Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Multiparametric modeling of mitochondrial DNA for cancer diagnosis

Master's Thesis

2026
Bc. Ondrej Škornák

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Multiparametric modeling of mitochondrial DNA for cancer diagnosis

Master's Thesis

Study Programme:	Computer Science
Field of Study:	Computer Science
Department:	Department of Computer Science
Supervisor:	Mgr. Jaroslav Budiš, PhD.

Bratislava, 2026
Bc. Ondrej Škornák

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Bc. Ondrej Škorňák
Študijný program:	informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor:	informatika
Typ záverečnej práce:	diplomová
Jazyk záverečnej práce:	anglický
Sekundárny jazyk:	slovenský

Názov: Multiparametric modeling of mitochondrial DNA for cancer diagnosis
Multiparametrické modelovanie mitochondriálnej DNA pre onkologickú diagnostiku

Anotácia: Súčasný rozvoj biomedicínskych technológií otvára nové možnosti pre detekciu, charakterizáciu a monitoring onkologických ochorení priamo z krvi pacienta. Veľký potenciál skrýva cielená analýza mitochondriálneho genómu, ktorý vykazuje špecifické zmeny spojené s prebiehajúcim onkologickým ochorením. Keďže fragmentačné a mutačné profily mtDNA môžu slúžiť ako citlivé biomarkery, hĺbková analýza cirkulujúcej mitochondriálnej DNA predstavuje veľký potenciál k rozšíreniu existujúcich metodík, ktoré sú typicky zamerané na analýzu fragmentov z jadrovej DNA.

Študent bude detailne analyzovať atribúty sekvenačných čítaní na úrovni jednotlivých DNA fragmentov, so zameraním na koincidenciu viacerých príznakov (dĺžka, somatické mutácie, lokalizácia v genóme). Navrhne a implementuje predikčný model, ktorý na rozdiel od tradičných, agregačných metód využije multiparametrický prístup na identifikáciu vysokorizikových onkologických fragmentov. Vyhodnotí presnosť klasifikácie porovnaním vzoriek onkologických pacientov a kontrolných jedincov.

Vedúci:	Mgr. Jaroslav Budiš, PhD.
Katedra:	FMFI.KI - Katedra informatiky
Vedúci katedry:	prof. RNDr. Martin Škoviera, PhD.
Dátum zadania:	02.01.2026

Dátum schválenia: 02.01.2026

prof. RNDr. Rastislav Kráľovič, PhD.
garant študijného programu

....
študent

....
vedúci práce

89056057

Comenius University Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Bc. Ondrej Škorňák
Study programme:	Computer Science (Single degree study, master II. deg., full time form)
Field of Study:	Computer Science
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Multiparametric modeling of mitochondrial DNA for cancer diagnosis

Annotation: The recent advancement of biomedical technologies opens new possibilities for the detection, characterization, and monitoring of oncological diseases directly from a patient's blood. Targeted analysis of the mitochondrial genome holds great potential, as it exhibits specific changes associated with ongoing cancer. Since mtDNA fragmentation and mutation profiles can serve as sensitive biomarkers, an in-depth analysis of circulating mitochondrial DNA represents an opportunity to enhance existing methodologies, which typically focus on nuclear DNA fragments.

The student will analyze the attributes of sequencing reads at the level of individual DNA fragments, focusing on the co-occurrence of multiple features (length, somatic mutations, genomic localization). He will design and implement a predictive model that would utilize this multiparametric approach to identify high-risk cancer-derived fragments. The classification accuracy will be evaluated by comparing samples from cancer patients and control subjects.

Supervisor:	Mgr. Jaroslav Budiš, PhD.	
Department:	FMFI.KI - Department of Computer Science	
Head of department:	prof. RNDr. Martin Škoviera, PhD.	
Assigned:	02.01.2026	
Approved:	02.01.2026	prof. RNDr. Rastislav Kráľovič, PhD. Guarantor of Study Programme

.... Student

.... Supervisor

Acknowledgments: I would like to thank my supervisor, Mgr. Jaroslav Budiš, PhD., for his expert guidance, valuable advice, and support throughout the writing of this Master's thesis. I also thank the Faculty of Mathematics, Physics and Informatics of Comenius University for providing quality education and creating proper conditions for research. Finally, I would like to thank my family, girlfriend, and all my friends for their continuous support and encouragement.

Abstrakt

Cirkulujúca bezbunková DNA (cfDNA) umožňuje minimálne invazívnu analýzu molekulárnych signálov spojených s ochorením a predstavuje základ moderných prístupov tekutej biopsie v onkológii. Väčšina cfDNA analýz sa zameriava na jadrovú DNA, avšak cirkulujúca bezbunková mitochondriálna DNA (ccf-mtDNA) môže poskytovať doplnkovú a menej preskúmanú diagnostickú informáciu. Onkologický signál v mtDNA sa môže prejavovať nielen v mutačných profiloch, ale aj vo fragmentačných vlastnostiach a v lokalizácii fragmentov pozdĺž mitochondriálneho genómu.

Táto práca skúma ccf-mtDNA na úrovni jednotlivých fragmentov a analyzuje koincidenciu (spolu-výskyt) kľúčových atribútov, najmä dĺžky fragmentu, jeho genomickej polohy a mutačných príznakov. Navrhujeme a implementujeme pipeline na rekonštrukciu mtDNA fragmentov zo sekvenačných dát a na extrakciu fragmentových príznakov vhodných pre strojové učenie. Na tomto základe predstavujeme multiparametrický predikčný rámec, ktorý priraďuje rizikové skóre jednotlivým fragmentom a následne agreguje fragmentové signály na klasifikáciu vzoriek do skupín onkologických pacientov a kontrol. Navrhnutý prístup vyhodnocujeme na kohorte pacientov a kontrol a porovnávame ho s jednoduchšími agregačnými baseline metódami, aby sme posúdili prínos spoločného, fragmentovo orientovaného modelovania.

Výsledky potvrdzujú použiteľnosť fragmentovo založeného multiparametrického prístupu pri analýze mtDNA v tekutej biopsii a prinášajú poznatky o kombináciách fragmentačných a mutačných signálov, ktoré sú najinformatívnejšie pre rozlíšenie pacientov a kontrol.

Kľúčové slová: tekutá biopsia, ccf-mtDNA, fragmentomika, multiparametrické modelovanie, detekcia rakoviny

Abstract

Circulating cell-free DNA (cfDNA) enables minimally invasive analysis of disease-related molecular signals and has become a central component of liquid biopsy approaches in oncology. While most cfDNA assays focus on nuclear DNA, circulating cell-free mitochondrial DNA (ccf-mtDNA) represents a complementary and comparatively underexplored source of information. Cancer-associated signals in mtDNA may appear not only in mutational profiles but also in fragmentation patterns and genomic localization of fragments along the mitochondrial genome.

This thesis investigates ccf-mtDNA at single-fragment resolution and studies the co-occurrence of key fragment attributes, including fragment length, genomic position, and mutation-related features. We design and implement a feature-extraction pipeline that reconstructs mtDNA fragments from sequencing data and produces a fragment-level representation suitable for machine learning. Building on this representation, we propose a multiparametric predictive framework that assigns risk scores to individual fragments and aggregates these signals to classify samples as oncological or control. The proposed approach is evaluated on patient and control cohorts and compared against simpler aggregative baselines to assess the added value of joint, fragment-level modeling.

The results demonstrate the feasibility of fragment-level multiparametric modeling for mtDNA-based liquid biopsy analysis and provide insights into which combinations of fragmentomic and mutation-related signals are most informative for cancer/control discrimination.

Keywords: liquid biopsy, ccf-mtDNA, fragmentomics, multiparametric modeling, cancer detection

Contents

List of Figures

List of Tables

Introduction

Recent advances in sequencing technologies have substantially improved our ability to digitize genomic information at scale. Next-generation sequencing (NGS) has become faster, more accurate, and more accessible, enabling its routine use in biomedical research and, increasingly, in clinical practice. Alongside laboratory progress, the development of computational methods for processing and interpreting large sequencing datasets has played a crucial role in translating raw reads into clinically meaningful information. As a result, genomics is now an essential component of modern precision medicine, with oncology being one of the most prominent areas of application [13].

Cancer remains a major global health challenge with a continuously increasing burden. According to global estimates for the year 2022, nearly 20 million new cancer cases were diagnosed worldwide and approximately 9.7 million people died from the disease [1]. Projections further suggest that the number of new cases will rise markedly over the coming decades, driven by demographic changes and population growth, emphasizing the need for scalable strategies for prevention, early detection, and treatment monitoring [18]. Early diagnosis is particularly important, as it can significantly improve treatment options and patient outcomes. Nevertheless, early-stage detection remains difficult for many cancer types due to subtle or non-specific symptoms and limitations of current screening modalities.

Traditional diagnostic workflows often rely on tissue biopsy, which provides direct access to tumor material and enables histopathological and molecular characterization. However, tissue biopsy is invasive, may not be feasible for repeated sampling, and can be associated with patient discomfort and procedure-related risks. Moreover, a single biopsy may not fully capture tumor heterogeneity and can be limited by sampling location and tumor accessibility. These constraints have motivated the development of minimally invasive alternatives, most notably liquid biopsy, which aims to infer disease-related information from biomolecules present in body fluids, typically blood plasma [6]. Among the most widely studied analytes is circulating tumor DNA (ctDNA), a subset of cell-free DNA (cfDNA) released into circulation from tumor cells. ctDNA-based assays have demonstrated utility across multiple clinical scenarios, including treatment response monitoring, detection of minimal residual disease, and relapse surveillance, while also showing potential for earlier detection [6].

Despite these advances, liquid biopsy data pose substantial analytical challenges. In many clinically relevant settings—especially in early-stage disease—tumor-derived signals can be extremely weak relative to background cfDNA originating from non-malignant tissues. This low signal-to-noise ratio complicates the reliable detection of cancer-associated molecular features and motivates the search for additional biomarkers and computational strategies that can improve sensitivity and robustness. In recent years, this effort has contributed to the emergence of cfDNA "fragmentomics," which studies fragmentation patterns of cfDNA, including fragment length distributions, genomic endpoint positioning, and sequence context at cleavage sites. Fragmentomic signals reflect biological processes related to nucleosome organization, chromatin accessibility, and cell death, and they have shown promise as complementary biomarkers for noninvasive cancer detection [15].

While most cfDNA-based approaches focus primarily on nuclear DNA, mitochondrial DNA (mtDNA) represents an important and comparatively underexplored source of information. The mitochondrial genome is compact and present in multiple copies per cell, which can make mtDNA-derived reads abundant under certain experimental conditions. Circulating cell-free mtDNA (ccf-mtDNA) has been studied as a biomarker in cancer and other diseases, with evidence suggesting that mtDNA may capture disease-associated alterations through both mutational patterns and fragmentation-related characteristics [12]. Recent work has further indicated that ccf-mtDNA exhibits distinct fragmentomic behavior compared to nuclear cfDNA, including region-dependent fragmentation profiles and cancer-associated deviations that can be leveraged for multi-cancer detection [5]. These findings highlight the potential of mtDNA fragmentomics as a complementary dimension of liquid biopsy analysis.

A common limitation of many existing diagnostic pipelines is the reliance on aggregated, sample-level descriptors (e.g., global fragment length histograms or overall mutation rates). Although aggregation can be effective and computationally convenient, it may obscure informative co-occurrence patterns that exist at the level of individual fragments. In other words, the joint presence of multiple attributes—such as fragment length, genomic location along mtDNA, and mutation-related signals—may contain discriminative structure that is weakened when features are summarized independently or averaged across all fragments. This thesis addresses this limitation by analyzing ccf-mtDNA at the resolution of single fragments and by explicitly modeling the co-occurrence of multiple fragment-level attributes.

The objectives of this thesis are threefold. First, we perform an in-depth characterization of ccf-mtDNA reads and reconstructed fragments, focusing on fragment length, genomic localization, and mutation-related features, and on how these properties interact. Second, we design and implement a multiparametric predictive framework that assigns risk scores to individual mtDNA fragments based on their joint feature representation. Finally, we evaluate the resulting classification performance by comparing oncological patient samples with control samples, and we benchmark the proposed approach against simpler aggregative baselines. By integrating fragment-level co-occurrence signals into a unified model, this work aims to improve the identification of high-risk, cancer-associated mtDNA fragments and to contribute to the broader effort of developing sensitive, noninvasive cancer detection methods.

The thesis is organized as follows. Chapter 1 summarizes the biological and computational background relevant to cfDNA analysis, mitochondrial genomics, and fragmentomics, and reviews related work. Chapter 2 describes the datasets and preprocessing pipeline, including alignment, quality control, and the reconstruction of mtDNA fragments. Chapter 3 introduces the fragment-level feature extraction procedure and presents an exploratory analysis of individual features and their co-occurrence patterns. Chapter 4 presents the proposed multiparametric modeling approach, including baselines, training protocol, and sample-level classification strategy. Chapter 5 reports experimental results and comparative evaluations. Chapter 6 discusses biological interpretation, limitations, and avenues for future work. Finally, Chapter 7 concludes the thesis.

Chapter 1

Background and Related Work

1.1 Liquid Biopsy and Cell-Free DNA

Cancer remains a major global health burden, and clinical outcomes are strongly influenced by the stage at which the disease is diagnosed. Conventional diagnostic workflows frequently rely on tissue biopsy and imaging. While tissue biopsy provides direct access to tumor material, it is invasive, not always feasible for repeated sampling, and may fail to capture spatial heterogeneity across primary and metastatic sites. These limitations have motivated the development of *liquid biopsy*, a minimally invasive strategy that aims to infer disease-related information from analytes circulating in body fluids, most commonly blood plasma.

Among liquid biopsy analytes, *cell-free DNA (cfDNA)* has become one of the most widely studied. cfDNA consists of short DNA fragments released into the bloodstream predominantly through cell death processes such as apoptosis and necrosis. In cancer patients, a fraction of cfDNA is derived from tumor cells and is referred to as circulating tumor DNA (ctDNA). The ability to profile ctDNA enables noninvasive characterization of tumor-associated molecular alterations and supports applications such as therapy selection, treatment monitoring, minimal residual disease detection, and relapse surveillance.

1.2 Mutation-Based Approaches and Their Limitations

Early liquid biopsy assays primarily focused on detecting tumor-specific genetic alterations in cfDNA, including point mutations, small insertions/deletions, structural variants, and copy number changes. Targeted mutation panels and ultra-deep sequencing can identify low-frequency variants, and such approaches have demonstrated clinical utility in multiple tumor types. Nevertheless, mutation-centric testing faces several well-known limitations.

First, ctDNA can be present at extremely low fractions, particularly in early-stage disease, which reduces sensitivity even with deep sequencing. Second, targeted panels are constrained to known loci and may miss tumors without canonical alterations in the targeted regions. Third, biological confounders can generate false-positive signals; for example, somatic variants arising from non-tumor sources may appear in plasma and complicate interpretation. Collectively, these challenges have encouraged the development of complementary signals and computational strategies that do not rely exclusively on mutation detection.

1.3 cfDNA Fragmentomics

A key complementary direction is *cfDNA fragmentomics*, which studies the physical and genomic properties of cfDNA fragments, including fragment length distributions, genomic coverage patterns, and sequence composition at fragment ends. These properties are shaped by chromatin organization and nuclease activity during DNA fragmentation, and they can therefore reflect tissue-of-origin and disease-associated changes.

Genome-wide fragmentation patterns have been shown to differ between cancer patients and healthy controls and can support cancer detection without requiring prior knowledge of specific mutations [2]. In addition to length-based features, fragmentation signals can be derived from positional patterns across the genome and from fragment-end motif frequencies. These features are often high-dimensional and correlated, making machine learning a natural choice for building predictive models. Fragmentomics-based classifiers have achieved strong performance in multi-cancer settings and have been extended to infer tissue-of-origin in some designs [9]. While most fragmentomics work has focused on nuclear cfDNA, the same conceptual framework can be applied to mitochondrial cfDNA, where fragmentation biology differs substantially from nucleosome-driven nuclear fragmentation.

1.4 Mitochondrial DNA and Its Relevance in Cancer

Mitochondrial DNA (mtDNA) is a compact circular genome of approximately 16.6 kb, present in multiple copies per cell. Unlike nuclear DNA, mtDNA is not packaged in nucleosomes, and it is exposed to distinct damage and repair dynamics. Across cancer types, tumors frequently harbor somatic mtDNA alterations, and mtDNA copy number changes have been reported in multiple malignancies. These observations have supported long-standing interest in mtDNA as a potential biomarker, including in the context of liquid biopsy.

A key potential advantage of mtDNA for plasma-based analysis is its copy number: a tumor clone carrying an mtDNA alteration may release many copies of the altered genome, at least in principle, increasing detectability relative to single-copy nuclear variants. However, mtDNA biology also introduces challenges, including heteroplasmy (mixed mutant and wild-type mtDNA within cells) and the presence of background mtDNA variants in normal tissues.

1.5 Circulating Cell-Free mtDNA as a Biomarker

A portion of plasma cfDNA originates from mitochondria and is referred to as circulating cell-free mitochondrial DNA (ccf-mtDNA). Multiple studies have reported that ccf-mtDNA levels and other mtDNA-related signals can be altered in cancer, motivating its investigation as a complementary liquid biopsy biomarker [12]. At the same time, ccf-mtDNA is also influenced by non-cancer processes such as tissue injury and inflammation, which may limit specificity if relying only on absolute concentration measures.

The *mutational* component of ccf-mtDNA has shown mixed results: tumor-specific mtDNA variants can be difficult to detect in plasma in many settings, likely due to dilution by background mtDNA from non-tumor sources and technical limitations at very low allele fractions. For example, tumor-specific mtDNA variants have been reported to be rarely detectable in cfDNA in some cohorts [17]. These findings suggest that mutation-only approaches may be insufficient in general and motivate the exploration of additional mtDNA-derived signals, including fragmentation patterns.

1.6 Fragmentation Profiles of ccf-mtDNA

ccf-mtDNA exhibits fragmentation behavior that differs markedly from nuclear cfDNA. Because mtDNA is not protected by nucleosomes, circulating mtDNA fragments tend to be substantially shorter and may show distinct end-motif composition and region-dependent fragmentation patterns. Studies that improved recovery of short cfDNA molecules demonstrated that very short mtDNA fragments are abundant in plasma and can dominate the mtDNA fragment length distribution [19].

Recent large-scale analyses indicate that cancer is associated with aberrant ccf-mtDNA fragmentomic features and that these features can support accurate multi-cancer detection [5]. Reported signals include shifts in fragment length distributions, altered fragment-end motif frequencies, and genomic localization patterns along the mitochondrial genome. These results provide strong motivation for fragment-level, multiparametric modeling of mtDNA, where multiple fragment attributes are considered jointly rather than in isolation.

1.7 Mutation Detection in ccf-mtDNA

Although plasma mtDNA mutation detection is challenging, it remains an important complementary signal. Detection sensitivity depends on the heteroplasmy level in tumor tissue, the extent of tumor DNA shedding, and the background abundance of wild-type mtDNA in circulation. Evidence suggests that in certain contexts, incorporating mtDNA-derived tumor signals can improve detection; for example, experimental models have shown that plasma mitochondrial tumor DNA can enhance detection performance in glioblastoma settings [8]. Nevertheless, across broader cohorts, the rarity of detectable tumor-specific mtDNA variants in cfDNA highlights the need for robust modeling strategies that leverage both mutational and non-mutational mtDNA features.

1.8 Machine Learning for mtDNA-Based Liquid Biopsy

The diversity and potential co-occurrence of mtDNA-derived signals (e.g., fragment length, genomic position, end motifs, and mutation-related attributes) naturally motivates machine learning approaches. Classical ML models (e.g., logistic regression, random forests, gradient boosting) can integrate multiple features and capture interactions that may be diluted by simple aggregation. Recent work demonstrates that mtDNA fragmentomics, analyzed via predictive modeling, can yield high diagnostic performance in multi-cancer detection tasks [5]. Beyond single-modality models, broader liquid biopsy research increasingly moves toward multi-modal integration, combining fragmentomics with mutations and other signals. Understanding how these paradigms translate to mtDNA is a central motivation for multiparametric, fragment-level modeling.

1.9 Related Work in Nuclear cfDNA Fragmentomics

Nuclear cfDNA fragmentomics provides essential methodological context for mtDNA-based approaches. Genome-wide fragmentation profiles have been used to detect cancer [2], and related studies extended fragmentomic analysis to characterization of tumor-associated fragmentation patterns and cancer classification [9]. Many feature engineering ideas from nuclear fragmentomics (e.g., regional fragmentation summaries, end-motif statistics, learned representations) can be adapted to mtDNA with careful consideration of mtDNA-specific biology and technical confounders (such as NUMTs and the short-fragment bias of library preparation).

1.10 Challenges and Opportunities

Despite strong recent progress, several challenges remain. From a technical perspective, mtDNA constitutes a small fraction of total plasma cfDNA in many settings, and mtDNA fragments are often extremely short, which can lead to biased recovery and reduced effective coverage. Accurate mapping is complicated by nuclear mitochondrial DNA segments (NUMTs), requiring careful alignment and filtering strategies. From a biological perspective, ccf-mtDNA signals are influenced by non-cancer processes (e.g., inflammation), which may reduce specificity unless models leverage more cancer-informative combinations of features.

These challenges also indicate clear opportunities. Fragment-level, multiparametric models can explicitly represent and learn co-occurrence patterns across length, localization, and mutation-related signals, potentially improving robustness relative to single-feature or purely aggregated approaches. In addition, mtDNA's compact genome makes ultra-deep targeted sequencing more feasible, enabling dense fragment-level characterization. Together, these factors motivate the methodological direction of this thesis: exploiting joint mtDNA fragment attributes for cancer-versus-control discrimination through multiparametric modeling.

Chapter 2

Data and Preprocessing

2.1 Datasets

This thesis builds on internal plasma cfDNA sequencing cohorts provided in collaboration with Geneton s.r.o., where mitochondrial reads (chrM) are available for downstream analysis. As an initial step, we conducted preliminary baseline experiments (developed originally as a course project) using only sample-level quality control (QC) summaries and insert-size statistics exported by `QualiMap` [10]. These experiments served two purposes: (i) to validate that mtDNA-derived, fragmentation-related information contains class-discriminative signal within a study, and (ii) to quantify the extent to which such signals generalize across independent cohorts.

We used two independent datasets:

- **PreveLynch**: $n = 976$ samples, including 753 healthy controls and 223 colorectal cancer patients.
- **GenoScan**: $n = 453$ samples, including 380 healthy controls and 73 prostate cancer patients.

Both datasets exhibit substantial class imbalance (approximately 3:1 controls to cases). Importantly, some patients contributed multiple samples collected at different timepoints. This motivates patient-level grouping in model evaluation to avoid data leakage (Section 5.3.1).

2.1.1 Baseline sample-level representation from QualiMap reports

For the preliminary baselines, each sample was represented by a vector of features parsed from `QualiMap BamQC` report outputs [10]. We define the following feature-set families (notation used consistently throughout the thesis):

- $\mathcal{F}_{QC}$ **(aggregated QC summaries; 14 features):** features from `genome_results.txt` including total reads, mapped reads, chrM coverage, error rate, mean mapping quality, GC content, and summary insert-size statistics (mean/median/SD), as well as derived ratios (e.g., mtDNA fraction, coverage coefficient of variation).

- $\mathcal{F}_{Frag}$ **(selected fragmentation descriptors; 26 features):** a compact set of fragmentation-related descriptors computed from `insert_size_histogram.txt` (percentiles, skewness, kurtosis, and proportions of fragments in selected size ranges) and coverage-uniformity summaries derived from `coverage_histogram.txt` (e.g., coverage inequality metrics).

- $\mathcal{F}_{Hist}$ **(full histogram-derived descriptors; 57 features):** an extended representation including the full set of histogram-derived descriptors extracted from insert-size and coverage histograms.

- $\mathcal{F}_{Hybrid}$ **(QC + fragmentation; 36 features):** a hybrid representation that combines $\mathcal{F}_{QC}$ with a curated subset of fragmentation descriptors, aiming to retain the strongest signals while avoiding excessive dimensionality.

These sample-level baselines are intentionally simpler than the fragment-level representation developed later in this thesis; they provide a reference point and highlight potential pitfalls such as batch effects and limited cross-cohort transfer.

2.2 Preprocessing Pipeline

2.3 Addressing Mitochondrial DNA Challenges

Chapter 3

Fragment-level Feature Extraction

3.1 Fragment Definition and Filtering Rules

3.2 Fragmentation Features

3.3 Mutation-related Features

3.4 Co-occurrence and Joint Patterns

Chapter 4

Multiparametric Modeling

4.1 Problem Formulation

4.2 Baseline Methods

4.3 Proposed Multiparametric Model

4.4 Training Protocol

Chapter 5

Experiments and Results

5.1 Experimental Setup and Metrics

5.2 Comparison of Baseline and Multiparametric Models

5.3 Preliminary baseline experiments: sample-level QC and fragmentation features

Before developing fragment-level multiparametric models, we performed preliminary experiments using only sample-level features exported by `QualiMap` [10]. The goal was to establish a baseline for cancer/control discrimination using mtDNA-associated QC and fragmentation summaries, and to assess generalization across independent cohorts. Feature sets are denoted as $\mathcal{F}_{QC}$, $\mathcal{F}_{Frag}$, $\mathcal{F}_{Hist}$, and $\mathcal{F}_{Hybrid}$ (defined in Section 2.1.1).

5.3.1 Evaluation protocol

All baselines were implemented in `scikit-learn` [11] using a pipeline that imputes missing values (median strategy), standardizes features, and trains a classifier. To prevent data leakage when multiple samples originate from the same patient, we used grouped cross-validation (5-fold GroupKFold), where the group identifier corresponds to patient ID. This ensures that all samples from a given patient are assigned exclusively to either training or test folds.

5.3.2 Models and metrics

We evaluated three standard classifiers with class-balancing due to the strong control/-case imbalance:

Table 5.1: Comparison of QualiMap-derived feature sets in preliminary baselines.

Feature set	# features	PreveLynch AUC	GenoScan AUC
$\mathcal{F}_{QC}$ (aggregated QC summaries)	14	0.760	0.630
$\mathcal{F}_{Frag}$ (selected fragmentation descriptors)	26	0.674	0.675
$\mathcal{F}_{Hist}$ (full histogram-derived descriptors)	57	0.664	0.703
$\mathcal{F}_{Hybrid}$ (QC + fragmentation)	36	0.814	0.693

Table 5.2: Performance on PreveLynch using the hybrid feature set $\mathcal{F}_{Hybrid}$ under 5-fold grouped cross-validation.

Model	ROC-AUC	Precision	Recall	F1
Random Forest	0.791 ± 0.019	0.708	0.336	0.453
SVM (RBF)	0.814 ± 0.021	0.484	0.659	0.557
Gradient Boosting	0.809 ± 0.021	0.678	0.440	0.530

- Random Forest (100 trees, max depth 10, balanced class weights),
- Support Vector Machine with RBF kernel (probability estimates enabled, balanced class weights),
- Gradient Boosting (100 estimators, max depth 5).

Performance is reported primarily using ROC-AUC, which is threshold-independent and appropriate under class imbalance.

5.3.3 Within-study performance and feature-set comparison

Table 5.1 summarizes the impact of feature engineering across both datasets. The histogram-heavy representation $\mathcal{F}_{Hist}$ did not consistently improve performance compared to simpler sets, while the hybrid representation $\mathcal{F}_{Hybrid}$ achieved the best within-study results on both cohorts.

For PreveLynch, $\mathcal{F}_{Hybrid}$ improved ROC-AUC by approximately 0.054 (from 0.760 to 0.814), indicating that combining coarse QC summaries with selected fragmentation descriptors captures more predictive signal than either alone.

5.3.4 Feature importance signals

Across both cohorts, several of the strongest predictors were technical QC metrics (e.g., error rate and mean mapping quality), raising concerns that some of the model signal may reflect dataset-specific technical differences rather than biology. The top-ranked features included:

Table 5.3: Cross-study validation results for preliminary baselines using $\mathcal{F}_{Hybrid}$ (train on one cohort, test on the other).

Training → Test	**SVM AUC**	**RF AUC**
PreveLynch → GenoScan	0.436	0.411
GenoScan → PreveLynch	0.382	0.386

Table 5.4: Direction of effect for selected fragmentation-variability features.

Feature	**PreveLynch (Cancer vs Control)**	**GenoScan (Cancer vs Control)**
frag_cv	Lower in cancer	Higher in cancer
frag_std	Lower in cancer	Higher in cancer
frag_pct_gt250	Lower in cancer	Higher in cancer

- **PreveLynch:** mapped_reads_total (10.3%), error_rate (8.5%), total_reads (7.0%), mean_mapping_quality (5.0%), frag_cv (3.5%).

- **GenoScan:** frag_cv (7.7%), mapped_reads_total (6.3%), total_reads (5.7%), mean_mapping_quality (4.7%), error_rate (4.3%).

5.3.5 Cross-study validation

A key stress test for clinical utility is generalization across cohorts. We therefore trained on one dataset and evaluated on the other, using the best-performing within-study representation $\mathcal{F}_{Hybrid}$. As shown in Table 5.3, performance collapsed, with ROC-AUC dropping well below 0.5 (worse than random guessing). This suggests strong dataset shift and/or cohort-specific artifacts.

5.3.6 Opposite-direction effects across cohorts

To understand the cross-study failure, we examined whether key features changed in consistent directions between cancer and control groups. Several fragmentation variability features exhibited *opposite* directions of effect, which can mechanically invert predictions when transferring models:

Together, these results motivate the need for careful confounder control, explicit batch-effect handling, and more robust representations (including fragment-level multiparametric modeling) that can be evaluated under cross-cohort settings.

5.4 Ablation Studies

5.5 Model Interpretation

Chapter 6

Discussion

6.1 Interpretation of Results

6.2 Limitations

6.3 Lessons from preliminary baselines: generalization and confounding

The preliminary baseline experiments based on sample-level `QualiMap` summaries revealed an important pattern: while within-study performance can appear promising (e.g., ROC-AUC ≈ 0.81 on PreveLynch using the hybrid representation $\mathcal{F}_{Hybrid}$), cross-study transfer can fail catastrophically (ROC-AUC < 0.5 when training on one cohort and testing on the other). This gap highlights that high accuracy on a single dataset is not sufficient evidence of clinically meaningful biomarkers.

6.3.1 Batch effects and technical confounders

A concerning finding was that several of the highest-importance predictors were technical QC metrics such as error rate and mean mapping quality. In principle, such metrics should reflect sequencing and alignment quality rather than cancer biology. If they predict labels, it suggests that cohort-specific protocols, processing time, site effects, or other hidden variables may be correlated with case/control status. This is consistent with the observed collapse under cross-study evaluation.

6.3.2 Cancer-type specificity and opposite-direction effects

Another plausible contributor is biological heterogeneity between cancer types. In the baseline analysis, several fragmentation variability features (e.g., frag_cv, frag_std) exhibited opposite directions of effect between the colorectal cancer cohort (Preve-Lynch) and the prostate cancer cohort (GenoScan). If true, such opposite trends can invert predictions in cross-cohort transfer and imply that a universal detector might require either explicit cancer-type conditioning or substantially more diverse multi-site training data.

6.3.3 Implications for the main thesis approach

These preliminary results directly motivate three design principles for the remainder of this thesis:

1. **Leakage-safe evaluation:** patient-level grouping must be enforced during model validation to avoid overestimation of performance.

2. **Robustness to batch effects:** study/site effects should be assessed explicitly, and batch-correction approaches such as empirical Bayes harmonization (ComBat) may be considered when appropriate [4].

3. **Richer, fragment-level representations:** sample-level QC summaries can conflate technical and biological variation; fragment-level multiparametric modeling provides a path toward capturing co-occurrence patterns that may be more biologically grounded and potentially more transferable.

6.4 Future Work

In addition to increasing cohort diversity, future extensions could include explicit domain adaptation (learning study-invariant representations), feature selection based on cross-cohort consistency, and systematic sensitivity analyses that quantify how much predictive signal remains after removing purely technical QC variables.

Chapter 7

Conclusion

Bibliography

[1] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024.

[2] Samir Cristiano, Angela Leal, Jillian Phallen, Justin Fiksel, Vilmos Adleff, Daniel C. Bruhm, Stine Ø. Jensen, Jose E. Medina, Chana Hruban, Jennifer R. White, and et al. Genome-wide cell-free dna fragmentation in patients with cancer. *Nature*, 570:385–389, 2019.

[3] Peiyong Jiang, K. C. Allen Chan, Y. M. Dennis Lo, and et al. Significantly higher plasma mitochondrial dna mutation load in hepatocellular carcinoma patients than in healthy individuals. *Proceedings of the National Academy of Sciences USA*, 112(11):E4285–E4293, 2015.

[4] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[5] Yang Liu, Fan Peng, Siyuan Wang, Huanmin Jiao, Miao Dang, Kaixiang Zhou, Wenjie Guo, Huanqin Zhang, Wenjie Song, Jinliang Xing, and et al. Aberrant fragmentomic features of circulating cell-free mitochondrial dna as novel biomarkers for multi-cancer detection. *EMBO Molecular Medicine*, 16(12):e16316, 2024.

[6] L. Ma et al. Liquid biopsy in cancer: current status, challenges and future prospects. *Signal Transduction and Targeted Therapy*, 2024.

[7] Ming-Lei Ma, Haoran Zhang, Peiyong Jiang, Ting Sin, Hon-Chiu Lam, Sam H. Cheng, and et al. Topologic analysis of plasma mitochondrial dna reveals the coexistence of both linear and circular molecules. *Clinical Chemistry*, 65(9):1161–1170, 2019.

[8] Russell Mair, Florent Mouliere, Charles G. Smith, Divya Chandrananda, Daniel Gale, Fabio Marass, David W. Y. Tsui, Thomas Watkins, Cindy Zhou, James

Morris, and et al. Measurement of plasma cell-free mitochondrial tumor dna improves detection of glioblastoma in patient-derived orthotopic xenograft models. *Cancer Research*, 79(1):220–230, 2019.

[9] Dimitrios Mathios, Jørgen S. Johansen, Samir Cristiano, Jose E. Medina, Jillian Phallen, Kasper R. Larsen, Daniel C. Bruhm, Noushin Niknafs, Lucas Ferreira, Vilmos Adleff, and et al. Detection and characterization of lung cancer using cell-free dna fragmentomes. *Nature Communications*, 12(1):5060, 2021.

[10] Konstantin Okonechnikov, Ana Conesa, and Fernando García-Alcalde. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32(2):292–294, 2016.

[11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] Fan Peng, Siyuan Wang, Zehui Feng, Kaixiang Zhou, Huanqin Zhang, Xu Guo, Jinliang Xing, and Yang Liu. Circulating cell-free mtdna as a new biomarker for cancer detection and management. *Cancer Biology & Medicine*, 21(2):105–110, 2024.

[13] Heena Satam, Kandarp Joshi, Upasana Mangrolia, et al. Next-generation sequencing technology: Current trends and advancements. *Biology*, 12(7):997, 2023.

[14] Ondrej Škorňák. Cancer detection from cell-free dna fragmentation patterns: A machine learning approach on mitochondrial dna, 2026. Course project report (internal).

[15] W. H. Adrian Tsui, Peiyong Jiang, and Y. M. Dennis Lo. Cell-free dna fragmentomics in cancer. *Cancer Cell*, 43(10):1792–1814, 2025.

[16] Ymke van der Pol, Norbert Moldovan, Jip Ramaker, Sanne Bootsma, Kristiaan J. Lenos, Louis Vermeulen, Shahneen Sandhu, Idris Bahce, D. Michiel Pegtel, Milena Čavić, and et al. The landscape of cell-free mitochondrial dna in liquid biopsy for cancer detection. *Genome Biology*, 24(229):1–22, 2023.

[17] Maximilian Weerts, Emiel C. Timmermans, Anja van de Stolpe, Roderick Vossen, Seyed Yahya Anvar, John A. Foekens, Joost B. EM Tuynman, Wilbert P.

Schrauwen, John W. Martens, Stefan Sleijfer, and et al. Tumor-specific mitochondrial dna variants are rarely detected in cell-free dna. *Neoplasia*, 20(7):687–696, 2018.

[18] World Health Organization. Global cancer burden growing, amidst mounting need for services. WHO News release, 2024. Accessed 2026-01-24.

[19] Ruoyu Zhang, Kiichi Nakahira, Xiaoxian Guo, Augustine M. K. Choi, and Zhenglong Gu. Very short mitochondrial dna fragments and heteroplasmy in human plasma. *Scientific Reports*, 6:36097, 2016.

Appendix A

In this appendix, we provide details on the implementation, pipeline parameters, dataset descriptions, and pseudocode.

Pipeline Parameters

Dataset Details

Pseudocode

Appendix B

This appendix contains supplementary results, including additional graphs and tables that support the findings presented in the main text.

Additional Figures

Supplementary Tables