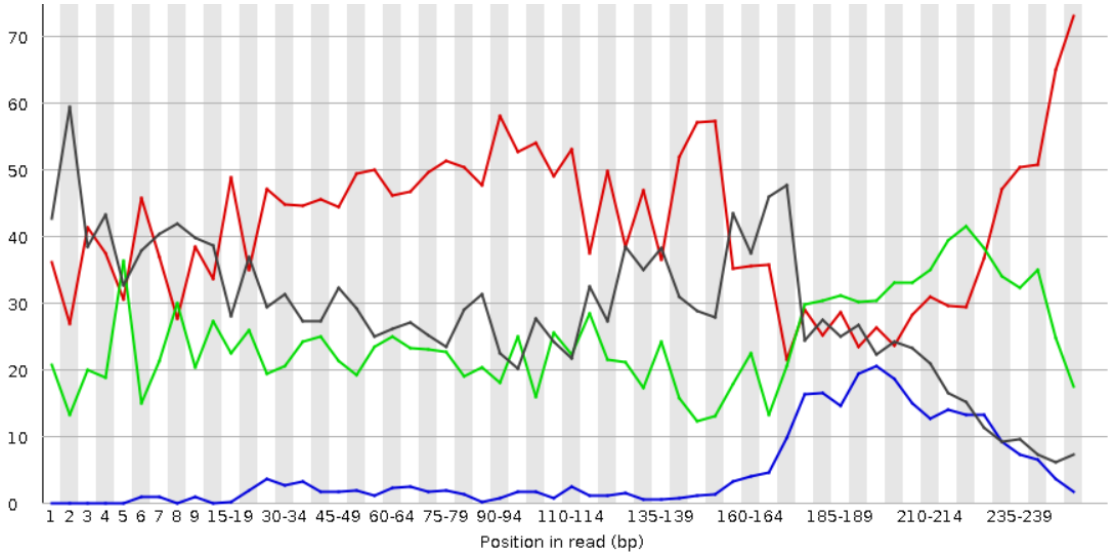

Figure 13. Distribution (in %) of sequence base content per position from one of the FASTQ files (black — G, red — T, green — A, blue — C). Cytosine-to-thymine conversion by bisulfite can be clearly seen in positions up to ~150. The anomaly beyond that position is of no concern since the average insert size (reported by the alignment step covered in the next section) is about 50, which means it will be effectively discarded.

4.4 Read mapping and methylation calling

Sequenced reads are of little use unless their location in the genome is determined. This is the job of aligners [30], which map reads onto a reference genome assembled from the DNA of many individuals (such as the one produced by the Human Genome Project [46]). Since reads are relatively short compared to the overall size of the genome (less than 1kbp vs more than 3Gbp), and reference genomes do not account for all genomic diversity found in humans, exact matches are rare [30]. Aligners consider between multiple possible alignments by assigning each one a score, calculated such that it rewards single nucleotide matches but penalizes mutations and gaps (insertions, deletions) [77]. Alignment with the highest score is then stored in an alignment report, often encoded as a SAM [56] (or its binary equivalent, BAM) file. Bisulfite alignment is based on the same principles, but with the added complexity of having to consider all purposefully introduced C $\rightarrow$ T (and G $\rightarrow$ A on the complementary strand) mutations [10].

Sun et al. [86] have evaluated different bisulfite alignment software, both on simulated and real data. Basing off of these results, the three with best precision and recall were picked for further use in this thesis: Bismark [10] (using the Bowtie 2 aligner), BS-Seeker2 [39] (Bowtie 2 aligner) and bwa-meth [70] (Burrow-Wheeler aligner). Furthermore, GRCh37 [17] (Genome Reference Consortium Human Build 37) was used as

the reference genome, since PCR primers had been previously designed specifically for this build. For all three software packages, the following tasks were performed: (i) reference genome preparation (indexing, C $\rightarrow$ T conversions); (ii) alignment of paired-end reads; and (iii) methylation calling. The output of this pipeline was a coverage report, one per sequencing experiment (a pair of FASTQ files), which for every available site counted the number of reads where cytosine in that position was methylated and where it was not.

Analysis of alignment files with SAMtools [56] revealed erroneous results for both bwa-meth and BS-Seeker2. For one pair of FASTQ files, for example, all reads mapped by Bismark were proper (paired-end) pairs, and were equally mapped to forward and reverse strands (paired-end sequencing produces pairs of complementary reads). However, BS-Seeker2 mapped a whole 96% of reads on the forward strand, and bwa-meth reported only 88% of reads being proper pairs. Due to these reasons, the following steps are all based on the results of Bismark.

4.5 Preprocessing

Coverage reports, which provided the counts of methylated (a) and unmethylated (b) cytosines per site, were used for calculating: (i) methylation proportion ($\frac{a}{a+b}$); and (ii) read depth ($a + b$). Methylation proportion can be interpreted as the proportion of cells in a blood sample for which a specific site was methylated, while read depth indicates the reliability of this value. These proportions were aggregated into a single dataset. Subjects whose DNA was sequenced more than once had conflicting methylation values available for some sites — in these cases, higher read depth was preferred.

The aggregated dataset was further filtered in two consecutive steps: (i) to remove low-coverage (read depth <300) methylation values; and (ii) to remove sites for which only <30% of subjects had methylation values available. The rationale behind the first step is that low coverage values are indicative of misalignment and would thus pollute the data with invalid information, while the second removes large "holes" of missing data. As a result, the amount of sites was narrowed from 17149 to 104. Figure 14 summarizes the read depths of the resulting dataset.

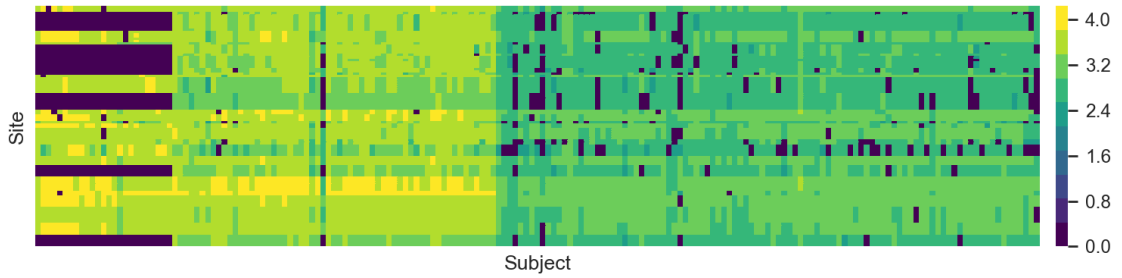

Figure 14. Heatmap of read depths of the filtered dataset (rows are sites, columns are subjects), transformed with the function $log_{10}(x + 1)$. Dark spots, which denote values either missing or removed during filtering, are especially prominent on the left side (all EDU subjects), likely because a smaller set of primers was used during PCR. The transition from light to dark green in the middle is due to a decrease in PCR cycles — according to the lab that performed the sequencing, the initial cycle count was unnecessarily high.

4.6 Imputation

By now, 9.47% of the values were missing from the dataset (dark spots in Figure 14), partly due to filtering performed in the previous step. To obtain a full dataset, a prerequisite for many machine learning methods, replacements for missing values should be derived from the existing data [27]. In the simplest case, this could be the mean or the most frequent value of each feature.

More sophisticated imputation methods, such as scikit-learn's IterativeImputer [71], attempt to capture the relationship between features into predictive models. IterativeImputer is an implementation of the MICE (multivariate imputation by chained equations) algorithm, which "changes the imputation problem to a series of estimations where each variable takes its turn in being regressed on the other variables" [94]. MICE performs imputation iteratively, incorporating previous imputations into each successive model for predicting the next feature [94]. This process is usually repeated for a number of times, which means that each feature undergoes multiple imputation cycles [94].

For the data at hand, IterativeImputer was run with a setting to constrain imputations to the interval $[\mathbf{0}, \mathbf{1}]$, for 10 cycles. The output of the final cycle was taken as the result of the imputation step for further use. Imputation accuracy can be estimated by comparing the unfiltered and imputed datasets, as is shown in Figure 15.

Figure 16 visualizes the imputed methylation dataset by correlating it with T_{EMRA} proportions from the FACS dataset. The gradients hint at the existence of methylation pattern for this lymphocyte subtype, making it a good candidate for a prediction model.

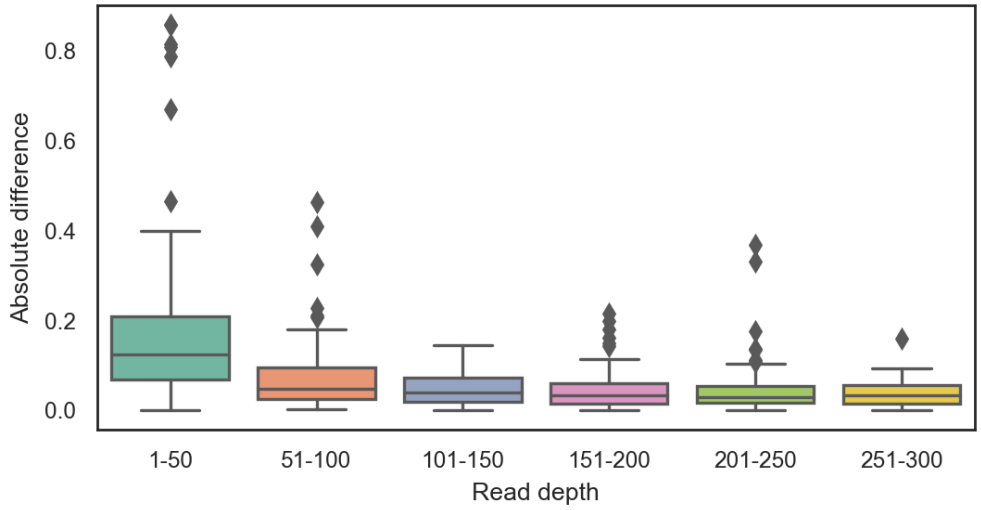

Figure 15. Absolute difference between the original low-coverage methylation proportions (the removal of which was described in Section 4.5) and their imputed counterparts. While the difference is much higher at lower read depths, this does not mean that imputation was inaccurate — the original values themselves are less reliable due to low read depth. Judging by the more reliable differences at read depths >100, imputation can be expected to be correct by an average margin of 5%.

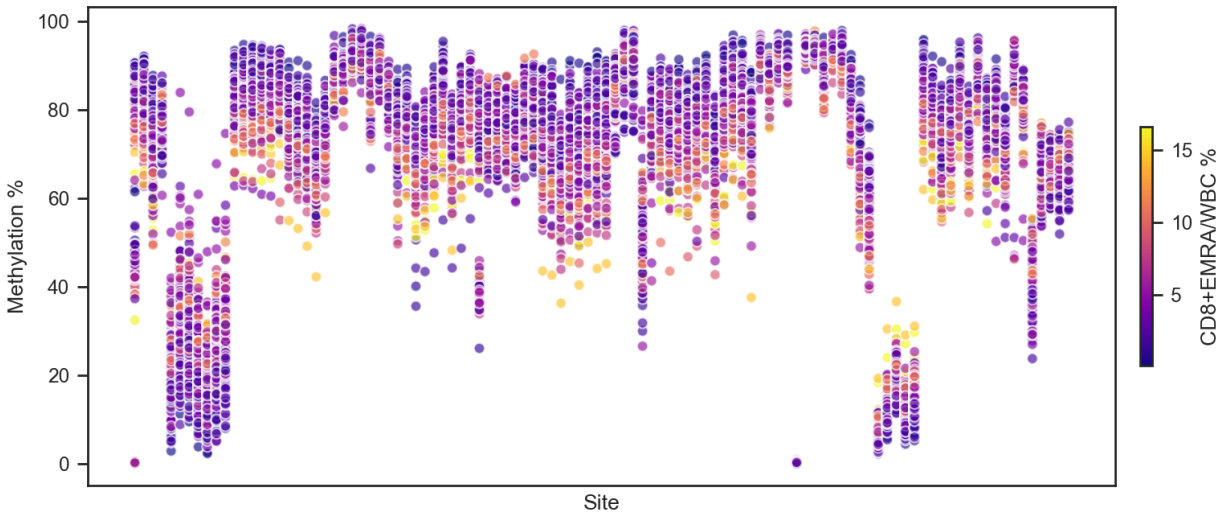

Figure 16. Plot of the imputed methylation dataset. Each column includes a dot for every subject, and each dot represents a subject's methylation at that site, color-coded by their measured $CD8^+$ T_{EMRA} proportion among all white blood cells.

4.7 Transformation for normality

Having covered the preprocessing performed on methylation proportions (the independent variables in the upcoming models), it is apt to turn the focus to lymphocyte proportions (the dependent variables). The only processing done here was transformation for normality, applied to proportions with strongly skewed (i.e. non-normal) distributions. Deviation from normality was quantitatively assessed with the Shapiro-Wilk test, whose null hypothesis is that the variable is normally distributed [37]. An example of a variable before and after transformation for normality can be seen in Figure 17.

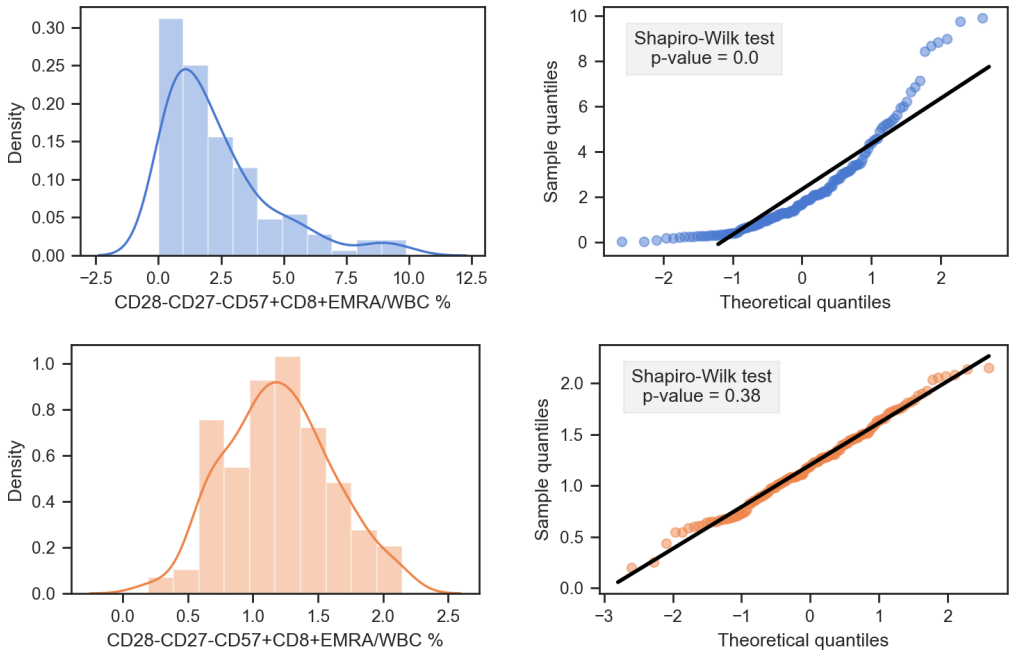

Figure 17. Distribution of $CD28^{-}CD27^{-}CD57^{+}CD8^{+}$ T_{EMRA} proportions among white blood cells (left) and its corresponding normality Q-Q plot (right), before (up) and after (down) a normalizing transformation. The transformation function was cube root. Q-Q plots can be used for qualitative assessment of normality — the distribution is considered normal if all points roughly fall onto the same line [37]. The p-values of the Shapiro-Wilk test are also given, showing that at an alpha level of 0.05, taking the cube root indeed normalized the distribution.

While linear regression models don't require the dependent variable to be normally distributed, they do so for the residuals (prediction errors) [67]. When residuals are not normally distributed, it might be indicative of a non-linear relationship, which can sometimes be linearized with a power transformation (Figure 18) [7].

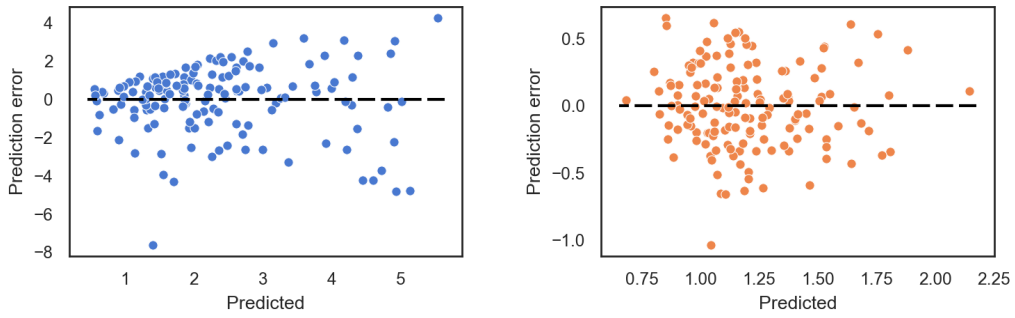

Figure 18. Residuals plots of a simple linear model trained to predict $CD28^{-}CD27^{-}CD57^{+}CD8^{+}$ T_{EMRA} (left) and its cube root (right). It can be seen that applying a normalizing transformation to the dependent variable improved the distribution of residuals by removing the fanning.

The normalizing transformations considered were: square root, cube root and logarithm (base 2, base 3, base 10). To emphasize, transformed variables were only used in model fitting and were inverse-transformed for subsequent analysis and presentation.

4.8 Model selection

Model selection was performed with a 5-fold (outer) + 3-fold (inner) nested cross-validation algorithm (Figure 19). The regression methods considered were simple linear, ridge, lasso and random forest. Each method was associated with a predefined set of hyperparameters, which were selected from the inner splits with scikit-learn's GridSearchCV [71]. The models were defined as a pipeline consisting of two steps: (i) feature selection with Python implementation of the Boruta algorithm [87]; and (ii) regression with the selected learning method.

Table 2 shows the regression methods and the associated hyperparameter values. Some parameters were given a range of values, forming the basis on hyperparameter selection, while others remained fixed throughout cross-validation. The hyperparameters' values (and ranges) were mostly chosen through trial and error, taking sklearn's defaults as the starting point.

This cross-validation algorithm was applied to all lymphocyte proportions of interest. Average (R)MSE, MAE and r were calculated over the combined predictions from all splits. Each regression method's performance was defined by its MSE, which formed the basis for selecting the most suitable method for predicting each cell proportion.

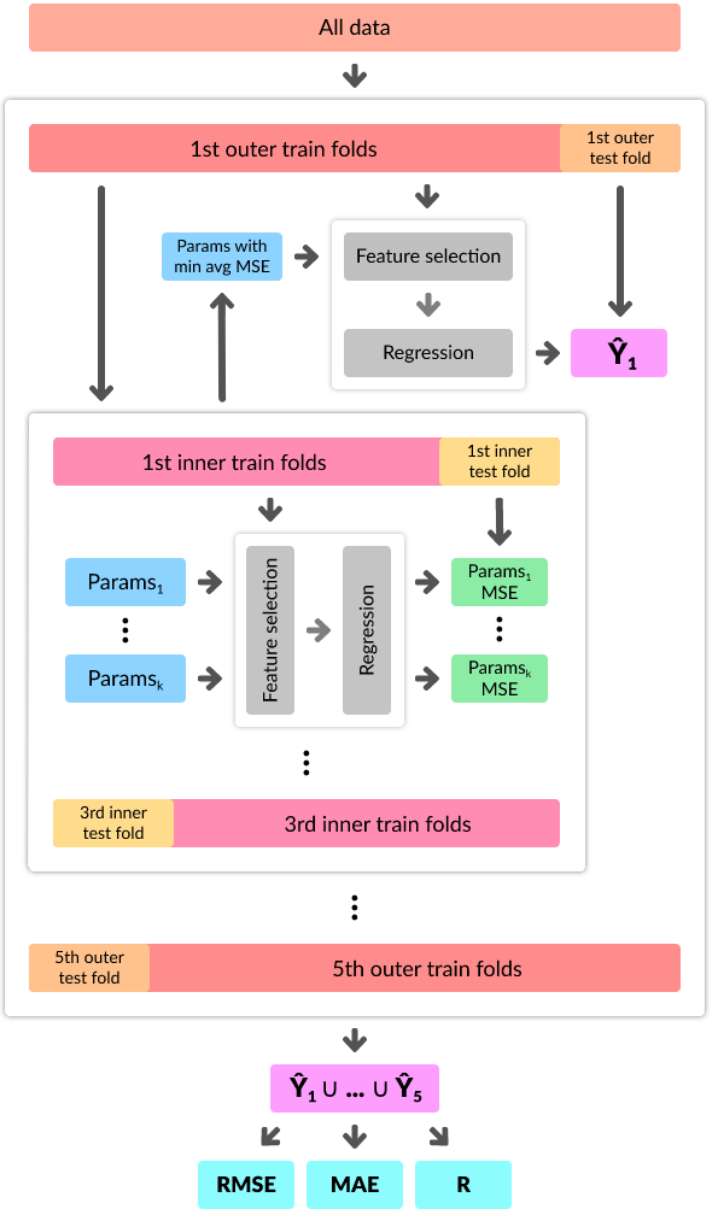

Figure 19. A 5-fold/3-fold nested cross-validation algorithm for model selection. Each set of outer training folds is first fed into the inner CV step for hyperparameter tuning. The best-performing hyperparameters are then used to learn a model on the outer training folds, which in turn is applied to the corresponding test fold to obtain predictions. After repeating it for each set of outer folds, the predictions are aggregated into a single set to calculate RMSE, MAE and r.

Table 2. Regression methods and their configuration parameters. Some were fixed for all models, while others were determined through hyperparameter selection.

Method	**Fixed hyperparameters**	**Selectable hyperparameters**
Linear	—	—
Ridge	—	alpha = $\{1^{-4}, 1^{-3}, 1^{-2}, 1^{-1}, 1\}$
Lasso	selection = random	alpha = $\{1^{-4}, 1^{-3}, 1^{-2}, 1^{-1}, 1\}$
RandomForest	min_samples_leaf = 2	n_estimators = {100, 325, 550, 775, 1000} max_depth = {10, 20, 30}

5 Results

Overall, models for predicting 9 different lymphocyte subtype proportions (among white blood cells) were trained and evaluated. These subtypes include: all lymphocytes (LYMP); T cells ($CD3^+$) and their primary subtypes — helper ($CD4^+$) and cytotoxic ($CD8^+$); plus cytotoxic T_{EMRA} ($CD8^+$ EMRA) with its senescent ($CD28^-$, $CD28^-CD27^-$, $CD28^-CD27^-CD57^+$) and exhausted ($PD1^+$) subtypes. Cross-validation results from model selection are covered first, followed by analysis of the final models, which were trained on all data using the best-performing regression method for each subtype.

Figure 20 visualizes the combined predictions from the five (outer) cross-validation splits along with the values of RMSE, MAE and r. In general, models predicting less specific proportions seem to outperform those predicting subtypes of T_{EMRA} — lymphocytes (A) and T cells (B) have r at 0.78–0.8, T cell subtypes (C — D) and T_{EMRA} (E) at 0.67–0.69, followed by senescent T_{EMRA} (G — H) at 0.59–0.63 and exhausted T_{EMRA} (I) at 0.1. RMSE and MAE decrease in the same direction, i.e. increasing in accuracy, but this is caused by a decrease in the values' ranges themselves and is therefore not indicative of increasing accuracy. When it comes to outliers, the models generally underestimate their actual proportions. It can be seen that all regression methods besides simple linear are represented, which means that regularized methods outperformed simple linear regression in all cases.

The anomalous predictions of exhausted T_{EMRA} (I) are an example of lasso's tendency to favour sparse solutions — in this case, the models of some cross-validation splits converged on a model that predicted near-constant values. It is clear that this subtype could not really be predicted from the available data.

The final models (Table 3) are diverse in terms of learning methods and their hyperparameters, but largely similar in the amount of selected sites (all less than 15% of the initial 104). An interesting pattern can be seen for the chosen normality transformation functions, for which the increase in strength follows the same pattern seen before for the models' r, RMSE, MAE and value scales.

Analysing the overlap of sites that each model was based on (Figure 21) provides an explanation for some of the results. Firstly, the 3 sites that were used for predicting $PD1^+CD8^+$EMRA are represented in almost all other models, which means that its prediction was based on a weak correlation with other subtypes. Secondly, the similar prediction results of $CD28^-CD8^+$EMRA and $CD28^-CD27^-CD8^+$EMRA are explained by the use of a largely identical set of sites. Thirdly, the fact that the sites for predicting most T_{EMRA} subtypes have no overlap with those for predicting LYMP and $CD3^+$, which they are also a part of, hints at a lack of correlation between the proportions of lymphocytes/T cells and T_{EMRA}. This makes sense because the numbers of T_{EMRA} increase through cell differentiation, which does not change the overall amount of T cells or lymphocytes.

Compared to the results of Bergstedt et al. [8], the models of this work achieved better performance in some cases and worse in others. The former include the models

predicting CD3$^+$ (r of 0.75 for them, 0.8 for this work) and lymphocytes (0.55 vs 0.78), while the latter include CD8$^+$ (0.8 vs 0.67) and CD8$^+$EMRA (0.8 vs 0.69). Despite these results, the differences in methodology, data and objectives do not allow any concrete conclusions to be drawn on the effectiveness of one approach over the other.

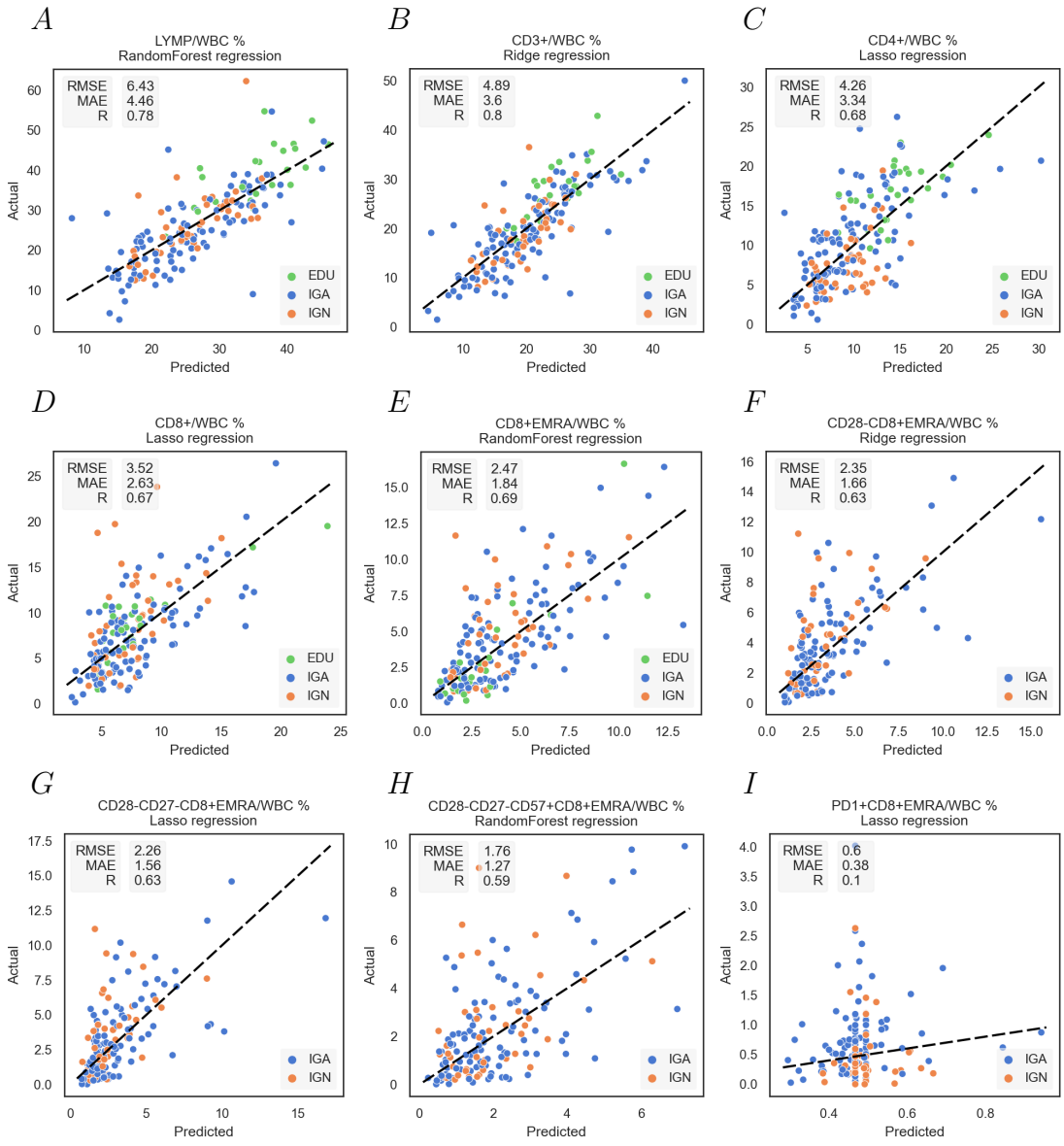

Figure 20. Cross-validation results of the best-performing method for predicting each lymphocyte subtype proportion (in %). The results were obtained by combining the predictions from 5 different models, one for each (outer) CV fold.

Table 3. Overview of the final regression models trained on all data, including: (i) the transformation for normality applied to the proportion; (ii) the number of sites remaining after feature selection; (iii) the regression method; and (iv) the hyperparameters selected via 3-fold cross-validation.

Subtype	**Transform.**	**Sites**	**Method**	**Hyperparameters**
LYMP	—	13	RandomForest	n_estimators = 325 max_depth = 20
$CD3^+$	—	12	Ridge	alpha = 0.1
$CD4^+$	$\sqrt{y}$	11	Lasso	alpha = 0.0001
$CD8^+$	$\sqrt{y}$	10	Lasso	alpha = 0.001
$CD8^+$ EMRA	$\sqrt[3]{y}$	14	RandomForest	n_estimators = 550 max_depth = 20
$CD28^-$ $CD8^+$ EMRA	$\sqrt[3]{y}$	9	Ridge	alpha = 0.1
$CD28^-CD27^-$ $CD8^+$ EMRA	$\sqrt[3]{y}$	9	Lasso	alpha = 0.001
$CD28^-CD27^-CD57^+$ $CD8^+$ EMRA	$\sqrt[3]{y}$	10	RandomForest	n_estimators = 325 max_depth = 10
$PD1^+$ $CD8^+$ EMRA	$\log_2 y$	3	Lasso	alpha = 0.01

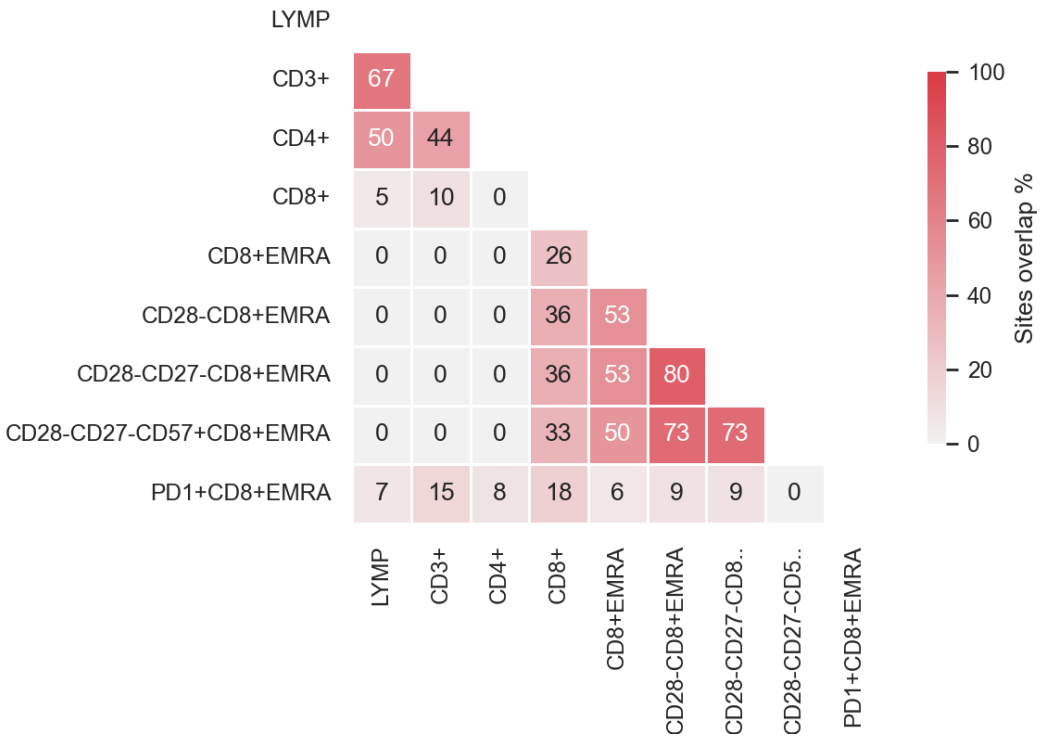

Figure 21. Overlap (in %) between sites that were used in the prediction of each lymphocyte subtype proportion, calculated for each pair of sites (S_i, S_j) as $\frac{|S_i \cap S_j|}{|S_i \cup S_j|}$.

6 Conclusion

In this thesis, the viability of using DNA methylation data for predicting the proportions of various lymphocytes in blood samples was explored. For this purpose, the DNA of 183 subjects was bisulfite-sequenced to obtain the blood samples' average methylation at a single site resolution. This data was further aggregated, filtered and imputed in preparation of machine learning against a reference dataset containing the actual proportions measured with FACS. A nested cross-validation model selection algorithm was devised to determine the most suitable regression method for predicting each lymphocyte subtype.

Overall, 9 different lymphocyte proportions were estimated, among them cytotoxic T_{EMRA} in various stages of differentiation. While the prediction of non-T_{EMRA} proved to be more successful, several subtypes of T_{EMRA} could be predicted with moderate accuracy. Since methylation-based cell type deconvolution is a cheaper and more scalable alternative to current methods of blood sample analysis, these results support the viability of this approach.

There is certainly room for improvement regarding the models' accuracy, especially if they were to be used in a clinical setting. The amount of subjects (183) and their profile (mostly elderly hospitalized women) introduces bias into the models, so the first area for improvement would be the inclusion of more subjects. Additionally, the sample should be more representative — both age-wise and in terms of the subjects' sex and health condition. In the primer design phase, the inclusion of new CpG sites that more specifically identify (or exclude) certain leukocyte subpopulations would also improve the accuracy of the models.

Recent advances in the relevant technologies, be it in flow cytometry (FACS) or bisulfite sequencing, would likely aid in achieving better results even if the experiment was repeated on the exact same sample. Specifically, the increasing precision of these technologies would reduce noise in the data and contribute to more accurate models. New methods for methylation sequencing, such as methylation arrays, are known to be cheaper [25] and more scalable than the laborious bisulfite sequencing protocol employed in this thesis, and would facilitate the application of the covered methodology at scale, e.g. in a clinical setting. More exotic machine learning methods, such as neural networks, would also be interesting to explore as an alternative to deconvolution methods based on linear regression.

References

[1] Nicholas M. Adams, Simon Grassmann, and Joseph C. Sun. Clonal expansion of innate and adaptive lymphocytes. *Nature Reviews Immunology*, 20(11):694–707, November 2020.

[2] Illumina. An introduction to next-generation sequencing technology [digital reference]. https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf, 2017. Accessed 2021-12-23.

[3] Vasily V. Ashapkin, Lyudmila I. Kutueva, and Boris F. Vanyushin. Aging as an epigenetic phenomenon. *Current Genomics*, 18(5):385–407, October 2017.

[4] David Barnett, Brooke Walker, Alan Landay, and Thomas N. Denny. CD4 immunophenotyping in HIV infection. *Nature Reviews Microbiology*, 6(11 Suppl):S7–15, November 2008.

[5] Udo Baron, Ivana Turbachova, Alexander Hellwag, Florian Eckhardt, Kurt Berlin, Ulrich Hoffmöller, Paula Gardina, and Sven Olek. DNA methylation analysis as a tool for cell typing. *Epigenetics*, 1(1):56–61, March 2006.

[6] Adam C. Bell and Gary Felsenfeld. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, 405(6785):482–485, May 2000.

[7] Kenneth Benoit. Linear regression models with logarithmic transformations [course material]. https://kenbenoit.net/assets/courses/ME104/logmodels2.pdf, March 2011. Accessed 2022-04-27.

[8] Jacob Bergstedt, Alejandra Urrutia, Darragh Duffy, Matthew L. Albert, Lluís Quintana-Murci, and Etienne Patin. Accurate prediction of cell composition, age, smoking consumption and infection serostatus based on blood DNA methylation profiles. Preprint at https://www.biorxiv.org/content/10.1101/456996v1, October 2018. Accessed 2022-05-05.

[9] Jules J. Berman. Understanding your data. In *Data Simplification*, chapter 4, pages 135–187. Morgan Kaufmann, Boston, 2016.

[10] Babraham Bioinformatics. Bismark: a tool to map bisulfite converted sequence reads and determine cytosine methylation states [software]. https://www.bioinformatics.babraham.ac.uk/projects/bismark. Accessed 2021-12-28.

[11] Alexei Botchkarev. A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, 14:45—79, January 2019.

[12] Max Bramer. Avoiding overfitting of decision trees. In *Principles of Data Mining*, pages 121–136. Springer London, London, 2016.

[13] Stefan Brunner, Dietmar Herndler-Brandstetter, Birgit Weinberger, and Beatrix Grubeck-Loebenstein. Persistent viral infections and immune aging. *Ageing Research Reviews*, 10(3):362–9, Jul 2011.

[14] Tianfeng Chai and Roland R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development*, 7(3):1247—-1250, June 2014.

[15] David D. Chaplin. Overview of the immune response. *The Journal of allergy and clinical immunology*, 125(2 Suppl 2):S3–23, February 2010.

[16] Talyn Chu, Aaron J. Tyznik, Sarah Roepke, Amy M. Berkley, Amanda Woodward-Davis, Laura Pattacini, Michael J. Bevan, ..., and Martin Prlic. Bystander-activated memory CD8 T cells control early pathogen load in an innate-like, NKG2D-dependent manner. *Cell Reports*, 3(3):701–708, March 2013.

[17] Deanna M. Church, Valerie A. Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, ..., and Tim Hubbard. Modernizing reference genome assemblies. *PLoS Biology*, 9(7):e1001091, July 2011.

[18] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, April 2010.

[19] Wikimedia Commons. DNA methylation landscape in mammals [digital image]. `https://commons.wikimedia.org/wiki/File:DNAme_landscape.png`. Accessed 2021-12-28.

[20] Rose Du, Vince Carey, and Scott T. Weiss. deconvSeq: deconvolution of cell mixture distribution in sequencing data. *Bioinformatics*, 35(24):5095—-5102, December 2019.

[21] Mark T. Esser, Rocio D. Marchese, Lisa S. Kierstead, Lynda G. Tussey, Fubao Wang, Narendra Chirmule, and Michael W. Washabaugh. Memory T cells and vaccines. *Vaccine*, 21(5–6):419–430, January 2003.

[22] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, October 2016.

[23] Gregory M. Fahy, Robert T. Brooke, James P. Watson, Zinaida Good, Shreyas S. Vasanawala, Holden Maecker, Michael D. Leipold, ..., and Steve Horvath. Reversal of epigenetic aging and immunosenescent trends in humans. *Aging Cell*, 18(6):e13028, September 2019.

[24] Babraham Bioinformatics. FastQC: a quality control tool for high throughput sequence data [software]. `https://www.bioinformatics.babraham.ac.uk/projects/fastqc`. Accessed 2021-12-28.

[25] Illumina. Field guide to methylation methods [digital reference]. `https://www.illumina.com/content/dam/illumina-marketing/documents/products/other/field_guide_methylation.pdf`, 2016. Accessed 2022-04-28.

[26] Peter Flach. Linear models. In *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, chapter 7, pages 194–230. Cambridge University Press, Cambridge, 2012.

[27] Peter Flach. Machine learning experiments. In *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, chapter 12, pages 343–359. Cambridge University Press, Cambridge, 2012.

[28] Peter Flach. Model ensembles. In *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, chapter 11, pages 330–342. Cambridge University Press, Cambridge, 2012.

[29] Peter Flach. Tree models. In *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, chapter 5, pages 129–156. Cambridge University Press, Cambridge, 2012.

[30] Nuno A. Fonseca, Johan Rung, Alvis Brazma, and John C. Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, December 2012.

[31] Mario F. Fraga, Esteban Ballestar, Maria F. Paz, Santiago Ropero, Fernando Setien, Maria L. Ballestar, Damia Heine-Suñer, ..., and Manel Esteller. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10604–10609, July 2005.

[32] Marianne Frommer, Louise E. McDonald, Douglas S. Millar, Christina M. Collis, Fujiko Watt, Geoffrey W. Grigg, Peter L. Molloy, and Cheryl L. Paul. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89(5):1827–1831, March 1992.

[33] Reinhold Förster, Ana C. Davalos-Misslitz, and Antal Rot. CCR7 and its ligands: balancing immunity and tolerance. *Nature Reviews Immunology*, 8(5):362–371, May 2008.

[34] Tünde Fülöp, Anis Larbi, and Graham Pawelec. Human T cell aging and the impact of persistent viral infections. *Frontiers in Immunology*, 4:271, September 2013.

[35] Jens Geginat, Antonio Lanzavecchia, and Federica Sallusto. Proliferation and differentiation potential of human CD8+ memory T-cell subsets in response to antigen or homeostatic cytokines. *Blood*, 101(11):4260–4266, June 2003.

[36] Wikimedia Commons. Genetic code [digital image]. `https://en.wikipedia.org/wiki/File:Genetic_code.svg`. Accessed 2021-10-23.

[37] Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2):486–489, April 2012.

[38] Abcam. Guide to human CD antigens [digital reference]. `https://docs.abcam.com/pdf/immunology/Guide-to-human-CD-antigens.pdf`, 2021. Accessed 2021-11-05.

[39] Weilong Guo, Petko Fiziev, Weihong Yan, Shawn Cokus, Xueguang Sun, Michael Q. Zhang, Pao-Yang Chen, and Matteo Pellegrini. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*, 14(1):774, November 2013.

[40] Isabelle Guyon and André Elisseeff. An introduction of variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.

[41] Bastiaan T. Heijmans, Elmar W. Tobi, Aryeh D. Stein, Hein Putter, Gerard J. Blauw, Ezra S. Susser, P. Eline Slagboom, and L. H. Lumey. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(44):17046–17049, November 2008.

[42] Julie B. Herbstman, Deliang Tang, Deguang Zhu, Lirong Qu, Andreas Sjödin, Zheng Li, David Camann, and Frederica P. Perera. Prenatal exposure to polycyclic

aromatic hydrocarbons, benzo[a]pyrene-DNA adducts, and genomic DNA methylation in cord blood. *Environmental Health Perspectives*, 120(5):733–738, May 2012.

[43] Sae R. Hong, Sang-Eun Jung, Eun H. Lee, Kyoung-Jin Shin, Woo I. Yang, and Hwan Y. Lee. DNA methylation-based age prediction from saliva: high age predictability by combination of 7 CpG markers. *Forensic Science International: Genetics*, 29:118–125, July 2017.

[44] Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):3156, December 2013.

[45] Eugene A. Houseman, William P. Accomando, Devin C. Koestler, Brock C. Christensen, Carmen J. Marsit, Heather H. Nelson, John K. Wiencke, and Karl T. Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13:86, May 2012.

[46] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[47] Peter A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13:484–492, May 2012.

[48] Naeem Khan, Naseer Shariff, Mark Cobbold, Rachel Bruton, Jenni A. Ainsworth, Alan J. Sinclair, Laxman Nayak, and Paul A. H. Moss. Cytomegalovirus seropositivity drives the CD8 T cell repertoire toward greater clonality in healthy elderly individuals. *The Journal of Immunology*, 169(4):1984–1992, August 2002.

[49] Stephan Kolassa. Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, 32(3):788–803, July 2016.

[50] Miron B. Kursa and Witold R. Rudnicki. Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11):1–13, September 2010.

[51] Anis Larbi and Tamas Fulop. From "truly naïve" to "exhausted senescent" T cells: when markers predict functionality. *Cytometry Part A*, 85(1):25–35, January 2014.

[52] Cheol-Koo Lee, Yoichiro Shibata, Bhargavi Rao, Brian D Strahl, and Jason D Lieb. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics*, 36(8):900–905, July 2004.

[53] Junghwa Lee, Eunseon Ahn, Haydn T. Kissick, and Rafi Ahmed. Reinvigorating exhausted T cells by blockade of the PD-1 pathway. *Forum on Immunopathological Diseases and Therapeutics*, 6(1–2):7–17, 2015.

[54] Violetta V. Leshchenko, Pei-Yu Kuo, Rita Shaknovich, David T. Yang, Tobias Gellen, Adam Petrich, Yiting Yu, ..., and Samir Parekh. Genomewide DNA methylation analysis reveals novel targets for drug development in mantle cell lymphoma. *Blood*, 116(7):1025–1034, August 2010.

[55] Morgan E. Levine, Ake T. Lu, David A. Bennett, and Steve Horvath. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. *Aging*, 7(12):1198–1211, December 2015.

[56] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, ..., and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[57] Andreas Lindhold, Niklas Wahlström, Fredrik Lindsten, and Thomas B. Schön. The regression problem and linear regression. In *Supervised Machine Learning: Lecture Notes for the Statistical Machine Learning Course*, pages 11–24. `http://web.archive.org/web/20191206202621/http://www.it.uu.se/edu/course/homepage/sml/literature/lecture_notes.pdf`, 2019. Accessed 2021-12-18.

[58] Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, January 2011.

[59] Carlos López-Otín, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, June 2013.

[60] Riccardo E. Marioni, Sonia Shah, Allan F. McRae, Brian H. Chen, Elena Colicino, Sarah E. Harris, Jude Gibson, ..., and Ian J. Deary. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biology*, 16(1):25, January 2015.

[61] Matthew D. Martin and Vladimir P. Badovinac. Defining memory CD8 T cell. *Frontiers in Immunology*, 9:2692, November 2018.

[62] Carmen Martin-Ruiz, Jedrzej Hoffmann, Evgeniya Shmeleva, Thomas von Zglinicki, Gavin Richardson, Lilia Draganova, Rachael Redgrave, ..., and Ioakim Spyridopoulos. CMV-independent increase in CD27-CD28+ CD8+ EMRA T cells is inversely related to mortality in octogenarians. *npj Aging and Mechanisms of Disease*, 6(1):3, January 2020.

[63] Lisa M. McEwen, Alexander M. Morin, Rachel D. Edgar, Julia L. MacIsaac, Meaghan J. Jones, William H. Dow, Luis Rosero-Bixby, ..., and David H. Rehkopf.

Differential DNA methylation and lymphocyte proportions in a Costa Rican high longevity region. *Epigenetics & Chromatin*, 10(1):21, April 2017.

[64] Katherine M. McKinnon. Flow cytometry: an overview. *Current Protocols in Immunology*, 120:5.1.1–5.1.11, February 2018.

[65] Michael J. Meaney and Moshe Szyf. Environmental programming of stress responses through DNA methylation: life at the interface between a dynamic environment and a fixed genome. *Dialogues in Clinical Neuroscience*, 7(2):103–123, June 2005.

[66] Ruslan Medzhitov. Recognition of microorganisms and activation of the immune response. *Nature*, 449(7164):819–826, October 2007.

[67] Märt Möls. Lihtne lineaarne regressioon. In *Biomeetria bioloogidele: statistiliste ja matemaatiliste meetodite rakendamisest eluteadustes*, chapter 3, pages 87–116. `http://www-1.ms.ut.ee/mart/TS2015/StatistikaAlgt6ed.pdf`, 2014. Accessed 2022-05-07.

[68] Roland Nilsson, José M. Peña, Johan Björkegren, and Jesper Tegner. Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research*, 8:589–612, March 2007.

[69] Donald B. Palmer. The effect of age on thymic function. *Frontiers in Immunology*, 4:316, October 2013.

[70] Brent S. Pedersen, Kenneth Eyring, Subhajyoti De, Ivana V. Yang, and David A. Schwartz. Fast and accurate alignment of long bisulfite-seq reads. Preprint at `https://arxiv.org/abs/1401.1129`, January 2014. Accessed 2022-05-05.

[71] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, ..., and Édouard Duchesnay. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, October 2011.

[72] Wikimedia Commons. Polymerase chain reaction [digital image]. `https://commons.wikimedia.org/wiki/File:Polymerase_chain_reaction.svg`. Accessed 2021-12-23.

[73] Aurora E. Pop-Vicas and Stefan Gravenstein. Influenza in the elderly: a mini-review. *Gerontology*, 57(5):397–404, August 2011.

[74] Mamatha Prabhakar, William Ershler, and Dan Longo. Bone marrow, thymus and blood: changes across the lifespan. *Aging Health*, 5(3):385–393, June 2009.

[75] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. Preprint at https://arxiv.org/abs/1811.12808, November 2018. Accessed 2022-05-07.

[76] Regression tree [digital image]. https://www.datacamp.com/community/tutorials/decision-trees-R. Accessed 2021-12-28.

[77] Maido Remm. Joonduse skoor. In *Bioinformaatika*, chapter 6, pages 95–122. Tartu Ülikooli Kirjastus, Tartu, 2015.

[78] William M. Rideout, Gerhard A. Coetzee, Aria F. Olumi, and Peter A. Jones. 5-methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science*, 249(4974):1288–1290, September 1990.

[79] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39, February 2010.

[80] Felix Sahm, Daniel Schrimpf, Damian Stichel, David T. W. Jones, Thomas Hielscher, Sebastian Schefzyk, Konstantin Okonechnikov, ..., and Andreas von Deimling. DNA methylation-based classification and grading system for meningioma: a multicentre, retrospective analysis. *The Lancet Oncology*, 18(5):682–694, May 2017.

[81] Zachary D. Smith and Alexander Meissner. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220, February 2013.

[82] Tom Strachan and Andrew P. Read. Amplifying DNA: cell-based DNA cloning and PCR. In *Human Molecular Genetics*, chapter 6, pages 163–190. Garland Science/Taylor & Francis Group, New York, 4th edition, 2011.

[83] Tom Strachan and Andrew P. Read. Chromosome structure and function. In *Human Molecular Genetics*, chapter 2, pages 29–60. Garland Science/Taylor & Francis Group, New York, 4th edition, 2011.

[84] Tom Strachan and Andrew P. Read. Human gene expression. In *Human Molecular Genetics*, chapter 11, pages 345–380. Garland Science/Taylor & Francis Group, New York, 4th edition, 2011.

[85] Tom Strachan and Andrew P. Read. Nucleic acid structure and gene expression. In *Human Molecular Genetics*, chapter 1, pages 1–28. Garland Science/Taylor & Francis Group, New York, 4th edition, 2011.

[86] Xiwei Sun, Yi Han, Liyuan Zhou, Enguo Chen, Bingjian Lu, Yong Liu, Xiaoqing Pan, ..., and Pengyuan Liu. A comprehensive evaluation of alignment software

for reduced representation bisulfite sequencing data. *Bioinformatics*, 34(16):2715–2733, August 2018.

[87] Simon Tamayo. Python implementations of the Boruta all-relevant feature selection method [software]. `https://github.com/scikit-learn-contrib/boruta_py`. Accessed 2022-05-10.

[88] Gaëlle Tilly, Tra-My Doan-Ngoc, Michelle Yap, Aurélie Caristan, Lola Jacquemont, Richard Danger, Marion Cadoux, ..., and Nicolas Degauque. IL-15 harnesses pro-inflammatory function of TEMRA CD8 in kidney-transplant recipients. *Frontiers in Immunology*, 8:778, June 2017.

[89] Babraham Bioinformatics. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries [software]. `https://www.bioinformatics.babraham.ac.uk/projects/trim_galore`. Accessed 2021-12-28.

[90] James W. Tung, Kartoosh Heydari, Rabin Tirouvanziam, Bita Sahaf, David R. Parks, Leonard A. Herzenberg, and Leonore A. Herzenberg. Modern flow cytometry: a practical approach. *Clinics in Laboratory Medicine*, 27(3):453–468, September 2007.

[91] Nataša Vasiljević, Amar S. Ahmad, Mangesh A. Thorat, Gabrielle Fisher, Daniel M. Berney, Henrik Müller, Christopher S. Foster, ..., and Attila T. Lorincz. DNA methylation gene-based models indicating independent poor outcome in prostate cancer. *BMC Cancer*, 14(1):655, September 2014.

[92] Lee Venolia and Stanley M. Gartler. Comparison of transformation efficiency of human active and inactive X-chromosomal DNA. *Nature*, 302(5903):82–83, March 1983.

[93] Wolfgang Wagner. The link between epigenetic clocks for aging and senescence. *Frontiers in Genetics*, 10:303, April 2019.

[94] Jesper N. Wulff and Linda E. Jeppesen. Multiple imputation by chained equations in praxis: guidelines and review. *Electronic Journal of Business Research Methods*, 15(1):41–56, April 2017.

[95] Michelle Yap, Françoise Boeffard, Emmanuel Clave, Annaick Pallier, Richard Danger, Magali Giral, Jacques Dantal, ..., and Nicolas Degauque. Expansion of highly differentiated cytotoxic terminally differentiated effector memory CD8+ T cells in a subset of clinically stable kidney transplant recipients: a potential

marker for late graft dysfunction. *Journal of the American Society of Nephrology*, 25(8):1856–1868, August 2014.

[96] Jeffrey A. Yoder, Colum P. Walsh, and Timothy H. Bestor. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 13(8):335–340, August 1997.

[97] James Zou, Christoph Lippert, David Heckerman, Martin Aryee, and Jennifer Listgarten. Epigenome-wide association studies without the need for cell-type composition. *Nature Methods*, 11(3):309–311, March 2014.

Appendix

I. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Simo Pähk**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Prediction of Cell Counts from DNA Methylation,
   *(*title of thesis)

   supervised by Ahto Salumets.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Simo Pähk
10/05/2022