Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Clinical and Molecular Medicine

Erlend Fauchald

Identifying a cross-cohort circulating microRNA signature for Lung Cancer prediction using Random Forests

Master's thesis in Informatics
Supervisor: Pål Sætrom
August 2021

NTNU
Norwegian University of
Science and Technology

Erlend Fauchald

Identifying a cross-cohort circulating microRNA signature for Lung Cancer prediction using Random Forests

Master's thesis in Informatics
Supervisor: Pål Sætrom
August 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Clinical and Molecular Medicine

NTNU
Norwegian University of
Science and Technology

Abstract

Lung cancer is a disease in which early diagnosis is of particular importance for patient survival. Current screening techniques are focused on at-risk populations because of their invasiveness, cost, and low specificity. MicroRNAs are small non-coding RNAs circulating in blood that hold potential as non-invasive biomarkers for many different diseases. These small RNAs have important regulatory functions in plants, viruses, and animals and have been proven to be differentially expressed in a wide range of human cancers - including lung cancer. Recent advances in sequencing technology has opened up for the identification and quantification of microRNA at massively parallel scales. By sequencing the microRNAs present in a patient's blood sample and pairing these with their diagnostic and prognostic outcomes, one could train supervised machine learning models that distinguish cancer patients from controls using only the microRNAs that are expressed in their blood samples. Using data of this kind from four European longitudinal cohort studies, three prediagnostic and one diagnostic, this study aimed to train such a model to create a cross-cohort lung cancer predictor that might be useful as a diagnostic tool. Random forests were found to be well suited for this task, as they can model the complex biological nature of the microRNA expression profiles while also adding a layer of interpretability; the most important features for doing predictions can be extracted directly from the model.

This project is of an exploratory nature, and as such, many different experiments for feature extraction and sampling were carried out. The main finding was a random forest model that when trained on linearly transformed prediagnostic training data could predict lung cancer in a separate diagnostic cohort with fair specificity and sensitivity. This model's top microRNAs were then further analysed for their role in biological and regulatory gene pathways, and these were found to be cancer related. Further work and more advanced statistical methods are needed to model lung cancer in the prediagnostic cohorts. Models that perform well on the prediagnostic cohorts would be capable of predicting lung cancer years before current diagnostic techniques, and this kind of model would be highly valuable in medical practice.

Sammendrag

Lungekreft er en sykdom hvor tidlig diagnose er spesielt viktig for pasientoverlevelse. Screeningteknikker som brukes i dag fokuserer ofte på pasienter i risikogrupper fordi de er invasive, kostbare og har lav spesifisitet. MikroRNA er små ikke-kodende RNA som sirkulerer i blod og har potensial som ikke-invasive biomarkører for en rekke ulike sykdommer. Disse små RNA-ene har viktige regulatoriske funskjoner i planter, virus og dyr, og det har blitt bevist at disse er differensielt uttrykt i mange ulike krefttyper, inkludert lungekreft. Nylige fremskritt innen sekvenseringsteknologi har muliggjort kvantifisering og identifisering av mikroRNA på massiv-parallel skala. Ved å sekvensere blodprøvene til pasienter for mikroRNA og deretter sammenstille disse dataene med deres prognostiske og diagnostiske utfall, kan man trene veiledete maskinlæringsmodeller som kan skille lungekreftspasienter fra kontroller ved å kun bruke mikroRNA som er uttrykt i pasientenes blodprøver. Denne type data fra fire europeiske longitudinelle kohorter, en diagnostisk og tre prediagnostiske, ble i denne studien brukt til å trene en slik modell som kan predikere lungekreft på tvers av kohorter og dermed brukes til diagnoseformål. Random forest er en maskinlæringsmetode som er spesielt velegnet til å modellere komplekse mikroRNA-ekspresjonsprofiler og som samtidig muligjør en grad av tolkbarhet i modellene: de viktigste mikroRNA-ene for prediksjoner kan hentes direkte ut fra modellen.

Dette prosjektet er av en utforskende art: flere ulike eksperimenter ble gjennomført vedrørende ekstrahering av de mest interessante forklaringsvariablene og samplingmetodene. Hovedresultatet ble en random forest modell trent på en linærtransformert versjon av de prediagnostiske kohortene som kunne predikere lungekreft i den diagnostiske kohorten med relativt god spesifisitet og sensitivitet. De viktigste mikroRNA-ene fra denne modellen ble deretter videre analysert for deres rolle i biologiske og regulatoriske gennettverk, og disse gennettverkene viste seg å korrelere med kreftrelaterte nettverk. Videre arbeid og mer avanserte statistiske metoder er likevel nødvendig for å kunne modellere lungekreft i de prediagnostiske kohortene. Modeller som skiller diagnoser fra kontroller i disse kohortene vil være i stand til å predikere lungekreft opptil flere år før diagnosen blir satt med klassiske metoder, og denne typen modell vil være svært verdifull i medisinsk praksis.

Preface

This document is my thesis for the master's program Informatics: Databases and Search at the Norwegian University of Science and Technology (NTNU). The main focus of this thesis is within the interdisciplinary field of bioinformatics. As such, it assumes some background knowledge in molecular biology, computer science and statistics, but the most important concepts are explained in the Background chapter of the report.

I would like to thank Pål Sætrom for his patience and excellent supervision during my work with this thesis. The reading lists he provided with subsequent discussions and conversations gave me an invaluable learning experience.

I would also like to thank Christina, Greta, Per, and Vera for reading my thesis, giving excellent suggestions and for supporting me throughout the process.

Erlend Fauchald
Trondheim
1st August 2021

Contents

Figures

Tables

Acronyms

C. elegans *Caenorhabditis elegans*.

AD Adenocarcinoma.

AUC Area Under the Curve.

CPM count-per-million.

CXR Chest X-Ray.

GCF Genomics Core Facility.

LC Lung Cancer.

LCLC Large Cell Lung Carcinoma.

LDCT Low Dose Computed Tomography.

MAE Mean Absolute Error.

miRNA MicroRNA.

ML Machine learning.

mRNA Messenger RNA.

MS Mass spectrometry.

MSE Mean Squared Error.

ncRNA non-coding RNA.

NGS Next-Generation Sequencing.

NSCLC Non-Small Cell Lung Carcinoma.

PCA Principal Component Analysis.

RCT Randomized Controlled Trial.

RF Random Forest.

RISC RNA induced silencing complex.

RNase Ribonuclease.

ROC Receiver operating characteristic.

SCLC Small Cell Lung Carcinoma.

SQ Squamous Cell Carcinoma.

TMM Trimmed Mean of M.

UTR Untranslated Region.

Chapter 1

Introduction

In medicine, the early diagnosis and outcome prediction of diseases are particularly interesting areas in which supervised Machine learning (ML) methods could have powerful applications. The large amount of historical data collected during routine medical care paired with already known patient outcomes can be leveraged to train ML models that can consider far more features and examples than any human clinician [1]. These models can then pick up statistical patterns that correlate with a diagnosis or patient outcome in other patient data, potentially aiding clinicians in carrying out appropriate treatment at an earlier disease stage. Recent advances in sequencing technology and Mass spectrometry (MS) have further expanded the latent dimensionality of patient data. By characterizing and quantifying the molecules in a patient's biological samples, such as blood or tissue, for either genomic, epigenomic, transcriptomic, proteomic, metabolomic, lipidomic or glycomic (collectively known as omics) - data [2], it is possible to augment patient records with their 'biochemical status' at a given time.

Omics data are good candidates for supervised ML, as the wide array of identified molecules through sequencing or MS often have complex and poorly understood biological pathways that no clinician can realistically consider simultaneously [2]. However, there are inherent limitations in using ML methods for diagnostic purposes; it is not necessarily possible to gain actionable insight from the model to potentially create new therapeutic procedures, only a given patient's risk of developing disease can be extracted. Additionally, as the dimensionality of the training data increases, so does the need for training examples, and omics data in particular are expensive (albeit decreasingly so) and subject to privacy laws. Therefore, the use of ML modelling techniques is still not widespread on omics data in medical practice [3]. However, longitudinal cohort health studies such as the Trøndelag health study (HUNT) can provide the needed training data, and in this study, transcriptomic data from multiple studies of this kind are used to train an ML model for diagnostic purposes.

1.1 Motivation

Lung Cancer (LC) is the leading cause of cancer death worldwide [4]. The majority of LC diagnosed patients have metastatic disease at the time of diagnosis, so despite the fact that early resection is effective on localized LC tumors, the cancer has most likely already spread once diagnosis is made, complicating treatment a great deal [5]. Low Dose Computed Tomography (LDCT) has shown promising results and is widely used in LC screening but is currently limited to at-risk populations because of its high cost and limited specificity [6]. A high quality blood-borne biomarker could address some of the issues with current imaging techniques, and circulating microRNAs (miRNA) are one such candidate.

miRNA are a class of small (about 22 nucleotides in length) non-coding RNA (ncRNA) that regulate translation of target protein-coding genes in Messenger RNA (mRNA) post transcription [7]. There are several reasons for why miRNAs could provide for robust blood-borne LC biomarkers:

- One specific miRNA can target hundreds of different protein-coding genes, and as such, they have been shown to have important roles in many aspects of eukaryotic development [7].
- Several studies have shown that miRNAs are deregulated in LC: some miRNA are upregulated in LC cases (oncomiRs) while some are downregulated [8–10].
- miRNA are present in blood, thought to be excreted from cells packaged in exosomes to perhaps mediate cell to cell communication. [7]
- miRNA have been shown to be quite stable, they do not degrade in clinical plasma samples, so storing the samples before doing the miRNA sequencing is viable. [11]

1.2 Problem description

Due to recent advances in sequencing technology, RNA expression profiles can be extracted from blood samples for a relatively low monetary cost. A sequencing technique used for this purpose is RNA-Seq [12], and this technique is the basis for the miRNA expression profiles used in this study. RNA-Seq uses Next-Generation Sequencing (NGS) to calculate the quantity of RNA in a given biological sample at a given moment, capturing the expressed genes (transcriptome) of the sample at that particular time. By applying RNA-Seq to a blood sample, mapping the results to the reference human genome and finally mapping these to already known and annotated miRNA, one ends up with the number of occurrences for observed known miRNAs in the sample. If paired with clinical and diagnostic patient data, these miRNA expression profiles could then be used to train an ML model that given a miRNA expression profile could predict the diagnosis status of the patient, and this is what the present study aimed to do.

To this end, data from four different longitudinal cohort studies was used: The Central Norway Lung Cancer Biobank (CNLCB), The Trøndelag Health Study (HUNT), The Norwegian Woman and Cancer Study (NOWAC) and the Northern Sweden Health and Disease Study (NSHDS). As part of the Id-Lung project [13], omics data extracted from blood samples of LC case-control pairs taken 0-12 years before cancer diagnosis from these four cohort studies have been made available for analysis. In this particular study, only the miRNA portion of the omics data is used, but Id-Lung also includes proteomics and DNA-methylation data. These omics data are also accompanied by diagnostic and clinical patient data that are relevant in a LC context (such as age, BMI, sex, smoking status, sample group and cancer stage, if any). CNLCB is a diagnostic cohort while the others are prediagnostic, meaning that the blood samples were taken some time before the actual diagnosis of cancer was made.

Several data pre-processing steps were made to the miRNA expression profiles and patient data (see section 3.1) before they were fed as training data to a Random forest model, which is well suited for biological sparse high dimensional data sets (see section 2.3). Experiments were carried out for different methods of unsupervised feature extraction, results from these are reported in sections 4.1.3 and 4.1.4. Different sampling techniques were employed, which helps in gaining an understanding of which cohorts could actually be predicted with the classifiers and regressors used in this project. Results from these are included in section 4.2.1. It became clear that only the diagnostic cohorts could be predicted reasonably without using more complex statistical methods. Consequently, the diagnostic cohort became the test set for the main models of this project, while a transformed version of the prediagnostic cohorts was chosen as the main training data. After being tuned for hyperparameters, the performance of the final Random forest model is reported in section 4.2.2. The most significant miRNAs from this model were then further investigated for the KEGG pathways of their target genes. These are reported in section 4.2.4 Finally, the aim was to answer these particular research questions:

1.2.1 Research Questions

RQ1: How well can LC cases be distinguished from controls in a cross-cohort manner with random forests based solely on miRNA expression profiles?

RQ2: In what way does the ML model's ability to distinguish LC cases from controls depend on the elapsed time between the blood sample date and the diagnosis date of patients?

RQ3: How can feature extraction methods help in improving the final model's predictive performance while also reducing model size?

RQ4: To what degree could the final model be used as a diagnostic tool?

Chapter 2

Background

This chapter aims to give the reader some background reading in both the biology interesting for the problem domain and some of the computational methods used in the modelling of the biological data. First, LC is covered with a focus on the efficacy of current screening techniques and their effect on survival. Then, miRNAs are explained in a general sense and their potential as circulating biomarkers for LC is discussed. Finally, the rationale behind some of the chosen computational methods are explained. Other methods were also used during the course of the project, these are outlined in the Methods chapter.

2.1 Lung Cancer

LC is the most prevalent cancer type found in men (about 1.4 million new cases in 2018) and the third most occurring cancer type in women (about 725 000 new cases in 2018) [4]. In 2020 alone, an estimated 2.2 million new cases of LC were found worldwide and were the cause of about 1.8 million deaths. This accounts for 11.4% of all new cases of cancer and 18% of all new cancer deaths for the year, making LC the leading cause of cancer death in 2020, followed by colorectum cancer at 9.4% of cancer deaths [14].

2.1.1 Histology

LC is often broadly divided into two main types: Small Cell Lung Carcinoma (SCLC), accounting for approx. 15% of LC cases, and Non-Small Cell Lung Carcinoma (NSCLC), accounting for approx. 85% of cases. The key differences between these types are patient outlook and the appearance of the cancer cells under a microscope, with SCLC cells being more aggressive and smaller in size than NSCLC cells [5]. NSCLC is often further divided into three major histological subtypes, each originating from different types of lung tissue: Squamous Cell Carcinoma (SQ), Adenocarcinoma (AD) and Large Cell Lung Carcinoma (LCLC) [15]. All LC types correlate well with the smoking status of the patient, but it is worth mentioning that AD is the most common LC type among non-smokers. [15] The data used in this study only distinguish cancer cases with labels for SCLC, AD and SQ, as these histological types together constitute the vast majority of diagnosed LC in the general population [15].

2.1.2 Survival

SCLC is generally both more aggressive and has a worse prognosis than NSCLC. The overall 5-year survival rate is only about 5% for this type of LC, which in part can be explained by the early and fast spreading that is characteristic for SCLC with 90% of patients presenting with either locally advanced or distant metastatic disease (stage III or IV cancer) at the time of diagnosis [5]. In the cases where early stage SCLC is diagnosed and treated, the 5-year survival rate looks much better than for the general SCLC patient; one study only consisting of patients diagnosed with stage I SCLC found a 40% survival rate with resection alone and 52% survival rate with resection in combination with chemo-/radio-therapy [5]. The survival rates for NSCLC are generally higher, with 5-year survival rates of 70-90% for stage I NSCLC. They however show a similar pattern in regards to cancer stage at the time of diagnosis. While stage I NSCLC at diagnosis gave 1-year survival rates of 81-85%, stage IV NSCLC had 1-year survival rates of 15-19%. Being less aggressive than SCLC, 75% of NSCLC patients have either stage III or IV cancer at the time of diagnosis [5]. Still, this means that most LC patients with either SCLC or NSCLC already have advanced cancer once the cancer is first detected.

2.1.3 Screening

The high percentage of LC patients presenting with advanced or metastatic disease and the improved survival rates associated with early discovery and treatment are strong motivations for improved LC screening. Today, screening is mostly done via imaging techniques such as Chest X-Ray (CXR) or Low Dose Computed Tomography (LDCT), with LDCT providing more detailed images of the chest at a higher monetary cost and a higher patient radiation exposure than CXR [5]. In addition, several Randomized Controlled Trials (RCT) have shown that screening with either CXR or sputum cytology is not associated with a reduction in either LC death or the number of patients with advanced disease compared to the general population over time, even though more cancer cases are detected [6]. LDCT has shown more promising results in similar RCTs, with a significant reduction in overall LC death: the National Lung Screening Trial (NLST) reported 356 LC deaths in the LDCT screened population vs 443 LC deaths in the control population [6]. However, LDCT identifies both malignant and benign non-calcified nodules as cancerous, and as such, the screening method is associated with variable false positive rate leading to overdiagnosis [6]. The rate at which LC is overdiagnosed because of LDCT is unknown, and varies from study to study, but according to the RCTs reviewed in [6], across all trials and cohorts, 20% of patients in each round had positive results from LDCT that required follow-up, while in reality only approx. 1% of them actually had LC [6]. The radiation dose, risk of false positives and the monetary cost of each LDCT scan makes screening today focused on at-risk populations only: for instance, there is no evidence that there is benefit in screening never-smokers for LC using LDCT [5].

Methods for LC screening that are cheaper, less invasive, that can detect malignancies earlier and have higher specificity (lower false positive rate) are therefore needed, and several potential blood-borne biomarkers have been investigated to this end [15]. As a small part of this effort, this thesis will look at how miRNA sequenced from blood samples can help in screening for LC, as these RNAs have been shown to be deregulated in LC and other forms of cancer [15].

2.2 MicroRNA

2.2.1 Discovery

miRNAs were first described in 1993 and then later in 2000 by molecular geneticists studying the *lin-4* and *let-7* genes, respectively, of the nematode *Caenorhabditis elegans* (*C. elegans*) [16, 17]. It was already known that these genes had important roles in properly timing the development of *C. elegans*. However, instead of transcribing to protein-encoding mRNAs as expected, these specific genes transcribed to non-coding RNAs (ncRNA), and among them were short ncRNAs of $\sim$22 nt length [16, 17]. These short ncRNAs were further shown to have partial complementarity to multiple well-conserved sites within the 3' Untranslated Regions (UTR) of specific mRNAs that were also known to influence the *C. elegans* development [16, 17]. It was therefore proposed that in some way relating to this partial complementarity and consequent hybridization, these short ncRNAs inhibited translation of the protein-coding mRNA with complementary sites in their 3'UTR [7]. At first, this was thought to be an isolated case unique to *C. elegans*, but in following studies, the *let-7* gene was also recognized in humans and other animals, with similar effects on gene expression over time [7]. Following this came discoveries of other short ncRNAs similar to *let-4* and *let-7* in both size and in originating from regions of RNA transcripts that fold onto themselves to form hairpin structures [7]. Eventually, these became the class of ncRNA known as MicroRNA, and today, more than 1900 different human miRNAs are contained in miRBase[18], the online archive of miRNA sequences and annotations.

2.2.2 Biogenesis

All miRNAs are similar in size and general function, but their biogenesis pathway sometimes deviates from the general case, which is why the distinction between canonical- and non-canonical-miRNAs is often made. miRNA genes are found across the whole genome, often as purely non-coding genes with miRNAs as the only product or in the introns or UTRs of protein coding genes. miRNA genes are seldom found in coding exons, as processing of the miRNA would destroy the protein coding sequence [19]. Many genes that code for miRNA have been found to be well conserved with comparable function across species [7]. The miR-1 in humans and its identified orthologs in other species, such as flies and worms, have been found to have similar important functions in the muscle and heart tissue of these animals [20].

A simplified version of the metazoan canonical miRNA biogenesis as described in [7] follows: The canonical miRNA biogenesis starts with RNA Polymerase II (Pol II) transcribing a miRNA gene to a pri-miRNA, which is a much longer sequence than the mature miRNA, sometimes over 1000 nts in length. The pri-miRNA then folds onto itself by the base pairing of two, often quite close, regions that are (partially) reverse complements of each other along the pri-miRNAs direction. This folding happens at one or several locations along the pri-miRNA and creates double helix structures ending in unpaired loops: often called hairpins (see fig. 2.1-A). Hairpins with stems of specific lengths (35 ± 1 base pairs) ending in a terminal unstructured loop and surrounded by single-stranded regions are what the Microprocessor complex [1] in turn recognizes as their targets and cleaves off from the pri-miRNA, creating independent hairpins called pre-miRNA (see fig. 2.1-B). These hairpins are then transported from the nucleus of the cell to the cytoplasm by the proteins Exportin-5 and RAN-GTP. In the cytoplasm, pre-miRNAs are further processed by the Dicer enzyme, which cuts up the strands near the terminal loops, effectively removing the loop and breaking the continuous sequence, producing a so-called miRNA duplex, consisting of the mature miRNA and its (sometimes partial) complement passenger strand (miRNA*) (see fig. 2.1-D). This duplex is then bound to a high energy state Argonaute (AGO) protein, which expels the miRNA* through relaxing back to a lower energy state, leaving only the miRNA, the AGO, and other proteins to form an RNA induced silencing complex (RISC) incorporated with miRNA. The RISC then uses the miRNA as a guide strand for targeting different mRNAs by binding to their 3'UTR, regulating their expression by different modes of translational silencing (see fig. 2.1-E).

Non-canonical miRNAs also form RISCs with AGO, but often bypass interaction with either the Microprocessor complex or Dicer [7]. Also worth noting is that one miRNA gene might not always produce the same mature miRNA because of deviations in how Drosha and Dicer cut the miRNA precursors or which of the sequences in the miRNA duplex is expelled by AGO [19]. These variations are referred to as isomiRs, and open for even wider targeting of different mRNAs from the same miRNA gene as each isomiR will have different gene targets [19].

[1] Microprocessor complex - consisting of one Drosha endonuclease accompanied by two DGCR8 proteins

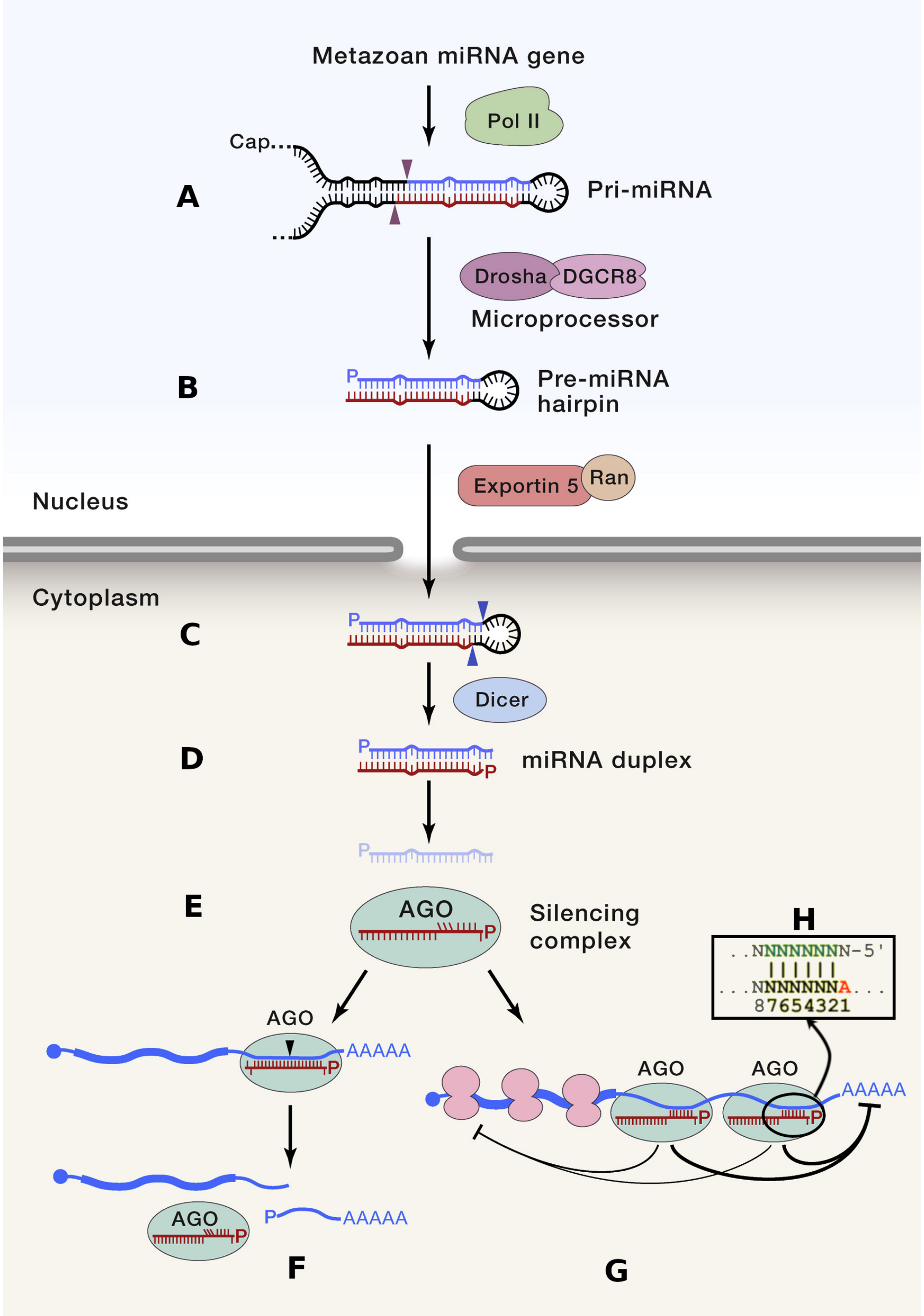

Figure 2.1: The canonical miRNA biogenesis and targeting pathway as seen in 'Metazoan microRNAs'[7] and seed region as seen in 'MicorRNAs-targeting and target prediction' [21].

2.2.3 Targeting and function

One of the key differences between miRNAs and other ncRNAs, like short interfering RNAs (siRNA), is that miRNAs do not require near-perfect complementarity to target mRNA to regulate their expression [21]. Once miRNAs are loaded into RISC, the silencing complex will seek out all mRNAs that are complementary by Wattson-Crick pairing. However, nucleotides at positions 2-7 from the 5' end of the miRNA, also known as the "seed" region (see fig. 2.1-H), have been identified as especially important because RISC uses these positions as nucleation signals for target recognition [21]. The importance of pairing in this relatively short region on the miRNA opens up for a wide range of potential mRNA targets by virtue of simple statistics. Some mRNAs will even have several target sites, most often in their 3'UTR as mentioned, resulting in additive silencing effects of the different matching miRNAs [21]. In humans, the most common effect of the miRNA mediated targeting is that RISC facilitates cleavage of the poly(A) tail from the mRNA, eventually leading to its degradation in the cytoplasm and in turn preventing its translation to proteins (see fig. 2.1-G) [19]. However, the nature of this silencing effect can vary and is often dependent on what kind of other complements exist between the miRNA and mRNA in question. In plants for instance, it is more common to see complementarity in the central region of the miRNA (nucleotides 9-11) in addition to the seed region, and this additional complement allows RISC to directly slice the target mRNA by endonuclease activity in AGO, causing an even stronger silencing effect (see fig. 2.1-F) [19].

More than half of human protein-coding mRNAs have one or multiple conserved miRNA target sites in their 3'UTR [22]. As such, any human disease or developmental process is more than likely subject to regulation by miRNAs at some point. Indeed, miRNA gene knockout studies in mice, in which one or more miRNA genes are made inoperable in the organism, sometimes provide for abnormal and unviable phenotypes, particularly when highly conserved miRNA families are knocked out [7]. The different miRNA knockout phenotypes in mice are remarkably diverse: some deteriorate in utero, while some have later severe developmental issues such as neurological disorders, infertility, blindness, deafness, immune disorders, cancers or other defects in a great variety of tissues [7].

miRNAs are therefore regarded as important regulatory ncRNAs in most animals, but mRNAs are already regulated extensively at the transcriptional level by chromatin, so why is their repression by miRNA in the cytoplasm necessary? miRNAs add another layer of regulation that extends and complements the transcriptional layer of control. They do this with mechanisms such as their wide mRNA targeting and their additive silencing effects on mRNAs with multiple targets. These mechanisms introduce complexities in the gene expression of each cell, which can help in development, differentiation and disease control in different tissues. David P. Bartel summarizes the importance of miRNAs in animals in 'Metazoan microRNAs'

[7]:

> ... the metazoan miRNAs can be thought of as the sculptors of the transcriptome. Transcription and other nuclear events set up a column of gene expression, and then the miRNAs, like a stone sculptor, chip away at this column. Occasionally they chip away enough to either trigger or sharpen a developmental transition, but more generally they produce a much more complex topology of gene expression, with more optimal levels of many proteins in each cell of each tissue.

2.2.4 Circulating MicroRNAs as Lung Cancer Biomarkers

miRNAs have been shown to be differentially expressed in cancer as both tumor suppressors and oncogenes (oncomiRs) [23]. miRNAs have also been shown to be deregulated in circulating blood for LC specifically [8–10]. This study uses blood serum and plasma samples as the main data source. miRNA patterns can be generated from serum, plasma and whole blood as they are present (to uncertain degrees) in exosomes, which are released in blood from almost all cell types, and some specific blood cells [24]. Although miRNAs have been found to be differentially expressed for LC cases and controls in individual studies, the biological mechanisms are complex and unclear. Differential expression studies often report different miRNA as being the most significantly differentially expressed [24, 25]. Expression of single miRNAs are accordingly not always disease specific - there are no specific circulating miRNAs found to date which can consistently separate LC cases from controls in a general sense across populations and cohorts, but there is reason to believe these miRNAs exist [24, 25]. As a consequence, the present study aimed to study miRNA signatures instead of single miRNAs for LC prediction across cohorts.

The process of sequencing miRNAs from serum and plasma blood samples was done by using a form of RNA-seq [12] where input material is enriched for small RNAs (commonly referred to as miRNA-seq). A detailed description of this process is included in section 3.1.1. miRNA expression profiles generated from NGS methods are more expensive, take a longer time, require a higher RNA contents in the sample, and generally requires a more complex infrastructure than qPCR and microarray methods [26]. However, they have a higher dynamic range and are not dependent on hybridization - one does not need to know the identity of the miRNAs to be sequenced, so NGS can combine de novo discovery and quantification of miRNAs. miRNA profiling with RNA-seq also introduces biases in the miRNA profiles, particularly for small and lowly expressed sequences [27, 28].

The differences in how miRNA profiling is done across studies, the differences in population characteristics of the source material used, and the batch effects that are introduced in the lab can all affect the reproducibility of miRNA differential expression studies. Therefore, the data preprocessing of miRNA profiles is an important step for identifying miRNAs that are actually differentially expressed in diseased patients, especially in multi-cohort studies such as this one. The data preprocessing step is outlined in section 3.1. Finally, to handle the inherent noisy and high dimensional nature of data amplified by PCR and sequenced by NGS, a machine learning method that can handle this kind of data was needed.

2.3 Random Forests

To deal with the complex biological nature and the high dimensionality of the miRNA expression profiles, random forests were chosen as the ML method for doing LC prediction. Random forests is an ensemble ML method that spawns a number of decision trees at training time [29]. By themselves, deep decision trees are prone to overfitting because of the rapidly decreasing sample sizes supporting each assumption as more splits are made, and these splits will model the training data exactly. Therefore, decision trees have low bias, but high variance - which can be particularly detrimental when modelling noisy transcriptomic data. Random forests try to address this high variance in individual trees by using bagging over several decision trees that are only trained on subsets of the original feature space, and sometimes a subset of the samples themselves[2] (bootstrapping). Both classification and regression tasks can be modelled with random forests - classifiers use the majority voted predicted class from the individual decision trees as output, while regressors take the average of the predictions from all the individual regression trees according to equation 2.1, where B is the number of bootstrapped samples, f_b is the decision tree trained on the bootstrapped training data and x' are the bootstrapped observations to be predicted.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{2.1}$$

Genomic data, such as miRNA expression matrices, are both high dimensional, having the property of large p and small n, and have features that are highly correlated. These kind of data are ill-suited to be modelled with classical statistical techniques, which often assume independent variables. Random forests can deal with high dimensional data because the individual trees will choose greedily among the bootstrapped features to split the data, and taking the average of these trees handles both the individual trees' high variance and the correlations between variables [30].

The most important features in the forest can also be conveniently extracted from the model by using the gini importance of each feature, which is used in the final experiment of this study. Gini importance is calculated by taking the decrease in node impurity and weighting it by the probability of reaching the node (node probability). The node probability is simply defined by the number of samples that

[2] Bootstrapping over samples was not used in the present study, see section 3.2.3.

reach the node divided by the total number of samples. Gini importance provides a good approximation of which features are most important for making a prediction, but it is not capable of totally excluding any features, and has also been shown to give biased representations in bioinformatic data specifically [31].

2.4 Autoencoders

An autoencoder was constructed during this project with the aim of creating a minimal representation of the miRNA matrices. Autoencoders have previously been used to extract latent miRNA features with promising results in miRNA-disease association prediction [32]. Autoencoders are feedforward, non-recurrent neural networks composed of two main parts - an encoder that transforms the data into the lower desired dimensionality code and a decoder that takes this code as input and tries reconstruct the original input of the encoder [33]. This general structure is outlined in figure 2.2. Backpropagation is done on all layers of the model and is based on the error in reproducing the original input data, meaning no target variables are fed to the model. By passing the data through the lower dimensional space, the autoencoder will hopefully learn to create efficient and minimal encodings for the training data; lower dimensional latent representations of the original data that retain only the 'most important' information. After training, the decoder part of the network can be decoupled, and the latent representations can be generated using the encoder only.

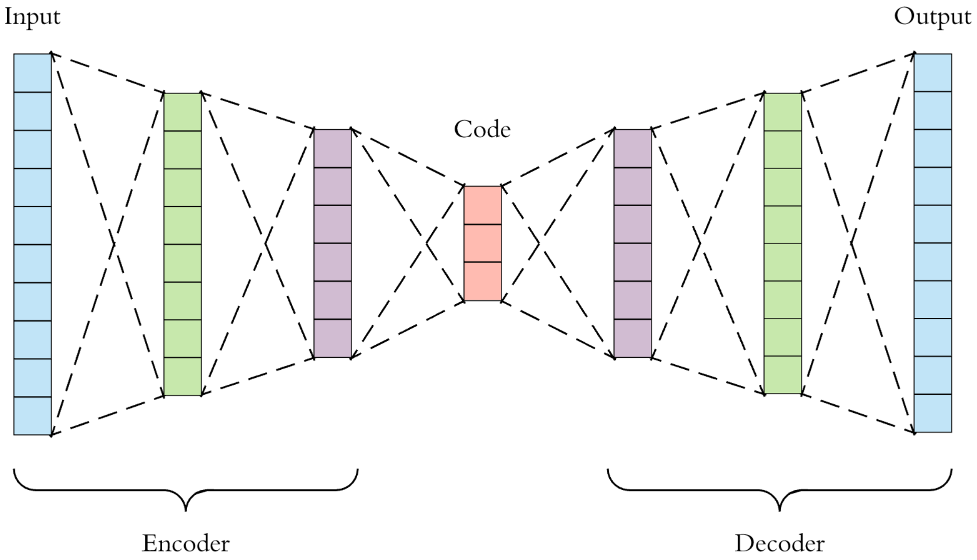

Figure 2.2: The general architecture of autoencoders. Figure taken from 'Applied deep learning - part 3: Autoencoders' [34]

.

Chapter 3

Methods

The following chapter describes the data preprocessing and modelling steps taken to create two random forest models, one classifier and one regressor, for predicting the LC diagnosis of a patient given a miRNA count vector. Both models were trained on prediagnostic cohorts - the classifier is trained on the miRNA count vectors in combination with their binary LC status, while the regressor is fed a linearly transformed LC status that is calculated with respect to the time elapsed between when the blood sample was taken and the LC diagnosis was made. Both models are finally tested on a diagnostic cohort, where the blood sample was taken on the same day as when the LC diagnosis was made. A flowchart visualizing the main steps taken to achieve this is included in figure 3.1. In addition to this main model, experiments were carried out to understand the miRNA profile data in a general sense. Some different dimensionality reduction and sampling techniques were performed to assess how they affected the predictive performance of models trained on miRNA data.

3.1 Data preprocessing

This section describes the steps taken to generate a robust multi-cohort miRNA-seq dataset used to train and validate models that try to predict for LC diagnosis given a miRNA count vector. Blood samples from The Central Norway Lung Cancer Biobank (CNLCB), The Trøndelag Health Study (HUNT), The Norwegian Woman and Cancer Study (NOWAC) and the Northern Sweden Health and Disease Study (NSHDS) were separately sequenced for miRNAs. The resulting data were then combined and processed, and the main patient cohort characteristics for the resulting data are included in section 4.1.1.

3.1.1 From blood samples to mature miRNA count matrices

The procedure for constructing mature miRNA count matrices from the patient and control blood samples of each cohort was carried out at the Department of

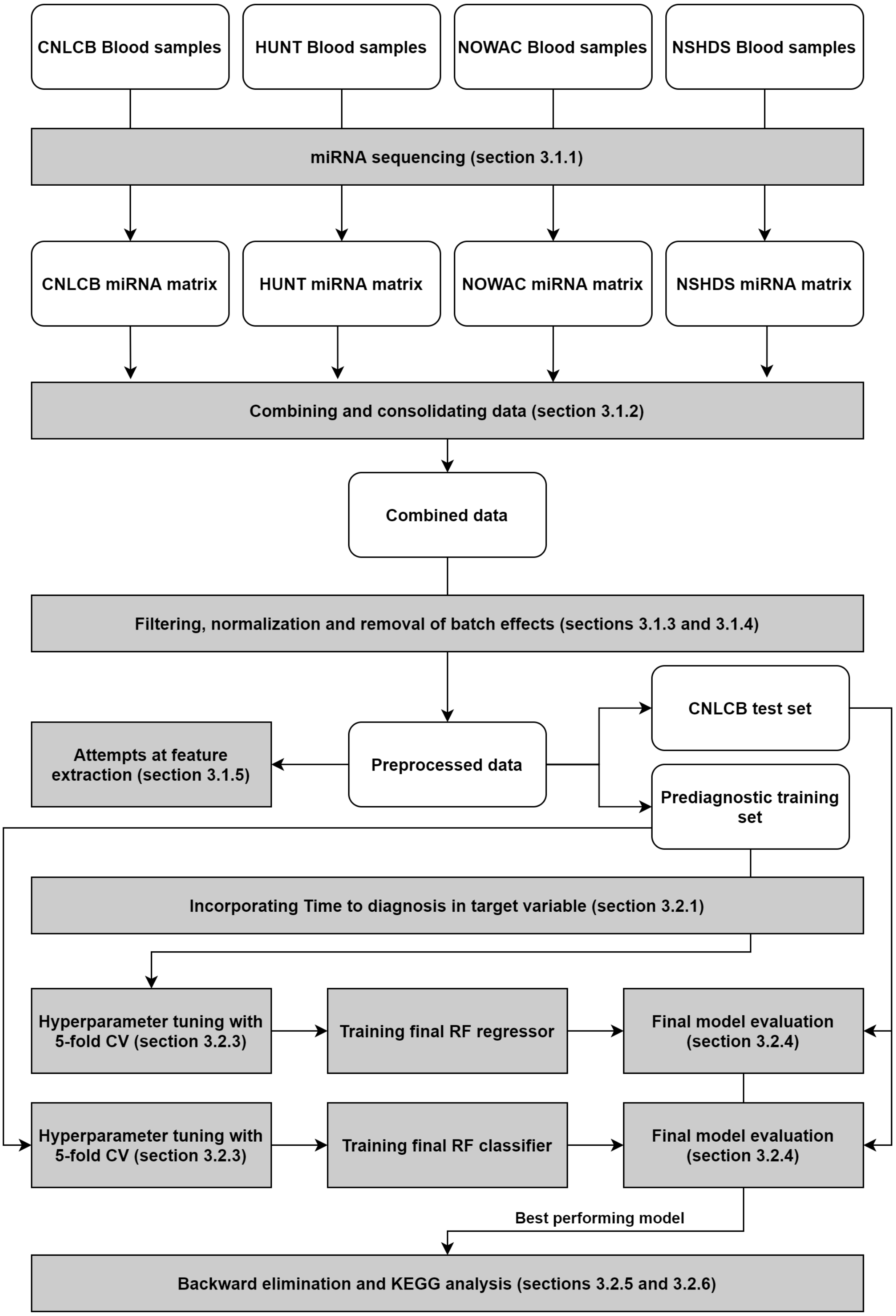
CNLCB Blood samples
HUNT Blood samples
NOWAC Blood samples
NSHDS Blood samples
miRNA sequencing (section 3.1.1)
CNLCB miRNA matrix
HUNT miRNA matrix
NOWAC miRNA matrix
NSHDS miRNA matrix
Combining and consolidating data (section 3.1.2)
Combined data
Filtering, normalization and removal of batch effects (sections 3.1.3 and 3.1.4)
CNLCB test set
Attempts at feature extraction (section 3.1.5)
Preprocessed data
Prediagnostic training set
Incorporating Time to diagnosis in target variable (section 3.2.1)
Hyperparameter tuning with 5-fold CV (section 3.2.3)
Training final RF regressor
Final model evaluation (section 3.2.4)
Hyperparameter tuning with 5-fold CV (section 3.2.3)
Training final RF classifier
Final model evaluation (section 3.2.4)
Best performing model
Backward elimination and KEGG analysis (sections 3.2.5 and 3.2.6)

Figure 3.1: A visual summary of the Methods chapter. White rounded boxes represent datasets, while the grey rectangles represent an abstract construct or methodology.

Clinical and Molecular Medicine at NTNU. Even though the four longitudinal cohorts collected different kinds of blood samples from the patients, either plasma or serum, they could be subjected to roughly the same procedure for miRNA quantification. A brief explanation of this process, described in more detail in [35, 36] with some minor deviations, follows.

After collection, the blood samples from the LC patients and control groups were initially stored at -80°C in a research Biobank for varying time intervals before being treated with the QIAGEN miRNeasy serum/plasma kit [37]. This kit is designed to isolate the RNA contents of both serum and plasma blood samples for subsequent sequencing. First, a reagent of this kit was added to the samples that promote the breakdown of the cell membranes (lysis), inactivate RNases, and extract most of the cellular DNA present in the sample. Then, the solution was centrifugated, forcing the broken down contaminants to the bottom of the solution, yielding an upper aqueous phase containing RNA, which was then extracted. Finally, these were added to the spin column included in the QIAGEN kit, which binds all the RNA to a membrane and washes away the final rest of contaminants. The RNA-containing solutions were then suspended in RNase-free water and stored away, again at -80°C, until quality control and the actual RNA sequencing stage of the process could begin.

Before starting library preparation, the RNA purity and concentration of the samples were assessed using the NanoDrop™ ND-1000 spectrophotometer [38]. Additionally, a random subsample of the samples were further quality controlled using Eukaryote total RNA pico assay on the 2100 Bioanalyzer [39]. Results from these controls were considered acceptable to continue with preparing the cDNA libraries. For this, the samples were prepared with NEXTFLEX® Small RNA-Seq Kit for Illumina® Platforms [40] according to instructions from the manufacturer. PCR Amplification was run for 13 cycles during this step, and ten different already known calibrator oligoribonucleotides were added to be used for internal controls. The resulting miRNA fragments were then sequenced on the Illumina® HiSeq system, with single read length of 50 base pairs, at the Genomics Core Facility (GCF) at NTNU.

The raw sequencing data were first quality controlled with fastQC [41], before they were trimmed for adapter sequences and then collapsed into single unique reads accompanied by their read counts. These reads were mapped to the human genome (hg38) with bowtie2 [42], and sequences that overlapped with mature miRNA loci were identified with htseq-count [43]. After filtering out reads with imperfect alignment to the genome, the total number of reads per miRNA was computed. The identified miRNAs were then annotated with miRBase v22 [18], and isomiR variants were identified using SeqBuster [44]. Finally, isomiRs with imperfect matching to the genome were removed, and the read counts were quality controlled with respect to the control oligoribonucleotides added before sequencing to ensure that each library had comparable sequencing runs. This whole process yielded mature miRNA matrices for each of the cohorts containing the raw read counts of all identified miRNA for each patient enrolled in the cohort.

3.1.2 Combining data from the different cohorts

As this study aimed to model a signal in the miRNA matrices that could predict LC across all the four cohorts, their data had to be carefully consolidated. Some differences existed between the cohorts in both the identities of the miRNAs included and in the kind of patient-specific data reported. miRNAs that were not identified in all the four cohorts were excluded from this study, decreasing the total amount of distinct miRNAs from 2272 to 1375. This inner join was done to ensure that the ML model would not be fed missing values during training or prediction and was therefore strictly necessary to produce viable cross-cohort models. Excluding this many miRNAs represents a significant loss of the feature space that could potentially hold information for making predictions. However, even stricter filtering for minimum expression is applied in the later steps. Also, the assumption can be made that miRNAs excluded in this way are more likely to only hold cohort-specific 'noise' than the remaining miRNAs. The resulting combined data included reads for 1375 different miRNAs for 1018 samples in total.

The patient-specific data were more complex to consolidate, as there were some differences in their level of detail and what kind of categories the different cohorts used. Cancer status, lane on the flow cell, sex, and age could all be extracted directly from the clinical datasheets. However, the histological subtypes, cancer stage groups, and smoking status had to be brought to the lowest common level of granularity. The histological subtypes were consolidated to either AD, SCLC, SQ or an 'Other' category. The cancer stage groups were consolidated to 'Early', 'Middle', or 'Advanced'. The smoking status became either 'Current', 'Former', or 'Never'. A column indicating what kind of blood sample the cohorts used was also added to the patient matrices. CNLCB and HUNT are based on blood serum samples, while NOWAC and NSHDS are based on blood plasma. Finally, the Time to diagnosis was calculated for each patient. This column represents the elapsed time between when the blood sample was drawn and when the LC diagnosis was established. HUNT, NOWAC, and NSHDS are all prediagnostic cohorts, meaning that a certain amount of time elapsed between the blood sample date and the diagnosis date for all patients. This was not the case for CNLCB, as blood samples in this cohort study were taken at the time of diagnosis. This fact was later leveraged in the design of the final model.

Some final touches were then done to remove all missing values and make presenting results easier. Missing values for patient age were imputed according to the mean (only a couple of added data points). Relevant continuous variables were made categorical for presentation of results. An overview of the resulting data is presented in 4.1.1.

3.1.3 Filtering and normalization

Once the data from the four different cohorts were properly combined, they were ready to be count filtered and normalized. This step is primarily done to reduce the noise in the mature miRNA count matrices that stem from either the blood samples themselves or from steps taken in section 3.1.1. First, utilizing the Bioconductor package edgeR [45], all miRNAs that did not have more than 2^6 count-per-million (CPM) in more than half the total samples (509) were filtered out of the data set. CPM was calculated according to equation 3.1, where X_i represents the number of reads for a particular miRNA in a sample and N is the total number of reads for that sample.

$$CPM_i = \frac{X_i}{N} * 10^6 \tag{3.1}$$

Filtering by CPM instead of raw read counts ensure that differences in per-sample library sizes are accounted for. In essence, CPM filtering makes it more likely that the remaining miRNAs in the data set are biologically relevant, as miRNAs generally need to be expressed above a certain threshold to have any real effect on gene expression (see section 2.2.3). Also, the discrete nature of the sequencing itself makes it so that the difference between a 0 and 1 read count, for example, can be very high, and this can interfere with the statistical methods used later (see also [46]). Lastly, this study looks at miRNAs expressed in circulating blood, and one could assume that lowly expressed miRNAs in this context are more likely to be related to unwanted background noise. This exact filtering was found by using a combination of searching for the optimal values in the prediagnostic cohorts[1] with regards to AUC and from recommendations in literature[2]. In total, this step further reduced the number of miRNAs in the combined miRNA matrix from 1375 to 193 unique miRNAs.

After filtering, the raw counts were transformed to $log_2(CPM)$, with CPM as defined in equation 3.1. CPM was used for the same reason as in the filtering stage - to normalize each read in a sample with the total miRNA content of the sample so that any two samples with different sequencing depth can be compared. With log transformation on top of this, the proportional rather than the additive differences between the reads are modeled, and proportional differences are often more interesting in this particular context. Also with base 2, the log transformation intuitively scales with PCR amplification, as a doubling in amount of reads corresponds to a change of exactly 1 $log_2(CPM)$. At this stage, further normalization is often done on RNA-seq data to reduce the variability between

[1] CNLCB was excluded in the search to avoid overfitting.

[2] Different studies vary in regards to recommendations on how strict this filtering should be, but 2^6 is within the ranges observed in multiple studies.

each identified transcript. An explanation for why this step was omitted is given in section 5.2.1.

3.1.4 Dealing with batch effects

Even though all cohorts were processed in a similar manner all the way from blood sample collection to the mature miRNA matrices, the presence of cohort-specific batch effects is highly probable. There are several steps involved in both sample taking, preparation, and sequencing that can potentially introduce these kinds of effects. Each of the four cohorts were sequenced within different time frames, and this means that the temperature and ozone levels in the lab and the staff performing the liquid handling could vary for each of them. These are well-known significant batch effects in NGS data [47]. Also, cohort-specific batch effects naturally include the biological differences between blood plasma and blood serum, as the cohorts vary in this regard as well. The existence of these cohort-specific batch effects were confirmed with PCA plots for the whole miRNA matrix with cohort coloring (see section 4.1.2). To reduce the impact these effects, the `removeBatchEffect()` function from the Bioconductor package limma [48] was used. This function works by fitting a linear model to the miRNA matrix including the batch vector, and subsequently removing the component in the matrix related to the batch. It is worth noting that data outputted by this method are ill advised to use with linear or statistical differential expression models at a later stage, as it is more robust to provide the batch vector as a covariate [49]. For the ML methods used in this study, including a categorical variable like this could not be done in an elegant way.

3.1.5 Further dimensionality reduction

The miRNA matrix produced by the preceding steps contained 193 normalized miRNA counts for 1018 patients. Producing a minimal version of the feature space could improve both the running time of the algorithms used and potentially increase the final model performance. However, many considered methods of dimensionality reduction, such as LASSO or feature importance in random forest, require a separate discovery dataset. This is because these methods filter out features that do not contribute to or decrease the accuracy of the model in a supervised manner, indirectly revealing the target variables to the model. Applying such methods lead to overfitting if the data is not isolated from the training or validation sets. Supervised feature extraction methods were consequently not used because retaining as many samples as possible in the training and validation sets was prioritized. Nevertheless, there exist methods for reducing dimensionality in an unsupervised manner that can be applied on training and validation data without overfitting as consequence, and some of these were tested. PCA was used primarily as a visualization tool. Autoencoders were tried out to see if an encoded version of the feature space could retain or even improve the model's predictive performance. Lastly, only using miRNAs with known LC associations in the literature was tried out to see if it could help in producing a minimal model with comparable or better performance.

PCA

2 component Principal Component Analysis (PCA) was used to visualize the main differences between samples in the miRNA matrices. PCA reduces the dimensionality of a matrix by finding the greatest variance by some scalar projection, and making this the first coordinate of the matrix. This step is then repeated to produce as many components as needed [50]. PCA was mainly used to verify the data preprocessing. In particular, the removal of batch effects had quite visible effects on the two principal components of the complete miRNA matrix (see section 4.1.2).

Autoencoders

In this study, an autoencoder with an architecture as outlined in figure 3.2 was constructed to reduce the dimensionality of the data from 193 to 32 dimensions. During training, the autoencoder's weights and biases were updated by backpropagation on its ability to recreate the miRNA matrix according to the Mean Squared Error (MSE) loss function. Adam [51] was used as the learning rate optimization algorithm. After 100 epochs over the training data, the model loss converged and the encoder part of the architecture was ready to be decoupled and used to create encoded versions of miRNA vectors. To test how the resulting encoding influences the predictive value of the miRNA data, an experiment was carried out on the whole dataset: two random forest classifiers with default hyperparameters from the scikit-learn implementation [52] were trained on the complete miRNA matrix and the encoded version respectively to predict the biological sex of each patient. It is well established that there is differential expression of miRNAs between biological sexes [53], and one could therefore assume that this experiment could serve as an adequate baseline for assessing data quality. Stratified 5-fold cross validation on the complete miRNA dataset was used to generate training and testing data. This experiment served as both a sanity check on the preprocessed complete data and as a test to check whether the encoding could have either positive or negative effect on predictive performance. ROC curves and AUC (see section 3.2.4) for each fold were used as evaluation metrics to compare the models trained on the original and encoded data respectively.

Known miRNAs

Several potential circulating miRNA biomarkers for LC have been identified in the scientific literature [54, 55]. Differential miRNA expression studies often report only the very best performing miRNAs for separating cases and controls, and the

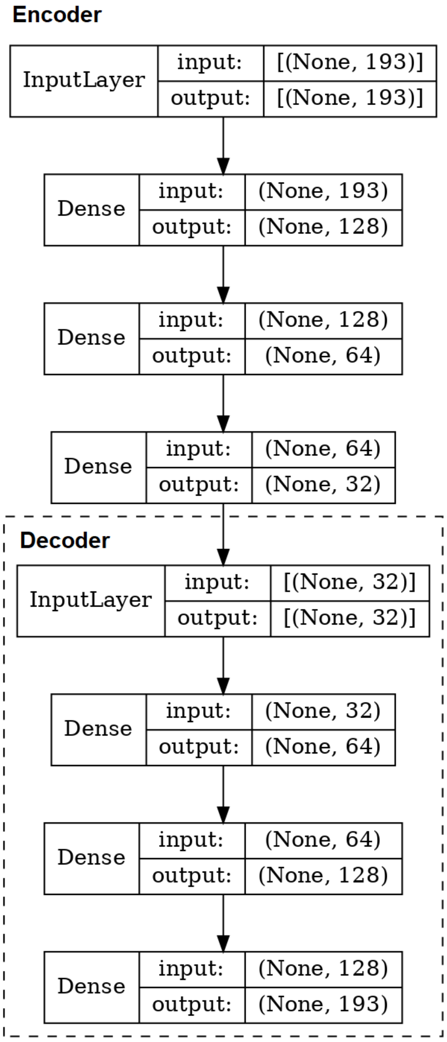

Figure 3.2: The architecture of the autoencoder tried out for dimensionality reduction. The dotted rectangle represents the decoder, which is decoupled after training to produce a 32-dimension version of the feature space. All hidden layers used ReLu as activation function.

Table 3.1: The panel of miRNAs identified in [55] having $log_2(CPM)$ values in the preprocessed data.

hsa-miR-10b-5p
hsa-miR-125b-5p
hsa-miR-126-3p
hsa-miR-126-5p
hsa-miR-146b-5p
hsa-miR-155-5p
hsa-miR-182-5p
hsa-miR-183-5p
hsa-miR-193a-5p
hsa-miR-19b-3p
hsa-miR-21-3p
hsa-miR-21-5p
hsa-miR-25-3p
hsa-miR-30c-5p
hsa-miR-335-3p
hsa-miR-335-5p
hsa-miR-483-3p
hsa-miR-483-5p
hsa-miR-486-3p
hsa-miR-486-5p
hsa-miR-7-5p

identified miRNAs frequently vary from study to study. To construct a panel of circulating miRNAs interesting in a LC context, a systematic review of differential expression studies in western populations was used [55]. This review screened differential expression studies according to several selection criteria and reported miRNAs with significant LC association in ≥ 2 of the selected studies. A dataset was constructed based on this panel by inner joining it on the complete dataset. The resulting dataset only included $log_2(CPM)$ for the miRNAs listed in table 3.1.

An experiment was then carried out to assess how this panel would fare in separating LC cases from controls in comparison with the complete data as prepared in sections 3.1.1-3.1.4. Again, two random forest classifiers were constructed with default hyperparameters from sklearn [52] and trained on the complete data and the data derived from the miRNA panel respectively. As this feature selection technique is focused on only retaining miRNAs with LC association, the models were trained and evaluated on their ability to separate LC cases from controls. To reduce the noise from LC cases with diagnosis date far into the future, all LC cases with more than 5 years in Time to diagnosis were excluded from this experiment, yielding an imbalanced dataset with 504 controls and 407 LC cases. Stratified 5-fold was used to evaluate the models, and ROC and AUC were reported for each fold.

3.2 Modelling

This section describes the steps taken to train and validate Random forest models on the data as prepared in the previous section. None of the further attempts at dimensionality reduction produced acceptable results in their respective experiments to warrant their use in further modelling. Throughout the modelling stage, both Random forest classifiers and Random forest regressors from scikit-learn [52] were trained and evaluated on the data as prepared in sections 3.1.1-3.1.4. Classifiers were trained to distinguish cases from controls in a binary fashion, with 0s representing controls and 1s representing LC cases. The random forest classifiers were trained without any consideration of Time to diagnosis, while the random forest regressors were trained to fit the miRNA expression profiles to a continuous target variable calculated by incorporating Time to diagnosis.

3.2.1 Incorporating Time to Diagnosis in the target variable

The time difference between the blood sample date and diagnosis date for patients in the prediagnostic cohorts could have significant effects on the performance of models trained on this data. The distribution for Time to diagnosis in the different cohorts is reported in figure 4.1. In some cases, the time difference is so significant that the blood samples could have been taken before any LC signal was present in the miRNA expression profile of the patient. To control for these time differences and also assess the degree to which they affected performance of models trained on this data, a linear transformation of the target variable in the prediagnostic cohorts was performed. The transformation was done according to equation 3.2 on all patients with an established LC diagnosis.

$$Y_{TTD} = 1 - \frac{\Delta T_{TTD}}{max(\Delta T_{TTD})} \tag{3.2}$$

Here, the elapsed time between drawing of blood samples and diagnosis date of each patient was divided by the maximum Time to diagnosis observed in the dataset and then subtracted from 1. All controls' target variables were set to 0. This transformation requires the assumption that any signal indicating LC will linearly decrease as the time between blood sample and diagnosis date increases. The exact shape of the function of time on this signal is not known, but this linear transformation produces a continuous target variable between 0 and 1 in which patients with a value closer to 1 have a higher probability of actually providing an LC signal in their miRNA expression profile. This transformation was only done on training data fed to the model, and functioned as a weighting of the examples from the prediagnostic cohorts. Time to diagnosis could have been included as an additional feature in the training data, but this would diminish the models capacity as a diagnostic tool, as for every example fed to the model, a Time to diagnosis feature would have to be given as well. Introducing this feature would also be complicated for the control groups, as it would not be applicable for these samples. Finally, models trained against this transformed target were only evaluated with binary targets, in practice making the model a probabilistic classifier, which when compared to the true binary targets could provide sensible ROC curves and other performance metrics relevant for any diagnostic tool.

3.2.2 Sampling

Two main sampling techniques for validation of trained models were utilized: hold-one-cohort-out and stratified 5-fold cross validation. As the goal of the study was to find a generalizable signal in the miRNA matrices that could predict for LC in any sample, the models' ability to predict LC across cohorts was emphasised. As patients in the CNLCB cohort had a more certain disease status due to the fact that blood samples were taken at the same time as a diagnosis was established and the fact that the prediagnostic cohorts' targets could be linearly transformed for training purposes as detailed in 3.2.1 - the CNLCB cohort was mainly used as the test cohort while the prediagnostic cohorts served as training data.

Hold-one-cohort-out was also used the other way around: to assess the predictive performance of models trained on one prediagnostic cohort validated on the remaining prediagnostic cohorts, but this was only used to inform decisions on the final design and to visualize how Time to diagnosis affects the training data. In the same vein, stratified 5-fold cross validation was used on the prediagnostic cohorts to gain understanding of how well models performed when trained and evaluated on only themselves. Some results from these hold-one-cohort-out experiments are included in section 4.2.1, while the full version is included in appendix A. It is important to note that all models from these experiments were validated on binary target variables, meaning that even if an LC diagnosis was set in the far future, it was classified as a case of LC. The hyperparameters used for this experiment were set to the default parameters for scikit-learn [52] random forests, except for `n_estimators`, which was set to 1000.

To see how models trained on CNLCB only could predict for LC on itself, classifiers for a stratified 5-fold cross validated CNLCB were trained and evaluated. Finally, to see how a model trained on CNLCB performed in the other cohorts, a classifier trained on the full CNLCB set was evaluated on the prediagnostic cohorts. As the CNLCB data only has binary targets, even when incorporating Time to diagnosis, only results from the classifiers trained on CNLCB are presented in the results section. For this experiment, the same hyperparameters from the classifier as in table 3.2 were used. Some results from this experiment are included in section 4.2.1, the full results are presented in appendix A.

3.2.3 Hyperparameter tuning

To find the optimal hyperparameters for the main random forest models used in this study, a mix of randomized and grid search on the most interesting parameters was done for both the regressor and classifier. The most interesting parameters for both were identified as: the number of trees used in the whole forest (`n_estimators`), the number of features to consider for each split in a tree (`max_features`), the maximum depth of each tree in the forest (`max_depth`), the minimum required samples for splitting a node in a tree (`min_samples_split`), the minimum num-

Table 3.2: The optimal hyperparameters for the RF classifier and regressor found by random- and grid- search.

Final hyperparameters		
Hyperparameter	**RF Classifier**	**RF Regressor**
`n_estimators`	1400	1500
`max_features`	log_2	log_2
`max_depth`	50	5
`min_samples_split`	5	2
`min_samples_leaf`	10	1
`bootstrap`	`False`	`False`

ber of samples required for each leaf in a tree (`min_samples_leaf`), and finally whether a bootstrapping of the samples or the whole dataset should be used when generating each tree in the forest (`bootstrap`). As CNLCB would serve as the final test set for the model, only the prediagnostic cohorts were used for evaluating the performance of each hyperparameter combination in the search.

First, random search with 5-fold cross validation was used to narrow down the search for the best performing hyperparameters. The `RandomizedSearchCV()` function from scikit-learn [52] was used for this purpose. The classifier and regressor each had to use different scoring functions because of the nature of their respective target function: as the regressor fits a model against the linearly transformed target calculated with the help of Time to diagnosis, negative Mean Absolute Error (MAE) was used. For the classifier, accuracy was used as the scoring function. The randomized search narrowed down the range of plausible values for each hyperparameter: some ranges had particularly poor results and could therefore be excluded. The narrowed search space was then searched exhaustively with the `GridSearchCV` function from scikit-learn[52], still using 5-fold cross validation and the same scoring functions for the classifier and regressor respectively. The final hyperparameters for both models were found using a combination of extracting the hyperparameters with the best results in cross validation overall, and an analysis of plots that visualized the mean score for each value of the hyperparameters in the search. The final hyperparameters for the random forest classifier and regressor are included in table 3.2.

3.2.4 Model evaluation

In diagnostic testing, simply outputting the binary predicted diagnosis status for a patient is often not sufficient, as the optimal balance between sensitivity versus specificity varies from disease to disease. In cancer diagnosis specifically, the cost of false positives are high, and so any diagnostic test should have adjustable prediction thresholds so that this trade-off can be tuned in an ad hoc manner. Consequently, the probability of a patient having LC was used as model output throughout. The RF regressor naturally outputs a probability that a given patient has LC, as the transformed target variable in the training data models this probability according to an approximation of the time decay on the miRNA signal. For the RF classifier, this probability has to be deduced from the proportion of trees in the forest that votes in favour of either a positive or negative diagnosis status, and this is done via the built-in `predict_proba()` function from the sci-kit learn implementation of random forests.

ROC curves were used as the main performance evaluation technique in this study. ROC curves visualize the sensitivity and specificity of a probabilistic binary classifier by plotting the true positive rate and the false positive rate for all the possible cutpoints of the predicted probability [56]. Area Under the Curve (AUC) is defined as the area under the ROC curve, and is useful as a summary of the ROC curve. Accuracy, specificity, recall and F1 score were also calculated for the main classifier and regressor respectively. These performance metrics require a set probability threshold, which was already set for the RF classifier. The probability threshold for the regressor was calculated by choosing the optimal value according to the F1 score.

3.2.5 Backward elimination of miRNAs

For the best performing model, which in this case was the Random forest regressor, a form of backward elimination was performed to assess which miRNAs contributed the most in predicting LC correctly. As gini importance has been shown to give biased representations of feature importance [31], this measure was only used as an ordering for which miRNAs to iteratively remove. Instead, for each removed miRNA, a model was trained and evaluated with AUC. A detailed explanation of the steps to perform backward elimination follows:

1. The feature importance of each miRNA was extracted from the best performing model using the `feature_importances_` attribute from scikit-learn [52]. This is an attribute that gives the gini importance for each feature in the random forest ensemble. The miRNAs were then sorted in ascending order according to the extracted feature importance.
2. For each miRNA in this ordered list, the following was performed iteratively until the list was empty:
    a. The first miRNA of the list was removed, ie. the least important miRNA.

   b. Training data from the prediagnostic cohorts and testing data from CNLCB containing only the remaining miRNAs in the list was generated.
   c. A random forest regressor was fitted to the training data (with transformed target variables).
   d. Prediction was done on the CNLCB test data with the fitted model.
   e. AUC for the predictions done in the previous step was calculated.
   f. The miRNA that was removed and the corresponding AUC was appended to a 2d array for analysis.

3. Finally, the resulting 2d array was plotted. The x-axis was set to the removed miRNAs and the y-axis was set to the AUC value for the model trained and evaluated on the resulting miRNA matrices.

The resulting plots were visually examined, and ten of the miRNAs were identified as having a particularly strong effect on AUC.

3.2.6 KEGG pathway analysis of most important miRNAs

The ten identified miRNAs that had the most significant effect on AUC after removal were further studied for the functional pathways and regulatory roles of their target genes. For this, DIANA-mirPath v3 [57] was used. TarBase v7.0 was used for miRNA target prediction, which provides experimentally validated miRNA targets. mirPath v3 outputted the KEGG pathways [58–60] associated with the identified target genes, and the P-value threshold for inclusion of a pathway was set to 0.05. The final KEGG pathway data outputted by DIANA-mirPath v3 were downloaded, and the $-log_{10}$ adjusted p-value for the 20 most significant pathways were plotted.

Chapter 4

Results

4.1 Results from data preprocessing

This section is designed to give an overview of the data as prepared in sections 3.1.1-3.1.4, and to provide results from the attempted dimensionality reduction described in section 3.1.5. First, some important characteristics of the cohorts are presented to give a brief overview of the datasets used in modelling. PCA plots before and after removal of batch effects are also included, which show how the miRNA matrices themselves were consolidated. Finally, results from models trained both on the encoded miRNA matrices and the known miRNAs with LC association in literature compared with the original data are presented.

4.1.1 Clinical characteristics

The most interesting clinical characteristics of the processed data is summarized in table 4.1. The final dataset was quite balanced in regards to cohorts, sample groups and blood sample types. Biological sex was skewed towards female, this is mainly because NOWAC is a female only cohort. Time to diagnosis distributions for all LC cases are visualized in figure 4.1. CNLCB is diagnostic, so Time to diagnosis is always 0. The other cohorts vary to some extent in their mean and max values of Time to diagnosis.

4.1.2 PCA for visualizing batch effects

The effects of using the limma `removeBatchEffect()` [48] function is visualized by PCA in figure 4.2. Before batch effect removal, there is a clear clustering of the cohorts in the two principal components. The type of blood sample used also correlates with the value of the principal components - the serum cohorts CNLCB and HUNT cluster together on both PCs, while the plasma cohorts NOWAC and NSHDS cluster along PC1. After batch effect removal, the cohorts are much more correlated in the two PCs, while apparent outliers are still kept.

Table 4.1: An overview of the most interesting clinical and histological characteristics of the combined and pre-processed data. Note that missing values are not included and some variables are only applicable to certain cohorts and sample groups.

Column	Variable	Number of patients
Cohort	CNLCB	222
	HUNT	246
	NOWAC	274
	NSHDS	276
Sample Group	Case	514
	Control	504
Histological subtype	AD	204
	SCLC	111
	SQ	126
Sex	Female	585
	Male	410
Age	<39	10
	40-49	81
	50-59	426
	60-69	353
	70-79	121
	80-89	27
	>90	0
Stage groups	Early	95
	Middle	138
	Advanced	248
Smoking status	Current	456
	Former	357
	Never	156
Blood sample type	Plasma	550
	Serum	468
Years to diagnosis	0-2 years	89
	2-4 years	107
	4-6 years	118
	6+ years	58

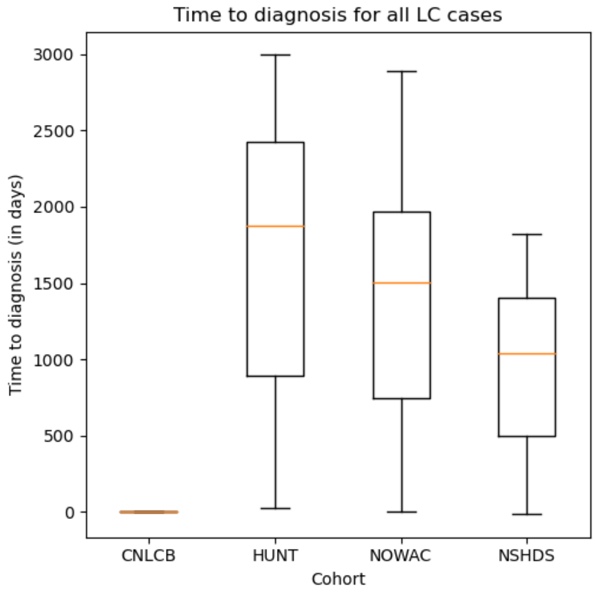

Figure 4.1: Box plot showing distribution of Time to diagnosis for each of the cohorts

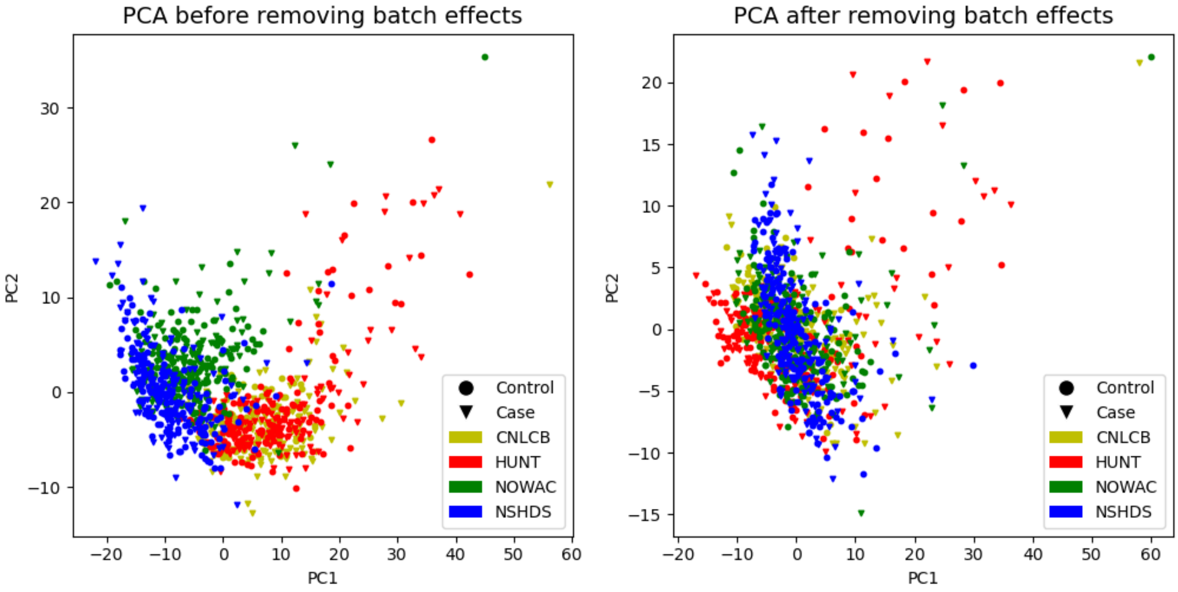

Figure 4.2: The effect of applying the limma removeBatchEffect() function to the combined $log_2(CPM)$ transformed miRNA matrix. Coloring is used to show cohort membership, and LC diagnosis is reflected in the shapes.

4.1.3 Autoencoders

The autoencoder's ability to reconstruct the complete miRNA matrix after passing it through a 32-dimensional space before and after training is visualized in figure 4.3 and 4.4 respectively. These figures show the $log_2(CPM)$ expression of each miRNA from three random patients in the training data: The orange plots represent the original read counts for each miRNA and the blue plots represent the results after passing it through the whole autoencoder architecture. Figure 4.4 shows that the encoder is able to recreate the data quite accurately, meaning that the 32 dimensional encoded representation has enough information to recreate the dataset. The original complete data and the encoded data were then subjected to an experiment - predicting the biological sex for each patient was used as a baseline. Figure 4.5 shows the ROC curves for predicting biological sex based on models fitted to the original and encoded data respectively. The models trained on the original data has a 0.1 higher mean AUC on the different folds than models trained on the encoded data, which is a significant difference.

4.1.4 Known miRNAs in literature

The performance of the models trained on the original data and on the panel of miRNAs listed in table 3.1 is presented in figure 4.6. It is worth noting that for this experiment, an upper bound of 5 years was set for Time to diagnosis, making controls overrepresented in the test data for all the folds. There was a small decrease of 0.05 in mean AUC going from the original dataset to the 21 miRNA panel from literature. This was deemed to be of sufficient significance to exclude the panel in further modelling.

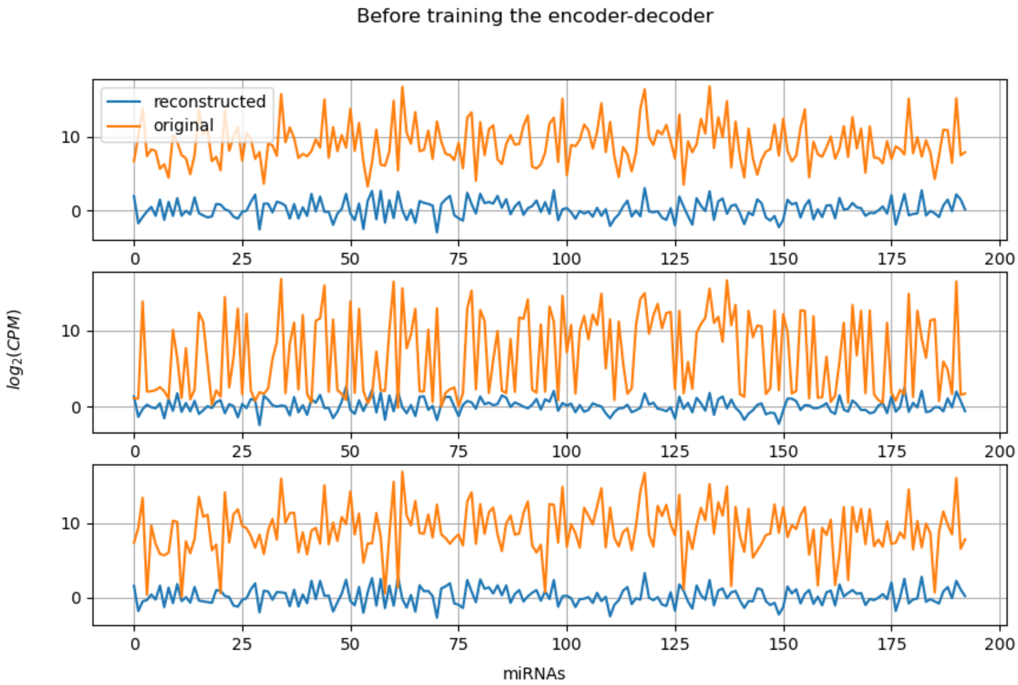

Figure 4.3: The autoencoder's ability to recreate the miRNA vectors of three random patients before training for 100 epochs. The x-axis represents each miRNA and the y-axis represents $log_2(CPM)$ values.

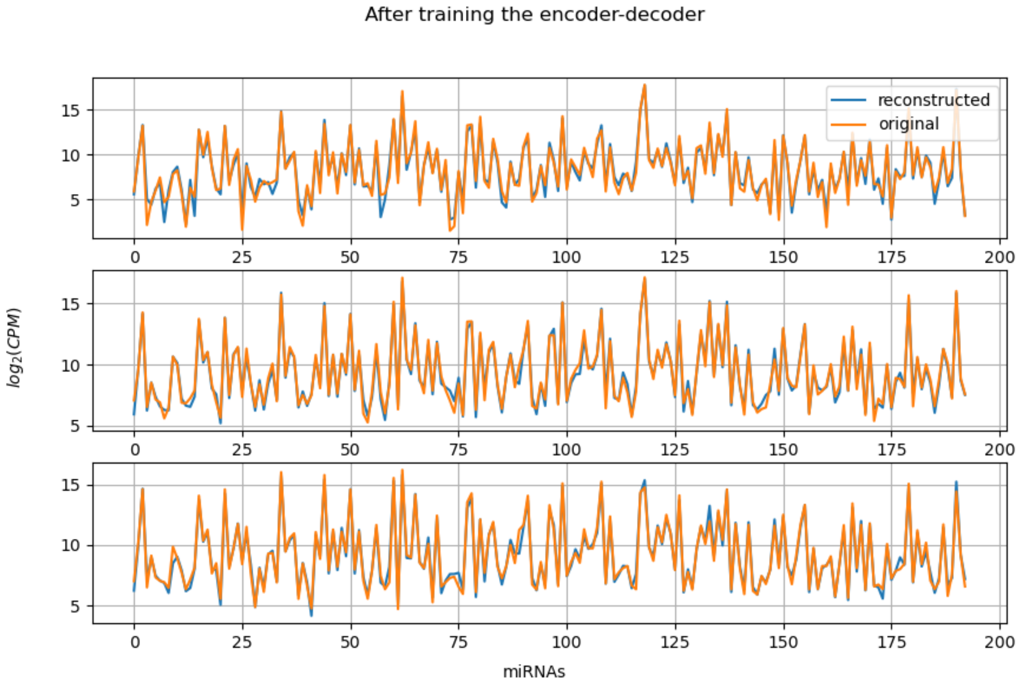

Figure 4.4: The autoencoder's ability to recreate the miRNA vectors of three random patients after training for 100 epochs. The x-axis represents each miRNA and the y-axis represents $log_2(CPM)$ values.

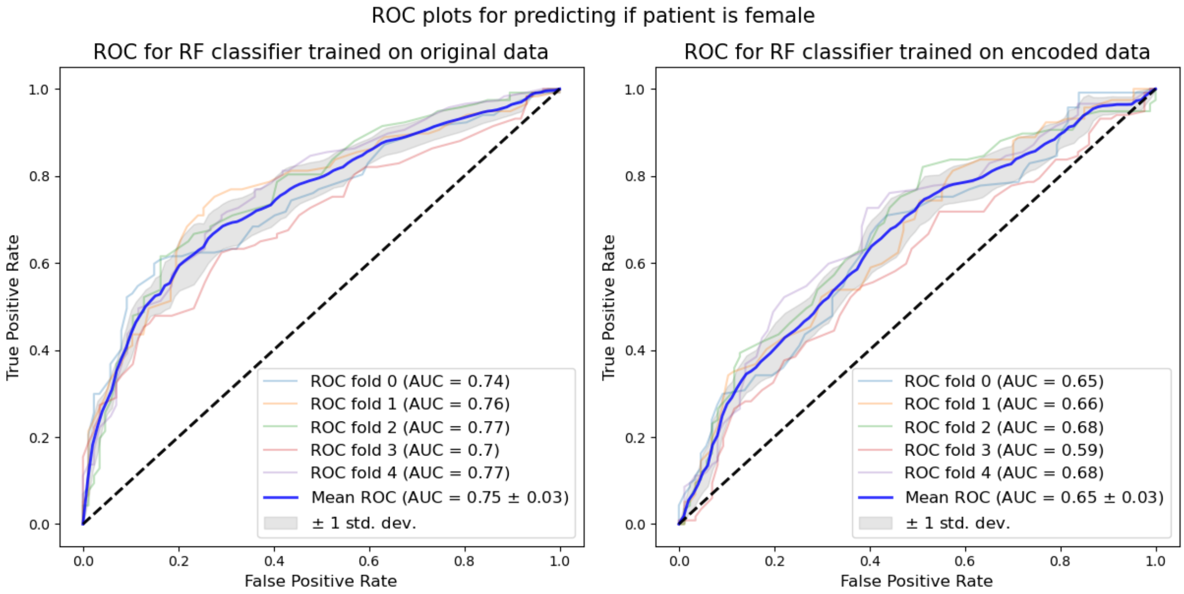

Figure 4.5: ROC curves for Random forest classifiers trained on the original miRNA matrices and the encoded 32-dimensional version respectively. Predicting the patient's biological sex was used as an experiment to assess the encoders ability to retain important information.

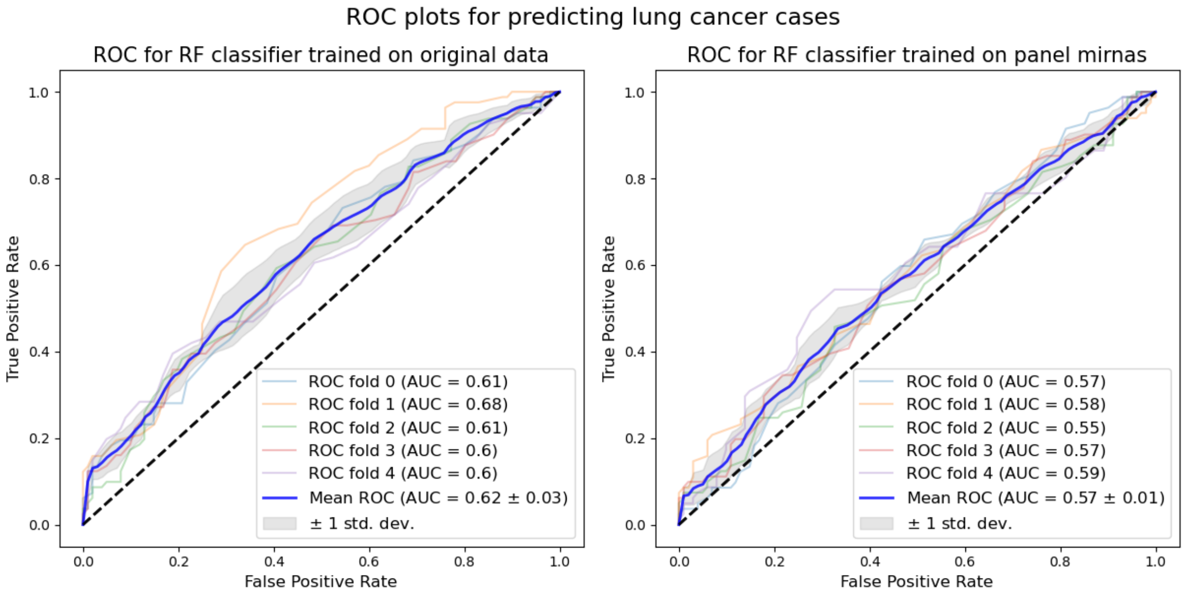

Figure 4.6: ROC curves for Random forest classifiers trained on the original miRNA matrices and the miRNA biomarker panel drawn from [55] respectively. Predicting lung cancer cases across 5 stratified folds was used as an experiment to gauge how the miRNA panel performed.

4.2 Results from modelling

This section presents results from section 3.2 - Modelling. First, results from the hold-one-cohort-out experiments are presented, which help in visualizing the problems with Time to diagnosis for validation of results. Then, results from the main models of this project are presented. The best performing models are investigated to generate a miRNA signature that was particularly significant in separating LC cases from controls in the diagnostic CNLCB cohort. Finally, the most significant biological regulatory pathways of these miRNAs are presented.

4.2.1 Hold-one-cohort-out validation

Prediagnostic cohorts

Hold-one-cohort-out validation was used as an experiment to assess the prediagnostic cohorts' predictive performance on each other. Additionally, each cohort was trained and evaluated on a stratified 5-fold cross validated version of itself. For each cohort, the leftmost plots in figure 4.7 show the ROC performance of both classifier and regressor models trained on the other cohorts and validated on the current cohort. The righmost plots of figures 4.7 show the performance of regressor models trained on a stratified 5-fold cross validated version of the current cohort. A more extensive version of these plots are included in appendix A. These plots show that models trained on the prediagnostic cohorts have poor predictive performance on each other and on themselves. The models were validated on the binary LC status of the patients, and no filtering on Time to diagnosis was done for this experiment. This means that any model will predict wrong in cases where the signal in the miRNAs of the blood samples is not yet present or too weak. Also, the varying time intervals for Time to diagnosis in the different cohorts (see figure 4.1) might make this effect even stronger. Despite these complications in the validation data, the regressor seems to perform marginally better than the classifier in all hold-one-cohort out experiments.

Models trained on CNLCB only

An experiment was carried out to see how classifiers trained on only CNCLB predicted LC on both itself and in the prediagnostic cohorts. ROC plots for 5-fold cross validation and a model trained on the whole CNLCB set are included in figure 4.8. It is evident in the leftmost plot of this figure that LC cases are far more separable in the diagnostic cohort than in the prediagnostic ones (see figure 4.7). Additionally, the classifier trained on the whole CNLCB set is significantly better at predicting LC cases in HUNT than in the other cohorts, which might be explained by the fact that both CNLCB and HUNT are based on serum blood samples. The performance on NSHDS and NOWAC for this classifier is close to that of a random classifier.

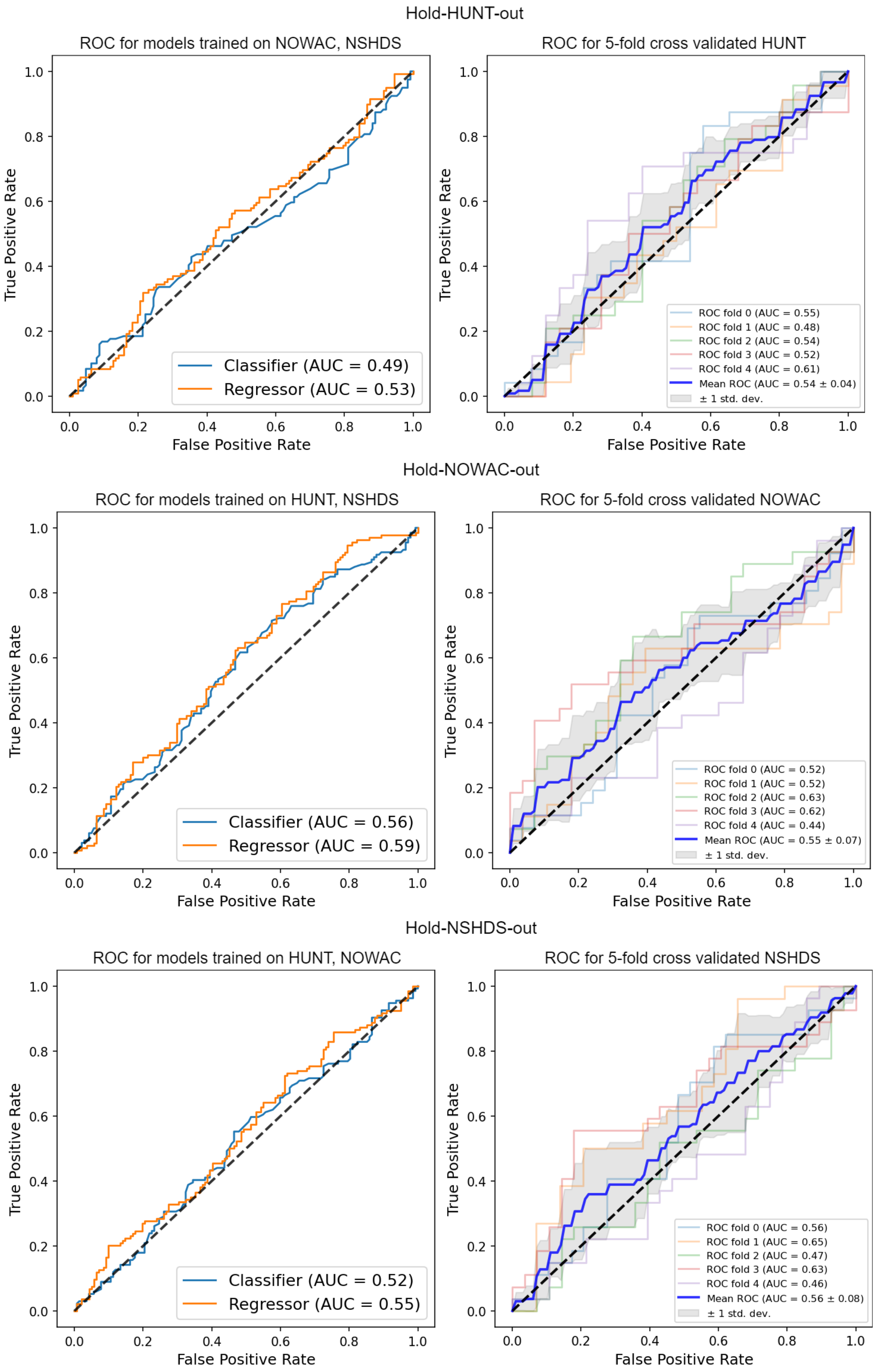

Figure 4.7: ROC plots showing results from the hold-one-cohort-out experiment on the prediagnostic cohorts.

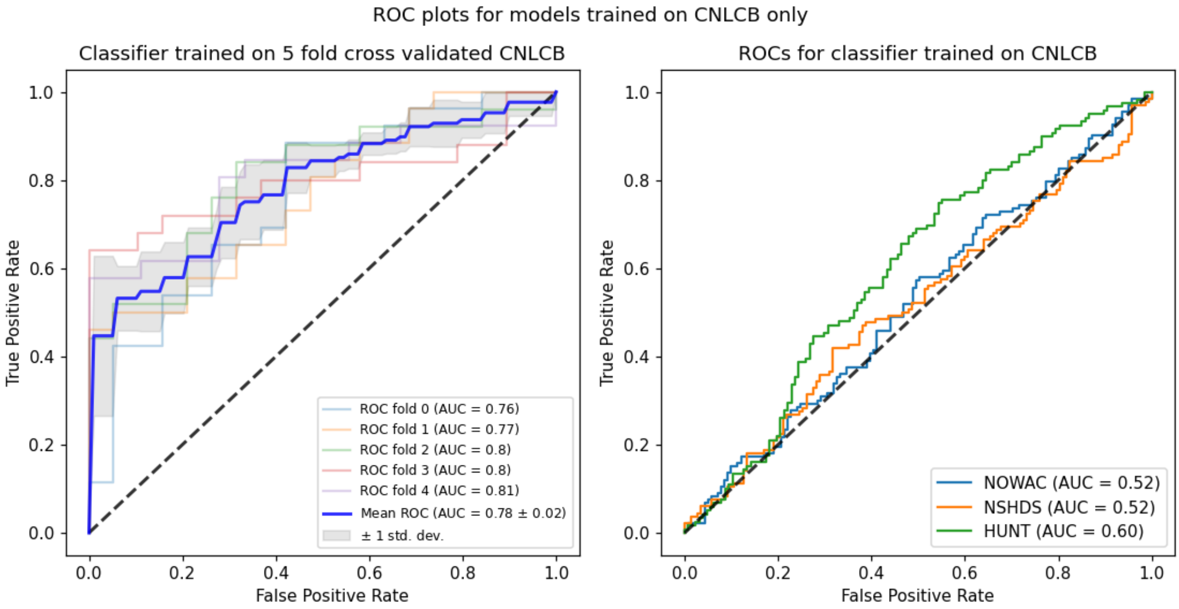

Figure 4.8: ROC curves for Random forest classifiers trained on CNLCB only. Left: Classifiers trained and tested on stratified 5-fold cross validated CNLCB. Right: Classifier trained on the whole CNLCB set evaluated with ROC plots for the different prediagnostic cohorts.

4.2.2 Time to diagnosis

Using the prediagnostic cohorts as training data and CNLCB as test data provided the best cross-cohort predictive performance. This can be explained in part by the fact that in CNLCB, the actual LC status of patients was known at the time of blood sampling. Also, this is the only experimental setup for which hyperpararameter tuning was done. The performance of models trained on the prediagnostic cohort and validated on CNLCB is reported in figures 4.9 and 4.10 for the classifier and regressor respectively. For both plots, the leftmost plot shows the model's performance on the test CNLCB set and the rightmost plot shows how the models generalized over the prediagnostic training data.

The classifier did not account for Time to diagnosis, while the regressor was trained on the linearly transformed targets from section 3.2.1. The regressor has a significantly higher AUC, F1-score and recall on the test set than the classifier. The classifier also overfits the training data noticeably, with an AUC of 1 and only 3 misclassified patients in total. This might be because of differences in hyperparameters chosen for the classifier, particularly a larger `max_depth` parameter will provide for more overfitted individual decision trees: for the classifier this parameter is set to 50 compared to 5 in the regressor. A smaller `max_depth` was also tried out on the final classifer, but provided worse results on the testing data. These results show that incorporating Time to diagnosis in the prediagnostic cohorts during training increased model performance on the diagnostic cohort. The regressor trained on the prediagnostic cohorts provides comparable results to that of the 5-fold cross validated models presented in figure 4.8.

4.2.3 Prediction performance across subgroups

The performance of the regressor model (figure 4.10) on different patient subgroups in CNLCB is reported in figure 4.11. The model is better at distinguishing late stage cancer than early stage cancer. Late stage LC ($n = 63$) with an AUC of 0.79 and early stage LC ($n = 23$) with an AUC of 0.65 when compared against all controls. ’LabCtrl’, the entirely healthy subset of the control groups, seem to be more separable from LC cases as well with an AUC of 0.94 compared to all LC cases, but this group has a low sample size ($n = 7$). The model is best at separating the histological subtype SCLC from controls ($n = 20$), with an AUC of 0.87 on this particular patient group. The model seems to perform roughly equally well across the other patient characteristics: smoking status is not strongly correlated with how well the model separates cases from control, neither is biological sex or age.

4.2.4 Most important miRNAs

The most important miRNAs in the best performing model were identified, and the biological functional pathways of these were found with KEGG [58–60]. The results from backwards elimination are included in figure 4.12. AUC for each removed miRNA is reported in figure 4.12. The yellow area of the top graph is focused in the bottom graph so that individual miRNAs are readable. The AUC for each removal is quite stable at first: none of the models perform significantly better than the starting model from figure 4.10. After about 125 removals the AUC starts to suffer, and at 183 removals the AUC has a sharp decline. Note that the uptick seen in the in the last miRNA removal is synthetic as this is a model trained on an empty dataset, representing a random chance classifier.

The miRNAs that produced this sharp decline in AUC were carried over for further examination. Figure 4.13 shows these 10 miRNAs’ gini importance in the original regressor model. Finally, the 20 most significant KEGG pathways associated with these top 10 miRNAs are reported in figure 4.14, which plots the $-log_{10}(p)$ value for each pathway. ’ECM-receptor interaction’, ’Hippo signaling pathway’ and ’Proteoglycans in cancer’ are quite separated from the other pathways in their significance. Some other cancer related pathways in this plot are: ’Chronic myeloid leukemia’, ’Renal cell carcinoma, ’Glioma’, and ’Pathways in cancer’.

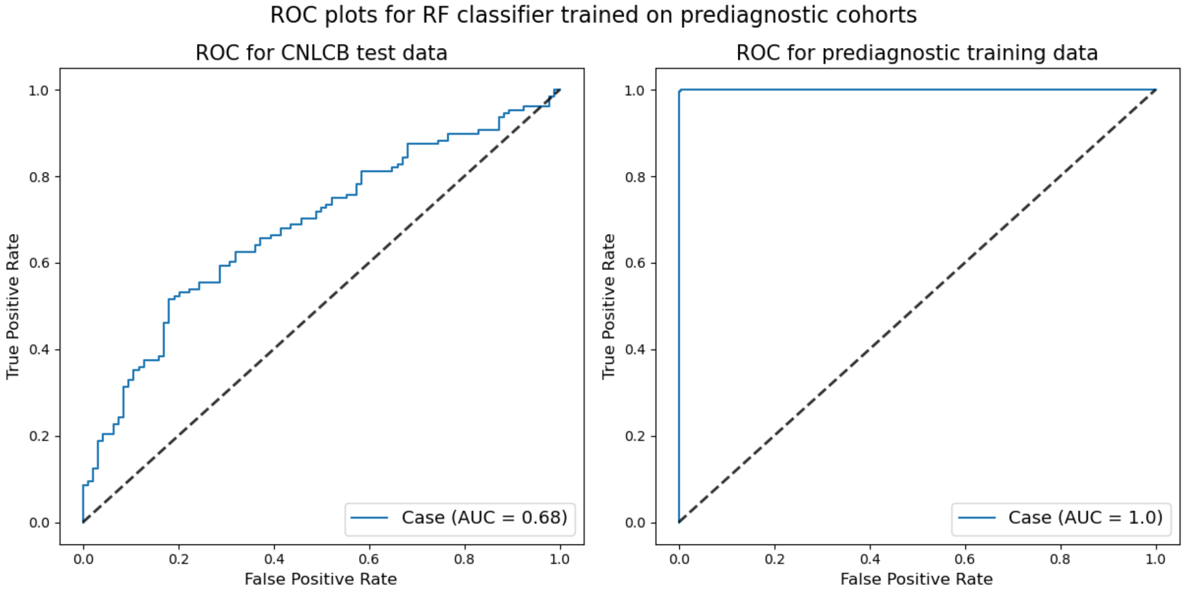

Figure 4.9: ROC plots for random forest classifier trained on prediagnostic cohorts. Here, nothing is done to address the Time to diagnosis variable. The model had an F1-score of 0.66, an accuracy of 0.64, precision of 0.72, and recall of 0.61 on the CNLCB test set.

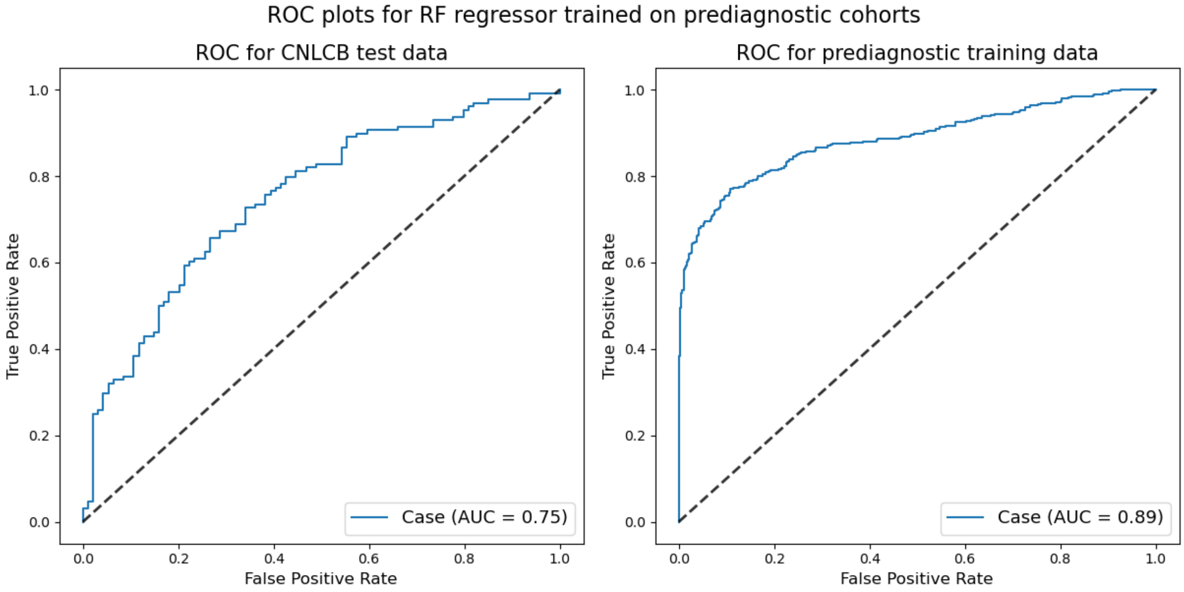

Figure 4.10: ROC plots for random forest regressor trained on targets where Time to diagnosis was accounted for (see section 3.2.1) in the prediagnostic cohorts. For calculating other metrics, the decision boundary was set so that F1 was maximized. This yielded an F1 of 0.78, accuracy of 0.70, precision of 0.69, and recall of 0.89 on the CNLCB test set.

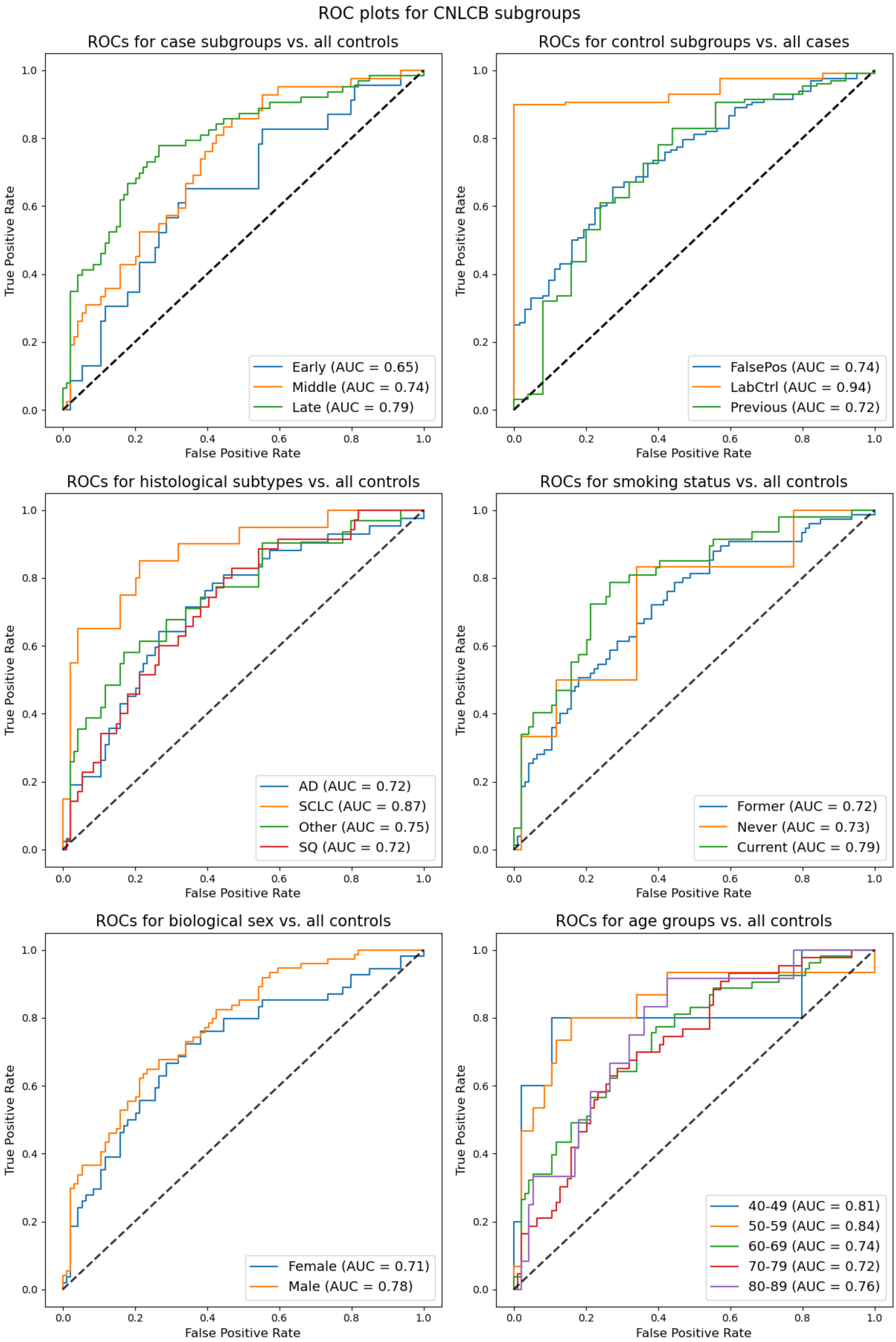

Figure 4.11: ROC for best performing model (figure 4.10) across CNLCB subgroups.

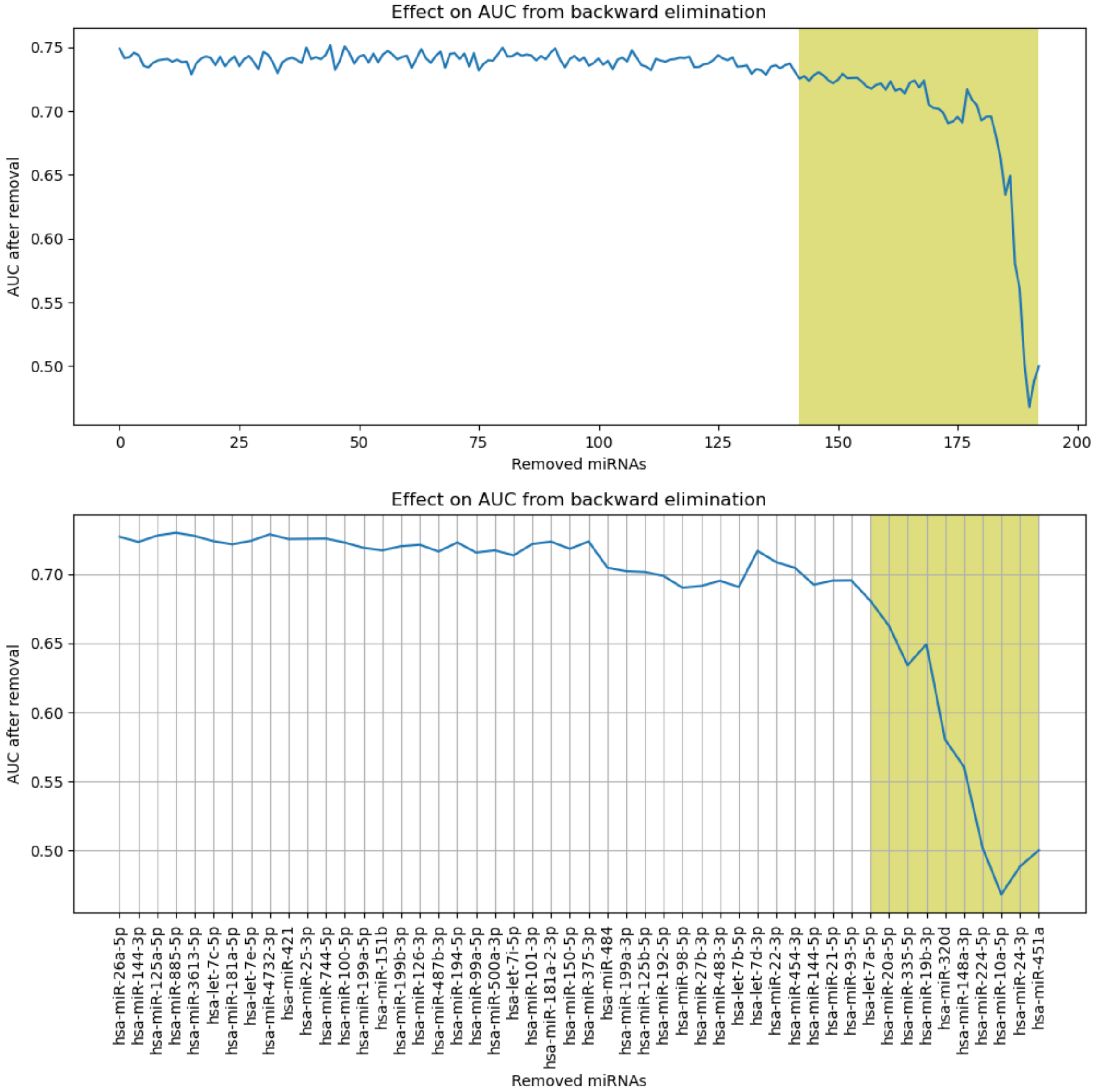

Figure 4.12: The effect of iteratively removing the 'most important' miRNAs from the final regressor model. AUC on the test CNLCB set is reported for each removal. The bottom graph presents the yellow area of the top graph zoomed in. The yellow area of the bottom graph are the identified miRNAs with most significant effect on AUC.

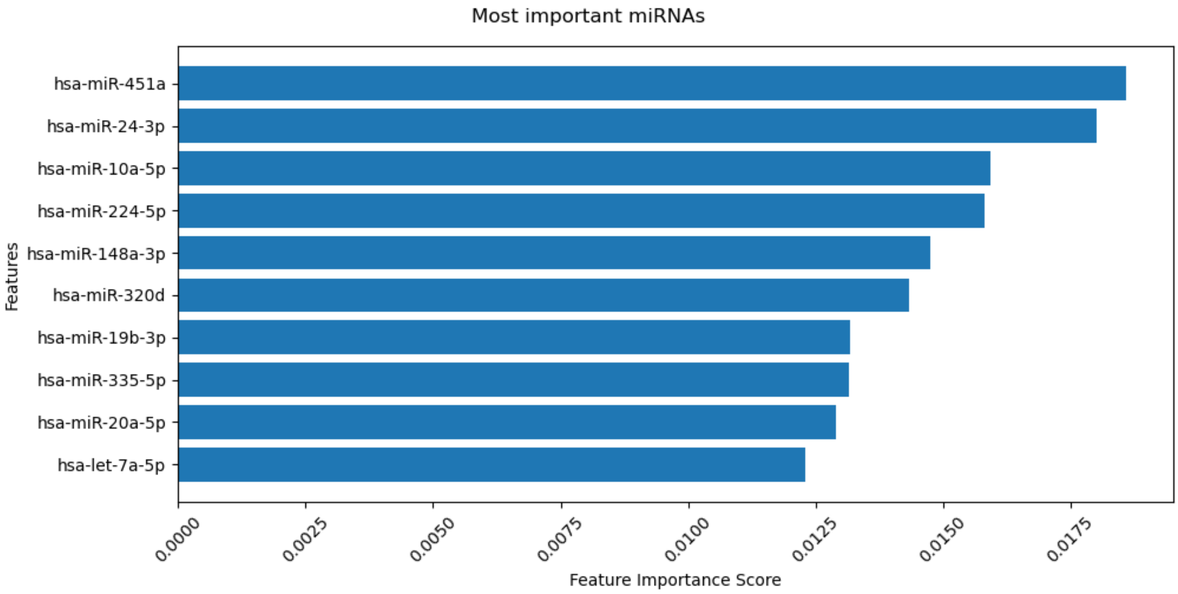

Figure 4.13: The gini feature importance of the best performing regression model for each of the top miRNAs as identified in figure 4.12

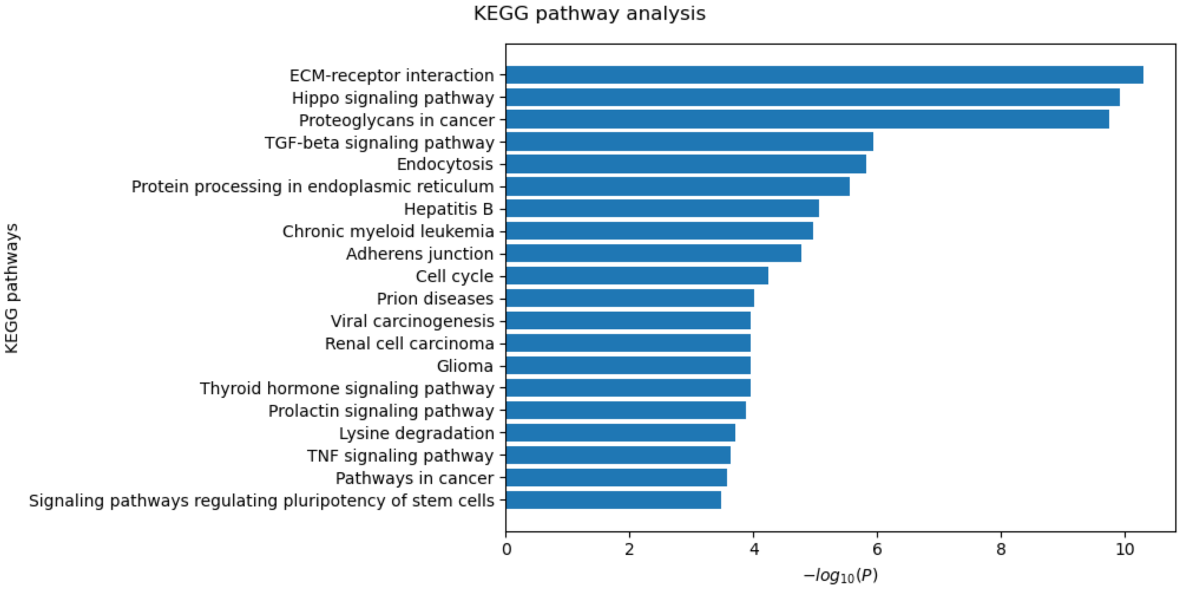

Figure 4.14: The 20 most significant KEGG pathways of the top miRNAs in figure 4.13.

Chapter 5

Discussion

This chapter is split into two parts: First, in section 5.1, the research questions posed in the introduction are answered with respect to the results in the previous chapter. Then, in section 5.2, a more in-depth discussion of potential flaws, strengths, and potential points of further work of the whole project are presented.

5.1 Answers to research questions

RQ1: How well can LC cases be distinguished from controls in a cross-cohort manner with random forests based solely on miRNA expression profiles?

When tested on the prediagnostic cohorts as in figures 4.7 and 4.8, none of the models could provide an acceptable separation of LC cases from controls. Further work in modelling of the prediagnostic miRNA signal is needed - a discussion of this is included in section 5.2.2. However, models trained on the transformed targets in the prediagnostic cohorts provided fair results on the diagnostic cohort: Figures 4.10 and 4.11 show how well the best performing random forest model could separate cases from controls across cohorts based solely on miRNA expression profiles of patients. On the whole CNLCB test set the model had an AUC of 0.75. With a decision boundary set to one that maximizes F1-score, the model had an F1-score of 0.78, accuracy of 0.70, precision of 0.69, and recall of 0.89. Worth noting is that the CNLCB test set included 128 LC cases and 94 controls in total, which is why F1 was used for setting the optimal threshold[1]. These are fair results considering prediction is done across cohorts. Maximizing F1 seemed to favor a model with higher recall (sensitivity) and lower precision. This might not be the optimal thresholds for doing prediction in a clinical setting, but can be tuned ad hoc.

[1] As F1 is often regarded as a better metric when evaluating performance on imbalanced data sets [61].

RQ2: In what way does the ML model's ability to distinguish LC cases from controls depend on the elapsed time between the blood sample date and the diagnosis date of patients?

Time to diagnosis directly influences both the training and validation of ML models that attempt to predict for LC. When comparing models trained on data where this variable is unaccounted for with models where the training data is transformed, the latter outperforms the former with a difference of 0.06 AUC and 0.12 in F1-score, as shown in figures 4.9 and 4.10. This is also evident in the single cohort 5-fold cross validation experiments visualized in figures 4.7 and 4.8. The models trained on the 5-fold cross validated diagnostic CNLCB data consistently performed better, with mean AUC of 0.78, than any of the other folds in the 5-fold cross validated versions of the prediagnostic data, which had varied mean AUC from 0.54 to 0.56 and much higher variance between the different folds.

RQ3: How can feature extraction methods help in improving the final model's predictive performance while also reducing model size?

The only kind of feature extraction that had a positive effect on the predictive performance of the random forest models was the filtering stage as outlined in section 3.1.3, in which miRNAs with less than 2^6 CPM in half of the total samples ($n = 509$) were filtered out. This step in particular could have been subject to further optimization, but the filtering used had a noticeable effect on predictive performance. This can be explained by the problems associated with lowly expressed miRNAs in differential expression studies, this rationale is also further explained in section 3.1.3. Two other methods of dimensionality reduction were tested: autoencoders and known miRNAs from literature. These were less successful (see figures 4.5 and 4.6) and consequently not used in further modelling. However, these results could not be deemed conclusive, and a further discussion of this is included in section 5.2.1.

RQ4: To what degree could the final model be used as a diagnostic tool?

Considering the results from the final model in figures 4.10 and 4.11 and its most important miRNAs with target gene pathways as outlined in figure 4.14, the model undoubtedly picks up a LC specific signature that could be useful in a clinical setting. However, the model does not have the precision required to be used directly in diagnosis of patients, as precision is important to avoid false positives. Additionally, even though the KEGG pathway analysis gives some insight into the most important miRNAs, the model might not be sufficiently explainable, which is an important characteristic for it to be useful in medical practice. This kind of ML model is probably more suited to be used as part of an ensemble of different methodologies for LC detection, in combination with methods such as LDCT, and possibly with other ML methods that use different biomarker data.

5.2 Strengths, weaknesses and further work

Following is a general discussion about the most important findings and lessons learned during the work on this thesis. Where relevant, proposals for particularly interesting areas of further research are also made.

5.2.1 Data preprocessing

Normalization

Normalization along the feature axis was omitted in data preprocessing. After the miRNA matrices had been filtered and normalized for per-sample sequencing depth, differential expression studies often have an additional data preprocessing step that was omitted in this particular study. As $log_2(CPM)$ only normalizes along sample library size [46], further normalization with methods such as Trimmed Mean of M (TMM) or quantile normalization is often performed on RNA-seq data to reduce the variability between each identified transcript. This is also recommended with miRNA-seq data when doing differential expression analysis [62]. However, this project utilized Random forests in the modelling stage, and with this kind of tree based method (see section 2.3), normalizing along the feature axis is at best unnecessary, and might even introduce unwanted biases at worst. The trees of a random forest are not influenced by the absolute value of any feature, they only look at the ordering of the values of each miRNA to decide the best thresholds for a split. There have also been discussions about whether TMM might be suited for miRNA-seq data at all [63]. Consequently, this additional normalization step was dropped in favor of keeping all variability observed between the miRNA features.

Autoencoders

Models trained on the encoded miRNA matrices did not provide for comparable or better predictive performance than models trained on the original data. The use of autoencoders to reduce the number of features in high dimensional data is a promising concept, as these kinds of techniques reduce dimensionality in an unsupervised manner. They can also reduce noise in the data, and provide for a more minimal feature space. The autoencoder as described in section 3.1.5 with its predictive performance on biological sex as reported in section 4.1.3, was not used in the main modelling stages of this project because of its poor performance. However, there is reason to believe that further development of this concept could provide for comparable or even better performance when applied on miRNA matrices and other high dimensional genomic data. First, the autoencoder used in this study was not extensively optimized, the architecture and setup described in 3.1.5 followed the generic autoencoder design. The structure of the network, the activation functions, number of hidden layers, the loss function, and the final latent dimensionality to be extracted from the encoder could all have been optimized for the miRNA domain to a greater extent.

miRNA expression profiles generated by NGS methodologies are inherently noisy, especially for lowly expressed miRNAs in the biological samples [28]. Adding a denoising component to the autoencoders could potentially help in solving this. Denoising autoencoders (DAE) have been used in medical imaging to remove noise and artifacts from diagnostic images with promising results [64]. By synthetically adding noise to the miRNA matrices that resemble the noise generated during sequencing[2], then feeding the noisy data to an autoencoder that backpropagates on its ability to recreate the original data - the encoder could learn how to remove the noise component in the miRNA matrices. Additionally, the loss function of the autoencoder could be tweaked so that precision is more important in highly expressed miRNAs than in lowly expressed miRNAs, placing a higher value on precision in the miRNAs that are sequenced more accurately. Hence, the design of the autoencoder presented in this study holds several potential points for further development, and its results should not be regarded as discouragement for further use in miRNA based disease prediction.

Known miRNAs in literature

Models trained on miRNAs with known LC association in literature [55] performed worse than models trained on the complete dataset in separating LC cases from controls. There are many potential reasons for this. As mentioned in section 2.2.4, there is little overlap in identified differentially expressed miRNAs in LC association studies [8–10]. This is often attributed to batch effects, population differences in the cohorts used, and differences in the actual procedures used for generating the miRNA data: there are a multitude of ways to isolate and sequence the miRNA contents of blood samples and there is no 'golden standard' method for this procedure [24, 25]. Additionally, in the experiment carried out in this study, the miRNA panel was tested on a mix of prediagnostic and diagnostic cohorts[3]. The miRNAs that are potentially differentially expressed at the prediagnostic stage are not necessarily the same as the ones that are differentially expressed at the time of diagnosis, and the panel of miRNAs were based on diagnostic cohorts. In summary, the performance of the panel miRNAs from [55] hints at a more general problem of reproducibility in miRNA disease association studies. However, the results presented here are in no way conclusive that the panel will not work in studies with exactly the same experimental procedure as used in the individual studies generating the miRNA panel.

[2] This might be possible as the characteristics of the biases in RNA-seq methods are systematic and highly reproducible [27].

[3] Albeit with filters on Time to diagnosis.

5.2.2 Modelling

Incorporating Time to diagnosis in the target variable

The linear transformation applied to the prediagnostic targets is simple and naive, but provides a reasonable linear approximation of the miRNA LC-signal decay. The transformation as presented in section 3.2.1 assumes that the decay of the LC-signal in the miRNA matrices are of a linear nature as Time to diagnosis grows. However, this is more than likely a simplification of the time-dependent variable: there is no reason to assume that as a patient approaches diagnosis, the expression of miRNAs with LC association increases/decreases in a linear fashion. In addition, the target approaches zero when Time to diagnosis reaches the maximum observed value for Time to diagnosis in the data. This max value was set for practical reasons, as it provided a continuous distribution of targets between 0 and 1, but it probably does not reflect reality accurately. A more exact approximation of this time decay could possibly provide for better adjusted target variables, and in turn better performance on the diagnostic cohort. Also, insight into how the miRNA signal develops before and during the LC disease progression could have useful and important ramifications for detecting LC with miRNA biomarkers at an earlier stage. One might be able to predict LC accurately even in the prediagnostic cohorts, which was not achieved in the present study (see figures 4.7 and 4.8).

One potential way to estimate the decay function more accurately could be to iteratively test a variety of different approximations and choose the one that produces the best fit in the training data. This approach would be somewhat prone to overfitting, but using a separate cohort for testing would minimize this risk. Using the predictive performance as the benchmark for the approximation is less complicated than trying to estimate the decay by looking at how the miRNAs are differently expressed in patients with varying Time to diagnosis. This is because individual variations in miRNA expression would more than likely drown out the signal related to the decay. Alternatively, follow-up miRNA profiling of each patient could help in modelling the decay directly by seeing how each miRNA is expressed as the LC progresses.

Survival analysis methods are another potential way of modelling the prediagnostic cohorts' miRNA-signal decay. Here, the target variable to be predicted is not a binary value or probability of whether a patient has a given disease, it instead takes the form of the time until an event occurs. In this context, the target to be predicted could be either Time to diagnosis or the time until patient loss of life. Random Survival Forest [65] is one such method that could potentially directly or indirectly model Time to diagnosis while simultaneously handling the high-dimensional, non-parametric, and noisy character of the miRNA expression profiles. An implementation of this method on the data as prepared here, with some additional data per patient, would serve as a natural continuation of this project.

Incorporating Time to diagnosis in the target variable of the prediagnostic training data yielded an increase in predictive performance on the diagnostic cohort. Even though the linear transformation is a simplification of reality, the transformation was sufficient to observe a significant increase in the AUC, F1 score, accuracy, and recall of the RF model in predicting the LC status of patients in the diagnostic CNLCB cohort. This means that there most likely is a time dependent signal in the prediagnostic cohorts that increases as Time to diagnosis approaches 0.

Performance across subgroups

The final model performed well across patient subgroups and seemed to generalize beyond other LC risk factors such as age and smoking status. This is evident in figure 4.11, as the AUC calculated for each subgroup is comparable. This could mean that the model is more or less independent of these risk factors and makes predictions based solely on miRNA expression profiles: it does not simply pick up statistical signals of whether a patient smokes or not[4] for instance. For the different histological subtypes of LC, SCLC seemed to be the most separable from the control groups, but this could also be because of its small sample size ($n = 20$). Interestingly, all histological subtypes were separated reasonably well from the control groups, meaning that the model might pick up a general signal of LC that is independent of which cells the cancer originates from. This was an unexpected result, as different LC types have rather different prognoses (see section 2.1.1).

Sampling

Performing miRNA-based LC prediction in a cross-cohort manner is complicated but provides more general and robust models. The mean AUC of models trained on the 5-fold cross validated CNLCB cohort still provided for the most accurate predictors (see figure 4.8). However, the final regressor presented in figures 4.10 and 4.11 should probably be regarded as the most statistically significant as this model was trained on a separate population and could still predict LC in another population with reasonable accuracy and sensitivity. The low reproducibility of differential expression studies for circulating miRNA-based disease prediction is still a problem that remains to be solved [24, 25]. However, the results from the final regressor shows that if miRNA profiles from different cohorts are generated in a similar manner, and then normalized and filtered appropriately, they could be used for creating ML models that can predict LC across populations.

[4] As smoking has been shown to produce differential expression of some circulating miRNAs [66].

Backward elimination and KEGG pathway analysis

The identified miRNAs in backward elimination seemed to have cancer related gene targets. Figure 4.14 validates that the model picks up a cancer related signal in the miRNA matrices, ’Proteoglycans in cancer’ is somewhat separated from the rest of the pathways in its $-log(p)$ value. It could be interesting to investigate the 10 miRNAs that were identified as especially important in this study (see figure 4.13) and see how they predict LC in other cohorts.

Chapter 6

Conclusion

During this project, a cross-cohort random forest model was created that could separate LC cases from controls in a diagnostic cohort with reasonable specificity and sensitivity. The best model considered the time difference between diagnosis date and blood sample date in the prediagnostic training data. A panel of miRNAs were then extracted from this model and these miRNAs were identified as having gene targets that correlated well with cancer related gene pathways. Further work and more advanced survival analysis methods are required to model the prediagnostic miRNA signals accurately, as this was not achieved with the methods presented here. If these prediagnostic cohorts could be predicted well, the models would have to be capable of predicting cancer years into the future - a promising concept for early discovery of LC. Several attempts at dimensionality reduction of the miRNA matrices were tested, but yielded fruitless results. However, further development of autoencoders, for instance to deal with miRNA-seq specific noise, could be promising areas of further research.

Bibliography

[1] A. Rajkomar, J. Dean and I. Kohane, ‘Machine learning in medicine,’ *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019, PMID: 30943338. DOI: 10.1056/NEJMra1814259. eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMra1814259. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMra1814259.

[2] B. Berger, J. Peng and M. Singh, ‘Computational solutions for omics data,’ *Nature Reviews Genetics*, vol. 14, no. 5, pp. 333–346, Apr. 2013. DOI: 10.1038/nrg3433. [Online]. Available: https://doi.org/10.1038/nrg3433.

[3] D. Ben-Israel, W. B. Jacobs, S. Casha, S. Lang, W. H. A. Ryu, M. de Lotbiniere-Bassett and D. W. Cadotte, ‘The impact of machine learning on patient care: A systematic review,’ *Artificial Intelligence in Medicine*, vol. 103, p. 101 785, Mar. 2020. DOI: 10.1016/j.artmed.2019.101785. [Online]. Available: https://doi.org/10.1016/j.artmed.2019.101785.

[4] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre and A. Jemal, ‘Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,’ *CA Cancer J Clin*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[5] B. K. Sean, C. P. A., B. Haval, C. Jakub, H. Tracy and D. Caroline, ‘Progress and prospects of early detection in lung cancer,’ vol. 7, Sep. 2017. DOI: https://doi.org/10.1098/rsob.170070. [Online]. Available: https://royalsocietypublishing.org/doi/full/10.1098/rsob.170070.

[6] P. B. Bach, J. N. Mirkin, T. K. Oliver, C. G. Azzoli, D. A. Berry, O. W. Brawley, T. Byers, G. A. Colditz, M. K. Gould, J. R. Jett, A. L. Sabichi, R. Smith-Bindman, D. E. Wood, A. Qaseem and F. C. Detterbeck, ‘Benefits and harms of CT screening for lung cancer: a systematic review,’ *JAMA*, vol. 307, no. 22, pp. 2418–2429, Jun. 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3709596/.

[7] D. P. Bartel, ‘Metazoan micrornas,’ *Cell*, vol. 173, no. 1, pp. 20–51, 2018, ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2018.03.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867418302861.

[8] Y. Niu, M. Su, Y. Wu, L. Fu, K. Kang, Q. Li, L. Li, G. Hui, F. Li and D. Gou, 'Circulating plasma miRNAs as potential biomarkers of non–small cell lung cancer obtained by high-throughput real-time PCR profiling,' *Cancer Epidemiology Biomarkers & Prevention*, vol. 28, no. 2, pp. 327–336, Oct. 2018. DOI: 10.1158/1055-9965.epi-18-0723. [Online]. Available: https://doi.org/10.1158/1055-9965.epi-18-0723.

[9] M. B. Wozniak, G. Scelo, D. C. Muller, A. Mukeria, D. Zaridze and P. Brennan, 'Circulating MicroRNAs as non-invasive biomarkers for early detection of non-small-cell lung cancer,' *PLOS ONE*, vol. 10, no. 5, J. D. Hoheisel, Ed., e0125026, May 2015. DOI: 10.1371/journal.pone.0125026. [Online]. Available: https://doi.org/10.1371/journal.pone.0125026.

[10] P. T. Hennessey, T. Sanford, A. Choudhary, W. W. Mydlarz, D. Brown, A. T. Adai, M. F. Ochs, S. A. Ahrendt, E. Mambo and J. A. Califano, 'Serum microRNA biomarkers for detection of non-small cell lung cancer,' *PLoS ONE*, vol. 7, no. 2, N. Vij, Ed., e32307, Feb. 2012. DOI: 10.1371/journal.pone.0032307. [Online]. Available: https://doi.org/10.1371/journal.pone.0032307.

[11] P. S. Mitchell, R. K. Parkin, E. M. Kroh, B. R. Fritz, S. K. Wyman, E. L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K. C. O'Briant, A. Allen, D. W. Lin, N. Urban, C. W. Drescher, B. S. Knudsen, D. L. Stirewalt, R. Gentleman, R. L. Vessella, P. S. Nelson, D. B. Martin and M. Tewari, 'Circulating microRNAs as stable blood-based markers for cancer detection,' *Proceedings of the National Academy of Sciences*, vol. 105, no. 30, pp. 10 513–10 518, Jul. 2008. DOI: 10.1073/pnas.0804549105. [Online]. Available: https://doi.org/10.1073/pnas.0804549105.

[12] Z. Wang, M. Gerstein and M. Snyder, 'RNA-seq: A revolutionary tool for transcriptomics,' *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, Jan. 2009. DOI: 10.1038/nrg2484. [Online]. Available: https://doi.org/10.1038/nrg2484.

[13] *Id-lung*. [Online]. Available: https://en.uit.no/forskning/forskningsgrupper/gruppe?p_document_id=507532 (visited on 29/07/2020).

[14] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal and F. Bray, 'Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,' *CA: A Cancer Journal for Clinicians*, DOI: https://doi.org/10.3322/caac.21660. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21660. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660.

[15] R. S. Herbst, J. V. Heymach and S. M. Lippman, 'Lung cancer,' *New England Journal of Medicine*, vol. 359, no. 13, pp. 1367–1380, 2008, PMID: 18815398. DOI: 10.1056/NEJMra0802714. eprint: https://doi.org/10.

1056/NEJMra0802714. [Online]. Available: https://doi.org/10.1056/NEJMra0802714.

[16] R. C. Lee, R. L. Feinbaum and V. Ambros, 'The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14,' *Cell*, vol. 75, no. 5, pp. 843–854, Dec. 1993. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/8252621/.

[17] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz and G. Ruvkun, 'The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans,' *Nature*, vol. 403, no. 6772, pp. 901–906, Feb. 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/8252621/.

[18] U. of Manchester. (2018). 'Mirbase,' [Online]. Available: http://www.mirbase.org/ (visited on 22/04/2020).

[19] S. M. Hammond, 'An overview of microRNAs,' *Adv Drug Deliv Rev*, vol. 87, pp. 3–14, Jun. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4504744/.

[20] P. Smibert and E. C. Lai, 'Lessons from microRNA mutants in worms, flies and mice,' *Cell Cycle*, vol. 7, no. 16, pp. 2500–2508, Aug. 2008. DOI: https://doi.org/10.4161/cc.7.16.6454.

[21] T. Saito and P. Sætrom, 'Micrornas – targeting and target prediction,' *New Biotechnology*, vol. 27, no. 3, pp. 243–249, 2010, Special Issue: Biotechnology Annual Review 2010, ISSN: 1871-6784. DOI: https://doi.org/10.1016/j.nbt.2010.02.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1871678410003729.

[22] D. P. Bartel, 'MicroRNAs: target recognition and regulatory functions,' *Cell*, vol. 136, no. 2, pp. 215–233, Jan. 2009.

[23] Y. Peng and C. M. Croce, 'The role of micrornas in human cancer,' *Signal Transduction and Targeted Therapy*, vol. 1, no. 1, p. 15 004, Jan. 2016, ISSN: 2059-3635. DOI: 10.1038/sigtrans.2015.4. [Online]. Available: https://doi.org/10.1038/sigtrans.2015.4.

[24] C. Backes, E. Meese and A. Keller, 'Specific miRNA disease biomarkers in blood, serum and plasma: Challenges and prospects,' *Molecular Diagnosis & Therapy*, vol. 20, no. 6, pp. 509–518, Jul. 2016. DOI: 10.1007/s40291-016-0221-4. [Online]. Available: https://doi.org/10.1007/s40291-016-0221-4.

[25] J. Wang, J. Chen and S. Sen, 'MicroRNA as biomarkers and diagnostics,' *Journal of Cellular Physiology*, vol. 231, no. 1, pp. 25–30, Sep. 2015. DOI: 10.1002/jcp.25056. [Online]. Available: https://doi.org/10.1002/jcp.25056.

[26] M. Baker, 'MicroRNA profiling: Separating signal from noise,' *Nature Methods*, vol. 7, no. 9, pp. 687–692, Sep. 2010. DOI: 10.1038/nmeth0910-687. [Online]. Available: https://doi.org/10.1038/nmeth0910-687.

[27] S. E. V. Linsen, E. de Wit, G. Janssens, S. Heater, L. Chapman, R. K. Parkin, B. Fritz, S. K. Wyman, E. de Bruijn, E. E. Voest, S. Kuersten, M. Tewari and E. Cuppen, 'Limitations and possibilities of small RNA digital gene expression profiling,' *Nature Methods*, vol. 6, no. 7, pp. 474–476, Jul. 2009. DOI: 10.1038/nmeth0709-474. [Online]. Available: https://doi.org/10.1038/nmeth0709-474.

[28] P. Chugh and D. P. Dittmer, 'Potential pitfalls in microRNA profiling,' *Wiley Interdisciplinary Reviews: RNA*, vol. 3, no. 5, pp. 601–616, May 2012. DOI: 10.1002/wrna.1120. [Online]. Available: https://doi.org/10.1002/wrna.1120.

[29] L. Breiman, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/a:1010933404324. [Online]. Available: https://doi.org/10.1023/a:1010933404324.

[30] X. Chen and H. Ishwaran, 'Random forests for genomic data analysis,' *Genomics*, vol. 99, no. 6, pp. 323–329, 2012, ISSN: 0888-7543. DOI: https://doi.org/10.1016/j.ygeno.2012.04.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888754312000626.

[31] C. Strobl, A.-L. Boulesteix, A. Zeileis and T. Hothorn, 'Bias in random forest variable importance measures: Illustrations, sources and a solution,' *BMC Bioinformatics*, vol. 8, no. 1, Jan. 2007. DOI: 10.1186/1471-2105-8-25. [Online]. Available: https://doi.org/10.1186/1471-2105-8-25.

[32] D. Liu, Y. Huang, W. Nie, J. Zhang and L. Deng, 'SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost,' *BMC Bioinformatics*, vol. 22, no. 1, Apr. 2021. DOI: 10.1186/s12859-021-04135-2. [Online]. Available: https://doi.org/10.1186/s12859-021-04135-2.

[33] Y. Wang, H. Yao and S. Zhao, 'Auto-encoder based dimensionality reduction,' *Neurocomputing*, vol. 184, pp. 232–242, Apr. 2016. DOI: 10.1016/j.neucom.2015.08.104. [Online]. Available: https://doi.org/10.1016/j.neucom.2015.08.104.

[34] A. Dertat, *Applied deep learning - part 3: Autoencoders*, Oct. 2017. [Online]. Available: https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798 (visited on 10/07/2020).

[35] R. Mjelle, K. Sellæg, P. Sætrom, L. Thommesen, W. Sjursen and E. Hofsli, 'Identification of metastasis-associated microRNAs in serum from rectal cancer patients,' *Oncotarget*, vol. 8, no. 52, pp. 90 077–90 089, Sep. 2017. DOI: 10.18632/oncotarget.21412. [Online]. Available: https://doi.org/10.18632/oncotarget.21412.

[36] R. Mjelle, S. O. Dima, N. Bacalbasa, K. Chawla, A. Sorop, D. Cucu, V. Herlea, P. Sætrom and I. Popescu, 'Comprehensive transcriptomic analyses of tissue, serum, and serum exosomes from hepatocellular carcinoma patients,' *BMC Cancer*, vol. 19, no. 1, Oct. 2019. DOI: `10.1186/s12885-019-6249-1`. [Online]. Available: `https://doi.org/10.1186/s12885-019-6249-1`.

[37] QIAGEN, *Mirneasy serum/plasma kit*. [Online]. Available: `https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/rna-purification/mirna/mirneasy-serumplasma-kit/?clear=true#orderinginformation` (visited on 19/05/2020).

[38] SelectScience, *Nanodrop™ 1000 spectrophotometer from thermo fisher scientific*. [Online]. Available: `https://www.selectscience.net/products/nanodrop-1000-spectrophotometer/?prodID=79482#tab-4` (visited on 19/05/2020).

[39] Agilent, *2100 bioanalyzer instrument*. [Online]. Available: `https://www.agilent.com/en/product/automated-electrophoresis/bioanalyzer-systems/bioanalyzer-instrument/2100-bioanalyzer-instrument-228250#productdetails` (visited on 19/05/2020).

[40] Illumina, *Nextflex small rna-seq kit v3*, Dec. 2020. [Online]. Available: `https://perkinelmer-appliedgenomics.com/home/products/library-preparation-kits/small-rna-library-prep/nextflex-small-rna-seq-kit-v3/` (visited on 19/05/2020).

[41] S. Andrews, *Fastqc: A quality control tool for high throughput sequence data*, 2010. [Online]. Available: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/` (visited on 23/05/2020).

[42] B. Langmead and S. L. Salzberg, 'Fast gapped-read alignment with bowtie 2,' *Nature Methods*, vol. 9, no. 4, pp. 357–359, Mar. 2012. DOI: `10.1038/nmeth.1923`. [Online]. Available: `https://doi.org/10.1038/nmeth.1923`.

[43] S. Anders, P. T. Pyl and W. Huber, 'HTSeq–a python framework to work with high-throughput sequencing data,' *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Sep. 2014. DOI: `10.1093/bioinformatics/btu638`. [Online]. Available: `https://doi.org/10.1093/bioinformatics/btu638`.

[44] L. Pantano, X. Estivill and E. Marti, 'SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells,' *Nucleic Acids Research*, vol. 38, no. 5, e34–e34, Dec. 2009. DOI: `10.1093/nar/gkp1127`. [Online]. Available: `https://doi.org/10.1093/nar/gkp1127`.

[45] M. D. Robinson, D. J. McCarthy and G. K. Smyth, 'Edger: A bioconductor package for differential expression analysis of digital gene expression data,' *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010. DOI: `10.1093/bioinformatics/btp616`.

[46] Y. Chen, A. T. Lun and G. K. Smyth, 'From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline,' *F1000Res*, vol. 5, p. 1438, 2016.

[47] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly and R. A. Irizarry, 'Nat Rev GenetTackling the widespread and critical impact of batch effects in high-throughput data,' *Nat Rev Genet*, vol. 11, no. 10, pp. 733–739, Oct. 2010.

[48] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth, 'limma powers differential expression analyses for RNA-sequencing and microarray studies,' *Nucleic Acids Research*, vol. 43, no. 7, e47, 2015. DOI: `10.1093/nar/gkv007`.

[49] G. Smyth, Nov. 2020. [Online]. Available: `https://rdrr.io/bioc/limma/man/removeBatchEffect.html` (visited on 19/06/2020).

[50] S. Wold, K. Esbensen and P. Geladi, 'Principal component analysis,' *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987, Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, ISSN: 0169-7439. DOI: `https://doi.org/10.1016/0169-7439(87)80084-9`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/0169743987800849`.

[51] D. P. Kingma and J. Ba, 'Adam: A method for stochastic optimization,' 2014. eprint: `arXiv:1412.6980`.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, 'Scikit-learn: Machine learning in Python,' *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[53] S. Sharma and M. Eghbali, 'Influence of sex differences on microRNA gene regulation in disease,' *Biology of Sex Differences*, vol. 5, no. 1, p. 3, 2014. DOI: `10.1186/2042-6410-5-3`. [Online]. Available: `https://doi.org/10.1186/2042-6410-5-3`.

[54] H. Yu, Z. Guan, K. Cuk, Y. Zhang and H. Brenner, 'Cancers (Basel)Circulating MicroRNA Biomarkers for Lung Cancer Detection in East Asian Populations,' *Cancers (Basel)*, vol. 11, no. 3, Mar. 2019.

[55] H. Yu, Z. Guan, K. Cuk, H. Brenner and Y. Zhang, 'Circulating microRNA biomarkers for lung cancer detection in western populations,' *Cancer Medicine*, vol. 7, no. 10, pp. 4849–4862, Sep. 2018. DOI: `10.1002/cam4.1782`. [Online]. Available: `https://doi.org/10.1002/cam4.1782`.

[56] J. N. Mandrekar, 'Receiver operating characteristic curve in diagnostic test assessment,' *Journal of Thoracic Oncology*, vol. 5, no. 9, pp. 1315–1316, 2010, ISSN: 1556-0864. DOI: `https://doi.org/10.1097/JTO.0b013e3181ec173d`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1556086415306043`.

[57] I. S. Vlachos, K. Zagganas, M. D. Paraskevopoulou, G. Georgakilas, D. Karagkouni, T. Vergoulis, T. Dalamagas and A. G. Hatzigeorgiou, 'DIANA-miRPath v3.0: deciphering microRNA function with experimental support,' *Nucleic Acids Res*, vol. 43, no. W1, W460–466, Jul. 2015.

[58] M. Kanehisa, 'KEGG: Kyoto encyclopedia of genes and genomes,' *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000. DOI: `10.1093/nar/28.1.27`. [Online]. Available: `https://doi.org/10.1093/nar/28.1.27`.

[59] M. Kanehisa, 'Toward understanding the origin and evolution of cellular organisms,' *Protein Science*, vol. 28, no. 11, pp. 1947–1951, Sep. 2019. DOI: `10.1002/pro.3715`. [Online]. Available: `https://doi.org/10.1002/pro.3715`.

[60] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe and M. Tanabe, 'KEGG: Integrating viruses and cellular organisms,' *Nucleic Acids Research*, vol. 49, no. D1, pp. D545–D551, Oct. 2020. DOI: `10.1093/nar/gkaa970`. [Online]. Available: `https://doi.org/10.1093/nar/gkaa970`.

[61] G. Forman and M. Scholz, 'Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement,' *SIGKDD Explor. Newsl.*, vol. 12, no. 1, pp. 49–57, Nov. 2010, ISSN: 1931-0145. DOI: `10.1145/1882471.1882479`. [Online]. Available: `https://doi.org/10.1145/1882471.1882479`.

[62] S. Tam, M. S. Tsao and J. D. McPherson, 'Optimization of miRNA-seq data preprocessing,' *Brief Bioinform*, vol. 16, no. 6, pp. 950–963, Nov. 2015.

[63] L. X. Garmire and S. Subramaniam, 'The poor performance of TMM on microRNA-Seq,' *RNA*, vol. 19, no. 6, pp. 735–736, Jun. 2013.

[64] L. Gondara, 'Medical image denoising using convolutional denoising autoencoders,' in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016, pp. 241–246. DOI: `10.1109/ICDMW.2016.0041`.

[65] H. Wang and G. Li, 'A Selective Review on Random Survival Forests for High Dimensional Data,' *Quant Biosci*, vol. 36, no. 2, pp. 85–96, 2017.

[66] C. M. Willinger, J. Rong, K. Tanriverdi, P. L. Courchesne, T. Huan, G. A. Wasserman, H. Lin, J. Dupuis, R. Joehanes, M. R. Jones, G. Chen, E. J. Benjamin, G. T. O'Connor, J. P. Mizgerd, J. E. Freedman, M. G. Larson and D. Levy, 'MicroRNA signature of cigarette smoking and evidence for a putative causal role of MicroRNAs in smoking-related inflammation and target organ damage,' *Circulation: Cardiovascular Genetics*, vol. 10, no. 5,

Oct. 2017. DOI: `10.1161/circgenetics.116.001678`. [Online]. Available: `https://doi.org/10.1161/circgenetics.116.001678`.

Appendix A

Additional Material

This appendix includes figures that were considered too detailed for the main report. First, the full results of hold-one-cohort-out validation on the prediagnostic cohorts are presented. Then, models trained on all the prediagnostic cohorts are evaluated, and finally models trained on CNLCB only are visualized.

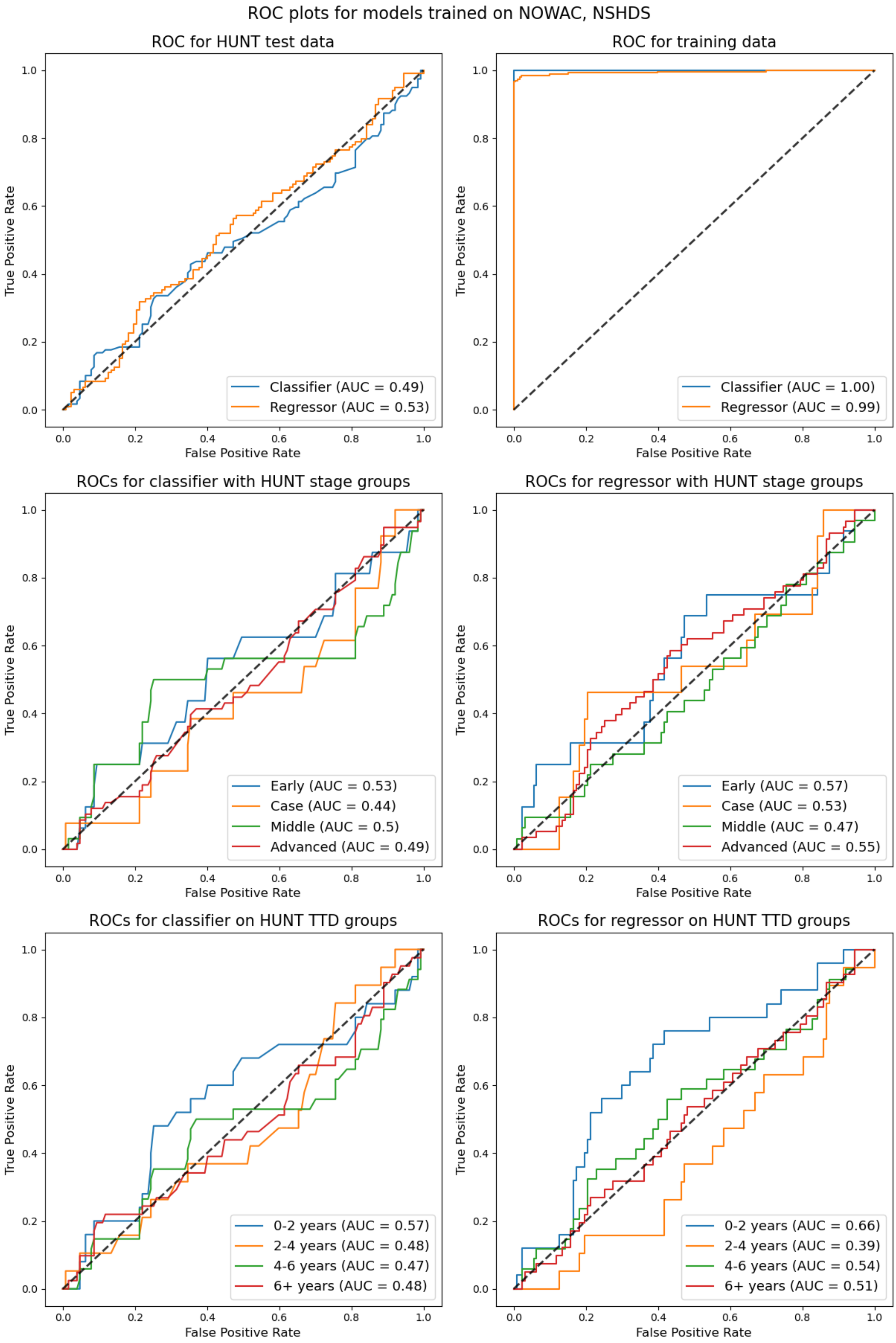

Figure A.1: ROC plots for models trained on NOWAC and NSHDS and tested on HUNT.

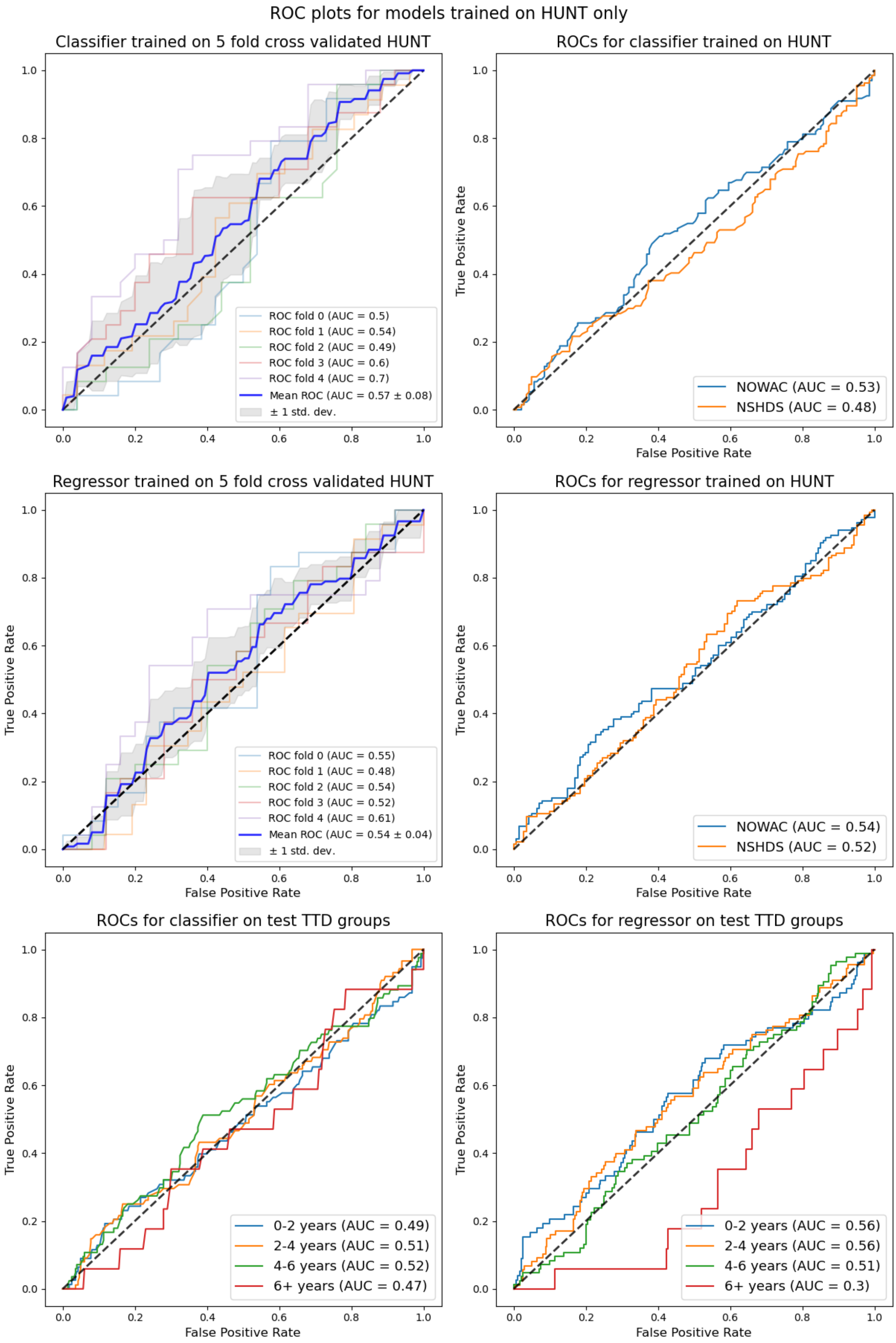

Figure A.2: ROC plots for models trained on HUNT and tested on NOWAC and NSHDS.

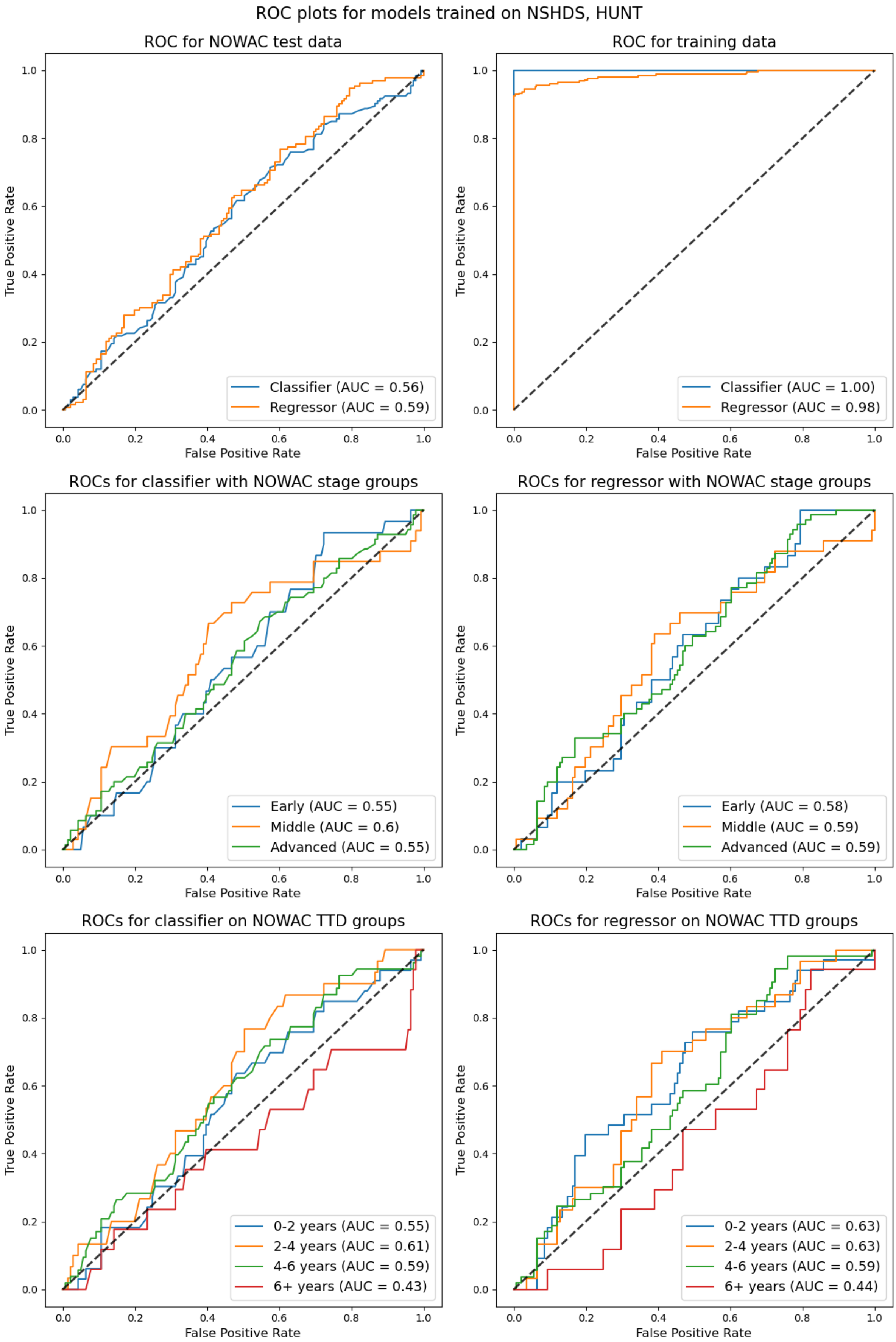

Figure A.3: ROC plots for models trained on HUNT and NSHDS and tested on NOWAC.

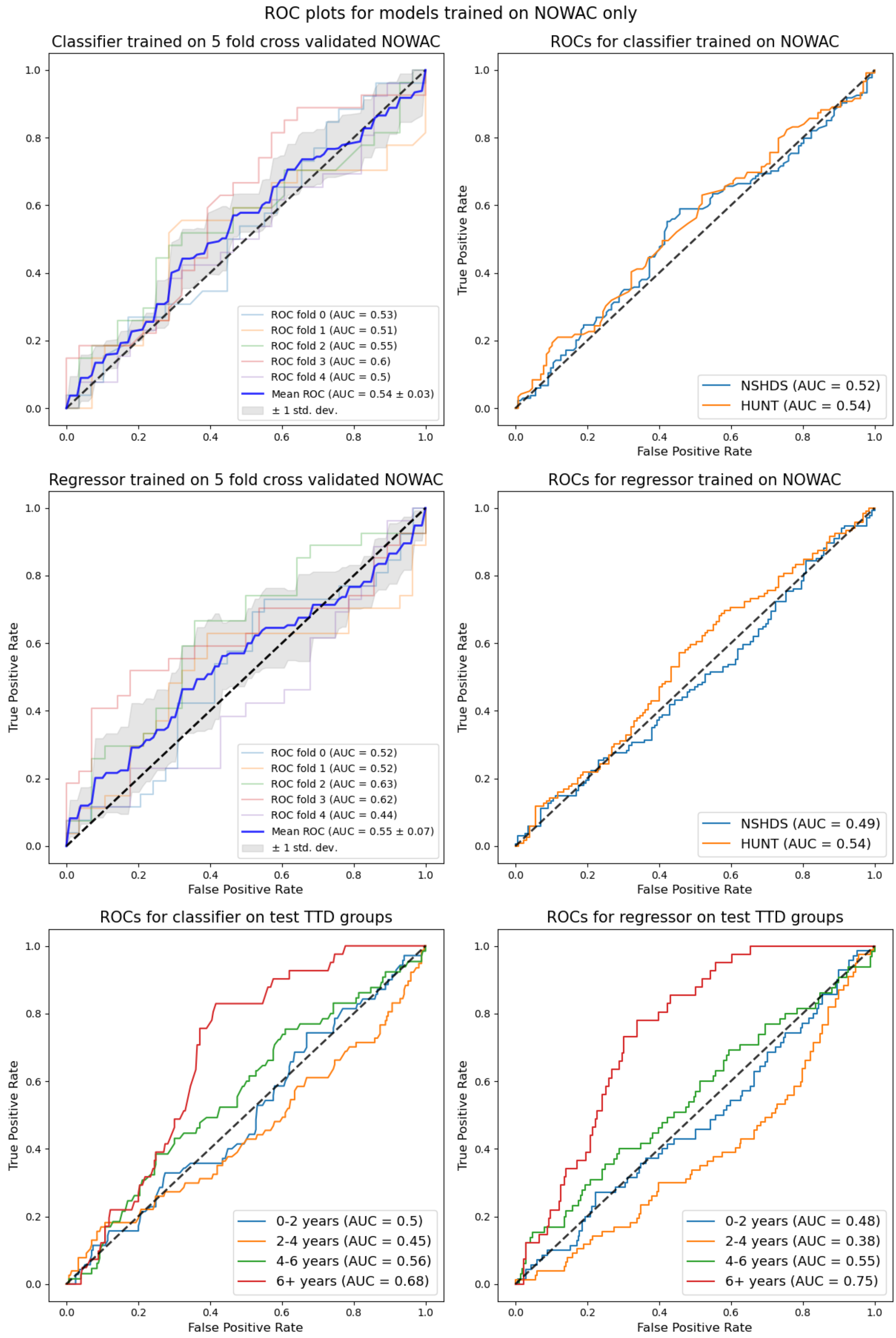

Figure A.4: ROC plots for models trained on NOWAC and tested on HUNT and NSHDS.

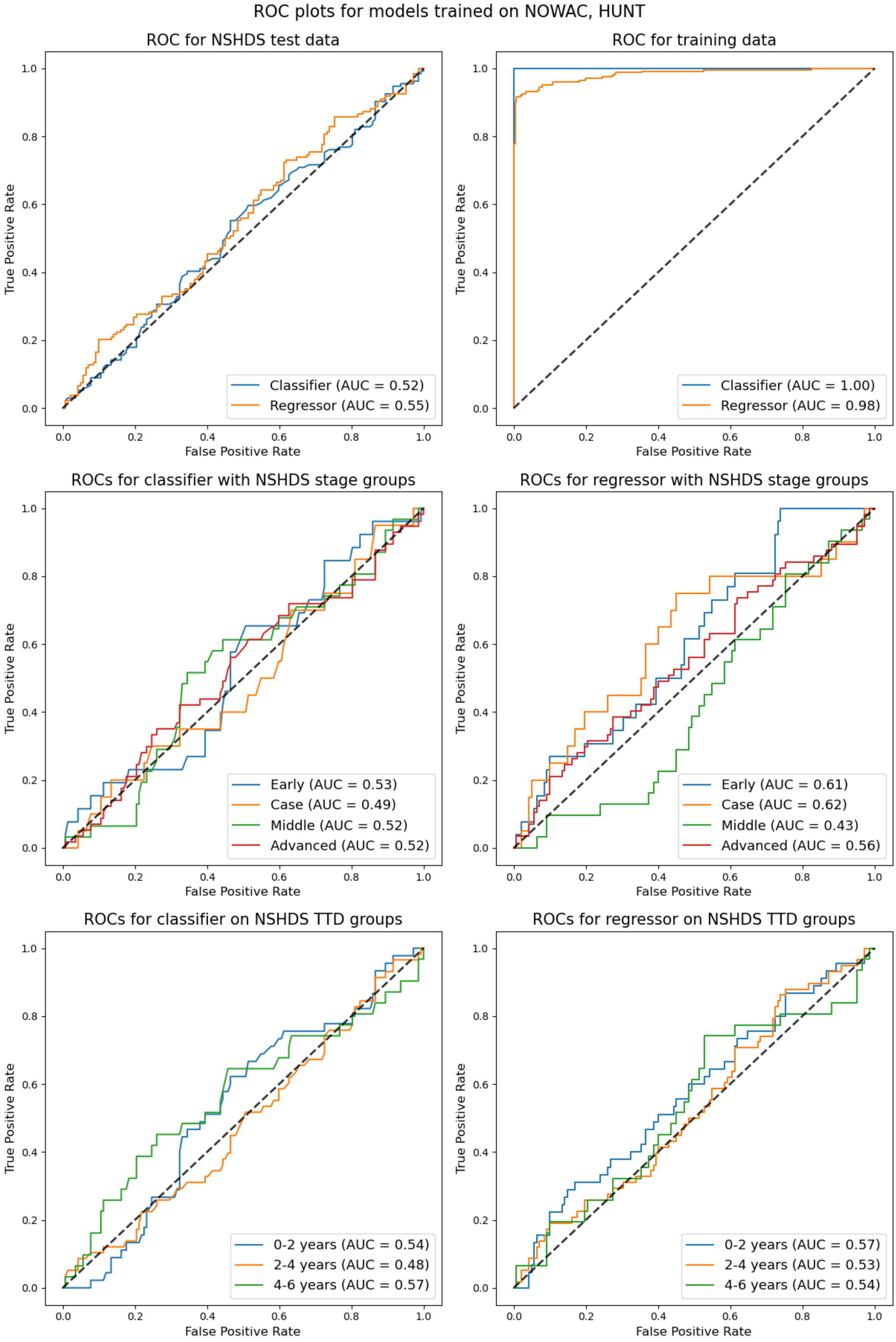

Figure A.5: ROC plots for models trained on HUNT and NOWAC and tested on NSHDS.

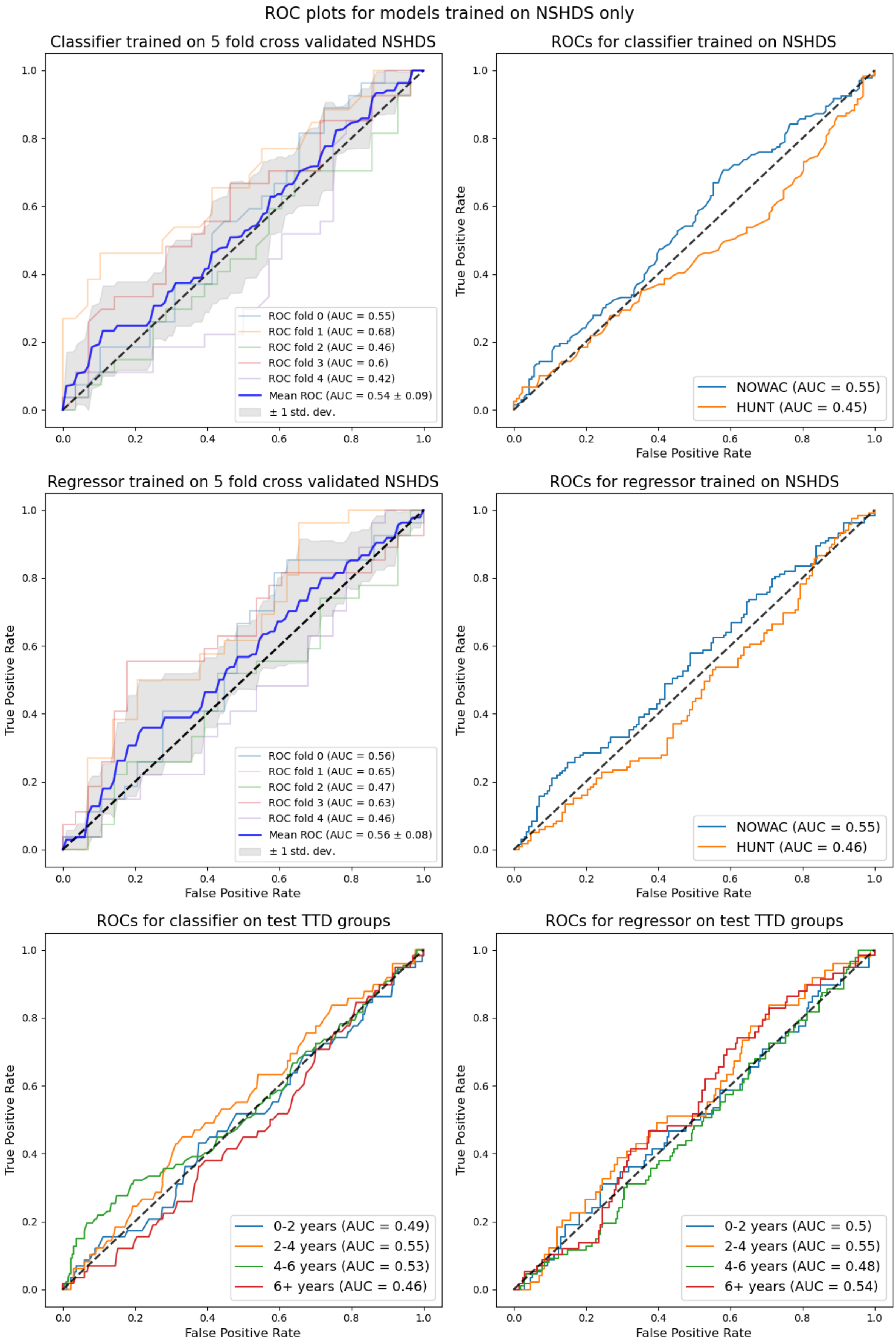

Figure A.6: ROC plots for models trained on NSHDS and tested on HUNT and NOWAC.

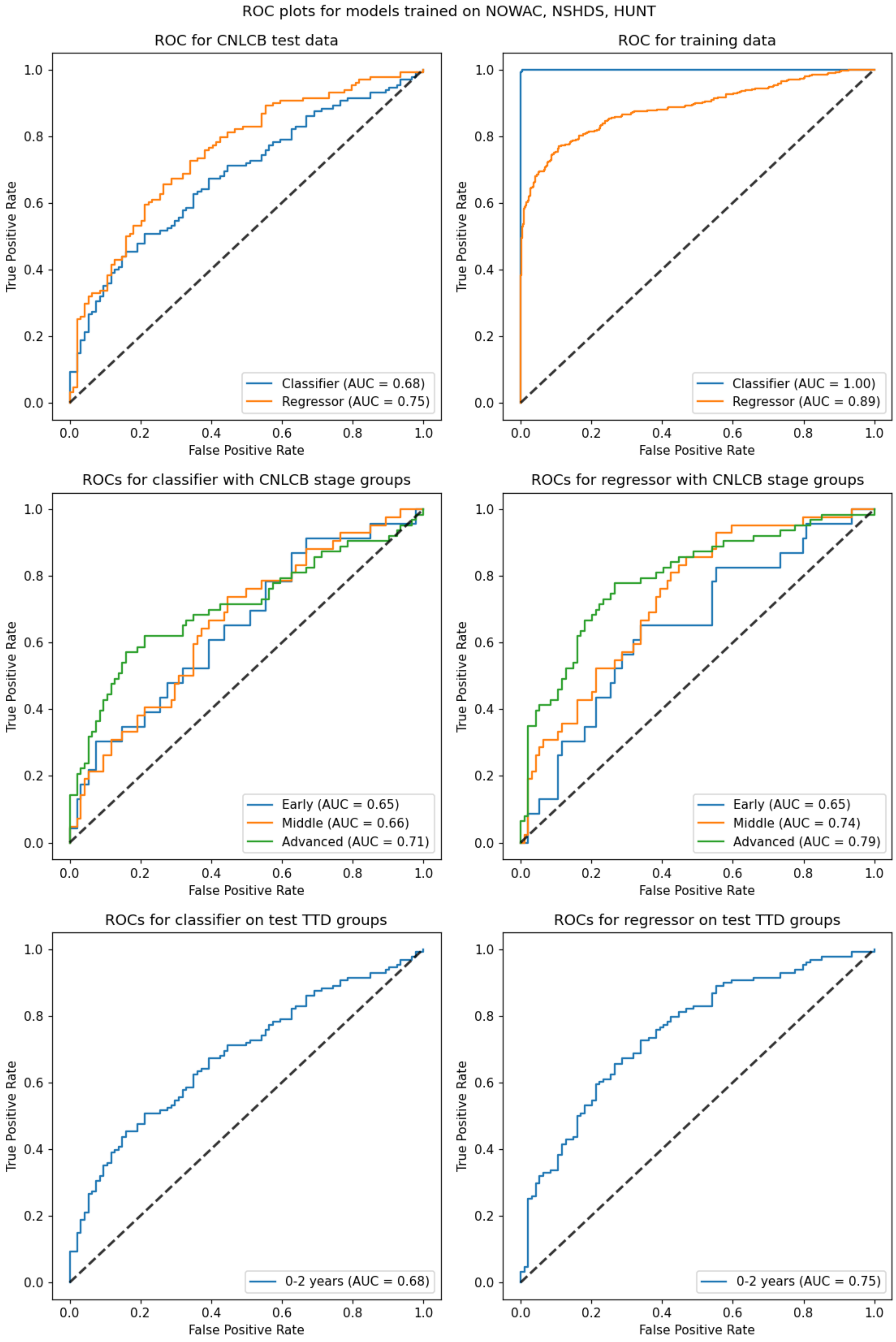

Figure A.7: ROC plots for models trained on HUNT, NOWAC and NSHDS and tested on CNLCB.

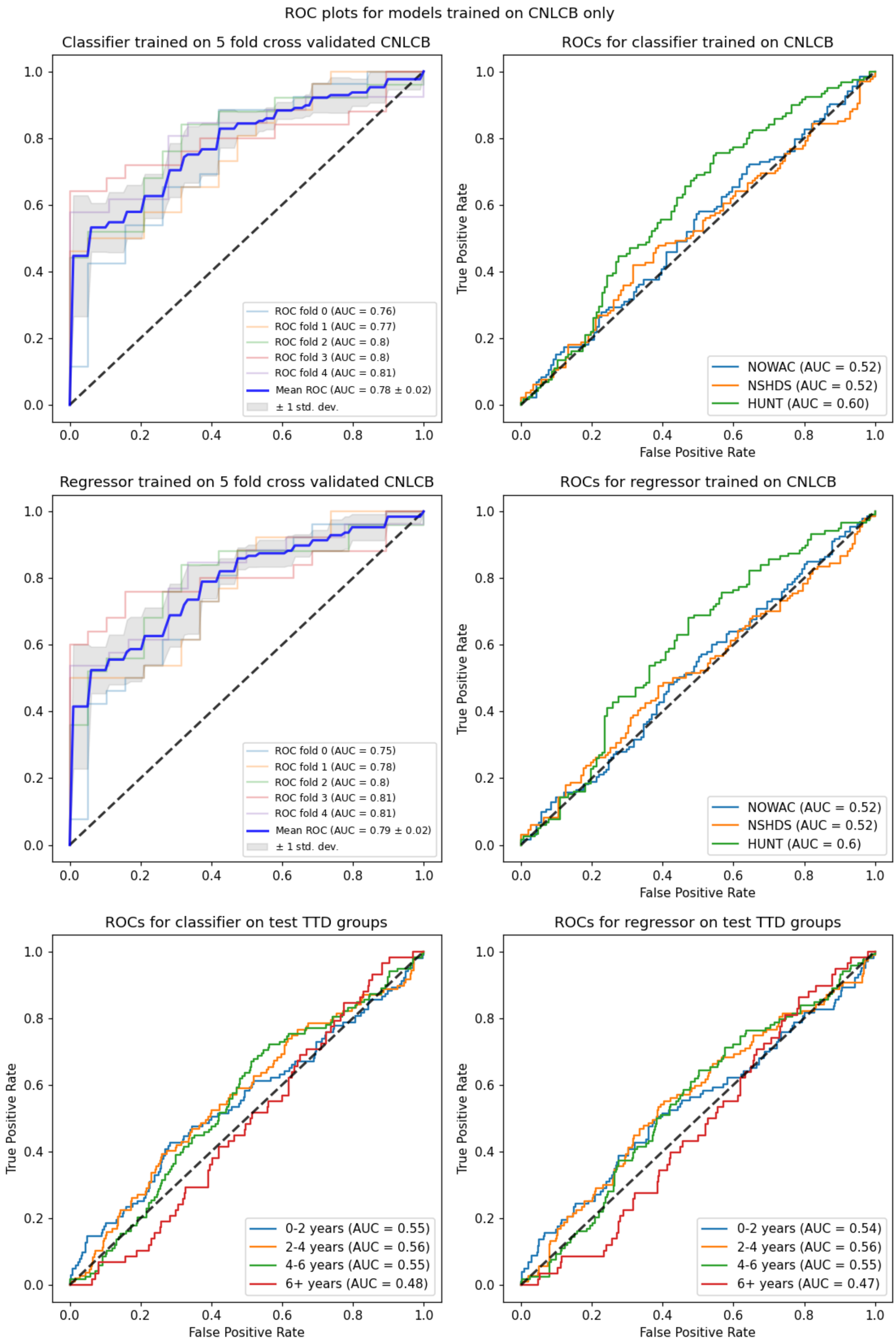

Figure A.8: ROC plots for models trained on CNLCB and tested on HUNT, NOWAC and NSHDS.

NTNU

Norwegian University of
Science and Technology