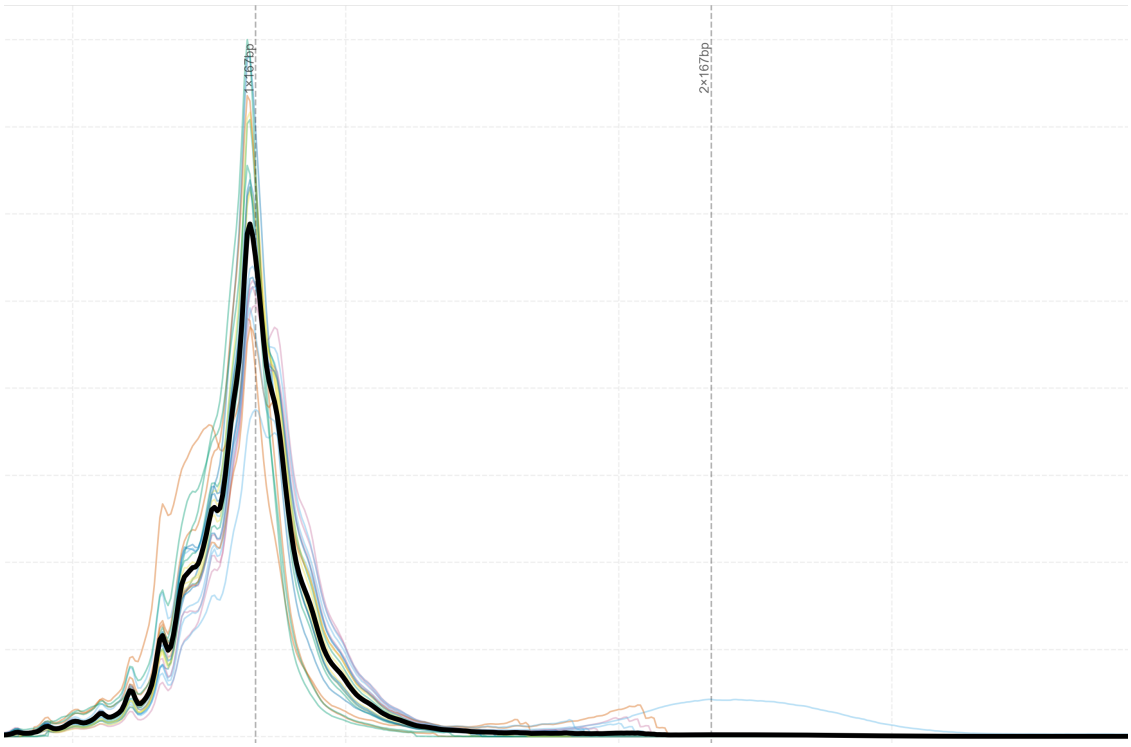

Learning the Fragment Size Distribution in Liquid Biopsy Sequencing

Master's thesis in Engineering Mathematics and Computational Science

SAMUEL KRONQUIST, ANNA SVENSSON FEHÉR

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Learning the Fragment Size Distribution in Liquid Biopsy Sequencing

SAMUEL KRONQUIST
ANNA SVENSSON FEHÉR

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Applied Mathematics and Statistics of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2025

Learning the Fragment Size Distribution in Liquid Biopsy Sequencing
SAMUEL KRONQUIST & ANNA SVENSSON FEHÉR

Supervisor: Eszter Lakatos, Assistant Professor at Applied Mathematics and Statistics
Examiner: Erik Kristiansson, Full Professor, Applied Mathematics and Statistics

Master's Thesis 2025
Department of Mathematical Sciences
Applied Mathematics and Statistics of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: The sample specific means of fragment length distributions, stacked.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Learning the Fragment Size Distribution in Liquid Biopsy Sequencing

SAMUEL KRONQUIST

ANNA SVENSSON FEHÉR

Department of Mathematical Sciences

Chalmers University of Technology

Abstract

Cancer as a disease affects thousands of patients every year. Earlier, cancer was analyzed through tissue biopsies derived during surgery. However, due to improved sequencing methods, liquid biopsies have become more common, as these are convenient and provide the opportunity to monitor cancer evolution in real time. The project aim was to employ computational methods to analyze fragment length distributions from liquid biopsies by finding characteristics related to cancer and then filter appropriately. We carried out the project by utilizing a combination of machine learning and statistical learning. The machine learning models were performed for two labels, where one was based on purity and one was generated through a minimalistic cell death model. We found characteristic information linked to cancer by evaluating the models based on feature importance. The project resulted in one label sufficient enough for usage, which led to several models outperforming relevant baselines. As the models were somewhat flawed due to comprised results and insufficient data, no filtering could be made with the guarantee of only removing healthy data. However, we still managed to find characteristic features because of synergistic results across the models.

Keywords: cancer, liquid biopsies, ovarian cancer, statistical learning, machine learning, statistics, Python, chromosome, necrosis

Acknowledgements

We sincerely and gratefully acknowledge the support of our supervisor, Eszter Lakatos. Many thanks for the continuous support and feedback during the project. Without your engagement and guidance, the quality of our project would not have been the same.

A thank you also to the colleagues of the research groups of Cvijovic and Polster labs for bi-weekly meetings and lunches with fruitful discussions and valuable inputs for our project.

Samuel Kronquist and Anna Svensson Fehér, Gothenburg, May 2025

List of Acronyms

Below is a list of acronyms that has been used throughout this thesis. The acronyms are listed in alphabetical order:

cfDNA	cell-free DNA
CNA	Copy Number Alterations
CNV	Copy Number Value
DNA	Deoxyribonucleic acid
NGS	Next Generation Sequencing

Contents

List of Figures

List of Tables

1
Introduction

The following section will introduce a thorough background to the project aim, the aim itself, and a description of the project.

1.1 Background

Cancer is a disease characterized by uncontrolled cell growth that can affect various parts of the body, including the ovaries. Ovarian cancer represents a significant challenge in oncology, with approximately 314,000 new cases and 207,000 deaths annually worldwide [45]. Despite accounting for only 3% of female cancer diagnoses, it ranks eighth in global cancer mortality.

Ovarian cancer is the umbrella term for any abnormal growth or tissue change located in the ovaries [14]. The reason why one can consider the diagnosis an umbrella term is due to the fact that there exist three types of tissue that cause the cancer. The first type originates from the cells covering the ovaries, the second from the cells producing the hormones, and the last from the ones ending up as the produced egg. As the symptoms for ovarian cancer overlap with many other woman-related health issues, diagnoses are often delayed, and approximately 70% of cases are identified at advanced stages [35, 45].

One technique that can revolutionize the healthcare and pharmaceutical industry is analyzing cell-free DNA (cfDNA) with liquid biopsies. Cell-free DNA consists of DNA fragments released into the bloodstream from dying cells, with a proportion originating from tumor cells (circulating tumor DNA or ctDNA) in cancer patients. As liquid biopsies are less invasive than tissue biopsies, they can be performed more frequently and thereby provide doctors with real-time data measured for the full cycle of cancer development in a patient. Furthermore, research has shown that fragment length distributions of cfDNA provide insight into the underlying mechanisms of cell death; apoptosis, necrosis, and autophagy, which produce characteristic fragmentation patterns [32].

These mechanisms serve as potential biomarkers for cancer detection, treatment response monitoring, and disease progression assessment [2]. For ovarian cancer specifically, where early detection methods remain insufficient, cfDNA fragment analysis presents an opportunity to develop more effective diagnostic and prognostic tools.

1.2 Project Description

This master's project focuses on developing and validating computational methods to analyze cfDNA fragment length distributions to learn about different characteristics and death mechanisms in patients with cancer.

The computational framework developed consists of two analytical and two machine learning approaches:

1. **Minimalistic cell death model**: A rule-based classifier utilizing fixed parameters and fragment length distributions to identify dominant cell death mechanisms. This method provides rapid analysis with minimal computational requirements, suitable for clinical settings requiring quick results.
2. **Bayesian Hierarchical Model**: A probabilistic framework providing uncertainty-aware analysis of cfDNA fragment distributions. This model captures relationships between fragment lengths, cell death mechanisms, and clinical variables such as tumor purity.
3. **Random Forest**: A machine learning method used to classify ranges of cancer purity, to allow for evaluation of feature importance. The model generates several decision trees and then lets the trees predict the class based on their majority vote.
4. **Gradient Boosting Machine**: Gradient Boost is another machine learning method used for classification and analysis of feature importance. The model still uses decision trees, but unlike random forest, it generates a tree every iteration who has a scaled contribution. The algorithm minimizes a loss function to get an initial prediction. It then generates residuals, which the next tree is trained on, meaning that each tree compensates for errors from the earlier iteration.

The project also includes visualizations to illustrate fragment length distributions, cell death mechanism proportions, and other results.

1.3 Project aim

The aim of the project is to enable filtering of fragment length distribution data based on cancer origin by a filtering software. This should be carried out by utilizing the computational frameworks disclosed in the project description.

2
Theory

In the following section, a thorough introduction is written for the ability to grasp the full context of the aim and key aspects of the project. As the project lies in the intersection of the areas of biology and mathematics, the theory section is divided into a biological background as well as a mathematical one.

2.1 Biological background

The human body is a complex biological system that is dependent on the symbiosis of several internal functions and anatomical components. It is the uniqueness of our bodies' biological material that contributes to the huge diversity within our species, and it all starts at a cellular level with the DNA.

2.1.1 DNA

Between individuals, only approximately 0.1 % of our DNA differs [21]. The biological variation in combination with life-altering factors such as our environment and relationships governs our differences in health and ability. This means, in reality, that a very small fraction of the genetic code dictates our uniqueness - but how does DNA actually carry genetic information?

The DNA possesses the complete instructions for each individual to develop and function [21]. The DNA molecule is composed of three key components: sugar, phosphate groups, and a base, which varies between four kinds. The two initial components create the structure of the backbone, while the base bonds with the base of the parallel strand, linking the two strands together into one cohesive unit. The role of the base in the final helix structure is essential, as the order in which the bases are lined up generates the individual genetic code.

2.1.2 Chromosomes and the genome

The DNA is found in the nucleus of the cell, tightly wrapped around a protein named histone [20]. The structure to which the previously mentioned components contribute is the foundational structure of the chromosomes and essential for the full genetic code to fit. In the human body, there are 23 pairs of chromosomes in total, of which one pair is specific to the sex of the individual. Each chromosome varies in size and contains a different number of genes [29]. The umbrella term that collects the

instructions described by the entirety of a cell's DNA is called the genome [22]. To enable accurate description of genes and their regions, each chromosome is analyzed by dividing the structure into three parts. The centromere is the middle part of the chromosome, splitting the structure into its two arms [29]. The region q refers to the longer arm, while p is used to specify the short arm. In the data set of the project, the regions will be noted as, e.g. 10q for the long arm of chromosome 10.

2.1.3 Mutations and CNAs

Several types of events can affect the biological information that the cell contains. When the event leads to a change in the sequence of bases in the DNA, a mutation has occurred [23]. How mutations affect our health depends on varying factors, such as the cell type of the affected cell, as well as the reason for the sequence alteration in the first place. Changes in the reproductive cells will affect the offspring majorly, as the mutation will be carried to all cells developed from the first [23]. For somatic cells, the majority of mutations will go unnoticed as the cell manages to handle the load of mutations by restoring them through repair mechanisms.

Mutations as events are caused due to changes in a single, a few, or larger regions of many nucleotides [1]. When many nucleotides are changed, the alteration will be on a chromosomal scale. The structural changes can lead to loss or amplification of segments, disrupted functions, segment insertion, or fusion of chromosomes. When the first of the previously mentioned instances occurs, the result becomes a deviating number of copies from the standard [30]. This phenomenon of copy changes is called copy number alterations (CNA). CNAs have a great value of interest, as they have been found to be influential for gene expression and may enable deeper understanding about cancer dynamics across different types of cancer. The full connection between the concepts disclosed in the earlier sections is visualized in Figure 2.1.

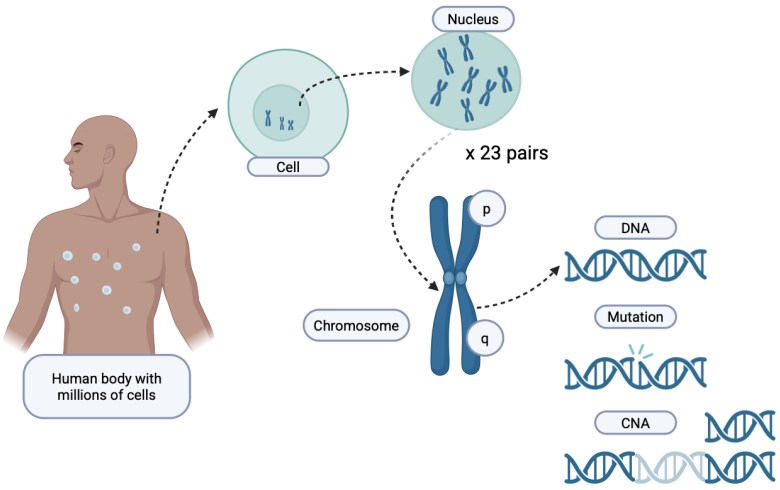

Figure 2.1: A visual representation of the life cycle of our genetic code. The image is created in the illustration software biorender.

2.1.4 Cell Death Mechanisms and Fragment Analysis

Cell death is a biological process that occurs through three distinct mechanisms: apoptosis, necrosis, and autophagy. Understanding these mechanisms gives us a better understanding of the fragment length analysis, as each type of cell death produces characteristic DNA fragmentation patterns [12].

Let us start with apoptosis, which is a highly regulated form of programmed cell death where cells naturally self-destruct. This process occurs when cells reach their maximum division potential (approximately 60 replications [17]) or when they become damaged beyond repair [12]. During apoptosis, cells undergo controlled cellular breakdown with systematic DNA fragmentation into specific lengths. A key characteristic of apoptotic cell death is the maintenance of membrane integrity until late stages, and notably, this process does not trigger an inflammatory response. The DNA is cleaved into characteristic fragments of specific lengths, typically multiples of 180-200 base pairs, creating a distinctive "ladder pattern" when analyzed [47].

Necrosis represents accidental or unprogrammed cell death, typically resulting from severe cellular trauma. Unlike apoptosis, necrosis is characterized by uncontrolled cell swelling and random DNA fragmentation. During necrotic cell death, the cell membrane ruptures, leading to cellular content leakage into the surrounding tissue. As described by Cleveland Clinic experts, this leakage triggers an inflammatory response, causing additional damage to neighboring cells [12]. When examining fragment analysis results, necrotic DNA typically shows a smear pattern rather than distinct bands [25], due to the random nature of DNA degradation. This process is irreversible, and the resulting tissue damage cannot be repaired.

Autophagy is a stress-response mechanism that occurs during periods of cellular

stress or nutrient deprivation [12]. This sophisticated process involves the self-digestion of cellular components and the recycling of cellular materials. Through autophagy, cells can maintain homeostasis by selectively degrading damaged or unnecessary organelles. The process serves as a survival mechanism during periods of stress, allowing cells to recycle components and maintain essential functions. During autophagy, cellular components are broken down and recycled, resulting in DNA fragmentation patterns that differ from both apoptosis and necrosis.

2.1.5 Cell death characteristics

In the following list of bullet points, the characteristic traits of each cell death type are described:

- **Apoptosis:** A regulated form of programmed cell death characterized by activation of caspases and endonucleases that cleave DNA at internucleosomal regions [8]. This produces a distinctive pattern with:
  - Fragment lengths centered around 167 bp (mono-nucleosomal)
  - Additional peaks at regular intervals of approximately 167 bp (di-nucleosomal at ∼335 bp, tri-nucleosomal at ∼500 bp)
  - Narrow peak widths due to precise enzymatic cleavage
  - Strong periodicity reflecting the nucleosome structure

- **Necrosis:** An unregulated form of cell death often resulting from acute injury [34], characterized by:
  - Longer fragment lengths (typically >250 bp, often centered around 300-500 bp) [48, 27]
  - Broader, less defined peaks
  - Minimal periodicity
  - Higher variability in fragment lengths

- **Autophagy:** A regulated process of cellular component recycling that shows intermediate characteristics [34]:
  - Fragment lengths intermediate between apoptosis and necrosis (typically 200-300 bp)
  - Moderate periodicity
  - Moderate peak width

2.1.6 Cell-free DNA and Circulating Tumor DNA

A result of cell death is that fragmented cell membranes can be detected in blood. This non-encapsulated DNA is known as cell-free DNA (cfDNA). A specific subset of cfDNA, known as circulating tumor DNA (ctDNA), represents a small portion of the total cfDNA and originates specifically from tumor cells [36].

The analysis of ctDNA through liquid biopsies - the sampling and analysis of bodily fluids such as blood, urine, or cerebrospinal fluid - has emerged as a powerful method for minimally invasive cancer detection and monitoring. This approach eliminates the need for invasive tissue biopsies while still providing crucial diagnostic information.

One significant characteristic of cfDNA analysis is that the extracted yields are typically very low, often less than one nanogram per milliliter of plasma. This presents a technical challenge, as traditional sequencing methods often require hundreds of nanograms of DNA [36]. However, advanced sequencing technologies have shown the capability to detect both low-abundance cfDNA and ctDNA with high accuracy and sensitivity, enabling the monitoring of cancer progression and treatment response through these blood-based markers. The analysis of cfDNA and ctDNA can reveal various cancer-specific features, including:

- Copy number aberrations
- Cell-of-origin information
- Cancer-associated fragmentation signatures
- Cancer-specific methylation features

Notably, research has shown that samples with high tumor fractions typically exhibit shorter DNA fragment sizes, suggesting that fragment length analysis could potentially be used as a method for classifying different cancer types [36].

2.1.7 DNA Sequencing Methods

The method used to extract this information from the cfDNA is DNA sequencing. There are currently two main approaches in use: traditional Sanger sequencing and Next-Generation Sequencing (NGS). Each method has its distinct advantages and applications in modern genetic analysis [19].

Sanger sequencing, considered the gold standard in sequencing technology, is optimal for analyzing small numbers of gene targets and samples. This method can complete analysis in a single day and is often used to verify NGS results due to its high reliability [19].

Next-Generation Sequencing, initially called "massively parallel sequencing" when introduced commercially in 2005, revolutionized genetic analysis through its ability to sequence many DNA strands simultaneously. NGS methods offer several advantages over traditional methods:

- Ability to analyze hundreds to thousands of genes simultaneously
- Lower sample input requirements
- Capability to detect variants at lower allele frequencies
- Cost-effective analysis of multiple samples

- Detection of various genomic features in a single run, including:
  - Single nucleotide variants, copy number variations, structural variants, and RNA fusions

The customizability of the NGS methods allows us to use different types that increase speed, throughput, and/or accuracy. These new sequencing methods have expanded their applications in multiple fields, including genomic research, clinical research, reproductive health, environmental science, agricultural studies, and forensic science [19]. Not only has next-generation sequencing been used across multiple fields, but it is also the method used in this project to extract the fragment length distributions.

2.2 Mathematical Background and Algorithms

To identify relevant cancer characteristics in the data, it was necessary to utilize both machine learning and statistical methods.

2.2.1 Random forest algorithm

To gain insight into which base pair lengths of fragments and which chromosome regions contribute more to higher cancer purity in our data, two machine learning methods were employed.

Random forest is an algorithm that can be applied either for regression or classification purposes. The method relies on the performance of several decision trees, and new observations are classified based on the majority vote for a class [5]. The main principle of the technique is defined by Leo Breiman in the printed paper *Random Forests*, and the citation reads as follows:

Definition 1 *"A random forest is a classifier consisting of a collection of tree-structured classifiers {h(**x**, Θ_k), k = 1, . . .} where the {Θ_k} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x**". [Breiman, 2001]*

Analyzing the feature importance for the random forest model enables a way to get insight into which variables dictate the dynamics of the cancer purity of the samples.

2.2.2 Gradient boosting machine

Gradient boosting machine (GBM) uses the prediction of several weak learners, in our case decision trees, to generate a better learner based on their joint effort [46].

In an online article called *All You Need to Know about Gradient Boosting Algorithm Part 1. Regression*, Tomonori Masui provides a thorough summary of the algorithm and its details, originally introduced by Jerome H. Friedman in *Greedy Function Approximation: A Gradient Boosting Machine* [28, 13].

The algorithm consists of two main steps, where the second one is made up of several substeps:

First, the model is initialized by a prediction of F_0:

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, x) \tag{2.1}$$

in the equation, L represents the loss function minimized by γ and i make up all the samples.

Secondly, for $m \in \{1, 2, \ldots, M\}$ iterations, one starts by computing the residuals for each sample:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n \tag{2.2}$$

This step is followed up by using the features x and the generated residuals to train the tree of the iteration. Hence, one create node regions R_{jm} for $j = 1, \ldots, J_m$, where J are the total number of leafs for a tree indexed by m.

Further, the procedure continues by finding a γ minimizing L at each node, aggregated over all samples assigned to the node region:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma) \quad \text{for } j = 1, \ldots, J_m \tag{2.3}$$

This means each tree node, the leaf will have a unique γ, scaling value. Lastly, the algorithm ends by updating the model accordingly to:

$$F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} \cdot 1(x \in R_{jm}) \tag{2.4}$$

where v is the learning rate, dictating the contribution of the newly introduced tree achieved during the current iteration m.

2.2.3 K-fold cross validation

K-fold cross-validation is a method that enables the user to check the overall performance of the model and the model's sensitivity to different splits of data. By split, one refers to the division between training and test data. It works by first splitting the full set of data into K smaller groups; these groups are then referred to as the folds of the cross-validation [41]. For each of the folds $k \in 1, 2, ..., K$ one then proceeds to use the k:th fold as test data and the rest as training data when running the model. When the procedure has been performed for all of the folds, the performance is evaluated by considering their joint average performance.

2.2.4 Kernel Density Estimation for Distribution Smoothing

Raw fragment length histograms often contain noise and binning artifacts. To overcome these limitations, we employ Kernel Density Estimation (KDE), a non-parametric technique for smoothing discrete data into a continuous probability density function [4, 10]:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \tag{2.5}$$

where K_h is the kernel function with bandwidth h, x_i are the observed fragment lengths, and n is the number of observations. We use a Gaussian kernel:

$$K_h(x) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{x^2}{2h^2}} \tag{2.6}$$

The bandwidth h controls the smoothness of the estimated density function and is selected using Scott's rule [10], which adapts to the data characteristics.

2.2.5 Expectation-Maximization for Parameter Learning in cfDNA Fragment Analysis

The analysis of cell-free DNA (cfDNA) fragment length distributions requires precise characterization of patterns associated with different cell death mechanisms. While fixed parameter approaches can be effective, they do not account for dataset-specific variations [27]. Our adaptive method employs Expectation-Maximization (EM) to learn optimal parameters directly from the data, resulting in more reliable classifications and higher confidence scores.

Expectation-Maximization is an iterative algorithm for finding maximum likelihood estimates of parameters in probabilistic models with latent variables [15]. In our context, the latent variables are the underlying cell death mechanisms, and the observable data are the extracted features from cfDNA fragment length distributions.

The algorithm alternates between two steps:

- **Expectation (E-step)**: Using current parameter estimates to calculate the probability of each sample belonging to each mechanism
- **Maximization (M-step)**: Updating parameters to maximize the expected log-likelihood given the current mechanism assignments

2.2.6 Bayesian Hierarchical Modeling

Bayesian hierarchical modeling provides a powerful framework for analyzing structured data with multiple sources of variability, making it well-suited for cfDNA fragment analysis across different samples, patients, and chromosome arms.

2.2.7 Markov Chain Monte Carlo (MCMC) for Posterior Inference

The posterior distribution of our model parameters θ given data D is:

$$p(\theta|D) \propto p(D|\theta) \cdot p(\theta) \tag{2.7}$$

Since this posterior is analytically intractable, we employ MCMC sampling, specifically the No-U-Turn Sampler (NUTS) [18], to approximate it. NUTS is an adaptive extension of Hamiltonian Monte Carlo (HMC) that automatically tunes step sizes and trajectory lengths, making it effective for high-dimensional models.

3
Methods

The chapter Methods explains the data introduced in the project, any processing of the data and the execution, implementation, and optimization of both the mathematical and machine learning approaches to solve the project aim.

3.1 The available data

In the initial stage of the project, three types of data were introduced, derived from patients with ovarian cancer. The first set of data was referred to as binned length data and included chromosome region-specific distributions of fragment sizes where each column was a separate distribution. For this data, both chromosome- and arm-specific data were available for many samples. The fragment lengths in the files ranged from 35 to 499 base pairs, and the chromosomes ranged from 1 to 22, neglecting the sex-specific chromosome. For the files with the chromosome arms, only 40 regions were represented per file, as the p-region was undecided for chromosomes 13, 14, 15, and 22 due to lack of quality. An overview of an example of the arm-specific files can be seen in Table 3.1.

Table 3.1: Example of the binned length data file structure

x	**10p**	**10q**	...	**9q**
35				
36				
...				
499				

The next set of data consisted of a map of reference CN-value data. These files were all patient-specific and included several region-specific values per chromosome arm of CN-values, derived from the best sample of a patient. These values were centered around a general baseline of 2, which corresponds to the copy number found in a healthy cell.

The last data available was a detailed sample file, breaking down the domain knowledge for each sample. The file included information regarding cancer purity generated from CN values, the context and time frame of taking the sample, and if it was a liquid biopsy from blood or a tissue biopsy from a tumor.

3.1.1 Pre-processing and normalization of the data

We selected chromosome arm-specific binned length files as our primary data source to maximize available data points per sample. This choice was strategic for two reasons:

1. It avoided training the model on the same data used to generate class labels (cancer purity).
2. It aligned with our project goal to evaluate feature characteristics of liquid biopsies for the filtering software.

During data preparation, we removed tissue-derived samples and chromosome files, leaving 22 liquid biopsy samples for analysis. As all data operated on the same scale, no further normalization was required.

3.1.2 General sample information

The 22 samples can be broken down in detail to their origin along with the specific numbers of samples per patient. The information can be seen in Table 3.2.

Table 3.2: Information about patients and their number of generated liquid biopsy samples of good enough quality to analyze.

Patient	Sample 1	Sample 2	Sample 3	Total
UP0003	HHE2	HHE4	HHE5	3
UP0008	HHE3	HHE5		2
UP0018	HHE1	HHE2		2
UP0026	HHE1	HHE3		2
UP0042	HHE1	HHE3		2
UP0043	HHE1	HHE3		2
UP0046	HHE1	HHE2		2
UP0052	HHE1			1
UP0053	HHE1	HHE3		2
UP0055	HHE1	HHE3		2
UP0056	HHE1	HHE2		2

From the table, one can distinguish that the subset of data is derived from a total of 11 patients. Here, patient UP0003 has contributed with three samples, while UP0052 only with one. Each of the samples has a corresponding cancer purity used for the labeling, which can be read from Table 3.3.

Table 3.3: Level of cancer purity for all samples that remains after pre-processing of the data. Empty cells in the table represent absences of a result for a derived sample or that no sample was collected. Absence of a result can occur in cases of quality issues, such as too little DNA in the collected blood sample. The columns of the table are named HHE1-HHE5 to be consistent to the other files and to consider the time aspect of when samples were derived for a patient.

Patient	HHE1	HHE2	HHE3	HHE4	HHE5
UP0003		0.060		0.255	0.160
UP0008			0.035		0.440
UP0018	0.160	0.085			
UP0026	0.075		0.10		
UP0042	0.045		0.035		
UP0043	0.025		0.040		
UP0046	0.020	0.050			
UP0052	0.040				
UP0053	0.520		0.300		
UP0055	0.240	0.240			
UP0056	0.065	0.130			

From table one can see that the level of purities ranges from 0.020 to 0.520.

3.2 Mechanistic modelling approach

The data used in this part are fragment length distributions, where each distribution for each chromosome part was analyzed independently.

3.2.1 Kernel density estimation

In our implementation, the discrete fragment length distributions were transformed into smooth probability density functions using Kernel Density Estimation (KDE). The implementation followed these steps:

1. For each sample-arm combination, extract fragment lengths and their counts from the preprocessed data.
2. Apply Gaussian kernel smoothing using SciPy's `gaussian_kde` function with weights proportional to fragment counts.
3. Generate smoothed distributions by evaluating the KDE model at 1000 equally spaced points spanning the range of observed fragment lengths.
4. Store both the KDE model and the smoothed distribution for further analysis and feature extraction.

The bandwidth parameter was determined adaptively using Scott's rule, which automatically selects an optimal smoothing bandwidth based on the statistical properties of the data. Scott's rule approximates the optimal bandwidth by $h_j = \hat{\sigma}_j n^{-1/(d+4)}$

for each dimension j, where $\hat{\sigma}_j$ is the sample standard deviation, n is the sample size, and d is the dimensionality of the data [44]. This approach provided appropriate resolution for detecting the nuanced peaks and patterns characteristic of different cell death mechanisms while minimizing noise effects.

From the smoothed distributions, we extracted features that include peak positions, heights, and widths using the `find_peaks` function from SciPy's signal processing module, with a prominence threshold of 5% of the maximum density. These features served as inputs to our classification algorithms.

3.2.2 Expectation-Maximization

Our implementation extends the fixed-parameter classification approach with an EM framework that learns optimal parameters from the data. The algorithm proceeds as follows:

Initialize with default parameters θ_0 based on biological knowledge.

- **E-step:** Classify samples using current parameters θ_{i-1}
- Calculate classification scores and confidence metrics.
- **M-step:** Update parameters θ_i based on current classifications
- Calculate parameter change $\Delta\theta = \|\theta_i - \theta_{i-1}\|$, if $\Delta\theta < convergenceThreshold$ then break

Returns best parameters θ^* and corresponding classifications.
Key parameters optimized in our model include:

$$\theta = \{\theta_{\text{apoptosis}}, \theta_{\text{necrosis}}, \theta_{\text{autophagy}}\} \tag{3.1}$$

$$\theta_{\text{mech}} = \{mean\_length, length\_tolerance, fwhm\_factor, periodicity\_weight, ...\} \tag{3.2}$$

, where $fwhm\_factor$ is the full width at half maximum of the highest peak.

3.2.3 Evaluation and Optimization

The quality of a parameter set is evaluated using a composite score function that incorporates:

$$Score(\theta) = \alpha \cdot Confidence(\theta) + \beta \cdot Separation(\theta) + \gamma \cdot Consistency(\theta) \tag{3.3}$$

where:

- $Confidence(\theta)$ measures the average classification confidence across all samples (difference between highest & second highest cell-mechanism).
- $Separation(\theta)$ quantifies the separation between mechanism clusters using silhouette scores
- $Consistency(\theta)$ evaluates the biological consistency of the results, such as correlation with sample purity

This multifaceted approach ensures that the learned parameters not only provide confident classifications but also maintain biological relevance [31, 8].

3.2.4 Parameter Update Mechanism

The parameter update step is critical to the EM algorithm's success. In our implementation, we update each parameter based on the weighted average of feature values for samples classified with high confidence:

$$\theta_{mech,j}^{new} = \frac{\sum_{i \in S_{mech}} w_i \cdot f_{i,j}}{\sum_{i \in S_{mech}} w_i} \tag{3.4}$$

where:

- S_{mech} is the set of samples classified as mechanism $mech$
- w_i is the confidence weight for sample i
- $f_{i,j}$ is the feature value corresponding to parameter j for sample i

To maintain biological plausibility, parameter updates are constrained to biologically reasonable ranges based on prior knowledge [39, 34].

3.2.5 Hierarchical Model Construction

The hierarchical structure allows for modeling at multiple levels:

1. **Population level:** Hyperpriors define general characteristics of cell death mechanisms.
2. **Chromosome arm level:** Each arm may display different propensities for various cell death mechanisms.
3. **Sample level:** Individual samples exhibit unique patterns influenced by factors like tumor purity and disease stage.

We implement the hierarchical model using PyMC [38], a Python library for probabilistic programming that allows for intuitive model specification:

1. **Population-level hyperpriors:** These encode prior knowledge about cell death mechanisms: (based on Section 2.1.5)

$$\begin{aligned}
\mu_{\text{apoptosis}} &\sim \mathcal{N}(167, 10^2) \\
\sigma_{\text{apoptosis}} &\sim \text{HalfNormal}(15) \\
\mu_{\text{necrosis}} &\sim \mathcal{N}(300, 20^2) \\
\sigma_{\text{necrosis}} &\sim \text{HalfNormal}(100) \\
\mu_{\text{autophagy}} &\sim \mathcal{N}(210, 15^2) \\
\sigma_{\text{autophagy}} &\sim \text{HalfNormal}(40)
\end{aligned} \tag{3.5}$$

2. **Chromosome arm level parameters:** For each arm a, we model mechanism weights:

$$\begin{aligned} w_{a,\text{apoptosis}} &\sim \text{Beta}(2,2) \\ w_{a,\text{necrosis}} &\sim \text{Beta}(2,2) \\ w_{a,\text{autophagy}} &= 1 - w_{a,\text{apoptosis}} - w_{a,\text{necrosis}} \end{aligned} \tag{3.6}$$

3. **Sample-specific parameters:** For each sample-arm combination, we incorporate sample-specific adjustments based on factors like tumor purity:

$$\begin{aligned} \text{purity\_effect}_{s,a} &\sim \mathcal{N}(0, 0.5^2) \\ w_{s,a,\text{apoptosis}} &= \text{sigmoid}(w_{a,\text{apoptosis}} + \text{purity}_s \cdot \text{purity\_effect}_{s,a}) \end{aligned} \tag{3.7}$$

3.2.6 MCMC Inference Implementation

We implement MCMC sampling with the following specifications:

- **Algorithm:** NUTS (No-U-Turn Sampler)
- **Samples:** 2,000 posterior samples (default)
- **Tuning:** 1,000 tuning steps for adaptation
- **Parallel chains:** Multiple chains (default: 2) for convergence checking
- **Target accept rate:** 0.9 (recommended for complex hierarchical models)

To improve computational efficiency, we employ a progress callback function that monitors the progress of the MCMC sampling and estimates the completion time.

3.2.7 Cell Death Mechanism Classification

After obtaining the posterior distribution through MCMC, we classify the dominant cell death mechanism for each sample-arm combination based on the posterior means of the mechanism proportions.

$$\text{dominant\_mechanism}_{s,a} = \arg\max_k \{w_{s,a,k}\} \tag{3.8}$$

where $k \in \{\text{apoptosis}, \text{necrosis}, \text{autophagy}\}$.

3.3 Machine learning-based approach

Both of the machine learning techniques discussed in Section 2.2 of the theory require a data frame with observations for several variables as well as a class to use as a label. To achieve this form of data input, the available data was evaluated and pre-processed according to the details in the upcoming section.

3.3.1 Data organization approaches

The three-dimensional data (sample identity, chromosome arm, and fragment length) required transformation into a tabular format. We explored two distinct approaches to do so:

3.3.1.1 Approach 1: Sample-centric organization

Each fragment length and chromosome region combination became a unique variable, with each patient sample representing an observation. This resulted in a data frame with dimensions of 22 × 18,600 (samples × variables, excluding class).

Table 3.4: Sample-centric organization of binned length data

	35_10p	35_10q	...	499_9q	Class
Sample ID 1					
Sample ID 2					
...					
Sample ID 22					

3.3.1.2 Approach 2: Chromosome region-centric organization

In this approach, we added the chromosome region to the sample index, creating 40 observations per sample (corresponding to the 40 chromosome regions per sample). This produced a data frame with dimensions of 880 × 465 (sample-chromosomes × fragment lengths, excluding class).

Table 3.5: Chromosome region-centric organization of binned length data

	35	36	...	499	Class
Sample ID 1_10p					
Sample ID 1_10q					
...					
Sample ID 22_9q					

We utilized both data organizations during model training and analysis, as each presented different advantages and challenges:

- The first approach risked overfitting due to the high variable-to-sample ratio.
- The second approach introduced potential correlation between different samples.

3.3.2 Class label assignment

We established class labels based on cancer purity values derived from copy number alteration (CNA) data analysis. Specifically, the purity measures deviations from the baseline copy number values, which we then used to categorize samples into three distinct classes:

These ranges were necessary due to the absence of true healthy baseline samples (without ovarian cancer) in our dataset. The levels were set to keep the quasi

Table 3.6: Cancer classification criteria for level of purity

Level of cancer	Ranges of purity
Quasi healthy	$0.00 \leq p \leq 0.05$
Low cancer	$0.05 < \mathrm{p} < 0.25$
Cancer	$\mathrm{p} \geq 0.25$

healthy baseline intact and to get as balanced classes as possible for the remaining two. The resulting class designations were added as a column to both data organization formats.

We also established class labels based on the necrosis mean derived from the mechanical model. Since the mechanical model generates one necrosis mean per chromosome region of a sample, the class labels were assigned based on the mean value of all regions for a sample.

Table 3.7: Cancer classification criterias for level of necrosis

Level of cancer	Ranges of necrosis
Quasi healthy	$0.00 \leq p \leq 0.40$
Low cancer	$0.40 < \mathrm{p} \leq 0.4077$
Cancer	$\mathrm{p} > 0.4077$

The ranges in the table were set to receive a balanced split between the three classes. The resulting class designations in this case were used similarly as above, but this time as an approach to increase the performance for the models.

3.3.3 Models and their evaluation

For GBM and random forest, the models were implemented to take fragment length distribution data of samples along with an associated class, train the model, check the model performance, and return feature importance. The implementation was performed according to the following:

1. Feed in the data frame with fragment length distribution data, chromosome region, sample id, and class
2. Split the data into X and y
3. Define training and test data with the function `train_test_split` from the module model_selection from Scikit-learn
4. Introduce the model and fit the training data to the model. The model classes both originate from Scikit-learn's ensemble module
   (a) For GBM: `GradientBoostingClassifier` is used
   (b) For Random forest: `RandomForestClassifier` is used
5. Evaluate the predicted labels vs. true labels and extract feature importances of the model with `feature_importances_`
6. Repeat but evaluate when implementing K-fold cross-validation into the model

A split of 85 and 15 % was used for the train and test split for the basic model to provide it with as many observations as possible during training. The parameters for the models were during the basic model and first evaluation with K-fold set to the input arguments of Table 3.8.

Table 3.8: Initial input arguments for GBM and random forest

Model	Estimators	Max depth	L.R
Random forest	100	15	-
GBM	100	5	0.1

The number of estimators was chosen to be the default setting for both random forest and GBM, just like the learning rate [42, 43]. The number of estimators can be considered low based on the possible range, however, as only 22 samples were left after filtering, the choice was logical to avoid the risk of overfitting. The maximal depth for random forest was chosen to 15 and for GBM to 5. The idea was to start in the middle of the ranges intended for tuning, as no domain knowledge was known to justify a different approach. For each model's final run, the parameters were set to the ones that promoted the optimized model per fold with hyperparameter tuning. This will be touched upon more in an upcoming section.

3.3.4 K-fold cross validation

To get an accurate representation of how much the split of data affects the performance, K-fold cross-validation was performed according to Section 2.2.3. For this, the `Kfold` function from Scikit-learn's model_selection module was applied to the data. In total, six folds were selected, and from each fold the feature importance and model performance were extracted. Further, both the average performance and the average importance were reviewed based on the collected results from all folds. When later including tuning in the cross-validation, the inner fold was set to ten. The number of folds was chosen to be 6 for the outer folds, aspiring to a relatively close split between training and test data to the one used during the basic model. For the inner fold, ten was chosen as it is a relatively common number and also promotes more data for training during the tuning compared to, e.g. 5.

3.3.5 Optimization of model performance

To optimize the performance for both the models, several aspects were taken into consideration. The list of techniques that were used to improve the model performances is the following:

- Variable correlation to further reduce the data set
- Data transformation
- Hyperparameter tuning
- Introduction of new labels

By checking the variable correlation, one can get a deeper understanding of possibilities to reduce the data set, which can support the model to avoid overfitting. A normal approach for data reduction is principal component analysis; however, as the project is dependent on variable identities to reach its aim, this is not suitable, as one loses the related information. Instead, the reduction was performed by excluding certain regions of data that behaved similarly according to correlation. The reduced set was given to the models to re-evaluate the model performance.

Another approach to improve the performance was trying to transform the data in ways that emphasize deviations in the dataset. The attempt at transformation included subtracting the median or the mean per region for the fraction values of the original files.

For GBM and random forest, there are several parameters that can be tuned to generate better performance without sacrificing fitting or producing high levels of variance [37]. To treat the tuning of the models equally and to keep down running time, both models were tuned on the number of estimators and max depth. For GBM, the estimator equals how many boosting stages are employed (the trees, one per iteration) and the depth parameter controls how many nodes each individual tree gets [42]. For random forest, the estimators symbolize the amount of trees in the model and the max depth the maximal depth of each tree [43]. In addition to this, GBM was tuned for its learning rate as well.

To perform the tuning, `RandomizedsearchCV` was employed from the Scikit-learn's model_selection module. `RandomizedsearchCV` was chosen based on it generally working well for data of high dimensions and that it is less computationally expensive compared to other solutions for tuning [37]. This is key, as the tuning is implemented as a part of the K-fold cross-validation procedure to get optimized parameters specific to each fold. The mixtures of distributions one will achieve for the defined parameter distribution are based on:

- Number of estimators: 100-1000, increment 50.
- Learning rate: 0.01-1.0, increasing with a factor of 10.
- Max depth: 3-7 for GBM and 10-30 for random forest with an increment of 1 and 5, respectively.

The ranges for the estimators were set to give both models room to get optimized. The larger range is important, as more trees usually result in better performance for GBM, which is quite robust to overfitting [42]. The learning rate was set to balance the estimators, as it governs the impact each tree has. A smaller learning rate means less contribution, which means there is a trade-off between the parameter values.

The end values of the range for the maximum depth of random forest were informed by David Paper, as he claims based on his experimentation that the parameter is vital to increased performance [37]. The lower end for GBM was set based on the

depth implemented by David Paper for the GradientBoostinRegressor. To complement this, the upper end was set based on an article breaking down the differences between random forest and GBM [11]. The article aligns with Paper's work, as the lower bound of the range matches his.

Finally, we also implemented classification with both random forests and GBM using labels obtained from the mathematical modelling approach, based on the generated proportion mean of necrosis. This could improve the result, as the new label is chromosome-sample-specific rather than only sample-specific, which therefore might better capture local regions of our samples. For this procedure, we repeated the steps described in Section 3.3.3 for the purity labels.

4

Results

This chapter presents our analysis of cfDNA fragment length distributions, detailing both mechanistic and machine learning approaches for identifying fragment length patterns associated with different cell death mechanisms and their relationship to tumor characteristics.

4.1 Exploration of data & distributions

The results are mainly based on the fragment distributions that we received. Analysis of these distributions revealed significant variations across different arms and chromosome samples.

Figure 4.1 & 4.2 shows two different fragment length distributions of the 1p chromosome arm. These plots illustrate how distribution profiles can differ significantly between samples. The first distribution has a relatively low purity, that is, low amount of cancer, (Figure 4.1) & shows a primary peak centered around 170 bp, with a small secondary peak around 310 bp. In contrast, the second distribution is from a high-purity sample (Figure 4.2) & shows a narrower and more pronounced peak centered around 160 bp, without a visible secondary peak.

These differences in fragment length distributions may reflect underlying biological variations, such as differences in cell-free DNA origin, tumor purity, or other sample-specific characteristics. The presence of secondary peaks, changes in the width of the peak, and shifts in the position of the primary peak are all characteristics that can provide valuable information about the biological context of the sample.

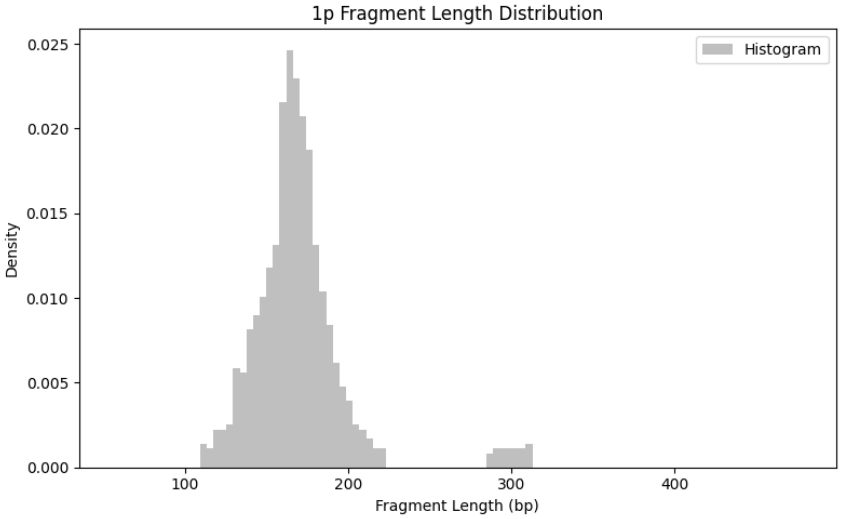

Figure 4.1: Fragment length distribution showing a primary peak with a small secondary peak (Sample UP0053_HHE1, 1p).

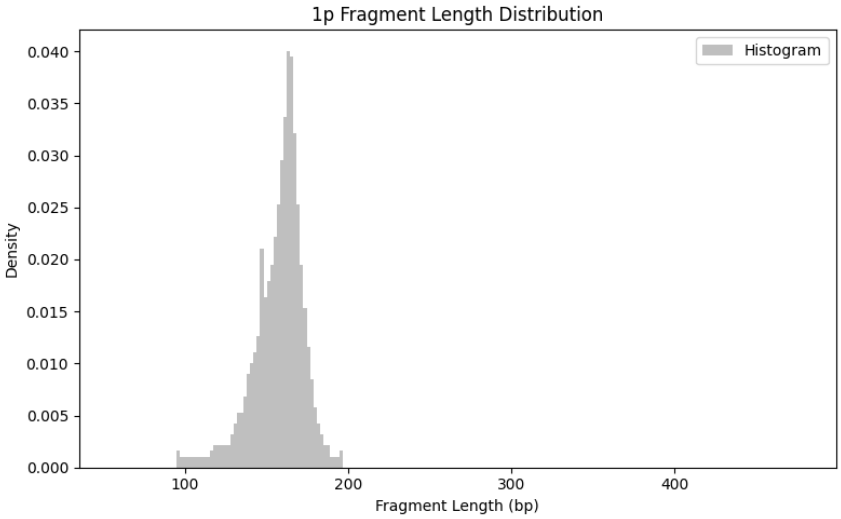

Figure 4.2: Fragment length distribution with a more pronounced single peak (Sample UP0046_HHE1, 1p).

The clear visual differences between these distribution patterns suggest significant potential for computational analysis. These variations are distinguishable even to the naked eye; however, the differences are more subtle between samples with similar purities. Therefore, we employed computational tools to extract more nuanced and quantitative information to better distinguish such samples.

4.2 Machine learning-based approach 1

To try to find patterns in the data, the first approach was begun by running classification on the fragment data labeled with purity, for random forest and gradient

boosting machine. In the upcoming tables, the abbreviations SC and CRC refer to the sample-centric layout 3.3.1.1 and the chromosome region-centric 3.3.1.2. From now on, these will only be referred to by their abbreviations.

4.2.1 Results using the purity label

The validation accuracy when using cancer purity ranges as class labels with input parameters from 3.3.3 can be seen in Table 4.1 as the initial performance.

Table 4.1: Initial model performances for both setups of data layout with the cancer purity label

Model	Layout	Initial perf.
RF	SC	0.500
RF	CRC	0.992
GBM	SC	0.250
GBM	CRC	0.992

Since the CRC layout in the case of random forest and gradient boosting had a notably better result compared to the SC layout, the models were re-run with the shuffle argument for splitting the data adjusted to false. The result can be seen in Table 4.2. This was performed to settle our suspicion of high correlation between arm-level information from the same samples e.g., UP003_HHE1_10q and UP003_HHE1_10p. By adjusting the splitting to false, we were able to guarantee that less data from the same sample occurred in the training and test set, unrealistically improving the performance.

Table 4.2: Model performances analyzing the effect of correlation between chromosome region-specific samples for the chromosome region-centric model layout

Model	Layout	Shuffle	Initial perf.
RF	CRC	False	0.098
GBM	CRC	False	0.098

Based on the results of the tables and the potential correlation in consideration, the SC layout had the highest accuracy with 50%. Due to the strong suspicion of correlation between observations for the CRC layout, the project was continued by using the latter settings for the CRC models. In order to see how the split of data affected the performance, each model was evaluated with K-fold cross-validation; see Table 4.3.

From the table we can determine that 3 out of 4 average fold performances outperformed the accuracies in Tables 4.1 and 4.2. Furthermore, the lowest and highest folds were notably different for all models, with the largest deviation found for random forest with the SC layout (0.750) and the smallest for GBM with the SC layout

Table 4.3: Model performances when implementing K-fold cross-validation for the chosen models and appropriate settings

Model	Layout	Shuffle	Lowest Fold	Highest Fold	Average Fold
RF	SC	True	0.000	0.750	0.389
RF	CRC	False	0.184	0.714	0.424
GBM	SC	True	0.250	0.500	0.361
GBM	CRC	False	0.171	0.823	0.492

(0.250). Since the difference between the folds varied and the majority of the models had better accuracies for the average K-fold, we decided to use this metric for evaluating models going forward. The choice of using average K-fold is necessary to deliver results and conclusions based on transparency rather than luck.

4.2.2 Approaches for improved model performance

Even though K-fold cross-validation improved the performance of many models, their overall performance was still too poor to be practically useful. To investigate the underlying reasons for this, we attempted three approaches: data reduction, data transformation, and hyperparameter tuning. Data reduction and transformation aimed to reveal whether the low performance was due to overfitting or inherent issues with the dataset (e.g., minimal differences between samples). Hyperparameter tuning was used to determine whether the poor performance stemmed from suboptimal model configurations or from deeper issues related to the data quality and labeling.

4.2.2.1 Reduce dataset according to correlation

To choose which data to exclude from the set, an average correlation matrix was generated for the 22 fragment distribution samples.

Based on the figure, a decision was made to cut down the dataset to 62.5% of the original data by removing chromosome regions 10p-18q. The reduction was attempted on random forest and GBM for the SC-layout but predicted identical results as the basic model in Table 4.1, of 0.500 and 0.250, respectively.

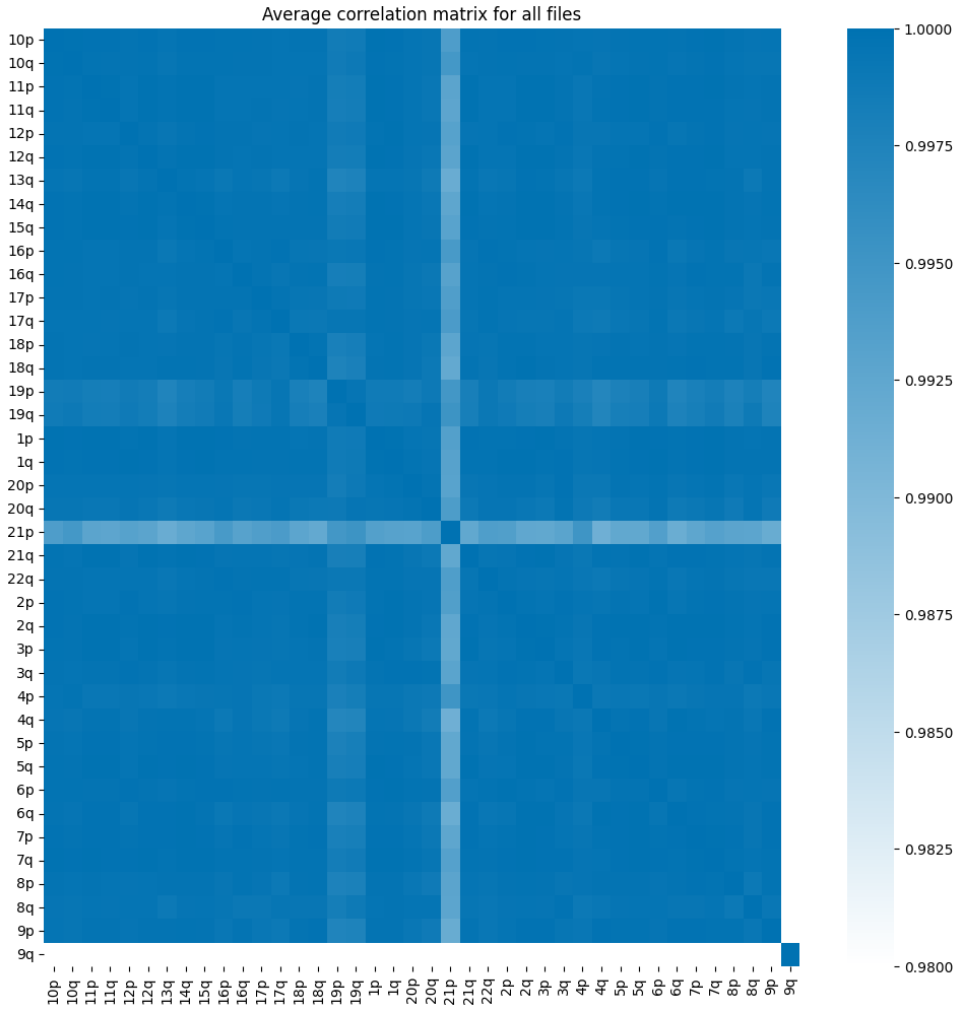

Figure 4.3: A correlation matrix based on the average of the 22 correlation matrices generated from each fragment distribution sample file.

4.2.2.2 Transforming the data

To highlight the differences in the fragment size distribution instead of focusing on shared characteristics, two new data frames were generated, one with the subtracted mean per column and one with the subtracted median. One example were such a difference can be seen, is when comparing the range 150-180 bp of Figure 4.1 and 4.2.

Prior to re-running the basic model for random forest and GBM for the SC-layout, the differences for a chromosome region between the samples were plotted. In the following figure, the distribution for two different samples for region 10p from the mean transformed data frame is plotted, see Figure 4.4.

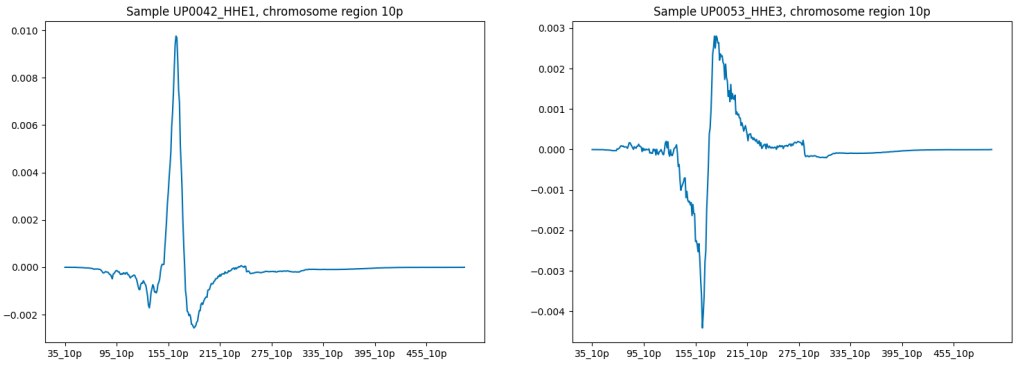

Figure 4.4: Distribution for the mean transformed data of samples UP0042_HHE1 and UP0053_HHE3 for chromosome region 10p.

Even if the figure visualizes completely different distributions, similarly to the plots during the exploration, the transformations still resulted in identical accuracies as the original data for the random forest and GBM with the SC layout.

4.2.2.3 Hyperparameter tuning

The last technique explored was tuning with random search. The tuning was integrated into the K-fold cross-validation and performed according to Section 3.3.5. The results compared to K-fold prior to tuning are evaluated in Table 4.4.

Table 4.4: Model performances with and without integrating tuning of hyperparameters in the K-fold cross validation for the chosen models

Model	Layout	Average	Tuned average
RF	SC	0.389	0.333
RF	CRC	0.424	0.446
GBM	SC	0.361	0.458
GBM	CRC	0.492	0.514

The biggest difference in tuning is seen for GBM with the SC layout, but higher performance was still achieved for all models except random forest with the SC layout.

4.3 Mechanistic approach

In contrast to machine learning methods, a mechanistic approach was employed to address the problem from a fundamentally different perspective. This strategy allows for independent validation, as the results obtained through mechanistic modeling can be compared against those derived from data-driven approaches.

4.3.1 Kernel density estimation

Fragment-length distributions can be better understood by applying smoothing techniques such as Kernel Density Estimation (KDE). This method helps visualize the underlying patterns in the raw histogram data.

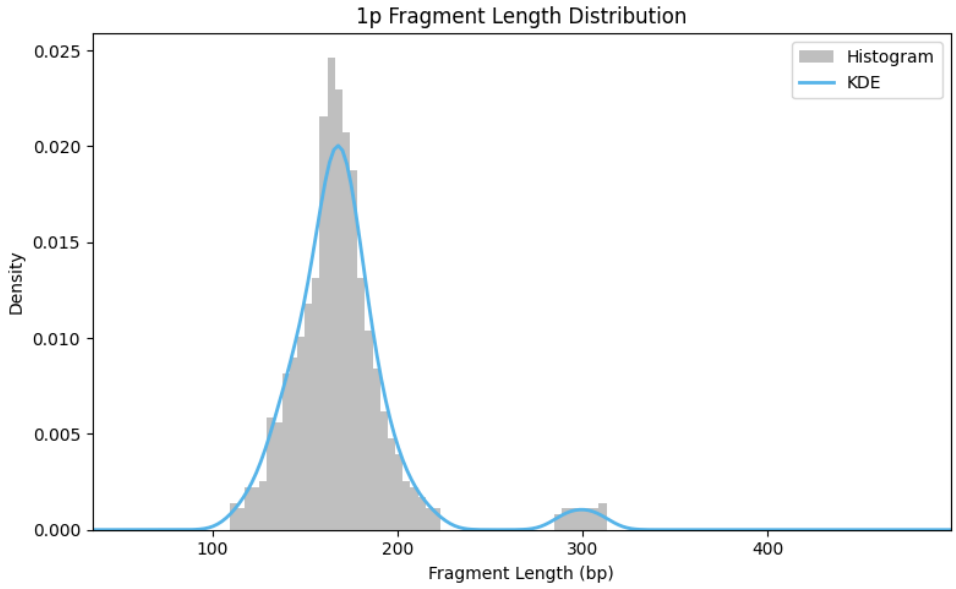

Figure 4.5: Fragment length distribution with KDE showing bimodal characteristics with a primary peak at 170 bp and a secondary peak around 310 bp.

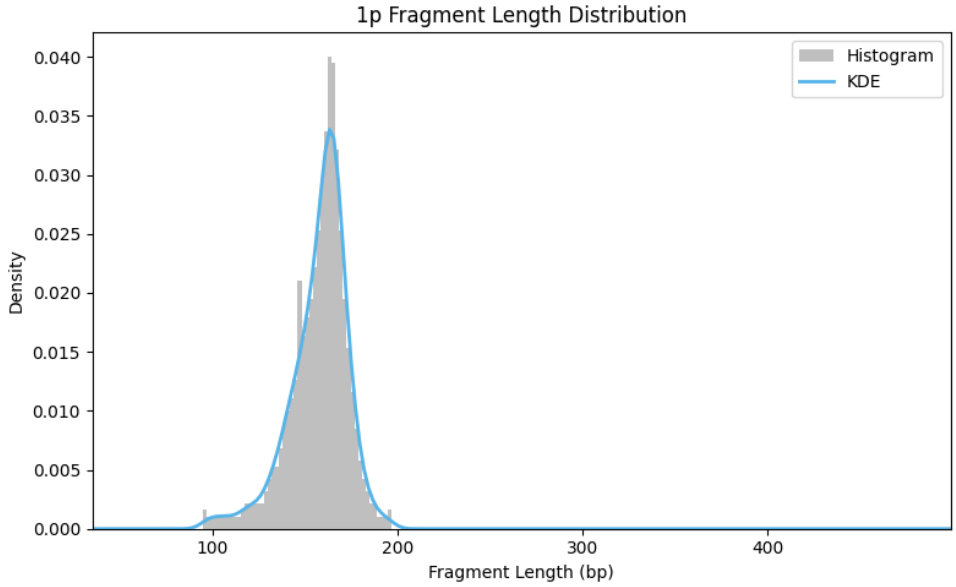

Figure 4.6: Fragment length distribution with KDE showing a single, narrower peak around 160 bp without secondary features.

The KDE visualization provides several advantages over simple histograms. The KDE curves (blue lines) in figures 4.5 & 4.6 smoothly capture the distribution pat-

terns while minimizing binning artifacts present in the raw histograms. Figure 4.5 illustrates how the KDE captures the primary peak at approximately 170 bp and a secondary peak at around 310 bp. Figure 4.6 shows how small artifacts—such as the 'peak' around 140 bp—are smoothed out.

4.3.2 Extraction of Cell-death Percentages

We implemented a minimalistic model to analyze fragment-length distributions from cfDNA sequencing data. The model applies Kernel Density Estimation (KDE) to fragment length distributions and extracts features characteristic of each cell death mechanism, including periodicity scores, mean fragment lengths, and peak widths measured by Full Width at Half Maximum (FWHM).

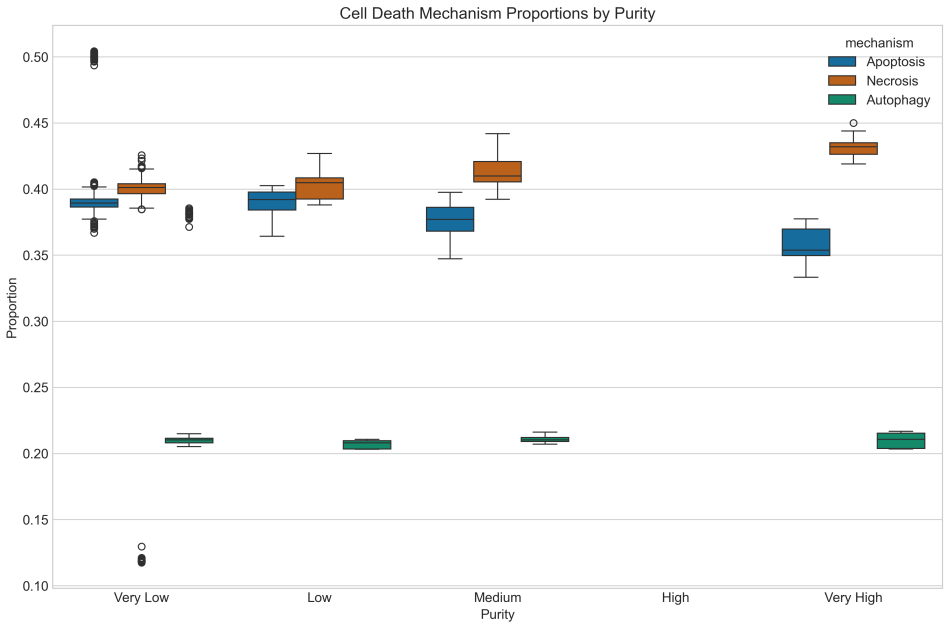

Figure 4.7: Distribution of cell death mechanism proportions across samples with different purity levels. The samples were divided into 5 equal boxes between purity 0.02 & 0.52. The boxplots show the proportion of apoptosis (blue), necrosis (red), and autophagy (green) detected in samples grouped by purity.

Figure 4.7 presents the quantification of relative contributions of different cell death mechanisms in samples of varying purity. The data shows a trend toward higher proportions of necrotic cell death in samples with greater purity.

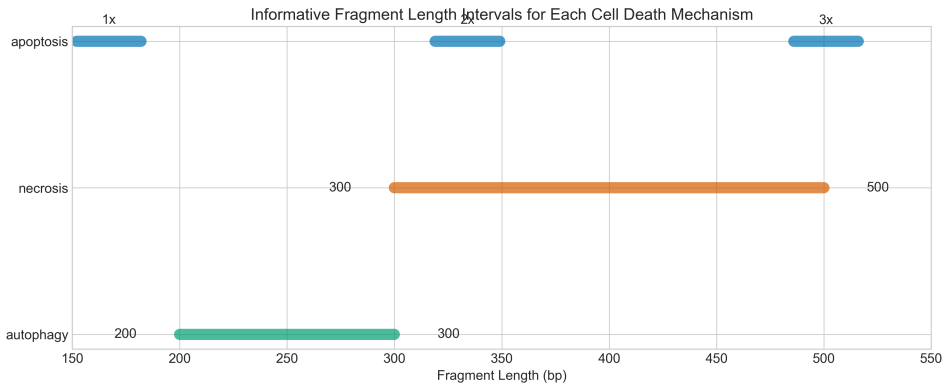

Figure 4.8: Informative fragment length intervals for each cell death mechanism. The horizontal bars indicate the fragment length ranges that are most characteristic of each mechanism.

The model identifies the most informative fragment length intervals for distinguishing between cell death mechanisms (Figure 4.8). For apoptosis, these intervals center around multiples of 167 bp. Necrosis is characterized by longer fragments in the 300-500 bp range, while autophagy typically shows enrichment in the 200-300 bp range.

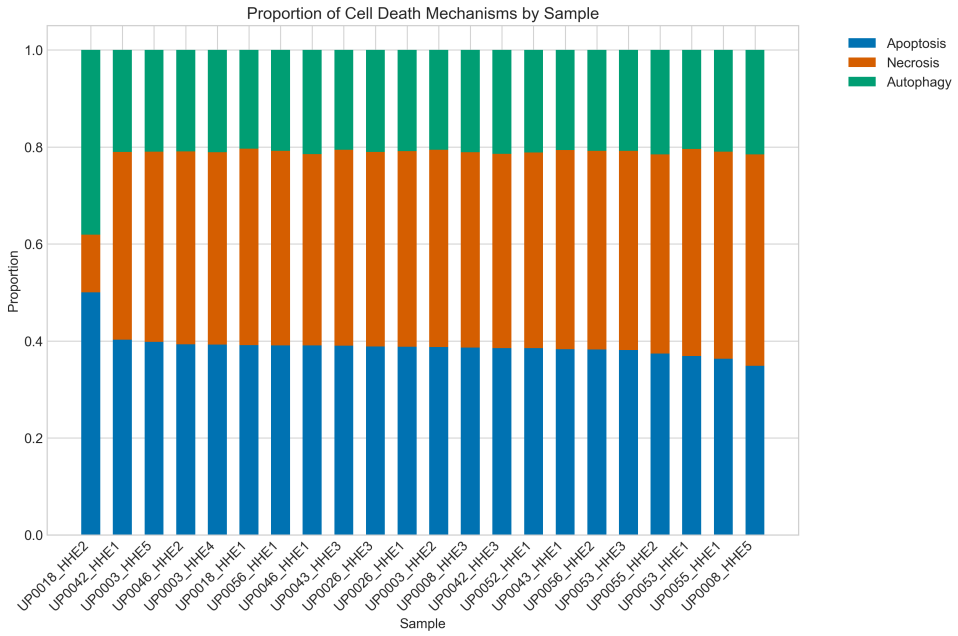

Figure 4.9: Proportion of cell death mechanisms by sample sort by proportion of apoptosis. The stacked bars show the relative contribution of apoptosis (blue), necrosis (red), and autophagy (green) in each sample.

Analysis of the cell death mechanism proportions (Figure 4.9) reveals considerable variation across samples. The distribution shows that samples exhibit different compositions of cell death mechanisms, ranging from those dominated by a single mechanism to others with a more balanced mixture.

4.3.2.1 Expectation-Maximization (EM) Parameter Learning

We explored enhancing our model using an Expectation-Maximization (EM) approach to learn optimal parameters from the data. This adaptive approach iteratively refines classification parameters based on the current assignments of samples to cell death mechanisms.

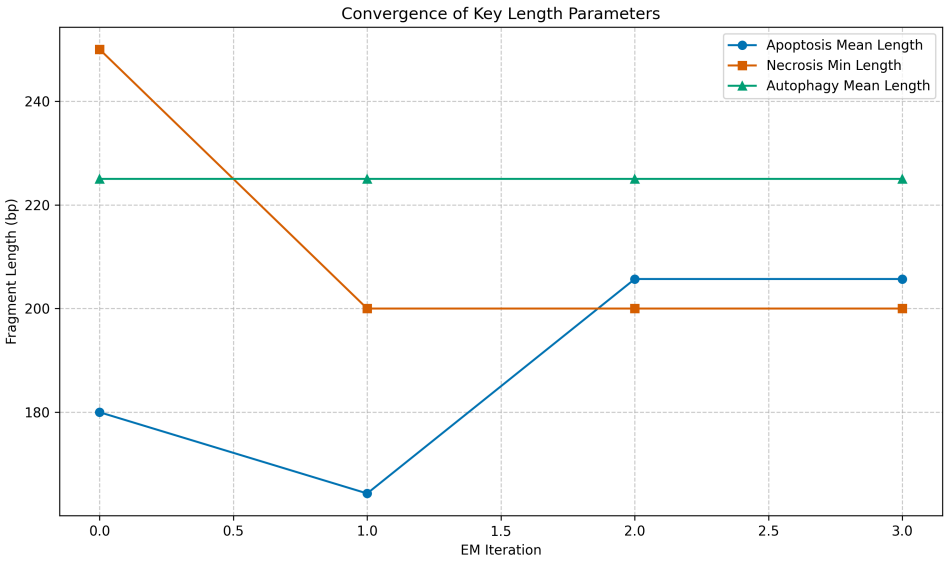

Figure 4.10: Convergence of key length parameters during EM iterations, showing the evolution of characteristic fragment lengths for each mechanism.

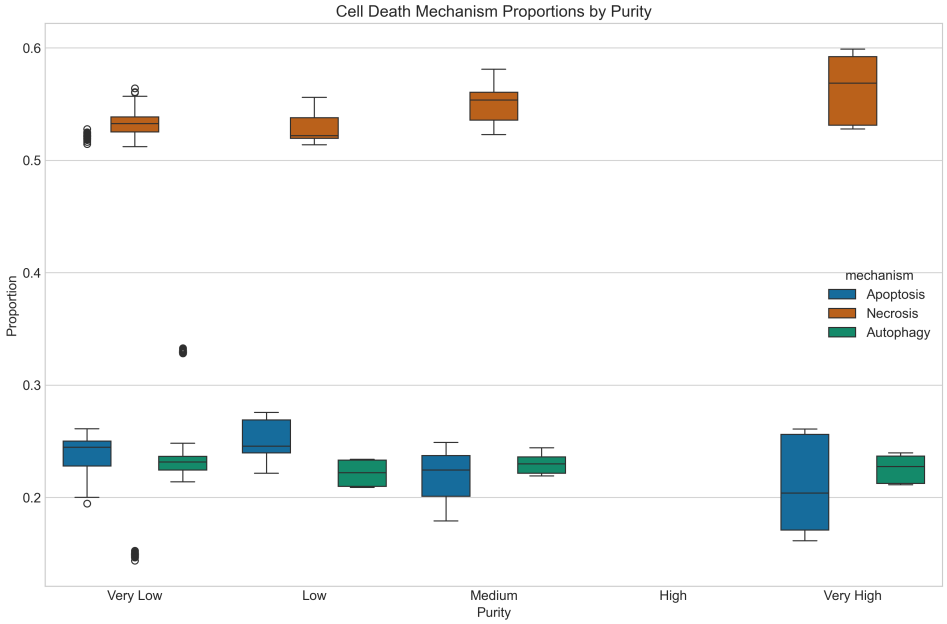

Figure 4.11: Distribution of cell death mechanism proportions using EM-derived parameters. Compared to Figure 4.7, this shows less biologically plausible associations with purity.

The EM algorithm converged to stable values as shown in Figure 4.10. The resulting mechanism distributions using EM-derived parameters (Figure 4.11) differ from those obtained using our standard model with hand-tuned parameters (Figure 4.7). Notably, the relationship between necrosis proportion and sample purity appears less pronounced in the EM-derived results.

4.3.3 Validation of Necrosis Measurements as Reliable Labels

To use our results as a new label for analysis, we examined two critical aspects of our necrosis measurements: the correlation between necrosis proportion and sample purity, and the consistency of necrosis signaling across different chromosomes within individual samples.

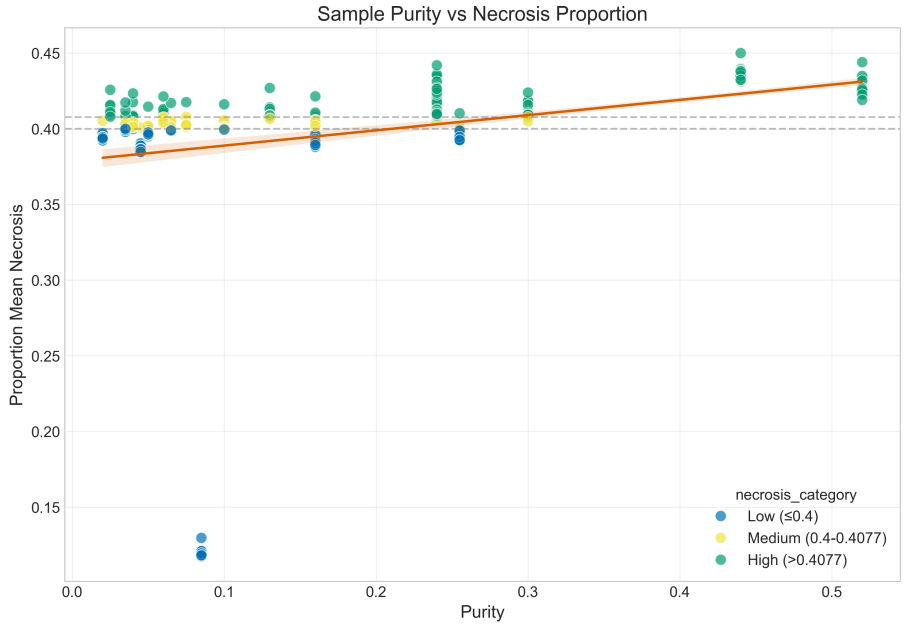

Figure 4.12: Correlation between sample purity and necrosis proportion. Samples are color-coded by necrosis category: low (blue), medium (yellow), and high (green). The linear trend (red line) indicates a positive association.

Figure 4.12 shows a positive correlation between sample purity and the proportion of necrosis-derived fragments.

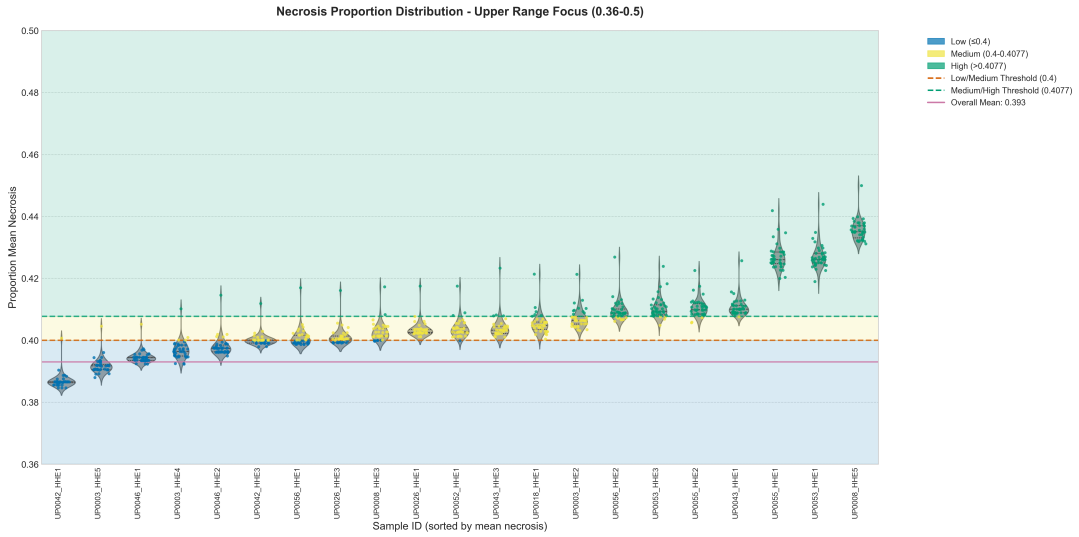

Figure 4.13: Violin plot showing the distribution of necrosis proportions across chromosomal arms for each sample, excluding sample UP0018_HHE2. Narrower distributions indicate more consistent necrosis signaling across the genome.

To assess measurement consistency across chromosomes, we examined the distribution of necrosis proportions in different visualization formats. Figure 4.13 displays the distribution of necrosis measurements across chromosomal arms for each sample, indicating the degree of internal consistency.

4.4 Machine learning-based approach 2

Due to the necrosis measurements being sufficient to use as reliable labels for the models, the entirety of the process in Section 4.2 was repeated but with labels based on the proportion mean necrosis.

4.4.1 Results using the necrosis label

To get an accurate comparison, the same setups were run, and the initial model results can be seen in Table 4.5.

Table 4.5: Initial model performances for both setups of data layout with the necrosis labels

Model	Layout	Shuffle	Initial perf.
RF	SC	True	1.000
RF	CRC	False	0.871
GBM	SC	True	0.750
GBM	CRC	False	0.856

Overall, all models performed better with the generated label rather than the purity.

The biggest difference is seen for the CRC layout, which had an improved accuracy of 77.3 and 75.8 percentage points for random forest and GBM, respectively. For the models with integrated K-fold, we could see the following values for the lowest, highest, and average folds; see Table 4.6.

Table 4.6: Model performances when implementing K-fold cross validation for the chosen models and appropriate settings, with updated class labeling using necrosis mean instead

Model	Layout	Shuffle	Lowest Fold	Highest Fold	Average Fold
RF	SC	True	0.333	1.000	0.708
RF	CRC	False	0.667	0.969	0.848
GBM	SC	True	0.000	1.000	0.458
GBM	CRC	False	0.575	0.939	0.761

For the new labeling, all average folds were remarkably improved. In addition to this, the highest folds were also higher than previously in Table 4.3. The only decrease is seen for the lowest fold of GBM with the SC-layout.

4.4.2 Hyperparameter tuning for improved model performance with necrosis label

In Section 4.2 hyperparameter tuning was the only approach to model optimization that resulted in higher accuracy. Therefore, the optimization was limited to tuning for the necrosis label, see results in Table 4.7.

Table 4.7: Model performances with and without tuning of hyperparameters for the K-fold cross validation, when employing the new labeling based on necrosis

Model	Layout	Average	Tuned average
RF	SC	0.708	0.764
RF	CRC	0.848	0.866
GBM	SC	0.458	0.583
GBM	CRC	0.761	0.834

Just like with the old labeling, the tuning improved the overall validation accuracy for all of them. The patterns of improvement remained similar to the improvements seen specific to each model with the purity label. The only exception was detected for random forest with the SC layout, which improved instead of decreasing like the prior case.

4.5 Final results and characteristics based on feature importance

The summarized results for the various variations of model setups, data layouts, and different labels are available in Table 4.8.

Table 4.8: Model performances for the final models when employing K-fold cross validation and hyperparameter tuning with random search. The table also contains adjusted model results to compensate for group-specific data leakage for the CRC layout.

Model	Layout	Shuffle	Label	Average Fold w. Tuning	Adjusted
RF	SC	True	Cancer purity	0.333	0.333
RF	CRC	False	Cancer purity	0.446	0.326
RF	SC	True	Necrosis	0.764	0.764
RF	CRC	False	Necrosis	0.866	0.837
GBM	SC	True	Cancer purity	0.458	0.458
GBM	CRC	False	Cancer purity	0.514	0.409
GBM	SC	True	Necrosis	0.583	0.583
GBM	CRC	False	Necrosis	0.834	0.798

For all of the CRC-layout results, one can expect a group leakage of approximately 17-18% per fold (26-27 observations out of 880/ 6 per fold, knowing each sample has 40 chromosome regions). This is expected based on the result seen in Table 4.2, indicating correlation. The leakage can be declared as each fold contains observations used for validation from samples the model will be trained on. The accuracies adjusted for the group leakage can be seen in the sixth column. The adjustment was made for a leakage of 26 observations in total.

Based on the table, the highest accuracies were found when using the necrosis label. Three model setups stand out as being sufficient to study in further detail. These are the random forests with SC and CRC layouts as well as the GBM with CRC layout. These are sufficient due to the performances being higher than a relevant baseline (e.g. a dummy classifier). As the classes were balanced for the necrosis label; 1: 32.84%, 2: 33.64% and 3: 33.52%, a dummy classifier would result in 33% accuracy, which is clearly lower than what the three models delivered.

To distinguish any findings of the models for base pair-related characteristics, the feature importances were visualized in Figures 4.14, 4.15 and 4.16.

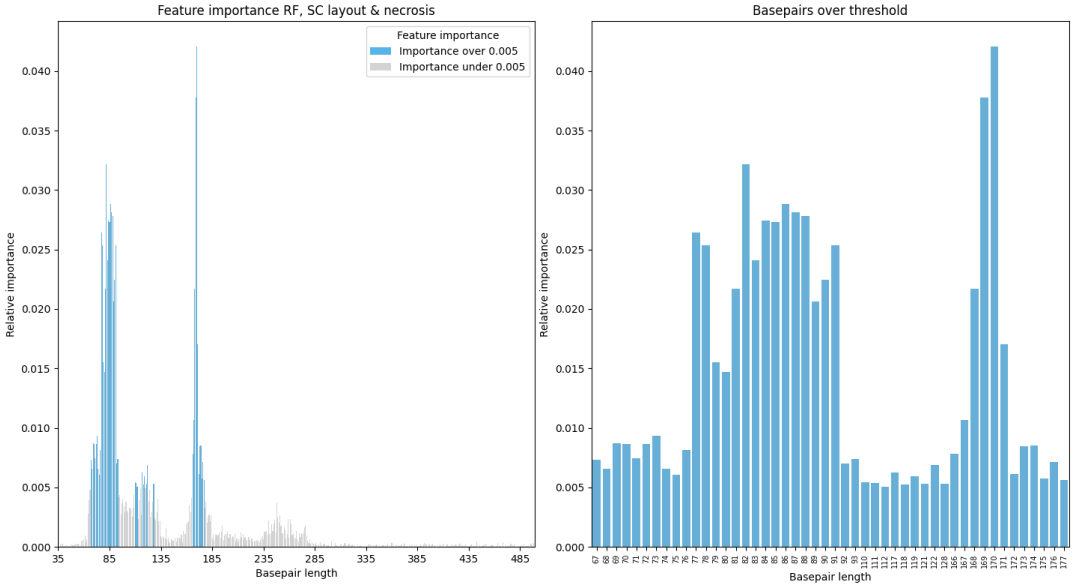

Figure 4.14: Feature importance visualized for random forest with the SC-layout. The x-axis on the plot to the left represents the full range of fragment length sizes in the dataset. The blue bars mark out base pair lengths passing the threshold of 0.005. The plot on the right is a close-up of the lengths passing the filtering and their corresponding feature importance.

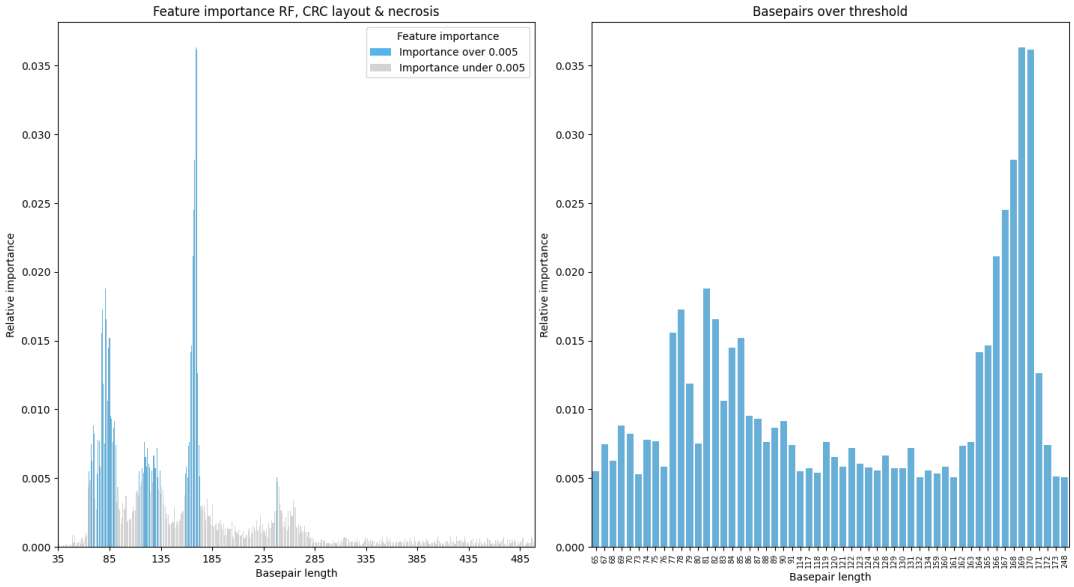

Figure 4.15: Feature importance visualized for random forest with the CRC layout. Similarly as before, the bars in blue represent the passing base pair lengths. The figure on the left is the zoom-in of the bars on the left.

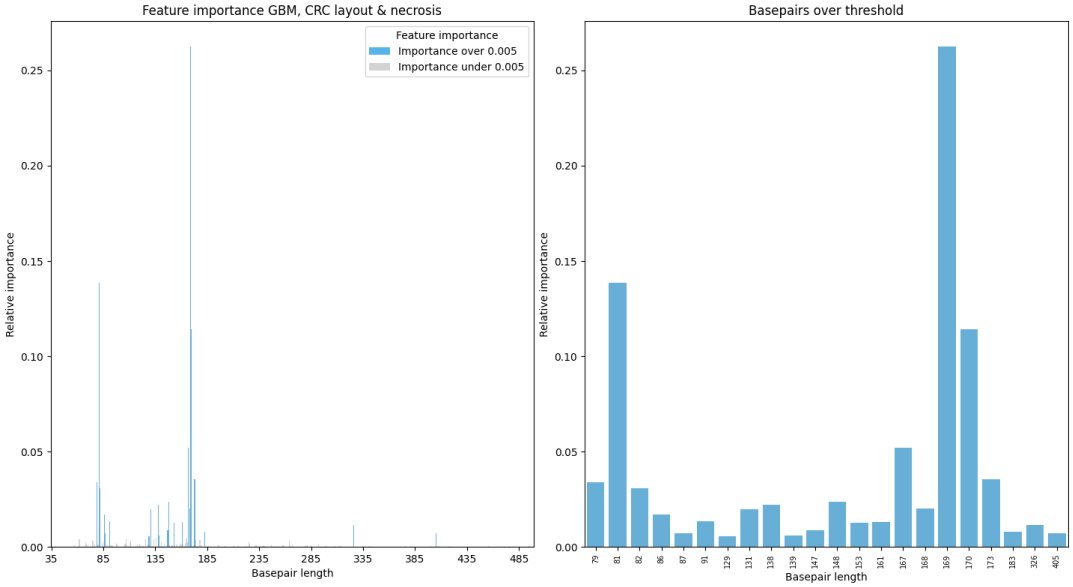

Figure 4.16: Feature importance based on gradient boosting machine with the CRC layout. Once again, the bars in blue represent the passing base pair lengths in relation to the whole range, while the figure on the left visualizes the identity of the passing bars.

From Figures 4.15 and 4.14 two quite distinct peaks can be seen for similar ranges. For Figure 4.15 the ranges for the peaks are quite similar in length, with the first peak in the range 77-85 and the second 164-171. In 4.14, the first peak is wider in range, ranging from 77 to 91, while the second is more narrow, with 168 to 171. For the last figure, Figure 4.16 the peaks are made up of much fewer bars with higher relative importance in comparison to the figures based on random forest. However, the few bars making up the peaks still lie within the same range as the ones for random forest, with 81-82 and 167-170.

5
Discussion

Our approaches to cfDNA fragment analysis reveal complementary insights into the fragment length distributions through both mechanistic modeling and machine learning techniques.

5.1 Mechanistic approach

Our mechanistic approach to cfDNA fragment analysis provides valuable biological insights beyond genomic alterations. This discussion examines the patterns observed and evaluates our model's performance.

5.1.1 Interpretation of Fragment Length Distributions

In this section, we analyzed cfDNA fragment length distributions using KDE-based visualizations to uncover patterns indicative of underlying biological processes. As demonstrated in figures 4.5 & 4.6, KDE visualization effectively reveals patterns that might be obscured in raw histogram data. The bimodal distribution seen in Figure 4.5, strongly suggests the presence of multiple cell death mechanisms contributing to the cfDNA pool. This pattern is consistent with previous findings that different cell death pathways produce characteristic fragment sizes [40, 24].

In contrast, the unimodal distribution in Figure 4.6 indicates a more homogeneous source of cfDNA fragments. This pattern likely reflects a predominance of a single cell death mechanism, possibly apoptosis, given the peak at approximately 160 bp, which corresponds to the mono-nucleosomal fragment size typically associated with apoptotic cell death [9, 6].

5.1.2 Mechanistic Model Performance and Biological Implications

In this section, we applied a mechanistic model to deconvolve cfDNA fragment patterns and interpret the biological contributions of different cell death mechanisms across samples. Our mechanistic model successfully quantified the relative contributions of different cell death mechanisms across samples. The relationship between tumor purity and cell death mechanisms, as shown in Figure 4.7, reveals a biologically plausible pattern where increased necrosis correlates with higher tumor purity. This finding aligns with established knowledge that solid tumors often experience

necrotic cell death due to hypoxic conditions in rapidly growing masses [7, 33].

The identified informative fragment length intervals (Figure 4.8) correspond well with the theoretical expectations for each cell death mechanism. The nucleosomal ladder pattern characteristic of apoptosis is reflected in intervals centered around multiples of 167 bp, while the longer fragments associated with necrosis (300-500 bp) demonstrate the less regulated nature of this cell death process [31].

The variability in mechanism proportions across samples (Figure 4.9) likely reflects the complex biology of cancer. Tumors can undergo multiple forms of cell death simultaneously, influenced by factors such as treatment status, microenvironment conditions, and tumor subtype [3]. This heterogeneity underscores the value of our approach in capturing the nuanced reality of tumor biology rather than assuming a single dominant mechanism.

5.1.3 Methodological Considerations

In this section, we evaluated the impact of different parameter estimation strategies on model performance, highlighting the trade-offs between data-driven methods and domain-informed approaches. The comparison between hand-tuned parameters and EM-derived parameters (Figures 4.7 and 4.11) presents an interesting methodological insight. Despite the theoretical advantage of letting the data inform parameter values, our findings suggest that domain knowledge-guided parameterization produced more biologically plausible results. This highlights the importance of incorporating prior biological understanding when developing models for complex biological systems [16].

The challenges encountered with the MCMC approach underscore the computational complexity of Bayesian methods in this context. While this approach offers the promise of uncertainty quantification, the balance between computational feasibility and model sophistication remains a consideration for future work.

5.1.4 Validation of Necrosis as a Biomarker

Lastly, in this section, we validated our mechanistic model by examining its biological relevance and consistency across the genome, reinforcing its robustness as a cfDNA analysis tool. The positive correlation between sample purity and necrosis proportion (Figure 4.12) provides compelling evidence for the biological relevance of our mechanistic classification. This relationship supports the hypothesis that tumors with higher malignant cell content exhibit increased necrotic cell death, consistent with the characteristic growth patterns of aggressive cancers [26].

The consistency analysis of necrosis measurements across chromosomal arms (Figures 4.13, A.1, and A.2) demonstrates that while some chromosomal variation exists, the overall signal is sufficiently stable within samples to serve as a reliable biomarker. This genomic consistency suggests that the cell death mechanisms we observe re-

flect systemic biological processes rather than technical artifacts or isolated genomic events.

5.1.5 Markov Chain Monte Carlo (MCMC) Exploration

We attempted to implement a Bayesian approach using MCMC methods to model cell death mechanisms with uncertainty estimates. Despite development efforts, we were unable to perfect this approach due to challenges with model convergence and computational efficiency. This approach was excluded from our final results but remains a potential avenue for future research.

5.2 Machine learning

For the machine learning approaches, the Table 4.8 suggested that the best working models were:

- Random forest with the SC-layout and necrosis label (0.764)
- Random forest with the CRC layout and necrosis label (0.837)
- Gradient boost with the CRC layout and necrosis label (0.798)

From the list, we can state that the necrosis label worked better than the purity label, that the CRC layout worked better than the SC, and finally that random forest performed better than GBM for the necrosis label.

There is one main advantage for the generated necrosis label compared to the purity, which may explain why it performed better. This is because the necrosis labeling produced chromosome region-specific identities rather than sample-specific ones. There are several reasons why it is useful to use a sample value generated from a mean of several data points instead of using a single value only. Some of these are that the necrosis label captures local variations for the chromosome arms and that it minimizes influence from data constituting noise, e.g., the dataset's outliers.

For the layouts, both setups have their own pros and cons. The SC has an imbalance between variables and observations, meaning the model easily overfits and that the accuracy is validated for very few observations. However, the layout is structured based on the nature of the data, as each sample is treated as a unique observation. The CRC has a much better balance between variables and observations but is drawn back by the data leakage. The leakage occurs as a result of treating chromosome region-specific observations from samples as unique observations when the samples have correlation between their chromosome regions (see Table 4.1 compared to 4.2). In 4 out of 4 cases prior to the tuning (based on 4.3 and 4.6), the CRC layout outperformed the SC. Looking at the tuned, final result after adjusting for the leakage, the CRC was still better in 3 out of 4 cases. Given that the adjustment is sufficient, the results indicate it is more favorable for the models to have more observations to learn from and to compensate for inaccuracies than to stick to the nature of the data by limiting the amount of data available for training.

Based on Tables 4.4 and 4.7 it is a fair statement that the tuning worked well, as 7 out of 8 performances increased by using fold-specific parameters. As the other approaches were evaluated for only one layout of both the models, it is not possible to state that the methods were unsuccessful in providing better results. However, there exist enough data to logically assume they were.

Lastly, from the plots visualizing the feature importance, we can see very similar trends between the figures. The peaks are quite dominant in Figures 4.14, 4.15 and 4.16, but what is interesting is the almost perfect overlap between each of the models in the identities of the bars producing the peaks. The higher peak per case is range-wise quite closely tied to the distinctive fragment lengths around 167 bp for apoptosis. For all of the cases, the highest importances were found specifically at 169-170 bp. A logical reason for this could be that the lengths around 167 bps are the most common during cell death, meaning all files will possess fraction values for the area being values other than 0.00. For the first occurring peak, the ranges instead were:

- Random forest with CRC 77-85
- Random forest with SC 77-91,
- Gradient boosting with CRC 81-82

Here, the highest relative importances at lengths 81 and 82 were slightly lower than for the other peak. Based on theory, a peak prior to 167 bp is expected, but what is interesting is that the overlap between the visuals is so big. This suggests that the models value the variables similarly in regard to importance, even if the steps of the algorithms vary. The highest feature importances in the machine learning models correspond to the centration of cfDNA fragment patterns associated with apoptosis (see Section 2.1.5). The initial peaks observed in the feature importance plots may be attributed to the broader, less distinct cfDNA fragment pattern of necrosis, which results from greater variability in fragment lengths. Due to uncertainties of the data leakage and the very few observations of the SC structure, the summarized results are not sufficient to justify neglecting data through filtering. However, what can be said is that the range of base pair lengths for the earlier peak has potential in being linked to cancer, as the peak is present for all of the best predicting models but also has base pair identities that are reasonable in regard of the theory for cell death in cancer and the cfDNA fragment patterns.

5.3 Future Improvements & Limitations

Our work presents several promising approaches for analyzing cfDNA fragment distributions, but certain limitations should be acknowledged:

- **MCMC Implementation:** The unsuccessful implementation of the Markov Chain Monte Carlo approach highlights computational challenges in applying Bayesian methods to this problem. Future work could focus on improving

model parameterization and computational efficiency to enable proper uncertainty quantification for cell death mechanism assignments.

- **Class Label Refinement:** Our three-class model (apoptosis, necrosis, autophagy) is a simplified representation of complex biological processes. More nuanced classifications could better capture the spectrum of cell death pathways and their intermediate states.
- **Limited Baseline Data:** The absence of cfDNA samples from healthy individuals represents a significant limitation of our study. Without baseline data from non-cancer subjects, it is difficult to establish normal fragment distribution patterns and definitively attribute observed features to cancer-specific processes. Future studies should incorporate samples from healthy controls to better distinguish cancer-specific signals from natural variation in cfDNA fragmentation patterns.
- **Sample Size:** Our analysis is based on a relatively small cohort. Expanding the dataset would increase statistical power and potentially reveal additional patterns or correlations not apparent in the current sample set.
- **Classification Threshold Optimization:** The choice of thresholds for categorizing necrosis levels (≤ 0.4, $0.4 - 0.4077$, > 0.4077) was based on observed data distributions. Further optimization of these thresholds using larger datasets could improve classification performance.

5.4 Conclusion

This study demonstrates the value of integrating mechanistic and machine learning approaches for analyzing cfDNA fragment distributions. Our mechanistic model successfully quantified the relative contributions of different cell death mechanisms and established necrosis proportion as a biologically meaningful parameter that correlates with tumor purity. This necrosis-based classification proved superior to direct tumor purity labels for machine learning models, achieving validation accuracies up to 83.7% with random forest on the chromosome region-centric layout.

The feature importance analysis revealed consistent patterns across different models, with fragment lengths around 80-90 bp and 167-170 bp emerging as particularly informative. These regions align with theoretical expectations for nucleosomal and inter-nucleosomal fragments associated with apoptotic cell death, suggesting that our models are capturing biologically relevant signals.

The comparison between parameter-learning approaches highlighted the value of domain knowledge in guiding model development. While data-driven parameter optimization through EM showed technical convergence, the hand-tuned parameters based on biological understanding produced more plausible results, particularly in maintaining the expected relationship between necrosis and tumor purity.

Our findings validate fragment length analysis as a valuable tool for liquid biopsy applications, capable of providing insights beyond genomic alterations. By characterizing cell death mechanisms through fragmentomic features, we offer a comple-

mentary approach that could enhance the diagnostic and monitoring capabilities of liquid biopsies, potentially contributing to more personalized cancer management strategies in the future.

A

Appendix 1

A.1 Code

Here is the link to the github containing the finished version of the project code: https://github.com/StAmirey/masterthesis-fragment-length-distribution

A.2 Extra figures

In this section lies extra figures for the result.

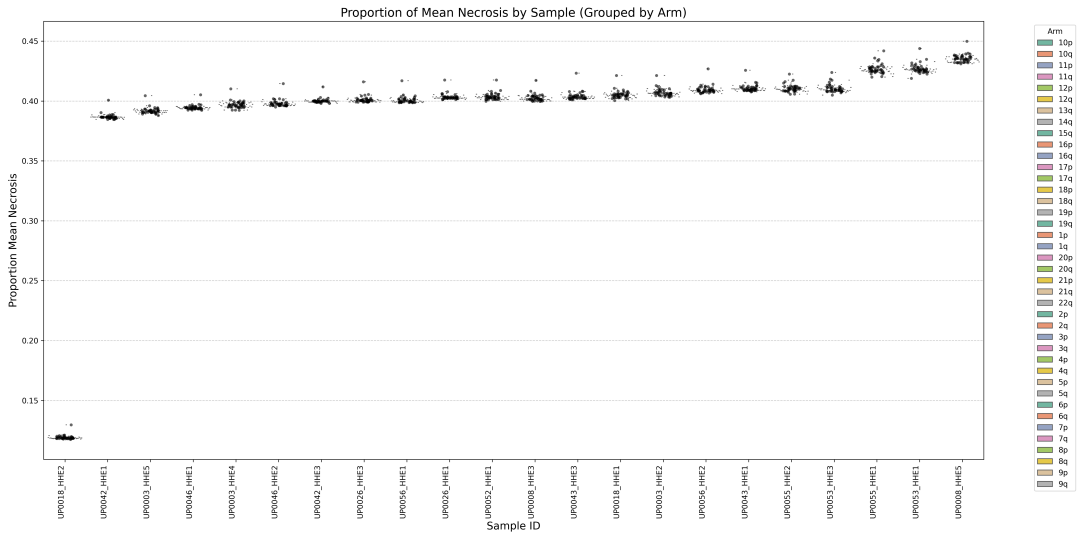

Figure A.1: Grouped representation of necrosis proportions across samples, highlighting the consistency of measurements across different chromosomal regions within each sample.

Figure A.1 provides an alternative visualization of chromosomal consistency, with samples plotted according to their measured necrosis values.

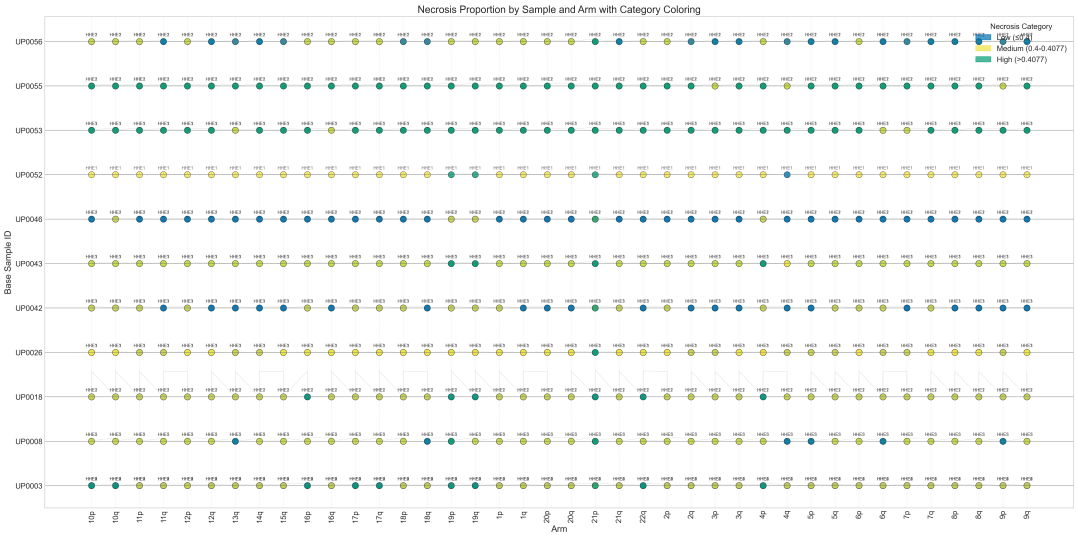

Figure A.2: Scatter plot showing necrosis categories (by color) across sample-arm combinations. Blue ≤ 0.4, yellow $0.4 - 0.4077$, and green > 0.4077. Consistent coloring within a sample row indicates uniform necrosis signaling across chromosomes.

For a more granular view of chromosomal variation, Figure A.2 maps necrosis categories across samples and chromosomal arms simultaneously. This visualization helps identify samples where certain chromosomal arms deviate from the overall pattern or where a sample contains measurements that fall into multiple categories.

Bibliography

[1] Khan Academy. *Types of mutations and their notations*. Accessed: 2025-03-21. 2025. URL: https://www.khanacademy.org/test-prep/mcat/biomolecules/genetic-mutations/a/types-of-mutations-and-their-notations.

[2] Said Assou et al. "Dynamic changes in gene expression during human early embryo development: from fundamental aspects to clinical applications". In: *Human Reproduction Update* 17.2 (2010), pp. 272–290. DOI: 10.1093/humupd/dmq036. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3189516/.

[3] Janine Aucamp et al. "Cell-free DNA: Preanalytical variables". In: *Clinica Chimica Acta* 515 (2021), pp. 48–54. DOI: 10.1016/j.cca.2020.12.031. URL: https://www.sciencedirect.com/science/article/pii/S0009898120305350.

[4] Z. I. Botev, J. F. Grotowski, and D. P. Kroese. "Kernel density estimation via diffusion". In: *Annals of Statistics* 38.5 (2010). Accessed: 2025-03-30, pp. 2916–2957. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-38/issue-5/Kernel-density-estimation-via-diffusion/10.1214/10-AOS799.full.

[5] Leo Breiman. "Random Forests". In: *Springer* 45 (2001). Accessed: 2025-04-03, pp. 5–32. URL: https://link.springer.com/article/10.1023/a:1010933404324.

[6] Abel Jacobus Bronkhorst et al. "Characterization of the cell-free DNA released by cultured cancer cells". In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1863.1 (2016), pp. 157–165. DOI: 10.1016/j.bbamcr.2015.10.022. URL: https://www.sciencedirect.com/science/article/pii/S0167488915003973.

[7] Wout Celus et al. "Circulating cell-free DNA fragment size analysis: A window into biology and pathology". In: *Molecular Oncology* 16.19 (2022), pp. 3501–3517. DOI: 10.1002/1878-0261.13274. URL: https://febs.onlinelibrary.wiley.com/doi/full/10.1002/1878-0261.13274.

[8] K. C. A. Chan et al. "Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing". In: *Proceedings of the National Academy of Sciences* 113.22 (2016). Accessed: 2025-03-30, E3106–E3114. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5678696/.

[9] Dineika Chandrananda, Natalie P Thorne, and Melanie Bahlo. "High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA". In: *BMC medical genomics* 8 (2015), pp. 1–19. DOI: 10.1186/

s12920-015-0107-z. URL: https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-015-0107-z.

[10] Y. C. Chen. *A Tutorial on Kernel Density Estimation and Recent Advances*. arXiv preprint arXiv:1704.03924. Accessed: 2025-03-30. 2017. URL: https://arxiv.org/pdf/1704.03924.pdf.

[11] Jason Chong. *Battle of the Ensemble – Random Forest vs Gradient Boosting*. Accessed: 2025-05-07. 2025. URL: https://towardsdatascience.com/battle-of-the-ensemble-random-forest-vs-gradient-boosting-6fbfed14cb7/.

[12] Cleveland Clinic. *Cell Death*. Accessed: 2025-03-16. Cleveland Clinic. 2023. URL: https://my.clevelandclinic.org/health/articles/cell-death (visited on 03/16/2025).

[13] Jerome H. Friedman. *Greedy Function Approximation: A Gradient Boosting Machine*. Accessed: 2025-04-04. 2001. URL: https://jerryfriedman.su.domains/ftp/trebst.pdf.

[14] World Cancer Research Fund. *Ovarian cancer*. Accessed: 2025-04-14. URL: https://www.wcrf.org/preventing-cancer/cancer-types/ovarian-cancer/.

[15] GeeksforGeeks. *ML | Expectation-Maximization Algorithm*. https://www.geeksforgeeks.org/ml-expectation-maximization-algorithm/. Accessed: 2025-05-01. 2023.

[16] Moritz Gerstung et al. "The evolutionary history of 2,658 cancers". In: *Nature* 578.7793 (2020), pp. 122–128. DOI: 10.1038/s41586-019-1907-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7054223/.

[17] Leonard Hayflick and Paul S. Moorhead. *The serial cultivation of human diploid cell strains*. Original paper introducing the Hayflick limit concept. 1961. DOI: 10.1016/0014-4827(61)90192-6.

[18] Matthew D. Hoffman and Andrew Gelman. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1593–1623. URL: https://www.jmlr.org/papers/volume15/hoffman14a/hoffman14a.pdf.

[19] Illumina. *Introduction to Illumina Sequencing*. Technical Document. Available at: https://emea.illumina.com/content/dam/illumina-marketing/documents/products/illumina Illumina Inc. URL: https://emea.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

[20] National Human Genome Reasearch Institue. *Chromosome*. Accessed: 2025-03-19. 2025. URL: https://www.genome.gov/genetics-glossary/Chromosome.

[21] National Human Genome Reasearch Institue. *Deoxyribonucleic acid (DNA)*. Accessed: 2025-03-19. 2025. URL: https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid-DNA.

[22] National Human Genome Reasearch Institue. *Genome*. Accessed: 2025-03-21. 2025. URL: https://www.genome.gov/genetics-glossary/Genome.

[23] National Human Genome Reasearch Institue. *Mutation*. Accessed: 2025-03-21. 2025. URL: https://www.genome.gov/genetics-glossary/Mutation.

[24] Peng Jiang and Y M Dennis Lo. "Plasma cell-free DNA fragment size estimation for non-invasive cancer screening: A systematic review". In: *Cancer Communications* 40.2-3 (2020), pp. 167–182. DOI: 10.1002/cac2.12044. URL: https://pubmed.ncbi.nlm.nih.gov/32133736/.

[25] Guido Kroemer et al. *Classification of Cell Death: Recommendations of the Nomenclature Committee on Cell Death 2009*. Accessed: 2025-03-16. 2009. DOI: 10.1038/cdd.2008.150. URL: https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/necrosis.

[26] Enrico Benedetti Lugli et al. "Enhanced diagnostic yield of tumor molecular profiling using cell-free DNA". In: *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* 38.15_suppl (2020), e13534–e13534. DOI: 10.1200/JCO.2020.38.15_suppl.e13534. URL: https://ascopubs.org/doi/abs/10.1200/JCO.2020.38.15_suppl.e13534.

[27] H. Markus et al. "Circulating Cell-Free DNA Fragmentation Patterns". In: *Frontiers in Genetics* 12 (2021). Accessed: 2025-03-30, p. 773459. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7764982/.

[28] Tomonori Masui. *All You Need to Know about Gradient Boosting Algorithm Part 1. Regression*. Accessed: 2025-04-04. 2022. URL: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502/.

[29] National Library of Medicine. *Chromosome Map*. Accessed: 2025-03-20. 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK22266/.

[30] National Library of Medicine. *Copy Number Alterations as Novel Biomarkers and Therapeutic Targets in Colorectal Cancer*. Accessed: 2025-03-21. 2022. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9101426/.

[31] F. Mouliere et al. "Fragmentation patterns and personalized sequencing of cell-free DNA in urine and plasma of glioma patients". In: *EMBO Molecular Medicine* 10.12 (2018). Accessed: 2025-03-30, e9323. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7162689/.

[32] Florent Mouliere et al. "Enhanced detection of circulating tumor DNA by fragment size analysis". In: *Science Translational Medicine* 10.466 (2018). Accessed: 2025-03-24, eaat4921. DOI: 10.1126/scitranslmed.aat4921. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6483061/.

[33] Takeshi Namekawa et al. "Detection of cell-free tumor DNA in plasma and urine from patients with localized prostate cancer by using methylation-specific PCR". In: *clinics and research in hepatology and gastroenterology* 42.6 (2018), pp. 1413–1425. DOI: 10.1002/pros.23533. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pros.23533.

[34] M. H. D. Neumann et al. "Cell-free DNA: Characteristics and Potential Clinical Applications". In: *Journal of Laboratory Medicine* 42.4 (2018). Accessed: 2025-03-30, pp. 173–183. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7066862/.

[35] National Health Service- NHS. *Overview- Ovarian cancer*. Accessed: 2025-04-14. 2022. URL: https://www.nhs.uk/conditions/ovarian-cancer/.

[36] Oxford Nanopore Technologies. *Cell-free DNA nanopore sequencing and methylation detection — promising potential for non-invasive cancer monitoring.*

Published: September 29 2022, Accessed: 2025-03-16. Oxford Nanopore Technologies. 2022. URL: https://nanoporetech.com/resource-centre/cell-free-dna-nanopore-sequencing-and-methylation-detection.

[37] David Paper. *Hands-on Scikit-Learn for Machine Learning Applications*. Apress Berkeley, CA, 2019, pp 137–163 (138), 177–179, 189, 204. DOI: https://doi.org/10.1007/978-1-4842-5373-1.

[38] PyMC Developers. *PyMC: Probabilistic Programming in Python*. Accessed: 2023-11-15. 2023. URL: https://www.pymc.io/.

[39] R. Sadeh et al. "Cell-free DNA Fragmentomics: The New "Omics" on the Block". In: *Molecular Cell* 81.18 (2021). Accessed: 2025-03-30, pp. 3683–3689. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8620333/.

[40] Carlos Sánchez et al. "Cell-free DNA reflects multiple cell death mechanisms in non-small cell lung cancer". In: *Biology* 10.7 (2021), p. 623. DOI: 10.3390/biology10070623. URL: https://www.mdpi.com/2079-7737/10/7/623.

[41] Scikit-learn. *3.1. Cross-validation: evaluating estimator performance*. Accessed: 2025-04-17. 2025. URL: https://scikit-learn.org/stable/modules/cross_validation.html.

[42] Scikit-learn. *GradientBoostingClassifier*. Accessed: 2025-05-02. 2025. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html.

[43] Scikit-learn. *RandomForestClassifier*. Accessed: 2025-05-02. 2025. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[44] David W. Scott. "Scott's rule". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4 (2010), pp. 497–502. DOI: 10.1002/wics.103. URL: https://onlinelibrary.wiley.com/doi/epdf/10.1002/wics.103.

[45] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249. URL: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21660.

[46] Bex Tuychiev. "A Guide to The Gradient Boosting Algorithm". In: *Datacamp* (2023). URL: https://www.datacamp.com/tutorial/guide-to-the-gradient-boosting-algorithm.

[47] Andrew H. Wyllie. *Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation*. Accessed: 2025-03-16. 1980. DOI: 10.1038/284555a0. URL: https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/dna-laddering.

[48] W. Zhou, M. Haber, and N. Bardeesy. "Cell-free DNA fragmentations and its applications". In: *Cell Genomics* 3.9 (2023). Accessed: 2025-03-30, p. 100478. DOI: 10.1073/pnas.2220982120. URL: https://pubmed.ncbi.nlm.nih.gov/37075072/.

Department of Mathematics and Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY