Ole Fredrik Borgundvåg Berg

Circulating miRNA and lung cancer:
- an analysis of available data

Master thesis, Spring 2022

Supervisor: Pål Sætrom
Artificial Intelligence Group
Department of Computer and Information Science
Faculty of Information Technology, Mathematics and Electrical Engineering

Abstract

Preface

This is a master thesis for a master's degree in informatics with specialization in artificial intelligence, conducted at NTNU and St. Olavs Hospital, supervised by Pål Sætrom. I want to thank friends and family for support.

Ole Fredrik Borgundvåg Berg
Trondheim, February 20, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The project will primarily be about using methods from machine learning and statistics to look at the diagnostic value of circulating miRNA when it comes to lung cancer.

1.1 Background and Motivation

Lung cancer is common type of cancer with a low survival rate (more statistics in 2.1.1). One of the major reasons for the low survival rate is the late diagnosis of lung cancer. However, several studies indicate that circulating miRNA could be a non-invasive way to diagnose lung cancer [**?**]. This could lead to earlier diagnosis, and thus a higher survival rate.

1.2 Goals and Research Questions

Different studies have pointed to different miRNA sequences for the diagnosis of lung cancer. The point of this project is to collect the datasets from the different studies and create a larger dataset. With that dataset I want to achieve the following overall goal:

Goal: Use algorithms from machine learning on to predict lung cancer from levels of circulating miRNA on a larger dataset

Most studies on the area use simple logistic regression on the data in order to predict lung cancer based on miRNA values (e.g. [**??**]), thus leading to the question:

Research question 1: Are there machine learning algorithms that generally performs better at diagnosing lung cancer based on miRNA values?

Logistic regression is a linear model, and thus is unable to find patterns in the data that are non-linear, which might be the case with the effect of lung cancer on miRNA levels. There have been attempts of using more advanced machine learning methods on miRNA and lung cancer (e.g. [**?**]). My project differs, as I try to collect all available datasets, which gives more statistical power and it will be more of a meta-analysis where datasets from different studies are compared.

The datasets are slightly different in what miRNA that are measured, and what technologies are used to measure miRNA levels. This begs the question:

Research question 2: Will a combined dataset lead have better diagnostic value than each of the datasets alone?

On one hand one might think that the more data, the more information the machine learning algorithm have, and thus, a combined dataset is better. However, it is possible that datasets of lower quality would confuse rather than help a machine learning algorithm.

Other minor questions that might be answered are:

- What are the respective quality of the different datasets?
- Do the same miRNA have the same diagnostic value across different datasets?
- What miRNAs are most important for diagnosing lung cancer?
- What is the effect of lung cancer on the miRNA levels?

However, these questions are more interesting from a medical/biological point of view, than they are from a machine learning point of view, and as such will be a lower priority.

1.3 Research Method

This project is primarily an experimental one, as one need to actually train models on the datasets in order to compare the outcomes. The outcomes of the machine learning model are quantitative, and thus an analytical approach will be used. The main theoretical parts of this project are the parts concerning miRNA and lung cancer, as the outcomes of the machine learning might help in understanding the effect of lung cancer on miRNA, but as these questions are not related to machine learning directly, they are not the main focus.

1.4 Report Structure

Chapter 2 will include some theory around lung cancer and miRNA, together with theory around the machine learning and statistical methods and concepts that are used in this project. Chapter 3 is about how the literature search was done, how the data was processed and how machine learning was applied. Chapter 4 is about how the experiments performed and their results. Finally, chapter 5 is about the conclusions that are made from the results, and their significance.

Chapter 2

Background Theory and Motivation

This project is a cross-disciplinary one, as it combines machine learning and medicine, and as such, some theory from both disciplines are necessary in order to understand the project. Most of the subsections here were originally found in **?**, as the necessary preliminaries are mostly the same as in the specialization project.

2.1 Biological Theory

The first major part of is the biological/medical part.

2.1.1 Lung Cancer

Lung cancer is the second most common type of cancer worldwide, and the the type of cancer with the highest total mortality worldwide, causing about 1.8 million deaths [**?**]. Lung cancer is also the cancer type leading to the most deaths in Norway, amounting to 1500 deaths per year [**?**]. The most important risk factor related to lung cancer is smoking. Smoking is estimated to explain about 90% of the risk of lung cancer in men, and 70% to 80% of the risk of lung cancer in women [**?**]. Furthermore, about 90% of lung cancer deaths in men, and 79% of lung cancer deaths in women are caused by smoking [**?**].

There are two main types of lung cancer, Small Cell Lung Cancers (SCLC) and Non-Small Cell Lung Cancers (NSCLC) [**?**]. Of lung cancer cases, about 80-85% are NSCLC, whilst 10-15% of the cases are SCLC, and a few percent are non-mayor types of lung cancer [**?**]. NSCLC cancers tend to grow slower than

the SCLC cancer types, and thus SCLC has usually already spread when it is diagnosed [**?**]. The NSCLC has three mayor subtypes, namely, adenocarcinoma (30-40%), squamous cell (30%) and large-cell undifferentiated carcinoma (10-15%) [**?**]. The treatment and prognosis for the different NSCLC subtypes are similar [**?**].

Lung cancer develops in different stages. According to **?**, the main four are:

1. The cancer is only situated in your lung
2. The cancer may have spread to the lymph nodes near the lung
3. The cancer has spread deeper into the lymph nodes and into the middle of your chest
4. Cancer is widespread throughout your body

The main advantage with diagnosing lung cancer early is that the cancer has not yet spread to other parts of the body, which means that it can be removed by surgery [**?**]. On the other hand, later stages might require chemotherapy, radiation therapy or immunotherapy, but as the cancer has spread widely, this cure will likely not remove the cancer [**?**].

2.1.2 MicroRNA

MicroRNA (miRNA) are short sequences of RNA, about 22 nucleotides each, that regulates the expression of mRNA by binding to the target mRNA sequence, and thus stopping it from being translated. Circulating miRNA has been found to be a biomarker for many diseases, including cancer, infectious diseases and mental illnesses [**????**]. miRNA-sequences are usually named with the prefix "miR-" and then a unique number that is incremented for each discovery of a miRNA-sequence. The most commonly used database with known miRNA-sequences is the miRBase database [**?**].

2.1.3 MicroRNA and Lung Cancer

The overall role role of miRNA in relation to lung cancer is not fully understood [**?**]. MicroRNA is thought to be both function as tumor suppressor genes and as oncogenes [**?**]. Anyways, there are multiple studies that report about differential expression of circulating miRNA-sequences in cancer patients compared to healthy controls, which is results in expression of miRNA being a promising method for diagnosing lung cancer [**?**].

2.1.4 MicroRNA profiling methods

There are several methods for measuring levels of miRNA. The most common ones are qRT-PCR, microarrays and sequencing. Here is a very high level description of the different methods. For more technical details see e.g. **?**. The different technologies typically have different issues.

qRT-PCR

Quantitative Reverse Transcription - Polymerase Chain Reaction (qRT-PCR) is the most common method in the studies used in this project. As the name implies, the process depend on reverse transcription, where miRNA are reverse transcripted, using the enzyme reverse transcriptase, into complementary DNA (cDNA). Then polymerase chain reactions are initiated and monitored in order to measure miRNA levels.

In qRT-PCR, one needs a primer for each miRNA-sequence that should be measured. Therefore it can only measure miRNA-sequences that are decided beforehand. The main advantage of qRT-PCR is that it is the most sensitive method of the different technologies [**?**], which means that the results are more accurate, and that it also works well when the concentration of miRNA is low.

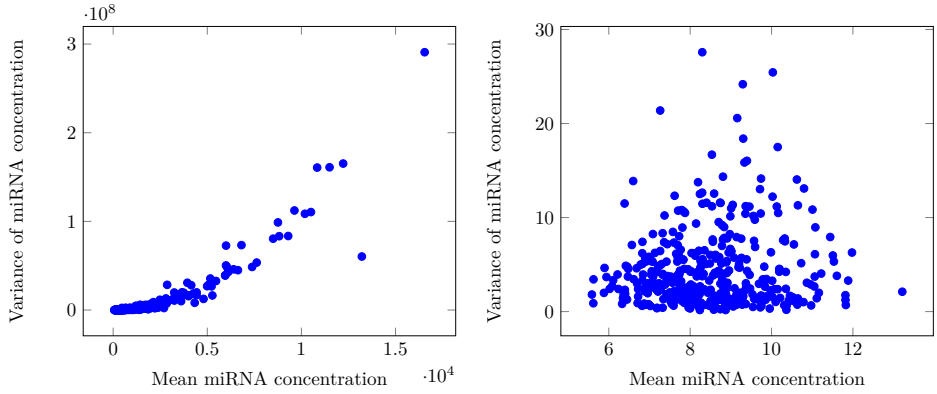
Microarrays

Micoroarrays are what is called a hybridization method. It starts out similarly to qRT-PCR, with converting miRNA into cDNA, only that the miRNA in this case are fluorescently labeled. The microarray has several spots, each with single-stranded DNA samples (called probes) that are mounted to the microarray. When the cDNA are added to the microarray, the cDNA will bind to the DNA samples that have the same sequence, in a process called hybridization. Afterwards, the microarray is washed clean, and only the cDNA that has managed to bind will remain. Thus, by checking for the fluorescence of the different spots, one can find which DNA-probes had cDNA bind to it, and which had not. The level of fluorescence can then be used as the concentration of the corresponding miRNA-sequence.

The main advantage of microarrays is that it is the cheapest of the main technologies [**?**]. The disadvantages is that it has low sensitivity, and that you have to decide beforehand what miRNA-sequences you want to measure, as you need to populate the microarray with the corresponding DNA-probes.

Sequencing

Sequencing also starts with converting miRNA into cDNA. A primer is then connected to the cDNA in one direction. The sequencing step works by adding



(a) Before log transformation

(b) After log transformation

Figure 2.1: Mean and variance in different miRNA sequences in [**?**]

fluorescent bases one by one, and then see if they adds to the sequence starting with the primer. Thus, one can read out the sequence of the cDNA.

The main disadvantage of sequencing is that it is expensive [**?**]. It is also less sensitive than qRT-PCR. The main advantage, however, is that you do not need to decide beforehand the miRNA-sequences you want to measure.

2.2 Machine learning theory

The second major part of this project is the machine learning.

2.2.1 Variance stabilizing transformation

In miRNA measurements, one often see that the variance in miRNA concentration is a function of the mean miRNA concentration. One possible transformation is the log transformation where one takes the logarithm of the data. That can change a curve where $\mathrm{Var}[y] \propto \mathrm{E}[y]^2$ into a curve where the variance of y is independent of the mean of y. One example of this can be seen in Figure 2.1.

Another advantage of a variance stabilizing transformation is to ensure that the data is not skewed. Other statistical tools like explained variance (subsection 2.2.5) assumes that the underlying data has a normal distribution. A normal distribution, however have no skew, therefore unskewing the data is necessary for ensuring that other methods are giving valid results. More formally, if we assume

that Y $g(X)$ for some function g and that $X \sim N(\mu, \sigma)$, then doing the transformation $y' = g^{-1}(y)$ ensures that our variables are normally distributed. In particular, if we assume that $g(X) = e^X$, then the log-transformation will ensure that our data is normally distributed.

2.2.2 Fold change

Fold change is defined as the ratio of a certain value between two different populations. In this project, the fold change used is typically the ratio levels of a certain miRNA-sequence between cases and controls. Log-fold change is the logarithm of the fold change (by convention $\log_2$ is used in this area of research). Furthermore:

$$\text{Fold change} = \frac{a}{b}$$
$$\text{Log-fold change} = \log_2\left(\frac{a}{b}\right) = \log_2 a - \log_2 b$$

In other words, the log-fold change is the difference in miRNA expression when the data are log-transformed.

2.2.3 Loess regression

Loess regression is also sometimes called local regression, and it is a type of regression that is made for smooting scatterplots [**?**]. The regression works by fitting a low degree polynomial for each datapoint. The fitting of each polynomial works by giving weight to nearby points, where more weight is given to points near the original datapoint. The regression value for each datapoint is thus the value of the corresponding polynomial evaluated in this point.

Loess regression is practical when mean and variance still are not independent after a log transformation. Using loess regression can ensure that they become independent as shown in **??**

2.2.4 Principal component analysis

Principal component analysis (PCA) is a method of data reduction, where a dataset in $\mathbb{R}^n$ is projected down on a lower dimensional vector space $\mathbb{R}^m$. The projection in PCA is the projection that ensures that the most of the variance of the original dataset is kept, whilst ensuring that the projection is not expanding the dataset. One of the main advantages of PCA is that you could project a dataset down to just two or three dimensions, which makes it possible to plot the dataset.

2.2.5 Explained variance

Explained variance is a way of analyzing the sources of variance in a dataset. Using linear regression, one assumes that the dependent variable y, covariates $\mathbf{X}$ and residuals $\epsilon \sim N(0, \sigma)$ have the relationship $y = \mathbf{X}\beta + \epsilon$, for some parameter vector β.

If one creates a linear regression model of the dataset one get a parameter vector $\hat{\beta}$ which is the maximum likelihood estimate of β, and predictions $\hat{y} = \mathbf{X}\hat{\beta}$ for y. Also define $\mathbf{SST} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares, $\mathbf{SSR} = \sum_i (\hat{y}_i - \bar{y})^2$ is the sum of squares due to regression and $\mathbf{SSE} = \sum_i (y_i - \hat{y}_i)^2$ is the sum of squared estimate of errors. Then we have the following relationship:

$$\mathbf{SST} = \mathbf{SSR} + \mathbf{SSE}$$

The proportion of the empirical variance that can be explained by the covariates is thus

$$R^2 = \frac{\mathbf{SSR}}{\mathbf{SST}}$$

2.2.6 Logistic regression

Assume that you have Bernulli trials where each $y_i \sim$ Bernoulli(p_i) with the relationship:

$$\frac{p_i}{1 - p_i} = e^{x_i^T \beta}$$

for some covariates x_i and some parameter vector β. Then one can show that

$$p_i = \frac{1}{1 + e^{-x_i^T \beta}}$$

A logistic regression model can find $\hat{\beta}$, the maximum likelihood estimate of β.

Logistic regression is a relatively simple classification model, and in the studies used in this project, logistic regression is the most commonly used model for diagnosing lung cancer based on miRNA levels.

2.2.7 Wilcoxon signed-rank test

Wilcoxon signed-rank test is a statistical test to find the location of a distribution [**??**]. More formally, if $\mathbf{X}_1$ and $\mathbf{X}_2$ are independently distributed with cumulative distribution function F, and

$$p = P\left(\frac{1}{2}(\mathbf{X}_1 + \mathbf{X}_2) > 0\right)$$

then the test is testing the null hypothesis $p = \frac{1}{2}$ against the alternative hypothesis $p \neq \frac{1}{2}$. The median of $\frac{1}{2}(\mathbf{X}_1 + \mathbf{X}_2)$ is called the pseudomedian of F. F is called symmetric about μ if $f(\mu + x) = f(\mu - x)$. Furthermore, F is called symmetric if there exists such a μ. If F is symmetric the median and the pseudomedian is the same. Thus, if F is assumed to be symmetric, the null hypothesis is equivalent to that the median of F is $\mu = 0$, against the alternative hypothesis that $\mu \neq 0$.

If one has pairs $(\mathbf{X}, \mathbf{Y})$ where $\mathbf{X} \sim F_X$ and $\mathbf{Y} \sim F_Y$, with the null hypothesis $F_X = F_Y$, then under the null hypothesis $\mathbf{X} - \mathbf{Y}$ both has a symmetric distribution (as the distribution is the same as $\mathbf{Y} - \mathbf{X}$ as $\mathbf{X}$ and $\mathbf{Y}$ are interchangeable) and it has a median of 0. Let $\mathbf{X} - \mathbf{Y}$ have a cumulative distribution function F_{X-Y}. If $F_X \neq F_Y$, it cannot be guaranteed that F_{X-Y} is symmetric. However, one can instead test the null hypothesis of F_{X-Y} being symmetric around 0 with the alternative hypothesis that F_{X-Y} having a median not equal 0 (but not necessarily being symmetric).

2.2.8 Bonferroni correction

A Bonferroni correction is a correction that is done when doing multiple testing in order to avoid false positives [**?**]. The correction is in order to control for family-wise error rate (FWER). FWER is the probability of at least one type I error when testing several hypotheses simultaneously. The correction is based on Boole's inequality:

$$P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$$

for all possible events A_i. Letting A_i be the event that one makes a type I error in hypothesis i, and letting n be the number of hypotheses gives that FWER $= P(\bigcup_{i=1}^{n} A_i)$. Setting a significance level α_0 for each hypothesis gives that $P(A_i) \leq \alpha_0$. Then:

$$\text{FWER} \leq P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i) \leq n \cdot \alpha_0$$

If we want FWER $\leq \alpha$, this can be achieved by setting $\alpha_0 = \frac{\alpha}{n}$.

2.2.9 Support Vector Machine

Skriv om SVM

A support vector machine (SVM) is a machine learning algorithm that works by creating boundaries in high dimensional vector space [**?**]. The easiest form of SVM is a binary linear SVM. Then the algorithm creates a hyperplane in the data

space that separates the two classes in an optimal way, where there are different loss functions that would lead to different hyperplanes being optimal. If the data is linearly separable, then a linear SVM would create a perfect separation.

However, the data is often not linearly separable, but might be separable with a more complex boundary. In these cases one can use a kernel SVM. A kernel SVM would map the data onto a higher dimensional space, where the data is linearly separable [**?**]. A common kernel, that is default in `scikit-learn` is the radial basis function (RBF) kernel [**?**]. RBF is a kernel that is based on radial distances between points, where points have exponentially less influence on each others as the distance between points grow.

2.2.10 Random Forest

A decision tree is a tree like model, where at each node the next node is further down the tree, and which node is next is based on the decision criterion in the current node. The decision criterion is a condition on your data point. Finally, leaf nodes have the classification of the data point. Decision tree learning is when one learns these criteria based on data, in order to find an optimal classification [**?**]. A random forest is a classifier where one trains multiple trees, where each three is trained on a subset of the training set. Having each tree being trained on only a subset of the training set is a way of reducing overfitting. The overall classification is then given by aggregating the results from the classification given by the different trees [**?**].

2.2.11 XGBoost

XGBoost is a machine learning algorithm that is based on gradient tree boosting [**?**]. Boosting algorithms are machine learning algorithms that combine weak models into stronger models, by using the combined output of several weak models. Gradient boosting is a type of boosting algorithm that uses an idea similar to gradient descent in order to find optimal weights given to each of the weaker models [**?**]. Instead of using the gradient directly, the algorithm just ensure that the weights are updated such that the loss function is lowered in each step. Gradient tree boosting is gradient boosting where the weak models are decision trees. XGBoost's decision trees have default directions of descending in the tree if there are missing data, thus good handling missing values is one of XGBoost's biggest advantages.

XGBoost is a popular machine learning algorithm on the machine learning contest site Kaggle[1], winning 17 of 29 contests in 2015 [**?**].

[1] `https://www.kaggle.com`

2.3 Structured Literature Review Protocol

The point of the literature search was to find studies relevant to miRNA and circulating lung cancer. The main search engine used was PubMed[2], which is a commonly used search engine for medical litterature. The search term used was:

```
(lung OR pulmonary OR NSCLC) and (tumor OR cancer OR carcinoma) and
(microRNA* OR miRNA* OR miR*) and (diagnosis OR biomarker OR detection)
and (serum or plasma or "whole blood")
```

In addition, I search in databases that have public gene expression data as shown in Table 2.1.

Database name	Search term
ArrayExpress[3]	`microrna lung cancer`
Gene Expression Omnibus (GEO)[4]	`(mirna OR microrna) AND "lung cancer" AND (diagnosis OR detection)`
OmicsDI[5]	`"lung cancer" AND TAXONOMY: 9606 AND "breast cancer" AND (mirna OR microrna) AND (serum OR plasma OR "whole blood")`

Table 2.1: Search in public gene expression databases

The inclusion criteria where based on what datasets I thought was relevant to this project:

- The paper is an experiment where circulating miRNA is measured.

Some of the studies measured miRNA levels in the lung tissue or in sputum, rather than measuring circulating miRNA. As the values are somewhat different between lung tissue miRNA and circulating miRNA [**?**], only the circulating miRNA ones was selected in order to have a consistent dataset. In addition, the research question was to look at the diagnostic value of circulating miRNA, which makes it most reasonable to base on circulating miRNA data.

- The study both have people diagnosed with lung cancer and controls not diagnosed with lung cancer.

[2] `pubmed.ncbi.nlm.nih.gov/`
[3] `https://www.ebi.ac.uk/arrayexpress/`
[4] `https://www.ncbi.nlm.nih.gov/gds`
[5] `https://www.omicsdi.org`

The controls in some of the studies are not healthy, but suffers from other kind of lung diseases. Other studies have both have both healthy controls, and controls with other lung illnesses. Both are relevant, as on one hand, one would like to see the difference between healthy controls and patients with lung cancer in order to remove miRNA changes due to other illnesses. On the other hand, people who are getting checked for lung cancer often have lung issues, which is the reason for their checkup.

Some studies were excluded as they did not have a control group like **?**.

- At least four different miRNA sequences were measured.

The point of this project is to combine and compare datasets. Having few miRNA sequences measured makes it hard to combine datasets, as there is a high likelihood that there are no overlapping miRNA sequences between the datasets. It is also hard to compare datasets measuring completely different miRNA sequences.

- Meta-analyses were used as source of relevant studies

Some of the studies found were meta-analyses. In that case relevant studies were retrieved from the references of the meta-analysis.

Chapter 3

Methodology

For a description of the literature review and the data processing, see **?**.

This project consists of the following steps:

- Finding one or more miRNA-sequences that have been found to be a consistent biomarker for lung cancer using meta analyses.
-

3.1 Technical setup

In Table 3.1, the main software used in this project is listed.

Software	Version	Usage
Python[1]	3.9.7	Programming language
NumPy[2]	1.20.3	Numerical calculations with vectors and matrices
scikit-learn[3]	0.24.2	Machine learning
XGBoost[4]	1.4.2	XGBoost machine learning algorithm
SciPy[5]	1.7.1	Scientific programming

Table 3.1: Software used in this project

[1] `https://www.python.org`
[2] `https://numpy.org`
[3] `https://scikit-learn.org/`
[4] `https://xgboost.readthedocs.io/`
[5] `https://scipy.org`

3.2 Log-fold-change correlation

? showed that there are little log-fold-change correlation in the data. Furthermore, it showed that even though the correlation direction was arbitrary, there was significant correlation between the datasets. However, the magnitude of the correlation was similar when randomizing the column corresponding to case-control, which means that the correlation could be partially be as a result of the covariance between different miRNA-sequences rather than due to case-control characteristics. The lack of correlation could be due to characteristics with the different studies, in particular technology used and what body fluid is measured. Therefore, a relevant experiment would be to see whether there is a larger correlation between datasets using the same technology and body fluid, contrasted with the correlation when the datasets differ in these characteristics.

Due to the limited amount of datasets in this project, grouping based on both characteristics would give too low statistical power to make any conclusions. Therefore, I will first group by technology and group by body fluid afterwards. I will use a t-test between the calculated correlations when both datasets are in the same category, to the correlation when only one of the datasets is in the certain group. To avoid spurious correlations, only pairs of datasets with at least 10 miRNAs in common are considered. The correlation is calculating using Pearson's r.

3.2.1 Using stages

There are some evidence that suggest that the diagnostic accuracy of microRNA is somewhat higher in later stages of lung cancer [**?**]. Thus, it might be valuable to check whether there is a higher log-fold-change correlation when only using cancer samples with advanced stages, in this case stage 3 and 4. I.e. the log-fold-change is the difference between the mean expression of the cancer samples in advance stages and the controls. If the correlation is better when considering later stages, that might suggest that later stages have more consistent expression, and thus are easier to diagnose. The test will be a t-test of the correlation coefficients when only stage 3 and 4 are considered, compared to the correlation coefficients when only stage 1 and 2 are considered.

3.3 Evidence for consistently differentially expressed miRNA-sequences

One question that has to be considered, especially given how **?** found that sign of the log-fold-correlation was virtually arbitrary, is whether there are evidence that

there exist any evidence that there exist any consistently differentially expressed miRNA-sequences at all. By consistently, I mean that it is generally up- or down-regulated. **??** shows that some miRNA-sequences are often differentially expressed, but that the direction is not consistent, which would make diagnosis hard and lead one to question whether the differential expression was primarily due to case-control characteristics. It is difficult to rule out that there exists any consistently differentially expressed miRNA-sequences, especially as many of the miRNA-sequences are only present in a few datasets, which would mean that it would be hard to say whether the differential expression is due to chance or study characteristics, or if it is due to case-control characteristics.

3.3.1 Paired sign test

The way I will try to do it is to estimate the following probability that a miRNA-sequence is differentially expressed in a certain way (either up- or down-regulated) in one dataset given that it is differentially expressed in the same way in another dataset. This will be done by looking at each pair of study where both have a certain miRNA. By looking at all these pairs, it is possible to calculate the wanted conditional probability. One question is whether one should only consider pairs where both miRNAs that are significantly differentially expressed, i.e. a p-value less than 0.05 on a t-test of the log-fold-change, not. One advantage of only considering significantly differentially expressed miRNAs is that when the difference is not significant, it is more likely that the sign of the difference is only due to chance. On the other hand, if a miRNA is significantly up-regulated in one study, but not significantly regulated in another study, this lowers the consistency of the differential expression. Due to advantages and disadvantages with both options, I will report using (1) only pairs where both are significant, (2) pairs where at least one is significant and (3) all pairs, and corresponding p-values that the conditional probability is larger than 0.50 using a binomial test.

3.3.2 Signed-rank-test with cross validation

Another way to find whether there are any consistency in the differential expression of miRNA is to use Wilcoxon signed-rank test (see subsection 2.2.7). This will be done by looking at the log-fold-change in a miRNA-sequence across different studies, and then use the signed-rank test to find whether the miRNA-sequence is significantly up- or down-regulated across studies, by looking at whether the signed-rank test of median differential expression of the miRNA-sequence.

Using t-test results as values for signed-rank-test

As seen in **?**, there is a large difference in the number of samples in the different datasets, which means that using raw log-fold-chance might lead to small datasets having a big impact on the signed-rank-test due to chance. Therefore, I will also do an experiment where instead of using log-fold-change in the signed-rank-test, I will use the p-value of a t-test instead. Then datasets with more samples get a larger impact as they have more statistical power. More formally, I will do a two-sided t-test of the log-fold-change and use

$$\frac{sign(t-value)}{p-value}$$

as the value for the signed-rank-test. Notice that the sign is the same as the log-fold-change, and that it is inverse proportional to the p-value. As the signed-rank-test only considers the rank of the value, and not the absolute value, any function decreasing in increasing p-values would work, including this.

Cross validation

Firstly, to ensure external validity of the results of the signed-rank test, I will do a test where I do a signed-rank test on all studies except two that are exempted. Then I will look at the 10 most and 10 least consistently differentially expressed miRNA-sequences based on the signed-rank test, using only miRNA that are in least ten of the studies, where these 20 miRNA-sequences are also in the two other studies. If two studies does not have at least 20 miRNA-sequences in common that are in least ten of the other datasets, the pair of them will not be used as left out datasets together. Otherwise, all pairs of two datasets will be tried as exempted datasets. If there is a larger consistency in the two exempted datasets in the expression of the miRNA that was had most consistency in the signed-rank test, that would suggest that the signed-rank test has external validity. The consistency in the two exempted dataset will be calculated similarly to subsection 3.3.1, i.e. the proportion of miRNAs that have the same direction of differential expression is compared between the 10 most and 10 least consistently differentially expressed miRNA-sequences in the signed-rank test.

Finding most consistently differentially expressed miRNAs

By using the signed-rank-test on all the datasets, one can find the miRNA-sequences that are the most consistently differentially expressed in the datasets. This will both be done using log-fold-change and using t-test results as the value in the signed-rank-test. Thereafter, I will find the 10 most consistently differentially expressed miRNAs using each of the two possible metrics for the change in

miRNA expression. The p-values will be adjusted using a Bonferroni correction, to adjust for the multiple testing.

3.4 Are datasets separable from each other?

One question arises when the consistency between the datasets is as poor as it has been shown to be in **?**, namely can one recognize what dataset a sample is from? Given that there are differences between the datasets, can one use these differences to recognize a dataset. The way I will test this is using logistic regression on a pair of datasets where one third of each dataset is used for testing and two thirds is used for training. The model will be trained to separate samples from the two datasets from each other. Only pairs of datasets with at least 10 miRNA-sequences in common are considered. The metric to evaluate the separation is AUC.

3.5 Hierarchical clustering of datasets

Hierarchical clustering of miRNA expressions is a common analysis in this field. As I want to find the comparability of the datasets, I will rather try to cluster the datasets. This would not only give information about what datasets are more comparable, but also whether there are clusters of datasets that are closer to each other, and in that case, what characterize them. This is somewhat similar to the analysis in **?** where **?** created a graph of what datasets were similar. The difference is that the hierarchical clustering will give clusters of datasets that are similar to each other, rather than just comparing pairs of datasets.

The clustering will be computed using `scipy.cluster.hierarchy.linkage` in SciPy with "ward" as method. The distance will be the mean of the squared difference in log-fold-change for each miRNA-sequence that the two datasets have in common. I.e. if x_i and y_i are the log-fold-change in miRNA i in the two datasets, and there are n miRNA-sequences in common between these datasets. Then the distance is

$$\text{dist(x, y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

The results will be visualized in a dendrogram.

3.6 Machine learning on single datasets

One question that was asked in section 1.2 was whether more advanced machine learning algorithms are better at diagnosing lung cancer based on miRNA-levels.

Therefore, I have chosen to do machine learning on single datasets. As seen in **?**, the results when doing machine learning across datasets were mostly poor. Furthermore, if the connection between miRNA expression and lung cancer is very sensitive to study characteristics, machine learning across different datasets might not be the best idea, compared to training on data where the characteristics are known to be the same. I have chosen four different types of machine learning algorithms to test.

- Logistic regression: Is natural to use as a baseline model to compare against, as it has been used in many of the studies that is included in this project
- Support vector machine: If the data is nearly linearly separable (which the PCA-plots in **?** suggest), this will find a such separation.
- Random forest:
- XGBoost: Has had the most success in tabled data with limited samples in Kaggle competitions (see subsection 2.2.11), and is thus a natural algorithm to test.

The models will be tested using AUC, and the AUC will be calculated using cross validation where the dataset is split into five equal parts and for each of the five parts, there will be a round where the model is trained on the four other parts of the dataset and tested on the last part. The resulting AUC will be the average over the five rounds.

3.6.1 Using stages

Another question is whether there are any difference in AUC between early and late stage cancer. Therefore, I will train and test a logistic regression classifier on datasets where stage is labeled, using either only late stage cancer or only early stage cancer. The training and testing will follow the same cross validation strategy as in section 3.6.

3.7 Baseline miRNA-sequence

One question, as asked in section 1.2, was what miRNA-sequences would be most successful in diagnosing lung cancer. This has not only clinical relevance, but is also important to note as a machine learning model would be more powerful than using one single miRNA-sequence. There will always be additional costs associated with measuring more miRNA-sequences, and therefore I need to show that a machine learning model will perform better than a model based on a single miRNA-sequence.

There are two main types of methods possible for finding such miRNA-sequences, each with some pros and cons:

1. Look at meta analyses for finding miRNA-sequences that are found to diagnose lung cancer well across studies.

   Pros:

   - The miRNA-sequences found would be based on more data, and thus they are likely better.
   - The miRNA-sequences are nearly[6] independent of the datasets used in this project, and are therefore mostly unbiased.

   Cons:

   - The miRNA-sequences that are reported in these meta analyses are often not in many of the datasets used in this project.

2. Look at the datasets used in this project to find miRNA-sequences that can separate cases and controls across the different datasets.

   Pros:

   - It is easier to limit the search to miRNA-sequences that are found in many of the datasets in this project.

   Cons:

   - The miRNA-sequences are biased, as the baseline is the ability of these miRNA-sequences to diagnose cancer on the datasets, but the miRNA-sequences were chosen because they diagnosed well on said datasets.

I want to try a hybrid strategy, in order to mitigate the cons of each method. That is, I want to try to find an intersection between microRNA-sequences that have been found in meta analyses to be consistently good at diagnosing lung cancer and microRNA-sequences that separate well in the studies used in this project.

There are several ways to measure to which degree a miRNA-sequence can be used to separate cases from controls. One possibility would be to use the t-statistic. The advantage with the t-statistic is that it has a known distribution (given the null hypothesis), and thus one could get to know whether a difference is plausibly a result of chance or not. The disadvantage of the t-statistic is that it does not only measure to which degree the miRNA-sequence separates well

[6] After all, the meta analyses might be based on some datasets used in this project

in the dataset, but also the statistical power of each dataset. Therefore, large datasets would be given more weight, and the value could hide to what degree the miRNA-sequence diagnoses correctly in the dataset.

Another alternative is to use Cohen's d. The advantage of Cohen's d is that it tells to what degree cases and controls are separated independent of the number of samples in the dataset. The disadvantage of Cohen's d is that it does not consider the statistical power at all, and thus one might expect a number of spurious results when using Cohen's d. A final statistic is to use AUC. The advantages and disadvantages are similar to Cohen's d, with the difference that AUC has the advantage that it is the metric that the results will be measured against in the end. However, Cohen's d has the advantage that it also looks at the absolute difference in expression, and not just whether there is a separation like AUC does.

After a consideration of the different statistics, I found that Cohen's d and AUC will be the most appropriate statistic for this purpose, as the t-statistic would give too much power to the large datasets (which might not be representative at all), and not tell the actual degree of separation.

Finally, I will go through the datasets to find what is the best biomarker for lung cancer in my datasets. This will be done by t-tests, looking at what are the

3.8 PCA analysis across datasets

PCA-plots of all the datasets are found in **?**. However, a PCA analysis with multiple datasets is still to be done. Doing a PCA-analysis, there might be several principal components that separate cases and controls well, but they might not be possible replicate across datasets, as there are study specific reasons that the principal components separate well. For exploration purposes and to get good statistical power, I will first do a study where I only look at **?** and **?**, as they have most samples. I will calculate the ten largest principal components in each dataset, using the miRNA-sequences that are in the both datasets. Then I will project every sample along the ten principal components, and using a t-test I will find whether there is a significant difference between cases and controls along the principal components for each dataset.

3.9 Machine learning using multiple datasets

As said in section 1.2, one of the goals of this project is to find whether combining multiple datasets will result in better diagnostic accuracy than using a single dataset. The result of training on one dataset and predict on another dataset was done in **?** with subpar results. However, it is possible that training on

multiple datasets will help the machine learning algorithm to find case-control patterns that transcend the patterns that are found internally in one dataset, leading to better generalizability.

3.9.1 Using the most replicated miRNA-sequences from the meta analyses

One option is to use the miRNA-sequences that are found in the meta analyses to be the best biomarkers for lung cancer across studies and train a model using these miRNA-sequences. One problem is that probably not all datasets have all the miRNA-sequences. Therefore, I will select the datasets that have all these miRNA-sequences and train a logistic regression model using leave one out cross validation, where one dataset is chosen in each iteration, which is used as test data, whilst the model is trained on the other datasets. The samples will be weighted so that each dataset has the same weight. The samples will be weighted so that the sum of weights in each dataset is the same, and the weight of all samples in the same dataset are the same.

3.9.2 Training on two datasets

I will train different machine learning models on two datasets and try to predict on a third dataset, and compare the results to the results that are found by training the model on only one of the models, and then predict on the third dataset. The results will only be considered if the three datasets have at least 10 miRNA-sequences in common in order to ensure the datasets are similar enough. The samples will be weighted so that the sum of weights in each dataset is the same, and the weight of all samples in the same dataset are the same.

Logistic Regression

The first model I will try is logistic regression as it is a basic classification model, and it is often used in the studies that tries to predict cancer based on miRNA, it therefore serves well as a baseline. The model will be trained on the miRNA-sequences that all the three datasets have in common.

XGBoost

It is plausible that a model like XGBoost will perform better on the datasets, as it has methods for handling missing data, and it can handle non-linear relationships in the data. In addition, it is a boosting algorithm, which usually performs well when data is sparse, as in this case. Here, I will make use of the way XGBoost

handles missing data and therefore train the model on all the miRNA-sequences that the two first datasets have in common.

3.9.3 Merging all datasets together

There are no miRNA-sequence that are in all datasets. Thus, if one tries to merge all the datasets together, there would not be any miRNA-sequences that all the datasets have. This would be a problem with logistic regression, but XGBoost has a way of handling missing data, which means that it can handle this. However, merging all datasets would not be useful, as one want to see whether one can train on some collection of datasets and test on another dataset, in order to check generalizability. Therefore, I will have a strategy similar to subsection 3.9.1, where I leave one dataset out which is used as a test set. Here, every sample would have the same weight, as there would not be any problem with have one dataset dominating as there are many datasets in the training set.

3.9.4 Maximal training set

Unfortunately, trying to train a model on all possible subsets of datasets is computationally infeasible. There are 25 datasets in this project, leading to $2^{25} = 33554432$ possible subsets of datasets. Thus, there has to be some sort of selection of what subsets are interesting to look at. The cases of training on two datasets, and training on all datasets except one, has been described above. There are some pros and cons associated with merging more datasets together:

Pros:

- More datasets lead to more samples, which gives greater statistical power.
- More datasets might lead to better generalizability for the algorithm, as it has to learn what is common across different datasets.

Cons:

- The more datasets are merged together, the larger the problems with different miRNA-sequences in each dataset. If one takes the intersection of miRNA-sequences, this intersection quickly becomes small; and if one take the union, one would end up with a lot of NaN-values.
- Using fewer datasets would mean that results would give information of the properties of the datasets included, which would be lost if many datasets are merged together.

One way to balance these conflicting concerns would be to find a compromise. I want the algorithm to be trained on at least 10 miRNA-sequences (similar to

subsection 3.9.2 and ?) in order to ensure that the algorithm have some different miRNA-sequences to consider. On the other hand, I want to merge the most datasets together. One possibility then is to generate all subsets that satisfy the following two criteria:

1. The datasets have at least 10 different miRNA-sequences in common.
2. If you add another dataset to the subset, you would end up with less than 10 miRNA-sequences in common.

These subsets might be called maximal subsets.

3.10 Stratification of the datasets

There are several possibilities to why the datasets are incompatible. One possibility is that some factors like what technology was used for measuring miRNA-levels play a role. There are other factors as well that differs between the datasets, like cancer stage and what body fluid was measured (plasma, serum, whole blood, etc.). If these factors play a role one would expect to see more consistency in datasets that are similar on these points. One way to test this hypothesis is to divide the datasets based these characteristics, and see if one sees a larger consistency between the datasets when the datasets are stratified in this way.

3.10.1 Training on two datasets, in-group vs. out-group

Here I will do something similar to subsection 3.9.2, only that the AUC will be compared when the datasets have the same characteristics to when they have different characteristics. I.e., I will compare the AUC when two datasets are using qRT-PCR to when one is using qRT-PCR and the other study is using a different technology. I will do this stratification for technology and for body fluid measured. Ideally, I would use it for stage of cancer too, but unfortunately the stages of cancer have many partial overlaps, which would make this kind of study hard and imprecise. Here I will use logistic regression and only use when the datasets have at least 10 miRNA-sequences in common (like subsection 3.9.2).

Using stages

It is possible that cancer stages is a covariate that hinders the replicability of the datasets. In order to check this hypothesis, I will do an analysis where I only use the datasets where samples are labeled, and only use cancer samples from stage 3 and 4. If there is higher consistency here, it would suggest that some of the lack of replicability is due to cancer stages.

3.10.2 Combining all except one

Another attempt will be to take all datasets with a certain characteristic, like technology or body fluid, and then train on all datasets except one that will be used for testing. This is similar to subsection 3.9.3, only with stratification of the datasets. I will use the union of the miRNAs in the datasets in each category to train on. To ensure that missing values will not be a problem, I will use XGBoost as model as it handles missing values by default. I will also try to do this using datasets where cancer stage is labeled, and try both using only early cancer samples and using only late cancer samples.

3.11 PCA for removing technical noise

The measured miRNA-levels will have some noise that is due to the technology that is used for measurement. One possible way to remove technical noise is to remove the first principal components from the data, with an assumption that the removed principal components correspond to technical noise rather than biology. It is difficult to say whether this is the case or not. If the datasets are more comparable when the principal components are removed that when they are not, then it would seem plausible that these principal components correspond to technical noise, or are at least have little to none connection with lung cancer.

3.11.1 Check comparability using PCA

One way to check if the datasets are more comparable is to check their joint PCA-plot. There are some problems with that. For once, the miRNA-sequences are not the same in the different datasets, which means that the principal components will not represent the datasets fully. Another problem is that there are many datasets. The more datasets that are plotted in one PCA-plot, the more chaotic the plot becomes, the fewer miRNA-sequences there are in common, and the less weight each dataset would have on the PCA on the joint dataset. On the other hand, I cannot plot all datasets against all datasets, as that would lead to too many PCA-plots. Therefore, I will plot two and two datasets in PCA-plots and see whether they became more comparable, using only the largest datasets.

3.11.2 Using machine learning models

The results from combining all datasets using sequencing generally lead to good results (see section 4.10.2), and will therefore be used to check if removing the principal components lead to better comparability, as you would assume that a dataset where noise is removed would be a dataset that would be comparable to

the sequencing datasets, as the sequencing datasets seem to have high external validity when compared to other sequencing datasets.

The way this will be done is for each dataset that is not using sequencing, I will find the miRNAs that they have in common with all the sequencing datasets. Thereafter, using these miRNAs I will do a leave-one-out cross validation on the sequencing datasets, similarly to subsection 3.10.2. I will also do a cross validation on the other dataset, similar to section 3.6. Finally, I will train a model on the other dataset and test on the sequencing datasets, and visa versa. All the machine learning models will be XGBoost, as it was that model that performed well in section 4.10.2.

Chapter 4

Experiments and Results

This section will contain the results from the experiments.

4.1 Studies included

The studies included in this project are **?**, **?**, **?**, **?**, **?**[1], **?**. An overview of the studies is found in **?**[2].

4.2 Log-fold-change correlation

The results from when checking the log-fold-change correlation when comparing the datasets where both datasets are in the same category, contrasted with when only one of the datasets are in the category, are shown in Table 4.1. The experiment is described in more detail in section 3.2. The in-group is when both datasets are in the group, and the out-group is when only one of the datasets are in the group. There was no significant change between the in-groups and the out-groups in any of the cases. This might be due to lack of statistical power as the correlations were generally better in the in-group, with serum as an outlier. However, according to **?**, the correlation seemed to be due to covariance between miRNA expressions rather than due to case-control characteristics. Therefore, I will replicate the case-control randomization done in **?**, but only for the in-groups.

The results from randomization of case-control characteristics is shown in Table 4.2. The results are similar to Table 4.1, which suggests that the correlation

[1] **?** is not the study where the dataset originated from, but it is a study using the dataset. The dataset is GSE71661 in the Gene Expression Omnibus, and has no citation listed: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71661`

[2] **?** is not described in **?** as the data arrived too late for it to be included.

Group	Mean IG	Mean OG	t-value	p-value
Microarray	0.027	−0.007	1.227	0.221
Sequencing	0.070	−0.008	1.099	0.275
qRT-PCR	0.068	−0.017	1.160	0.249
Serum	−0.018	0.008	−0.423	0.673
Whole blood	0.115	−0.013	2.242	0.027
Plasma	0.024	0.011	0.305	0.761

Table 4.1: Pearson's r of the log-fold-change between pairs of datasets.
Note: IG = in-group, OG = out-group

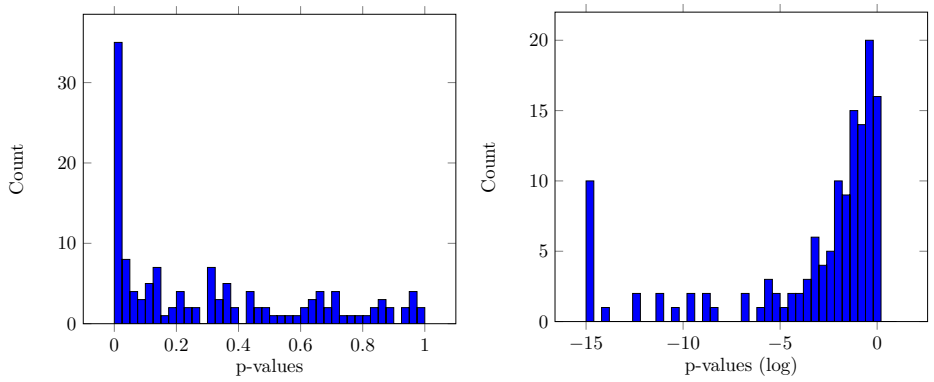
Group	Mean Non-shuffled	Mean Shuffled	t-value	p-value
qRT-PCR	0.068	0.033	0.394	0.698
Microarray	0.027	0.026	0.047	0.962
Sequencing	0.070	−0.005	1.330	0.213
Plasma	0.024	0.036	−0.184	0.855
Serum	−0.018	0.024	−0.557	0.584
Whole blood	0.115	−0.017	1.758	0.090

Table 4.2: Pearson's r of the log-fold-change between pairs of datasets inside each group when case-controls characteristics are shuffled and not shuffled

in each group is not primarily a result of covariance between miRNAs. There is a difference between the experiment done here and the experiment in **?**, namely that here I look at the direction of the correlation, while **?** primarily looked at the significance of the correlation. It might be that the case-control characteristics are the cause of the direction of the correlation, whilst covariance between the miRNA-sequences is the cause of the significance of the correlation.

It is hard to test the last hypothesis as the p-values do not have a known distribution. They are not uniformly distributed as they are significant correlation between the datasets. They are also far from normally distributed, due to a very strong left skew, which would make a t-test give misleading results. One possibility might be to log-transform the p-values. The results are shown in **??**. Neither are normally distributed, but the distribution of the log-transformed p-values seem closer to a normal distribution than the non-transformed p-values.

The experiment thus becomes to look at the log transformed p-values when randomizing the case-control assignment with the log transformed p-values without randomization. Then a t-test will be performed to check for possible difference in the p-values between the randomized and the non-randomized case. The



(a) Non-transformed (b) Log-transformed

Figure 4.1: p-values of the log-fold-change correlation between each pair of studies with the same technology or the same body fluid

t-test showed that the correlation was more significant when the case-control assignment was not randomized ($p = 0.026$), which suggests that case-control characteristics is the cause of some of the significance in the correlations, rather than it being only due to covariance between miRNA expressions.

When the datasets were grouped, there was a skew towards better correlations in in-groups than out-groups, but the difference was not significant. As it might have been significant with more statistical power, one test is to test all correlations that are in in-groups to all correlations that are only in out-groups. This would have more statistical power, with the disadvantage that it does not say anything about which groups have more internal consistency, as it seems like it varies between the groups.

The result was that there was still no significant difference between the groups ($p = 0.138$), which suggests that the increased correlation in in-groups is either non-existent or too small to be found with the current number of datasets. Either way, the mean correlation in the in-group was $r = 0.025$, which is a very small correlation that would suggest that case-control characteristics' effect on log-fold-change is either much smaller than the other effects, or the effects cannot be replicated across datasets. Indeed, **?** shows using linear regression that the proportion of variance in the miRNA expression that is due to case-control characteristics varies and is generally quite small. The highest proportion is 0.446 in **?** and the smallest is 0.009 in **?**. The proportion found by linear regression is probably overstating the actual proportion due to overfitting to the data, as the proportion of explained variance was smaller in datasets with larger sample sizes

Mean Late	Mean Early	t-value	p-value
0.037	−0.041	0.757	0.457

Table 4.3: Pearson's r of the log-fold-change between pairs of datasets when only stages 3 and 4 are considered compared to when only stages 1 and 2 are considered.

[**?**].

4.2.1 Using stages

The datasets where stage is marked are **?????????**. However, **?** only have samples in stages 1 and 2, and is therefore not be used in analysis. The results when only using advanced stages, as described in subsection 3.2.1, are shown in Table 4.3. This shows that there are no significant difference in the log-fold-change correlation when only considering late stages compared to when only considering early stages. There was a small insignificant improvement, though, which makes it hard to say that there are no improvement at all, especially as the sample size here was quite small. However, the results suggests that there is no large improvement in log-fold-change correlation when using only late stage cancer. Indeed, **?** suggested that the improvement in diagnostic value of miRNA in late stage cancer was relatively small.

4.3 Evidence for consistently differentially expressed miRNA-sequences

Here are the results from trying to find evidence for consistently differentially expressed miRNA-sequences.

4.3.1 Paired sign test

The calculated probabilities that a miRNA-sequence is differentially expressed in the same way in two different studies are shown in Table 4.4. All probabilities are significantly higher than 0.50, which gives evidence that there are miRNA-sequences that have a differential expression that have a differential expression that is more consistent than chance levels. Especially the case when only considering pairs when both were significant one achieves a probability that is somewhat better than chance. On one hand, this suggests that there are some consistencies between datasets, which could mean that there might be possible to find a model

Pairs	Probability	p-value
All pairs	0.504	0.035
One significant	0.511	8.938×10^{-5}
Both significant	0.560	4.885×10^{-19}

Table 4.4: Conditional probabilities as explained in subsection 3.3.1 and p-values corresponding to the alternative hypothesis that this conditional probability are greater than 0.50

that diagnose better than chance across datasets. On the other hand, in 44% of the cases there are inconsistencies in the direction of differential expression of the miRNA-sequences. One problem is that this would limit the results from naïve machine learning, as a machine learning algorithm might find one miRNA to be a good separator in one dataset, and if this miRNA is differentially expressed in the test dataset, there is a 44% chance that the miRNA-sequence will separate the cases and controls in the wrong way! One important question that arises is why this happens in the pairs where both are significant? Are there differences between cases and controls across studies? Are they due to differences in technology or body fluid? Are they due to chance?

Stratification of the datasets

If the differences are due to technology or body fluid, one might assume that the consistency is higher when only checking against datasets where these characteristics are similar. An analysis checking only significant pairs where both are in studies with the same characteristic, and the results are reported in Table 4.5. Microarrays and whole blood give significantly better than chance levels. Still the results are worse than when doing no such grouping as in Table 4.4, which suggests that differences in technology or body fluid are not the cause that significantly differentially expressed miRNA are often differentially expressed in different direction, at least on this crude level of grouping. It is possible that more subtle differences in body fluid or technology is the cause, but that it will only be apparent with a finer level of grouping. However, if that was the case one would still expect a coarse level of grouping to give an improvement over no grouping at all, and thus this is evidence that weakly suggest that this hypothesis would not be true.

Group	Probability	p-value
qRT-PCR	0.436	0.985
Sequencing	0.550	0.171
Microarray	0.524	4.48×10^{-9}
Serum	0.517	0.143
Plasma	0.477	0.870
Whole blood	0.551	1.37×10^{-5}

Table 4.5: Conditional probabilities as explained in subsection 3.3.1 and p-values corresponding to the alternative hypothesis that this conditional probability are greater than 0.50, when only using pairs where both miRNA-sequences in a pair is in the same group

Possible significance levels

The differences might be due to chance. I sat a significance level of $p = 0.05$, but due to the number of miRNAs, this will lead to many false positives, especially if the number of miRNA that are differentially expressed due to case-control characteristics is low. Due to that, I tried with different significance levels, and the results are shown in Table 4.6. It seems like this hypothesis was unlikely as the probability does not seem to increase when the significance level decreases. Therefore, one could conclude that chance is not the reason for these differences.

4.3.2 Signed-rank-test with cross validation

Here are the results from the signed-rank-test including when cross validation was used.

Cross validation

The experiment is described in section 3.3.2. The results are shown in Table 4.7. The results suggest that the miRNA-sequences that were the least consistently differentially expressed in the signed-rank-test were somewhat more consistently differentially expressed in the two excluded datasets. This contrasts with the assumption that the miRNA-sequences most consistently differentially expressed in the signed-rank-test would be the most consistently differentially expressed in the two excluded datasets. Why does this happen? One hint may lay in subsection 4.3.1, which suggest that the consistency in significantly better if only looking at pairs where both are significantly differentially expressed.

Significance level	**Probability**	**p-value**
5×10^{-2}	0.560	4.88×10^{-19}
5×10^{-3}	0.578	2.38×10^{-11}
5×10^{-4}	0.51	0.308
5×10^{-5}	0.533	0.0862
5×10^{-6}	0.524	0.222
5×10^{-7}	0.549	0.0811
5×10^{-8}	0.549	0.113
5×10^{-9}	0.554	0.117
5×10^{-10}	0.586	0.0435
5×10^{-11}	0.583	0.0572

Table 4.6: Conditional probabilities as explained in subsection 3.3.1 and p-values corresponding to the alternative hypothesis that this conditional probability are greater than 0.50, when only using pairs where the differential expression is significant on the given significance level

Most significant	**Least significant**	**t-value**	**p-value**
0.486	0.534	-3.25	0.001 14

Table 4.7: The proportion of pairs that had the same direction of differential expression in the two excluded datasets, among the miRNAs that were shown to be most and least consistently differential expressed in the signed-rank test as described in section 3.3.2. The t-value is the t-value for the difference between the two proportions, and the p-value is the corresponding p-value.

Most significant	Least significant	t-value	p-value
0.482	0.502	−0.423	0.672

Table 4.8: The proportion of pairs that had the same direction of differential expression in the two excluded datasets, among the miRNAs that were shown to be most and least consistently differential expressed in the signed-rank test as described in section 3.3.2. The miRNA sequence had to be significantly differentially expressed in the two excluded datasets in a t-test. The t-value in the table is the t-value for the difference between the two proportions, and the p-value is the corresponding p-value.

Most significant	Least significant	t-value	p-value
0.505	0.495	0.709	0.478

Table 4.9: The proportion of pairs that had the same direction of differential expression in the two excluded datasets, among the miRNAs that were shown to be most and least consistently differential expressed in the signed-rank test as described in section 3.3.2, with the difference is that the p-value of a t-test of the log-fold-change was used instead of the log-fold-change. The t-value in the table is the t-value for the difference between the two proportions, and the p-value is the corresponding p-value.

Only pairs with significantly differentially expressed miRNAs Now I will only consider pairs where both are significantly differentially expressed in the two excluded datasets using a t-test and a significance level of $p = 0.05$. The results are shown Table 4.8. The results show that there are no significant difference in the proportion of the pairs with same direction of differential expression. This suggests that the lack of improvement in the proportion in section 4.3.2 was not the result of insignificance in the differential expression in the pairs.

Using t-test results instead of log-fold-change in signed-rank-test One possible reason for the results in section 4.3.2 could be that small studies have a big log-fold-change due to chance. Therefore, I will also try using t-test results in the signed-rank-test as explained in section 3.3.2. The results are shown in Table 4.9. Neither here was there any signs of external validity in the results from the signed-rank-test. Thus, it might seem that the signed-rank-test is a subpar way to find what miRNAs are the most consistently differentially expressed.

MiRNA	p-value	Direction
miR-769	0.319	Up
miR-532	2.63	Up
miR-30e	2.73	Up
miR-548d-5p	3.19	Up
miR-630	3.19	Up
miR-1183	3.19	Up
miR-326-3p	3.19	Up
miR-200c-3p	3.43	Up
let-7f-3p	4.47	Up
miR-2110	4.47	Up

(a) Using log-fold-change

MiRNA	p-value	Direction
miR-769	0.319	Up
miR-30e	1.76	Up
let-7f-3p	2.24	Up
miR-133a-3p	2.73	Up
miR-2110	3.19	Up
miR-548d-5p	4.47	Up
miR-191-3p	4.47	Up
miR-205-5p	4.47	Up
miR-664a-5p	6.07	Up
miR-630	6.07	Up

(b) Using t-test results

Table 4.10: The most consistently differentially expressed miRNA-sequences according to a signed-rank-test as described in section 3.3.2. p-values are adjusted using Bonferroni correction.

Signed-rank-test

A signed-rank-test was done using all datasets. The signed-rank-test was done both using log-fold-change and using t-test results. Using t-test results is explained in section 3.3.2, and this experiment is explained more detailed in section 3.3.2. The results are in Table 4.10. As section 4.3.2 suggests that this test has low external validity, one should be careful when using this data for conclusions. None of the miRNA-sequences were significantly consistently differentially expressed in the datasets when adjusted for multiple testing, which could either be because there are no such miRNA-sequence, or there are too few datasets included in this study to get enough statistical power to find any.

4.4 Are datasets separable from each other?

Learning a logistic regression model to separate samples from two datasets lead to a mean AUC of 0.223 and a standard deviation of 0.213. The experiment is explained in more detail in section 3.4. The results were very poor, and somewhat suspicious, as a random separation would give an AUC of 0.50. On the other hand, in some sense an AUC of 0.223 is good, because it separates well, only that it mixes up the two categories. I do not know what caused this result, but there was strong evidence that the datasets could be separated, and therefore I tried using XGBoost instead. XGBoost gave a mean AUC of 0.801 with a standard deviation of 0.245. This suggests that the datasets can be separated from each other, but far from perfectly in general. However, looking at the AUC values,

several had AUC > 0.99, which means that there were datasets that were easier to distinguish than others. One question that remains is whether this knowledge can be used to adjust the datasets so that they are more comparable. One possibility would be to use linear regression to find expression patterns that are characteristic for that dataset and then adjust for it. However, this would not work as the miRNA expressions are already standardized to a mean of 0, so the linear regression would not find any mean effects of any dataset. Furthermore, logistic regression performed poorly compared to XGBoost when trying to find dataset characteristics, which suggests that what distinguish a dataset are non-linear patterns, which is hard to adjust for.

4.5 Hierarchical clustering of datasets

The results from doing hierarchical clustering of the datasets is shown in Figure 4.2. The clustering was based on the difference in log-fold-change as described in section 3.5. There seem to be a close cluster that includes **?????**. There might be that these datasets are close to each other, and that models trained on one of the datasets would do well on the other datasets. Testing this hypothesis by training on pairs of datasets using logistic regression resulted in a mean AUC of 0.510 with a standard deviation of 0.085. Trying to test on one dataset while training on the others using XGBoost resulted in AUCs with mean 0.470 and standard deviation 0.147. Overall, even though this is a cluster, the diagnostic value between these datasets are low, at least when using standard models.

4.6 Machine learning on single datasets

The results from machine learning on simple datasets is shown in Table 4.11. The results are from using cross validation on the datasets with the given machine learning algorithms. A more detailed explanation is found in section 3.6. There seem to be little difference in the overall results, except for XGBoost preforming somewhat worse than the other three, however the difference is small and the number of datasets is small. Therefore, XGBoost performing worse might be due to chance alone. No hyperparameter tuning was done, and hyperparameter tuning might would have changed the results somewhat, but overall it seems like the model used have little impact on the results. When looking at the results for a single dataset there is somewhat variance, but as the overall results are similar, the variance in results for a single dataset is likely the result of chance rather than intrinsic advantage of some models to some datasets. As the linear model, logistic regression, performs as well as non-linear models, SVM, random forest and XGBoost, there is limited evidence for non-linear relationships between

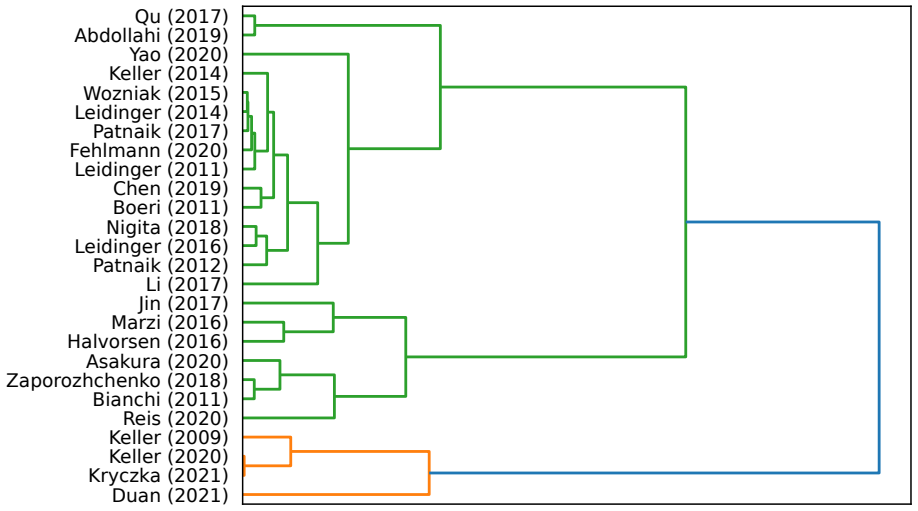


Figure 4.2: Hierarchical clustering

miRNA expression and lung cancer. However, there might be that possible non-linear relationships are either too subtle to find at the current sample size, or the non-linear relationship might be a type that cannot be modelled with random forest or XGBoost.

4.6.1 Using stages

The resulting AUCs when training and testing a logistic regression model on only late stage cancer or only on early stage cancer is shown in **??**. The results might suggest that there might be a higher AUC when only using late stage cancer, however the statistical power here is too small to conclude with any certainty.

4.7 Baseline miRNA-sequence

There were several candidate miRNA-sequences that could be used as a baseline in this project.

4.7.1 Meta analyses

Meta analyses gave an overview of the possible miRNA-sequences that can be used as baselines in this project [**????**], whereas **?** was the most thorough of

Study	Logistic Regression	SVM	Random Forest	XGBoost
?	0.670	0.895	0.918	0.853
?	0.723	0.911	0.970	0.939
?	0.775	0.835	0.832	0.813
?	0.895	0.793	0.790	0.623
?	0.995	0.994	0.979	0.991
?	0.922	0.988	0.959	0.947
?	0.858	0.892	0.866	0.815
?	0.972	0.950	0.909	0.972
?	0.749	0.715	0.657	0.606
?	0.660	0.519	0.491	0.462
?	0.932	0.936	0.939	0.931
?	0.976	0.971	0.952	0.968
?	0.608	0.474	0.549	0.543
?	0.535	0.641	0.700	0.685
Mean	0.805	0.822	0.822	0.796

Table 4.11: The mean AUC when using cross validation on the given studies with the given models, as described in section 3.6

Study	Mean early	Mean late
?	0.586	0.726
?	0.735	0.634
?	0.500	0.250
?	1.000	*
?	0.850	1.000
?	0.083	0.525
?	*	1.000
?	*	0.333
?	0.350	0.600
Mean	0.586	0.634

Table 4.12: The mean AUC when using cross validation on the given studies when only using early stage cancer samples or only using late stage cancer samples, as described in subsection 3.6.1. Empty fields mean that there were less than two cancer samples in the dataset, thus any inference would be impossible.

the meta analyses. These meta analyses suggest that the miRNA-sequences that have been shown to be able to diagnose lung cancer in most studies are miR-21 and miR-210, with **?** suggesting that miR-182, miR-155 and miR-17 are in third, forth and fifth place, respectively. All of these miRNA-sequences were reported to be up-regulated in cases compared to controls. However, these results were not representative for the studies used in this project.

? found that of all studies that they went through that miR-21 was significantly up-regulated in cases in 48 studies, and down-regulated in two studies. However, among the studies used in this project, **????** all reported that miR-21 was down-regulated in cases compared to controls, which suggests that miR-21 might not be as good of a biomarker for lung cancer as the meta analyses suggest. An overview of the reported up- and down-regulation of the aforementioned miRNA-sequences in the studies in this project is shown in Table 4.13.

Table 4.13 shows that there are no certain direction that either of the sequences is consistently up-regulated. However, miR-17 was consistently down-regulated in the sample, which contrasts **?** which reported that miR-17 had been up-regulated in 7 studies and down-regulated in one study, if only looking at the studies using circulating miRNA. Everything considered, this points to very inconsistent results across datasets, which suggest that there might be little consistency, and hard to replicate results. Indeed, **?** found little consistency in the datasets that are considered in this study.

4.7.2 Using datasets

The meta analyses gave some candidate miRNA-sequences that can be used as baselines in this project, namely miR-21, miR-210, miR-182, miR-155 and miR-17. The Cohen's d and AUC of the miRNA-sequences in the different datasets are shown in Table 4.14 and Table 4.15 respectively.

Interestingly, the average Cohen's d of three of the miRNA-sequences were negative, even though **?** found that they were consistently up-regulated in cancer compared to healthy controls, which again suggest that these miRNA-sequences are not as good biomarkers for cancer as the meta analysis suggests. Overall miR-210 scored best on both Cohen's d and AUC, and is therefore what I chose as baseline.

4.8 PCA analysis across datasets

The results when finding the 10 largest principal components in **?** and projecting **?** and **?** onto the principal components can be found in Table 4.16. The t-values and the p-values are from the t-test of projecting cases and controls along the principal component. Similarly, the results when finding the 10 largest principal

Study	miR-21	miR-210	miR-182	miR-155	miR-17
?	Up				
?					
?					Down
?	Up	Up			
?					
?					
?	Down	Up	Down		Down
?		Down			
?	Down			Down	
?		Up	Up		Down
?				Down	Down
?					
?					
?					Down
?	Up				
?	Down				Down
?					
?					
?					
?	Down	Down			Down
?					
?					
?		Up	Up	Up	
?			Down	Up	
?					
?		Up			

Table 4.13: Whether the miRNA-sequences were reported to be significantly up- or down-regulated ($p < 0.05$) in the studies.
Note: **?** only reports miR-210 and miR-182 to be up-regulated in adenocarcinoma. In **?** and **?** abu-miR-155 was measured instead of hsa-miR-155.

Study	miR-21	miR-210	miR-182	miR-155	miR-17
?	-0.784				
?	0.496	0.719	0.427	0.592	0.690
?					0.811
?	-0.300			-0.025	-0.158
?	0.165		0.610		0.147
?				-1.631	
?	-0.411	0.046	-0.272	-0.051	-0.504
?	0.307	-1.634		-0.547	0.676
?	-1.509		0.112	-1.109	0.152
?	0.321	1.499	0.617		-0.678
?	-0.067	-0.208	0.008		1.303
?	-0.193				
?					
?	0.165	-0.102	0.221		-0.663
?	0.377	0.132	-0.066	-0.014	0.203
?	-0.804		-0.442		-0.756
?		-0.318			-0.242
?					
?	-0.215	-0.341			-0.309
?	-1.009	-0.836			-1.044
?	-0.044	-0.070	0.217	0.254	-0.211
?	-1.183				-0.954
?	-0.374	1.436	1.265		
?	0.221	0.013	-0.357		0.429
?	-0.660	1.370	-0.039	-0.096	-0.283
?		0.758	0.630	0.287	0.609
Average	-0.275	0.164	0.209	-0.234	-0.039

Table 4.14: Cohen's d of the different miRNAs in the different datasets. Difference case - controls

Study	**miR-21**	**miR-210**	**miR-182**	**miR-155**	**miR-17**
?	0.345				
?	0.630	0.742	0.601	0.660	0.690
?					0.744
?	0.421			0.491	0.412
?	0.506		0.608		0.443
?				0.111	
?	0.366	0.490	0.405	0.463	0.354
?	0.653	0.105		0.289	0.739
?	0.167		0.574	0.189	0.521
?	0.418	0.814	0.672		0.296
?	0.491	0.446	0.501		0.833
?	0.449				
?					
?	0.564	0.481	0.598		0.323
?	0.649	0.554	0.487	0.490	0.564
?	0.218		0.334		0.253
?		0.444			0.333
?					
?	0.425	0.421			0.421
?	0.217	0.260			0.230
?	0.471	0.477	0.541	0.576	0.449
?	0.194				0.250
?	0.441	0.910	0.918		
?	0.541	0.533	0.421		0.639
?	0.440	0.800	0.440	0.440	0.480
?		0.694	0.703	0.612	0.676
Average	0.430	0.545	0.557	0.432	0.483

Table 4.15: AUC of the different miRNAs in the different datasets

#	PVE	t-value A	p-value A	t-value B	p-value B
1	0.274	45.974	0.0	0.543	5.87×10^{-1}
2	0.036	−13.585	4.56×10^{-41}	−0.659	5.10×10^{-1}
3	0.032	13.023	5.95×10^{-38}	−3.600	3.28×10^{-4}
4	0.024	−14.710	1.17×10^{-47}	−16.965	2.13×10^{-59}
5	0.022	−3.558	3.79×10^{-4}	−10.135	2.01×10^{-23}
6	0.018	−3.386	7.16×10^{-4}	−8.523	3.59×10^{-17}
7	0.017	0.346	7.30×10^{-1}	2.030	4.26×10^{-2}
8	0.012	2.329	1.99×10^{-2}	11.784	8.97×10^{-31}
9	0.010	1.369	1.71×10^{-1}	−3.753	1.81×10^{-4}
10	0.010	−1.727	8.42×10^{-2}	4.915	9.81×10^{-7}

Table 4.16: The result from PCA analysis looking at the 10 largest principal components in **?**
Note: A = **?**, F = **?**, PVE = proportion of variance explained (i.e. the proportion of variance in **?** that is explained by the principal component)

components in **?** is found in Table 4.17. Many of the components with significant separation in one dataset also have a good separation in the other dataset. Interestingly, the component sometimes separates well, but in different directions in the two datasets. One candidate principal component as separator is the forth principal component in **?** which both separates well in both datasets, and it separates in the right direction in both datasets. What remains is to test this principal component in other datasets to see whether it separates well beyond **?** and **?**.

Unfortunately, few datasets have all miRNAs in the chosen principal component. Therefore, I will use only datasets that have at least half of the miRNAs in the principal component, and replace missing values with 0. The results are shown in Table 4.18. There are four other datasets that have a significant difference between cases and controls along the principal component. However, the sign of the difference is positive in two cases (**?** and **?**) and negative in two cases (**?** and **?**). It is hard to say what is the reason that there is a difference in the opposite direction, but it is notable that there is a significant difference along the component anyway. If the component had no connection with case-control characteristics, one would not expect that there should be a significant difference along the axis in several datasets. However, it is possible that this component corresponds to common batch effects or to demographic effects, which could lead to a difference between cases and controls where the direction would depend on the study. In that case, the component would not be of any diagnostic value.

#	PVE	t-value A	p-value A	t-value B	p-value B
1	0.580	−40.510	7.82×10^{-298}	−1.371	1.71×10^{-1}
2	0.107	−13.928	5.01×10^{-43}	13.864	2.56×10^{-41}
3	0.060	1.716	8.62×10^{-2}	5.946	3.39×10^{-9}
4	0.046	19.658	5.87×10^{-82}	−6.316	3.49×10^{-10}
5	0.024	−0.322	7.48×10^{-1}	−8.652	1.24×10^{-17}
6	0.019	−11.674	5.94×10^{-31}	1.473	1.41×10^{-1}
7	0.014	23.020	8.92×10^{-110}	3.474	5.27×10^{-4}
8	0.013	22.401	2.12×10^{-104}	3.333	8.78×10^{-4}
9	0.011	2.889	3.89×10^{-3}	17.967	8.87×10^{-66}
10	0.010	9.795	2.20×10^{-22}	1.755	7.95×10^{-2}

Table 4.17: The result from PCA analysis looking at the 10 largest principal components in **?**
Note: A = **?**, F = **?**, PVE = proportion of variance explained (i.e. the proportion of variance in **?** that is explained by the principal component)

Study	t-value	p-value	Proportion miRNA
?	−14.710	1.17×10^{-47}	1.000
?	−16.965	2.13×10^{-59}	1.000
?	−2.239	0.032	0.593
?	3.631	3.76×10^{-4}	0.709
?	−0.617	0.540	0.773
?	−0.335	0.738	0.974
?	−0.112	0.911	0.724
?	−0.158	0.875	0.995
?	−2.083	0.044	0.660
?	2.695	7.65×10^{-3}	0.515
?	−2.111	0.068	0.541

Table 4.18: The result from a t-test when projecting cases and controls along the fourth largest principal component in **?**. The proportion miRNA is the proportion of miRNA-sequences in the principal component that was also in the dataset

Test set	AUC
?	0.537
?	0.465
?	0.461
?	0.494
?	0.680

Table 4.19: AUC when using cross validation and the most replicated miRNA-sequences as described in subsection 3.9.1

4.9 Machine learning based on several datasets

Here are the results from training on several datasets at once:

4.9.1 Using the most replicated miRNA-sequences from the meta analyses

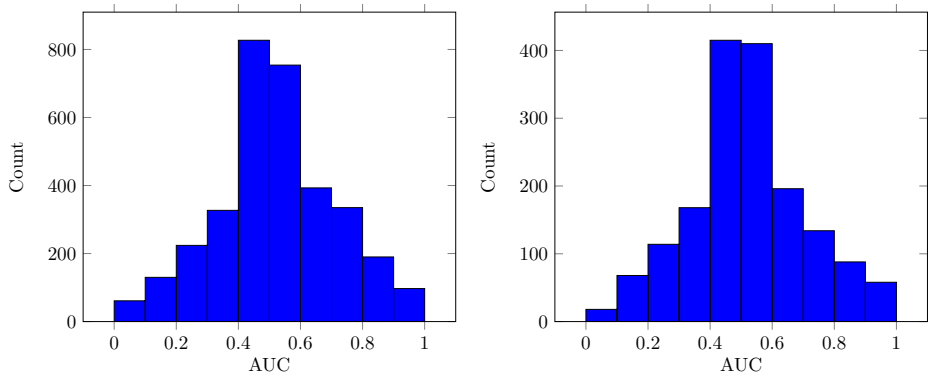
The most replicated miRNA-sequences from the meta analyses were miR-21, miR-210, miR-182, miR-155 and miR-17 (see subsection 4.7.1). Furthermore, the datasets that have all these miRNA-sequences are **?????**, which means that they are the studies that was used in the cross validation. More details are in subsection 3.9.1. The resulting AUC values are in Table 4.19. Seemingly, the results are poor except when using **?** as the test set. It should be noted, however, that **?** has only 10 samples, which means that one should be careful about concluding based on that AUC value, especially as it is an outlier.

4.9.2 Training on two datasets

Firstly, the results from training on two datasets and testing on a third, as described in subsection 3.9.2.

Logistic Regression

The AUC values from training on one of the datasets and testing on the third dataset using logistic regression are shown in Figure 4.3a. The AUC values from training on two datasets and testing on a third are shown in Figure 4.3b. The histograms are very similar, and this can be confirmed by other statistical measures. When training on just one of the datasets the mean AUC was 0.521 and the standard deviation was 0.194. When training on both datasets, the mean



(a) Histogram over AUC values when training on one dataset and test on another according to subsection 3.9.2

(b) Histogram over AUC values when training on two datasets and test on a third dataset according to subsection 3.9.2

Figure 4.3: AUC values when training on one or two datasets using logistic regression

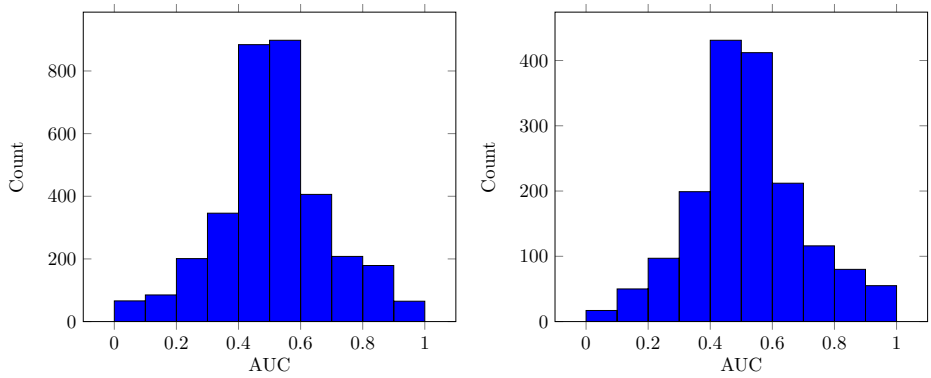
AUC was 0.519 and the standard deviation was 0.187. This is worse than the baseline miR-210, which had a mean AUC of 0.595 (see Table 4.15).

XGBoost

The AUC values from training on one of the datasets and testing on the third dataset using logistic regression are shown in Figure 4.4a. The AUC values from training on two datasets and testing on a third are shown in Figure 4.4b. The mean and standard deviation in AUC values when training on one dataset were 0.509 and 0.175 respectively. The mean and standard deviation when training on two datasets were 0.519 and 0.181 respectively. Again both the histograms and statistics are similar in the two cases, which suggest that combining two datasets have little to none effect. Furthermore, the results were very similar to the ones achieved with logistic regression, which suggest that the problem is not the model.

4.9.3 Merging all datasets together

The AUC values from merging all datasets except one which is used for testing, are shown in Table 4.20. A more thorough description of the process is in subsection 3.9.3. The mean of the AUC values is 0.511 and the standard deviation of the AUC values is 0.149, which is not only lower than the AUC of miR-210



(a) Histogram over AUC values when training on one dataset and test on another according to subsection 3.9.2

(b) Histogram over AUC values when training on two datasets and test on a third dataset according to subsection 3.9.2

Figure 4.4: AUC values when training on one or two datasets using XGBoost

(0.595), but it also worse than chance, which means that this was a poor way to learn a model to diagnose lung cancer.

4.9.4 Maximal training sets

Here are the results for the generated maximal training sets, as defined in subsection 3.9.4. The AUC values are shown in Figure 4.5. The mean AUC value is 0.517 and the standard deviation is 0.172, which is still worse than the baseline of miR-210.

4.10 Stratification of the datasets

Here are the results when the datasets are stratified.

4.10.1 Training on two datasets, in-group vs. out-group

There are several ways to stratify the datasets. I will report the results for the different ways. A description of the method is in subsection 3.10.1.

Technology

The results when training on one dataset and testing on another dataset when stratifing using technology are shown in Table 4.21. The in-group is when both

Test set	AUC
?	0.621
?	0.571
?	0.352
?	0.491
?	0.483
?	0.861
?	0.579
?	0.650
?	0.500
?	0.495
?	0.400
?	0.530
?	0.263
?	0.335
?	0.613
?	0.580
?	0.333
?	0.240
?	0.624
?	0.497
?	0.501
?	0.833
?	0.465
?	0.371
?	0.560
?	0.541

Table 4.20: AUC when merging all datasets except one which is used for testing, as described in subsection 3.9.3

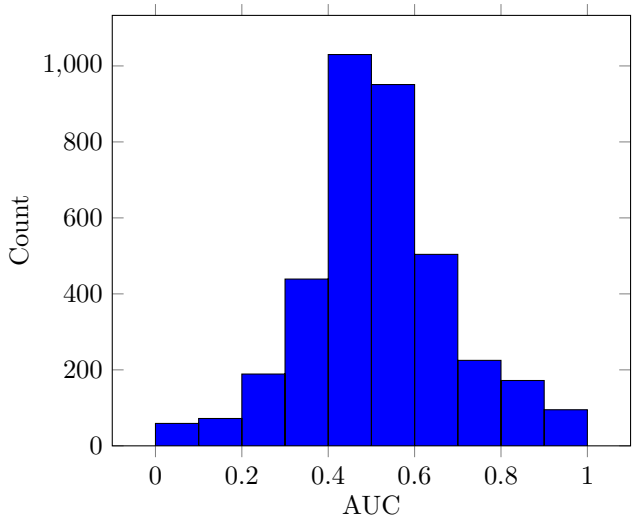


Figure 4.5: Histogram over AUC values when training maximal datasets as described in subsection 3.9.4

datasets have the same technology, and the out-group is when only one of the datasets use the technology. As the table shows, the AUC was generally somewhat better on in-group training and testing than out-group training and testing. However, the improvement in AUC was only significant for sequencing. There was least improvement in microarrays, which might be a result of the category "microarray" hiding heterogenity, as the microarrays in the studies varied a lot.

In order to check whether the hypothesis that the problems are due to heterogenity in the microarray-technology, I wanted to do an experiment with a finer stratification of the microarray technology. Of the microarray-technologies that have been used multiple times, there were three that used *Exiqon microarrays* (**???**), three that used *Agilent microarrays* (**???**), three that used *Geniom microarrays* (**???**) and two that used *SurePrint microarrays* (**??**). This experiment will only consider pair of studies where both use microarrays. The in-group here is when the pair of studies have the same type of microarray, and the out-group is when they use different type of microarrays. The results were that the in-group had a mean AUC of 0.445 while the out-group had a mean of 0.534. The difference were not significant using a t-test (p=0.081). One would assume that the in-group would have a higher AUC than the out-group, but if anything there seems like the out-group has a somewhat higher AUC. This suggest that heterogenity in micorarray-technology is not the cause of the poor results from

Technology	Mean IG	Std. IG	Mean OG	Std. OG	t-value	p-value
Sequencing	0.615	0.153	0.467	0.206	2.447	0.015
Microarray	0.524	0.216	0.484	0.181	2.096	0.037
qRT-PCR	0.545	0.194	0.494	0.168	1.404	0.162

Table 4.21: The results when training on one dataset and testing on another, when stratifing by technology as described in subsection 3.10.1
Note: IG = in-group, OG = out-group, mean and standard deviation are of AUC values, and t-values are in-group minus out-group and p-values correspond to the t-values

Body fluid	Mean IG	Std. IG	Mean OG	Std. OG	t-value	p-value
Plasma	0.457	0.204	0.513	0.197	−1.878	0.061
Serum	0.476	0.205	0.512	0.206	−0.784	0.434
Whole blood	0.526	0.088	0.507	0.188	0.564	0.573

Table 4.22: The results when training on one dataset and testing on another, when stratifing by body fluid as described in subsection 3.10.1
Note: IG = in-group, OG = out-group, mean and standard deviation are of AUC values, and t-values are in-group minus out-group and p-values correspond to the t-values

microarrays.

Body fluid

The results when training on one dataset and testing on another dataset when stratifing using body fluid are shown in Table 4.22. Again, the in-group is when both datasets have the same technology, and the out-group is when only one of the datasets use the technology. In contrast to when stratifing by technology, it seems that there are no use in stratifying by body fluid. None of the changes in AUC are significant, and only one of the changes are positive. It might suggest that contributes to more variance in the resulting data, then what body fluid does.

Using stages

The result from training and one dataset and testing on another dataset, when only using late or only using early stage cancer from datasets where stage is

Mean Late	Std. Late	Mean Early	Std. Early	t-value	p-value
0.453	0.179	0.517	0.173	−1.081	0.287

Table 4.23: The results when training on one dataset and testing on another, when using only late or only early stage cancer samples from datasets where stage is labeled.
Note: mean and standard deviation are of AUC values, and t-values are late minus early and p-values correspond to the t-values

Technology	Mean AUC	Std. AUC	t-value	p-value
qRT-PCR	0.424	0.100	−2.134	0.965
Sequencing	0.795	0.133	4.444	0.011
Microarray	0.459	0.210	−0.724	0.759

Table 4.24: The results when training on all datasets except one in a certain category, when stratifing by technology as described in subsection 3.10.2

labled, are in Table 4.23. There are no significant difference between the AUCs in the two cases, and both are close to 0.50, which suggests that stage does not explain the low AUC scores in the previous results.

4.10.2 Combining all except one

Here are the results when training on all datasets except one in a certain category, and then using the last dataset as testset. For checking whether the AUC values are better than chance I took a one sided hypothesis of AUC > 0.50 using a t-test.

Technology

The results from stratifing by technology are shown in Table 4.24. Similarly to section 4.10.1 sequencing is an outlier where the AUC in the category is significantly better than chance. Notably, an AUC of 0.795 is quite higher than any of the other AUCs achieved so far when testing on a different dataset than testing on. Some caution should be noted, as it is a clear outlier when compared to other results, and it is based on only 6 datasets. Still, it gives some preliminary evidence that technology plays a role in why the datasets are incomparable, and that replicating results across datasets could be possible for some technologies.

Also here I want to see whether stratifing by subtypes of microarrays will

Body fluid	Mean AUC	Std. AUC	t-value	p-value
Serum	0.417	0.216	−0.939	0.805
Plasma	0.529	0.225	0.363	0.364
Whole blood	0.592	0.113	2.151	0.038

Table 4.25: The results when training on all datasets except one in a certain category, when stratifing by body fluid as described in subsection 3.10.2

be beneficial. The results might be assumed to be similar to section 4.10.1 as the subcategories are small, with the largest ones have three datasets. Thus training will be done on maximally two datasets. The resulting mean AUC was 0.381 and the resulting standard deviation was 0.200, which was not significantly better than 0.50 (p=0.961). This suggests that neither here heterogenity in the micorarray-technology was the reason for the poor results for the microarrays.

Body fluid

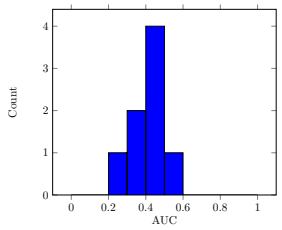
The results from stratifing by body fluid are shown in Table 4.25. Here the AUC when training and testing on whole blood is significantly larger than 0.05. This contrasts section 4.10.1 where the internal consistency was not significantly higher than the external consistency for the whole blood-datasets.

Distribution of AUC values

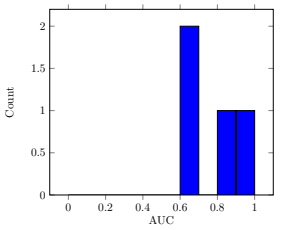
In the subsections above, only summary statistics were reported. However, mean and variance can hide a lot of information about the distribution, e.g. whether the distribution is unimodal or bimodal. As t-values have been used to check for statistical significance, there has been an implicit assumption that AUC values have been approximately normally distributed. To be sure, some checks are done. Histograms of the AUCs are shown in Figure 4.6. It is hard to judge the normality of the plots due to the low sample sizes. Therefore I have also plotted histogram and Q-Q plot combining all the AUC values from the different categories. Those are in Figure 4.7. The Q-Q plots shows that the distribution of AUC values follows the normal distribution quite nicely, except in the tails of the distribution, where the disparency in the right tail is partially caused by the sequencing datasets.

Using stages

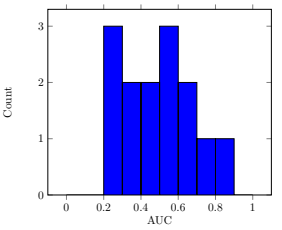
The results from stratifing using cancer stages can be found in Table 4.26. There mean AUCs, both when only using early stage cancer and only using late stage



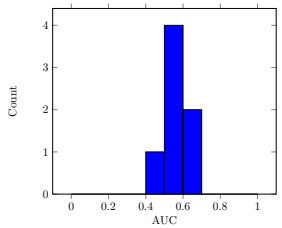
(a) Histogram over AUC values when training on all datasets except one in the qRT-PCR category and test on the last dataset.



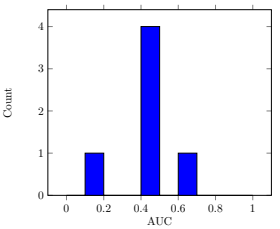
(b) Histogram over AUC values when training on all datasets except one in the sequencing category and test on the last dataset.



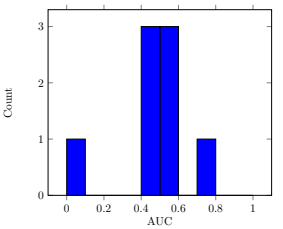
(c) Histogram over AUC values when training on all datasets except one in the microarray category and test on the last dataset.



(d) Histogram over AUC values when training on all datasets except one in the whole blood category and test on the last dataset.

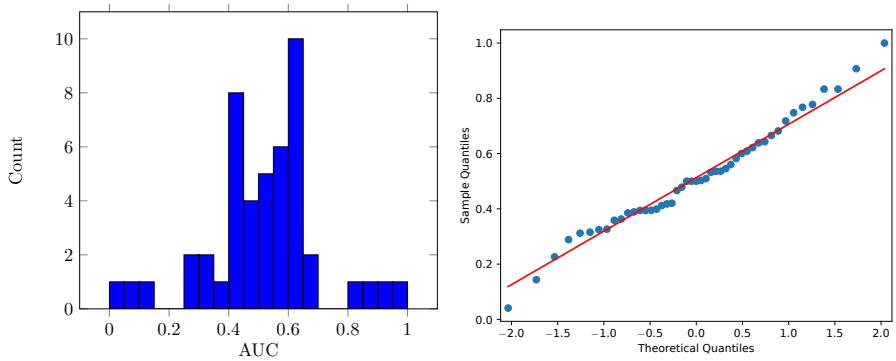


(e) Histogram over AUC values when training on all datasets except one in the serum category and test on the last dataset.



(f) Histogram over AUC values when training on all datasets except one in the plasma category and test on the last dataset.

Figure 4.6: Histogram over AUC values when training on all datasets except one in a category and test on the last dataset as explained in subsection 3.10.2



(a) Histogram over AUC values when training on all datasets except one in one category and test on the last dataset, aggregated over all categories.

(b) Q-Q plot over AUC values when training on all datasets except one in one category and test on the last dataset, aggregated over all categories.

Figure 4.7: Histogram over AUC values when training on all datasets except one in a category and test on the last dataset as explained in subsection 3.10.2

Cancer stage	Mean	Std.	t-value	p-value
Early	0.390	0.165	−1.762	0.129
Late	0.410	0.141	−1.798	0.115

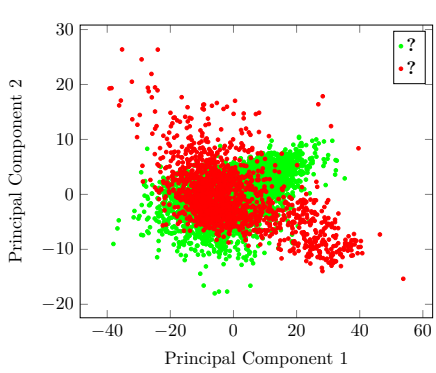
Table 4.26: The results when training on all datasets except one using datasets where cancer stage is labeled, when stratifing by cancer stage as described in subsection 3.10.2

cancer, were below 0.50, which suggest that there are no improvement in AUC by stratifing by stage.

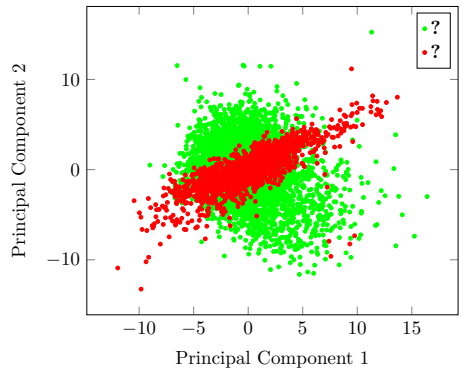
4.11 PCA for removing artifacts

4.11.1 Check comparability using PCA

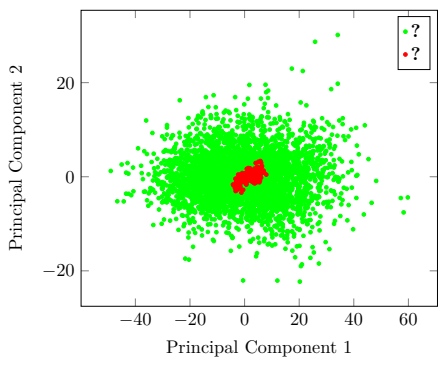
The resulting PCA-plots when the two first principal components are removed, as described in subsection 3.11.1, are shown in Figure 4.8. It is hard to say whether the PCA adjustments made the datasets more similar using these PCA-plots, but it seems like the spreads are more equal after the removal of the principal components, with the exception of **?** and **?**.



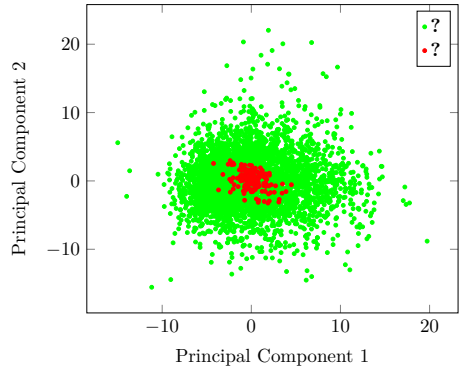
(a) PCA of ? and ? using two the principal components of the joint dataset, without any principal components removed



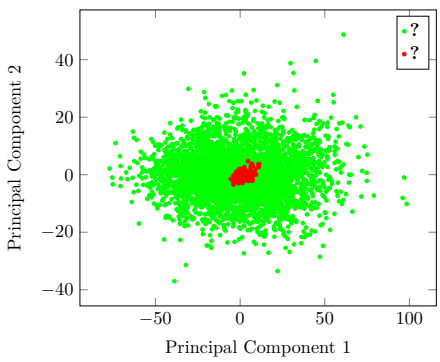
(b) PCA of ? and ? using two the principal components of the joint dataset, after the two first principal components of each dataset are removed



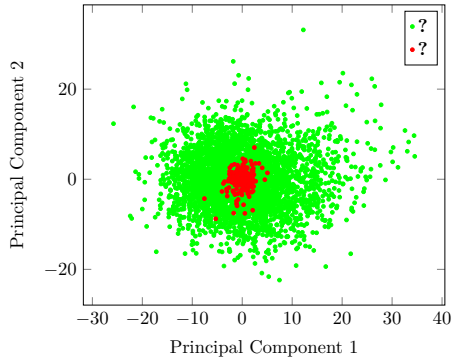
(c) PCA of ? and ? using two the principal components of the joint dataset, without any principal components removed



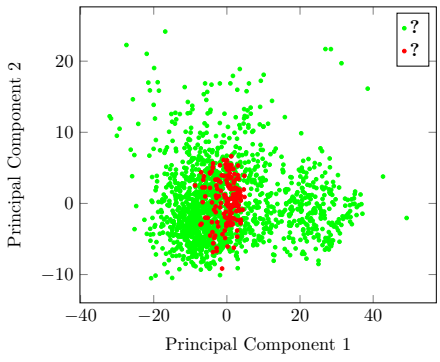
(d) PCA of ? and ? using two the principal components of the joint dataset, after the two first principal components of each dataset are removed



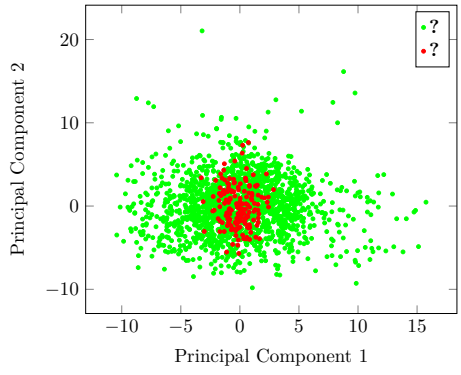
(e) PCA of ? and ? using two the principal components of the joint dataset, without any principal components removed



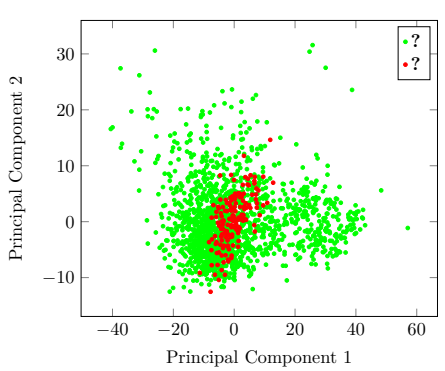
(f) PCA of ? and ? using two the principal components of the joint dataset, after the two first principal components of each dataset are removed



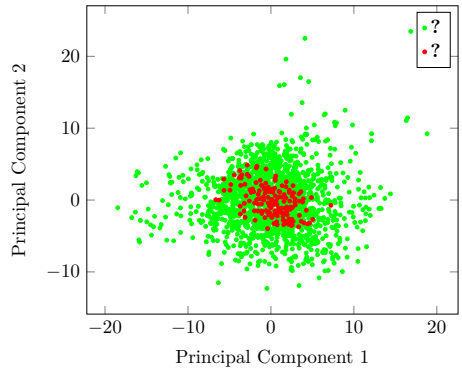
(g) PCA of ? and ? using two the principal components of the joint dataset, without any principal components removed



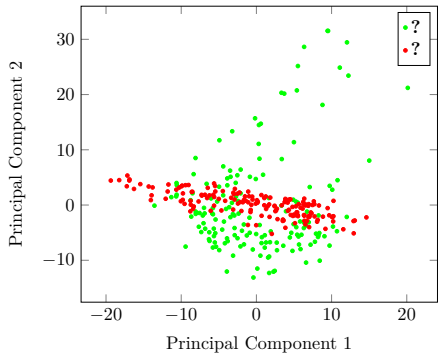
(h) PCA of ? and ? using two the principal components of the joint dataset, after the two first principal components of each dataset are removed



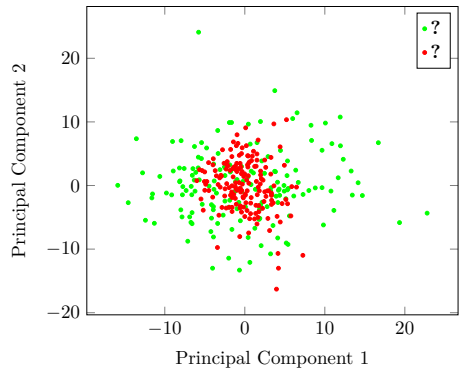
(i) PCA of ? and ? using two the principal components of the joint dataset, without any principal components removed



(j) PCA of ? and ? using two the principal components of the joint dataset, after the two first principal components of each dataset are removed



(k) PCA of ? and ? using two the principal components of the joint dataset, without any principal components removed



(l) PCA of ? and ? using two the principal components of the joint dataset, after the two first principal components of each dataset are removed

Figure 4.8: PCA of datasets with and without removing the two first principal components in each individual dataset

4.11.2 Using machine learning models

The resulting AUC-values from doing the experiment described in subsection 3.11.2 are shown in Table 4.28 and **??**. The internal results were poorer both in the sequencing datasets and in the non-sequencing datasets, which means that some information about case-control was lost when removing the two first principal components. This does not mean that these principal components were due to biological factors involved with cancer, it might still be technical artifacts like batch effects, or it might be due to demographic differences between cases and controls. The external validity, here represented by to which degree it is possible to get good results when training or testing on sequence data, was low in both cases. Everything considered the data give little indication on whether removing the two first principal components had any effect on the comparability.

Study	I. study	I. seq	To seq	From seq
?	0.613	0.670	0.602	0.608
?	0.700	0.613	0.476	0.446
?	0.821	0.451	0.318	0.488
?	0.947	0.636	0.405	0.570
?	0.950	0.735	0.437	0.559
?	0.924	0.826	0.385	0.592
?	0.450	0.548	0.444	0.525
?	0.515	0.826	0.504	0.342
?	0.950	0.741	0.300	0.485
?	0.559	0.826	0.370	0.269
?	0.684	0.770	0.591	0.605
?	0.969	0.799	0.414	0.458
?	0.847	0.619	0.227	0.501
?	0.897	0.672	0.671	0.529
?	0.600	0.640	0.088	0.470
?	0.491	0.707	0.415	0.513
?	0.854	0.581	0.917	0.429
?	0.500	0.746	0.556	0.452
?	0.850	0.740	0.647	0.599
?	0.322	0.409	0.492	0.618
?	0.634	0.646	0.299	0.364
?	0.851	0.586	0.566	0.576
Average	0.724	0.672	0.460	0.500

Table 4.27: The resulting AUC-values when doing the experiment as described in subsection 3.11.2, without removing two first principal components.
Note: I. = internal, Seq = sequencing, To seq = training model on study, testing on sequencing datasets, From seq = training model on sequencing datasets, testing on study

Study	I. study	I. seq	To seq	From seq
?	0.546	0.676	0.558	0.574
?	0.564	0.620	0.511	0.496
?	0.700	0.578	0.529	0.482
?	0.796	0.644	0.451	0.365
?	0.600	0.607	0.661	0.552
?	0.936	0.574	0.499	0.508
?	0.250	0.534	0.361	0.566
?	0.594	0.574	0.494	0.545
?	0.742	0.604	0.591	0.486
?	0.477	0.574	0.395	0.441
?	0.348	0.570	0.505	0.477
?	0.964	0.623	0.296	0.462
?	0.737	0.664	0.535	0.561
?	0.901	0.630	0.550	0.563
?	0.650	0.605	0.535	0.560
?	0.393	0.607	0.639	0.572
?	0.042	0.455	0.722	0.539
?	0.500	0.451	0.500	0.533
?	0.532	0.593	0.509	0.520
?	0.612	0.329	0.602	0.503
?	0.410	0.491	0.601	0.547
?	0.744	0.534	0.622	0.610
Average	0.593	0.570	0.530	0.521

Table 4.28: The resulting AUC-values when doing the experiment as described in subsection 3.11.2, removing two first principal components.
Note: I. = internal, Seq = sequencing, To seq = training model on study, testing on sequencing datasets, From seq = training model on sequencing datasets, testing on study

Chapter 5

Evaluation and Conclusion

This chapter contains the conclusions inferred from the results in this project.

5.1 Evaluation

5.2 Discussion

As there was an inconsistency between what was reported in the meta analyses and

5.3 Contributions

5.4 Future Work

Appendices