# Computational methods for cell-free DNA based diagnostics

By

Mariyam Siraj

Under the supervision of

Dr Vibhor Kumar

Indraprastha Institute of Information Technology Delhi

July 2025

# Computational methods for cell-free DNA based diagnostics

**By**

**Mariyam Siraj**

**Submitted**

**In partial fulfilment of the requirements for the degree of**

**Master of Technology**

**To**

**Indraprastha Institute of Information Technology Delhi**

**July,2025**

## Certificate

This is to certify that the thesis titled "**Computational Methods for cell-free DNA based diagnostics**" being submitted by **Mariyam Siraj** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

**July,2025**

Dr Vibhor Kumar

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi 110020

## Acknowledgements

I am sincerely grateful to **Dr. Vibhor Kumar**, my thesis supervisor, for his invaluable mentorship and unwavering support throughout this endeavor. His expertise and encouragement played a pivotal role in the successful completion of this thesis. I am also indebted to all the faculty and staff at IIIT Delhi for their consistent assistance, which greatly facilitated the entire process. Thank you all for your guidance and support.

I am indebted to my fellow batchmates for their encouragement and support, and I wish to express particular thanks to **Niharika Dubey** from the Ph.D. 2022 batch for her significant support in this work.

Lastly, heartfelt thanks are extended to my family, whose enduring support and inspiration have been pivotal throughout my academic journey. Their unwavering belief in me has made this achievement possible.

Mariyam Siraj
MT23248

**Abstract**

Cell-free DNA (cfDNA) has emerged as a promising biomarker for non-invasive cancer diagnostics, offering a window into tumor-derived genetic and epigenetic information through simple blood sampling. However, the biological signals embedded in cfDNA, such as end motifs and nucleosome positioning, remain underutilized compared to conventional mutation-based assays. This thesis develops and validates computational frameworks that harness fragmentomic and chromatin features of cfDNA to enhance cancer detection, tissue-of-origin classification, and cross-platform model generalizability.

Firstly, the study demonstrates that short sequence motifs at cfDNA fragment ends, particularly 6-mer patterns encode robust cancer-specific signatures. By applying machine learning classifiers to these motifs, the models achieved high accuracy in distinguishing cancer patients from healthy individuals and exhibited strong performance in differentiating among multiple cancer types. Secondly, to address platform variability and limited sample sizes, a model-based transfer learning strategy was implemented. Using decision tree adaptation techniques (SER and STRUT, with class imbalance-aware variants), models trained on Illumina sequencing platform were successfully transferred to data generated by alternative sequencing platforms, such as Nanopore sequencing, improving cross-domain prediction without retraining from scratch.

Thirdly, this work investigates nucleosome occupancy patterns around transcription factor binding sites as an informative layer for tissue-of-origin inference. A custom computational pipeline quantified nucleosome positioning and chromatin accessibility signatures, which, when integrated with machine learning and feature reduction, provided accurate tumor-type classification and revealed biologically interpretable chromatin features relevant to cancer progression.

Collectively, this thesis advances the field of cfDNA diagnostics by demonstrating that shallow, cost-effective sequencing combined with robust computational pipelines can deliver high diagnostic accuracy, interpretability, and adaptability across sequencing technologies. The developed approaches lay the groundwork for scalable, minimally invasive multi-cancer early detection and tissue-specific monitoring, supporting future implementation of cfDNA-based liquid biopsies in precision oncology.

# Contents

## List of figures

**List of tables**

# Chapter 1

# General introduction and related works

## 1.1 Thesis structure

The thesis is organized as follows:

**Chapter 1:** Introduces the importance of computational methods for the diagnosis of different disorders with the aid of cell-free DNA. Also, related works and their associated drawbacks have been discussed.

**Chapter 2:** The role of cell-free DNA end motif as a diagnostic biomarker in cancer has been discussed.

**Chapter 3:** Transfer learning-based approach has been implemented which adapts on different sequencing platform with scarce sample size.

**Chapter 4:** The role of nucleosome occupancy around transcription factor binding sites as the basis for obtaining the tissue of origin for different cancer types has been discussed.

**Chapter 5:** A discussion on the outcomes obtained from the analyses conducted in the thesis and the future scope of these findings.

## 1.2 Introduction

Cancer continues to be a global health challenge, with increasing incidence and mortality despite advances in therapeutic approaches [1]. In recent years, the analysis of cell-free DNA (cfDNA) has emerged as a transformative strategy in non-invasive diagnostics, particularly in oncology. cfDNA refers to short fragments of nucleic acids released into the bloodstream and other bodily fluids as a consequence of cellular apoptosis, necrosis, or active secretion. Since its discovery by Mandel and Metais in the late 1940s, cfDNA went from being from a biological curiosity to a therapeutically significant biomarker with numerous uses in organ transplant monitoring, cancer, infectious disease diagnosis, and prenatal testing [2]. A subfraction of circulating tumor DNA (ctDNA), that hosts genetic and epigenetic changes unique to tumors, is exploited by the use of cfDNA in cancer diagnostics. cfDNA-based testing, or liquid biopsy, offers several advantages over traditional tissue biopsies [3]. It is minimally invasive, enables repeated sampling, and captures tumor heterogeneity across spatial and temporal dimensions [3]. These characteristics make it particularly suited for early detection, monitoring treatment responses, and identifying resistance mutations [3], [4].

The expanding field of cfDNA diagnostics is being shaped by innovations in computational modelling [1]. Systems biology, probabilistic modeling, and machine learning are some of the analytical methods that have enabled researchers to examine cfDNA data at a level of resolution never before possible [1]. Subtle fragmentation patterns, mutation signatures, and epigenetic

profiles that influence the tissue-of-origin, forecast prognosis, and group patients for targeted treatments can all be found using these models [1]. Such computational frameworks have a huge role in distinguishing between healthy and pathological cfDNA based on features like fragment length distribution and nucleosome positioning [5].

One of the most significant technical challenges in cfDNA diagnostics is the detection of low-frequency ctDNA molecules amidst a background of abundant non-tumor cfDNA [4]. This issue is particularly pressing in early-stage cancers where ctDNA concentration can fall below 0.01% of total cfDNA [4]. To overcome this, advanced sequencing technologies like digital droplet PCR, next-generation sequencing (NGS), and molecular barcoding are complemented by sophisticated error correction algorithms [4]. Song et al. (2022) discuss how these computational approaches enhance sensitivity and specificity, improving diagnostic accuracy [6]. The fragmentation patterns of cfDNA are also of significant diagnostic relevance. Tumor-derived cfDNA is typically shorter (132–145 bp) than cfDNA from normal apoptotic cells (~167 bp), and these differences can be exploited through size selection techniques to enrich for ctDNA [2]. Moreover, cfDNA fragmentation profiles reflect the chromatin landscape of the tissue of origin, providing a unique opportunity for tissue-specific disease detection [2].

Beyond oncology, cfDNA has found applications in prenatal diagnostics, infectious disease monitoring, and even parasitology. In prenatal care, cfDNA derived from the fetus circulates in maternal blood and has enabled non-invasive prenatal diagnosis for conditions like Down syndrome, Rhesus factor incompatibility, and monogenic disorders [7]. These applications have prompted regulatory bodies to consider cfDNA-based tests for routine clinical use. In parasitology, cfDNA detection has enabled non-invasive diagnosis of diseases such as malaria, schistosomiasis, and leishmaniasis using urine and saliva samples [8].

Similarly, in infectious diseases, cfDNA analysis through NGS is revolutionizing pathogen detection. Long et al. (2016) demonstrated that cfDNA sequencing significantly outperforms traditional blood cultures in identifying pathogens responsible for sepsis, thereby improving clinical decision-making in intensive care units [9]. These examples underscore the versatility of cfDNA and the critical role computational methods play in extracting meaningful signals from complex datasets.

Another important domain of cfDNA diagnostics is data pre-processing and standardization. Pre-analytical variables such as blood collection tubes, storage conditions, and centrifugation protocols can significantly impact cfDNA quality and yield [2]. Computational methods are increasingly being employed to normalize these variables and correct for batch effects, ensuring data consistency across clinical sites [6]. Integration of multi-omics data is also becoming central to cfDNA analysis [1]. Combining genomic, epigenomic, transcriptomic, and proteomic signals from cfDNA enhances predictive power and enables a more comprehensive understanding of disease mechanisms [1]. Machine learning models trained on multi-modal datasets can identify minute correlations and give an estimate of disease state and progression.

Ethical considerations and regulatory challenges also accompany the rise of cfDNA diagnostics. Incidental findings, data privacy concerns, and the potential for overdiagnosis necessitate robust bioethical frameworks [6]. Moreover, the transition of cfDNA assays from

research settings to clinical practice requires rigorous validation, standardization, and regulatory oversight [6]. Moreover, the utility of cfDNA in chronic inflammatory and autoimmune conditions, such as systemic lupus erythematosus and inflammatory bowel disease, adds to its diagnostic capabilities [8]. These applications make use of cfDNA's ability to reflect cell turnover rates and tissue damage, offering insight into disease flares and progression. Research also indicates that physiological states such as pregnancy, strenuous exercise, or even psychological stress can affect cfDNA levels [8], emphasizing the importance of computational normalization to accurately interpret biological signals.

Recent advances in sequencing platforms and bioinformatics have enabled real-time, high-throughput analysis of cfDNA [6]. Computational pipelines that include steps for read alignment, variant calling, and noise suppression are essential in producing clinically relevant interpretations [6]. Moreover, the development of cloud-based platforms and AI-driven diagnostic tools promises to provide easy access to cfDNA testing, particularly in resource-limited settings [1]. Importantly, cfDNA analysis has shown great promise in the identification of minimal residual disease post-treatment [1]. In transplantation medicine, cfDNA is being explored as a biomarker of organ rejection. Donor-derived cfDNA (dd-cfDNA) in the recipient's circulation serves as an early indicator of graft injury, often preceding clinical symptoms [10]. This has important implications for post-transplant surveillance and immunosuppressive therapy management [10].

In conclusion, the integration of computational methodologies with cfDNA analysis is reshaping the diagnostic landscape across multiple domains. From cancer to infectious diseases and prenatal testing, cfDNA serves as a versatile biomarker offering minimally invasive, real-time insights into health and disease. As this thesis will explore in depth, computational methods are not merely supportive tools but are foundational to the extraction, interpretation, and clinical translation of cfDNA data.

### 1.3 Related works

Computational methods have become essential for processing cfDNA data due to its low concentration and fragmentation, enabling improved sensitivity and specificity in cancer detection [11]. This chapter explores computational approaches in cfDNA analysis, focusing on cancer diagnostics. It reviews the state-of-the-art in fragmentomics, machine learning models, methylation analysis, and transfer learning. We also incorporate recent findings from studies applying cfDNA analysis to pediatric cancers, nanopore-based methylation profiling, and ultrasensitive detection methods to provide a broader context for the methods used in this thesis.

#### 1.3.1 Computational Methods in cfDNA Diagnostics

Traditional cfDNA analysis pipelines involve sequencing, alignment, error correction, feature extraction, and classification. Wan et al. (2017) emphasized that effective cfDNA diagnostics depend heavily on computational sensitivity, especially for early-stage tumors with low ctDNA fractions [11]. Sharma et al. (2021) identified the challenges of cfDNA methylation analysis, such as signal dilution and batch effects, which require advanced modelling [12] .

In a study conducted by Bedin et al., cfDNA quantification was found to be adequately capable of distinguishing colorectal cancer (CRC) patients from healthy controls and from patients with polyps. Furthermore, the prognosis of CRC patients was poorer for those with higher cfDNA levels, indicating that cfDNA dosage predicts survival [13].

Wan et al. (2019) used an ML classifier trained on cfDNA biomarkers for colorectal cancer detection, achieving high accuracy despite data sparsity [14]. Similarly, Peneder et al. (2021) applied an integrated ML model combining multiple fragmentation metrics in pediatric sarcomas, demonstrating cfDNA's diagnostic value even in tumors with low mutation burdens [15].

#### 1.3.2 Fragmentomics and End Motif Analysis

Fragmentation profiles of cfDNA reflect chromatin organization and nucleosome positioning. These patterns, particularly end motifs, offer diagnostic potential. Hou et al. (2024) systematically evaluated ten distinct cell-free DNA (cfDNA) fragmentation patterns to assess their diagnostic potential for cancer detection and tissue-of-origin identification [16]. Shen et al. (2024) introduced EMIT, a transformer-based deep learning model that improved cancer detection using end motif signals [17].

Peneder et al. (2021) benchmarked cfDNA fragmentation metrics and introduced the LIQUORICE algorithm, designed to detect ctDNA via cancer-specific chromatin signatures [15]. Their work focused on pediatric cancers, notably Ewing sarcoma, showing that end motif-based features remain informative even without recurrent genetic mutations [15]. The existence of tumors was linked to fragment size distributions and short fragment enrichment. All these demonstrate the usefulness of fragmentomics in challenging clinical settings.

### 1.3.3 Transfer Learning and Cross-Platform Generalization

Cross-platform variability poses a major challenge in cfDNA diagnostics. Han et al. (2021) addressed this with AECT, a domain adaptation model for estimating transcription start site coverage from shallow cfDNA data [18]. It preserved performance across sequencing protocols. This challenge is mirrored in Katsman et al. (2022), who used Oxford Nanopore Technologies to perform methylation and fragmentation profiling from shallow cfDNA sequencing [19]. Their work confirmed that cfDNA features, including methylation and fragment size, could be reliably extracted from ONT data [19]. They further demonstrated consistency in tumor fraction estimates between ONT and Illumina WGS platforms, supporting the feasibility of cross-platform models.

### 1.3.4 Methylation-Based Detection and Epigenomics

cfDNA methylation changes provide another diagnostic layer. Unlike mutations, aberrant methylation occurs early in tumorigenesis and is tissue-specific. Li et al. (2018) proposed CancerDetector, a probabilistic model that exploits joint methylation states of adjacent CpGs to detect cancer-derived reads in cfDNA at low sequencing depths [20]. Their approach amplified subtle signals by analyzing read-level methylation signatures, achieving sensitivity at single-read resolution.

Katsman et al. (2022) applied Nanopore sequencing to simultaneously assess methylation and fragmentation features of cfDNA [19]. Their method supported and confirmed methylation's diagnostic potential at shallow coverage. Their results validate nanopore sequencing for multi-feature cfDNA analysis and point toward future real-time diagnostic applications.

### 1.3.5 Non-Cancer Applications and Broader Implications

Beyond oncology, cfDNA has transformative implications in prenatal diagnostics. Wright and Burton (2008) reviewed the use of cell-free fetal nucleic acids (cffNA) for non-invasive prenatal diagnosis (NIPT). They discussed enrichment techniques, such as fragment size selection and molecular assays targeting paternally inherited alleles [7]. Norton et al. (2015) conducted a large study evaluating the clinical performance of cell-free DNA analysis for NIPT of common trisomies, specifically trisomy 21 (Down syndrome), trisomy 18, and trisomy 13 in a population of pregnant women at average risk [21]. Oellerich et al. (2022) reviewed the use of donor-derived cell-free DNA (dd-cfDNA) as a non-invasive biomarker to monitor graft integrity and detect rejection in solid organ transplantation. It can be quantified using methods like droplet digital PCR or next-generation sequencing (NGS), typically targeting donor-specific SNPs [22]. The authors emphasized that fractional dd-cfDNA measurement can be influenced by variations in recipient cfDNA (e.g., due to infection), while absolute quantification offers more reliable interpretation [22]. De Vlaminck et al. (2014) introduced a noninvasive method for detecting heart transplant rejection using donor-derived circulating cell-free DNA. Elevated cfdDNA levels were detected up to five months prior to biopsy-confirmed rejection, highlighting its potential for early detection [23].

*Table 1.1 Related publications in computational methods for cell-free DNA diagnostics*

| Study | Focus Area | Methodology | Cancer Type / Context | Key Contributions |
|---|---|---|---|---|
| **Ju et al. (2024)** | cfDNA Fragmentomics | Analysis of cfDNA end motifs; Machine Learning models | Multi-cancer detection | Developed diagnostic models by analyzing cfDNA fragmentation patterns. |
| **Peneder et al. (2021)** | Fragmentomics in pediatric cancers | WGS + ML classifier (LIQUORICE) | Ewing sarcoma, pediatric sarcomas | Demonstrated diagnostic potential in low-mutation tumors using fragmentation signatures. |
| **Li et al. (2018)** | Methylation-based cfDNA detection | Probabilistic model (CancerDetector) | Liver cancer | Achieved ultrasensitive cancer detection at individual-read level. |
| **Katsman et al. (2022)** | Real-time cfDNA profiling | Nanopore WGS + methylation + fragmentation analysis | Lung adenocarcinoma | Validated shallow sequencing for methylation and CNA detection; platform versatility. |
| **Wright & Burton (2008)** | Non-invasive prenatal testing (NIPT) | Fragment size analysis + PCR | Fetal cfDNA | Laid groundwork for sensitive cfDNA analysis under low-signal conditions. |
| **Mathur et al. (2025)** | ML on cfDNA biomarkers | Machine learning classifier | Colorectal cancer | Demonstrated ML utility in early detection using cfDNA features. |
| **Long et al. (2016)** | Pathogen detection in ICU | NGS + cfDNA from plasma | Sepsis diagnosis | cfDNA-based NGS increased diagnostic yield over blood culture, detecting bacterial and viral infections |
| **Wan et al. (2019)** | ML-based CRC detection | cfDNA WGS + ML models with confounder control | Early-stage colorectal cancer | Achieved AUC of 0.92 in stage I/II CRC, showing strong early detection potential |
| **Tan et al. (2023)** | cfDNA Methylation Profiling | Enzyme-mediated methylation sequencing integrating genome-wide methylation, fragmentation, and CNA characteristics | Early cancer detection | Enhanced performance in blood-based early cancer detection through multimodal epigenetic characterization. |
| **Ji et al. (2023)** | Single-Molecule Methylation Profiling | Nanopore-based single-molecule sequencing generating up to 200 million reads per cfDNA sample | Cancer monitoring | Provided significant improvement over existing methods, allowing for longitudinal monitoring during cancer treatment. |

### 1.4 Datasets used

The information regarding all the datasets (along with their accession ids) used in this thesis is presented in the table below.

***Table 1.2 Datasets used in thesis***

| S. No. | Dataset type | Accession ID | Cancer Types |
|---|---|---|---|
| **1.** | Illumina (WGS) | dbGaP study ID 34536 | Bile duct cancer,breast cancer,colorectal cancer,gastric cancer,lung cancer,ovarian cancer,pancreatic cancer |
| **2.** | Illumina (WGS) | GEO ID: GSE71378 | Ovarian cancer,skin cancer, breast cancer,lung cancer,uterine cancer,colorectal cancer,prostate cancer,head and neck cancer,bladder cancer,liver cancer,kidney cancer,pancreatic cancer |
| **3.** | Illumina (hmC-seal pulldown) | GEO ID: GSE202988 | Bladder cancer,breast cancer,colorectal cancer,kidney cancer,lung cancer,prostate cancer |
| **4.** | Oxford Nanopore (WGS) | GEO ID: GSE185307 | Lung cancer |
| **5.** | Illumina (WGS) | EGA ID: EGAS00001001024 | Hepatitis B, cirrhosis |

# Chapter 2

# End motif as a biomarker for cfDNA based diagnostics

## 2.1 Introduction

Over the last few years, cell-free DNA (cfDNA) has become increasingly recognized as a viable biomarker for non-invasive diagnosis. cfDNA is composed of short DNA fragments that are shed into the circulation through cellular apoptosis and necrosis from tissues including normal somatic cells, tumors, fetal tissues during pregnancy, and transplanted organs [24]. cfDNA is present in the blood as fragmented DNA from apoptotic cells and is increasingly being investigated as an agent for liquid biopsy purposes [25]. The molecular properties of the fragments, including their length, genomic site of cleavage, and sequence properties constitute the foundation of a novel field of investigation known as **fragmentomics** [26]. This field of study has allowed scientists to obtain useful knowledge on the biology of cfDNA and its clinical diagnostic potential. Although significant advancements have been made in its use in the clinic, the biological processes controlling the fragmentation of cfDNA are still poorly understood.

One of the important areas of fragmentomics is the study of plasma DNA end motifs, which are short nucleotide sequences, normally occurring at the 5′ termini of cfDNA fragments [27], [28]. The focus is the site of cleavage, irrespective of where in the genome it occured [27]. Such motifs are thought to be influenced by the activity of nucleases like DNASE1L3, whose function and expression could be modified under various pathophysiologic conditions [29], [30]. Therefore, end motifs could represent both the tissue of origin of cfDNA and underlying biological processes like enzyme dysregulation during disease.

The tissue-specificity of end motifs was additionally confirmed with liver transplantation and pregnancy models [10]. These models suggested that sequence signatures contain tissue-of-origin information since end motif profiles of cfDNA originating from the liver or placenta could be clearly distinguished from those of hematopoietic origin [31]. This has outcomes for numerous uses of end motif analysis, including noninvasive prenatal diagnostics, transplant rejection monitoring, and cancer diagnoses.

Subsequent research has also shown that cfDNA fragmentation is non-random but determined by chromatin accessibility and nucleosome positioning [31]. For example, cfDNA fragments tend to end at preferred ends, which are genomic coordinates preferentially cleaved during cfDNA formation [30]. Such preferred ends are tissue specific and have been found in a number of contexts, such as fetal and liver-specific cfDNA [30]. In hepatocellular carcinoma patients, tumor-specific preferred ends have been used to separate them from controls, indicating that they may act as cancer biomarkers [32]. The composition of cfDNA end motifs also depends upon the action of certain nucleases. Experiments on DNASE1L3 knockout mice have revealed a considerable lowering of the frequency of certain 4-mer end motifs, especially those

beginning with CC, i.e., CCCA, CCTG, and CCAG [30]. This indicates the important role of DNASE1L3 in producing these motifs. The levels of these nucleases may be changed in many diseases, such as cancer, and it results in alterations in cfDNA end motif patterns [33]. This end motif diversity variation could be used as a biomarker for detecting cancer. In addition, cfDNA molecules contain single-stranded 5′ DNA overhangs, which are called jagged ends. The occurrence and properties of these jagged ends are determined by nucleases such as DNASE1 and DNASE1L3 and can reveal further information regarding the tissue of origin and disease status [33].

Complex cfDNA fragmentation patterns can now be deciphered thanks to developments in computer techniques like machine learning models. Even in the early stages, it has been shown that models that consider cfDNA end motif information may correctly distinguish cancer patients from non-cancerous ones [32]. For example, by learning feature representations from cfDNA end motifs, a deep learning model known as EMIT (end-motif inspection via transformer) was able to categorize those with and without cancer with excellent accuracy [32].

## 2.2 Rationale and motivation

The motivation behind this study stems from the pressing need for non-invasive, accurate, and cost-effective biomarkers for early cancer detection and tissue-of-origin classification. Circulating cell-free DNA (cfDNA), released into the bloodstream from apoptotic and necrotic cells, carries structural and sequence signatures including fragmentation patterns and end motifs that reflect the pathological status of the originating tissue. Among these features, short sequence motifs at the ends of cfDNA fragments (end motifs) have emerged as a promising source of diagnostic and prognostic information, especially in oncology.

The primary objective of this study was to determine whether cfDNA end motif profiles specifically 6-mer patterns at fragment end contain sufficient information to distinguish between healthy individuals and cancer patients, and further, to differentiate among various cancer types. In addition, we aimed to evaluate the impact of sequencing depth on model performance by examining classification accuracy across varying numbers of input reads.

## 2.3 Methodology

Three datasets were utilized for the analysis. The first dataset (accession: PRJNA543358; dbGaP ID: phs001827) included samples from six different cancer types (bile duct cancer, breast cancer, colorectal cancer, gastric cancer, lung cancer and ovarian cancer) along with healthy controls. It had high tumor burden. The second dataset (GSE202988) also contained samples from six different cancer types but it had low tumor burden. The third dataset (EBI accession: EGAS00001001024) contained samples from patients with hepatitis B, liver cirrhosis, and healthy controls, and was used as a negative control to demonstrate that high classification accuracy based on end-motif signatures is specifically achievable in distinguishing cancer from healthy samples.

Cell-free DNA end motifs carry signatures that help to distinguish cell-free DNA originating from cancer versus normal tissues. For feature engineering, **k-mer motifs** (i.e., all possible combinations of four nucleotide bases: A, G, C, and T) were used. A k-mer length of 6 was selected after initial exploratory analysis suggested it offered a balance between motif complexity and computational feasibility (out of k=4 and k=8). These were generated using the Cartesian product of the nucleotide set. From each FASTQ file, only the DNA sequence lines (every 4th line starting from the second line in the file) were extracted. Before k-mer extraction, the raw FASTQ files were processed to extract only high-quality sequences. Sequences with ambiguous bases (e.g., N) or reads shorter than the desired k-mer length were excluded. A random shuffle was applied to the sequences to minimize positional bias during sampling. To assess how sequencing depth impacts k-mer distributions and downstream classification, read subsets of increasing size were generated at certain intervals to a maximum of 5 million reads. For each read in a given subset, only the first k-mer (prefix) was extracted from the sequence. If the extracted k-mer matched a valid DNA pattern (i.e., contained only A, G, C, or T with exact length specified), its frequency was counted.

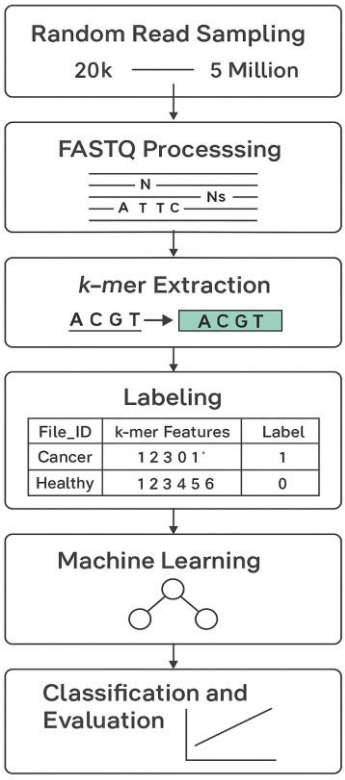


*Figure 2.1 Workflow for end-motif based cfDNA classification*

### 2.3.1 Labeling and Data Output

Each input file was associated with a binary label, representing a specific class (e.g., healthy vs. cancer). After computing k-mer counts for each read subset, the results were stored in CSV format. Each row in the output CSV represented a sample with the following structure:

- **File_ID**: The id of the patient sample file.
- **k-mer features**: Frequency of each 6-mer in the sampled reads.
- **Label**: Class label of the sample (0 or 1).

Separate CSV files were generated for each file and read-depth combination.

### 2.3.2 Machine Learning

To ensure robustness and reproducibility, a **5-fold Stratified Cross-Validation** was implemented using StratifiedKFold. This approach maintains the original class distribution across training and validation sets for each fold, ensuring balanced performance assessment. For each fold the dataset was split into training and validation sets. A **Random Forest classifier** was trained on the training data to compute **feature importances**. The top 50 most important features were selected based on these importance scores. These selected features were then used for model training and evaluation within that fold. All analyses were performed using Python 3.8, with key libraries including scikit-learn, NumPy, and pandas.

### 2.3.3 Classification and Evaluation

A second instance of the Random Forest classifier was trained using only the selected top 50 features from the training data. This model was then evaluated on the corresponding validation data of each fold. Predictions (y_pred) were compared against the true labels using **accuracy** as the primary performance metric. The accuracy for each fold was recorded. This step helped identify the discriminative potential of the k-mer features and demonstrated the robustness of the model across multiple data splits.

## 2.4 Results

Fig. 2.2 (a) and (b) illustrate the distribution of test accuracies across different sequencing depths. A clear trend is observed where the median test accuracy improves with increasing read depth. At lower read counts (20k and 50k), the accuracy is relatively modest and exhibits higher variability, suggesting limited and unstable predictive power. As the number of reads increases, both the median accuracy and consistency improve significantly. the classification accuracy saturates at 2 million reads for the first dataset. These results indicate that increasing sequencing depth enhances the reliability and accuracy of end-motif-based classification models. However, it may be true for samples with high tumor burden.

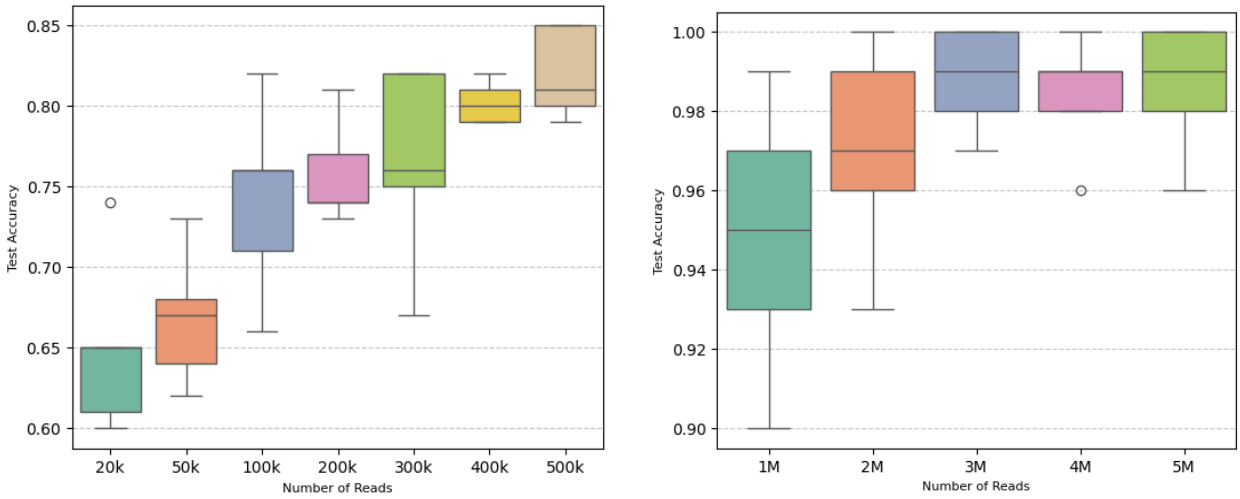


*Figure 2.2 Variation in classification accuracy with increasing read depth for dataset with high tumor burden*

Figure 2.3 (a) and (b) illustrate the classification accuracy achieved using end-motif features at different read depths, ranging from 20,000 to 5 million reads. The blue line represents accuracy obtained using the original (non-randomized) labels, while the red dashed line corresponds to accuracy using label-shuffled data, serving as a baseline. As the number of reads increases, the model's accuracy with original data improves steadily, reaching approximately 95% at 5 million reads. In contrast, the accuracy on shuffled labels remains near random chance (~50%) across all read depths. This indicates that the end-motif signal is biologically meaningful and not a result of overfitting or noise, with predictive power increasing as more sequencing data becomes available.

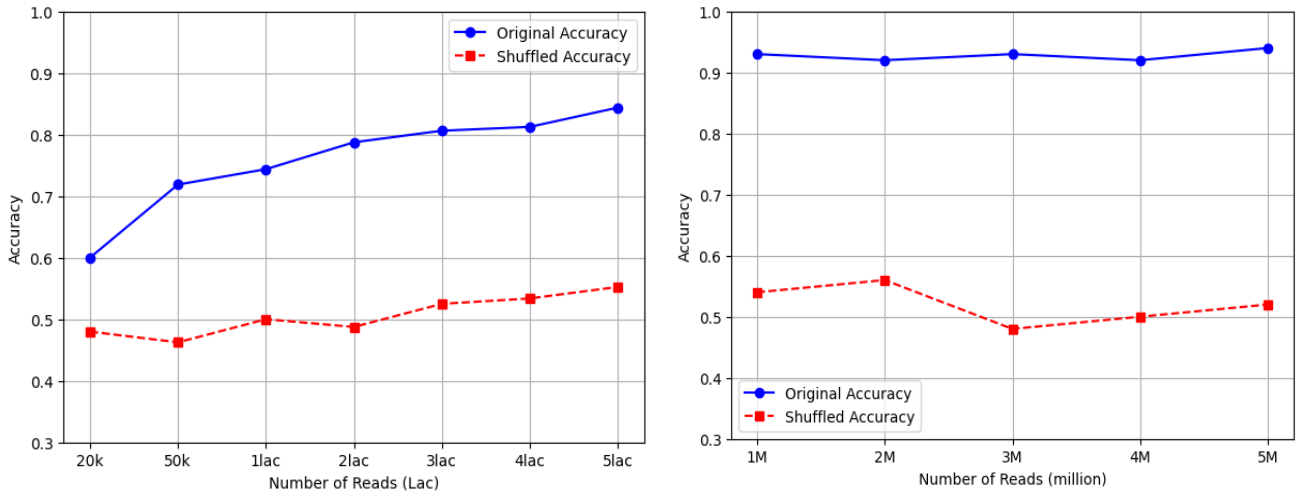


*Figure 2.3 Accuracy comparison between original and shuffled labels across varying sequencing depths.*

Figure 2.4 illustrates that the model achieved a median accuracy of about 0.62–0.63 at 1M–3M reads, with only modest gains as sequencing depth increased to 5M–11M reads. At 7M and 9M reads, median accuracy rose to ~0.75–0.8, but this remained notably lower than the performance seen for high tumor burden samples at comparable depths.

This result indicates that when the fraction of tumor-derived cfDNA is low, the cancer-specific end motif signal becomes diluted by background cfDNA from normal tissues. Consequently, deeper sequencing partly compensates for this reduced signal but does not fully overcome the intrinsic biological limitation. In addition to tumor burden, technical factors in the sequencing process likely influenced the observed results. Differences in library preparation methods, end repair steps, and sequencing chemistries can affect the preservation of natural fragment ends and motifs. Such effects may be negligible for samples with abundant tumor cfDNA but become more impactful when the signal is weak, as in low tumor burden cases. Together, these findings underscore that both biological tumor fraction and sequencing methodology are critical considerations for designing cfDNA-based diagnostic assays using end motif features.

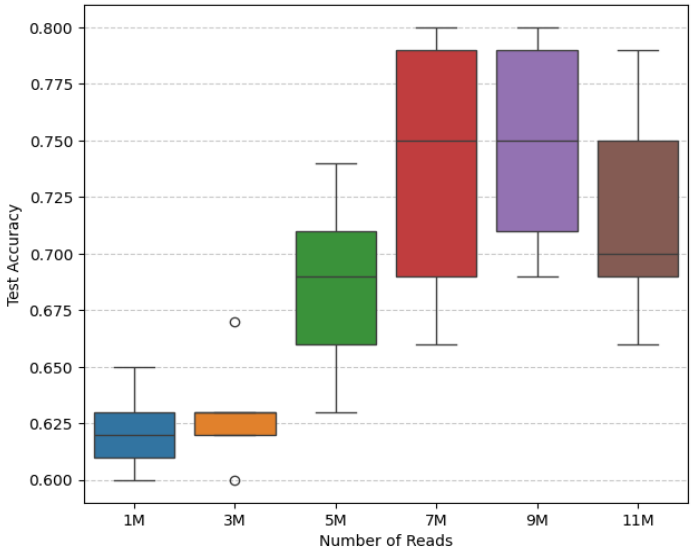


*Figure 2.4 Test accuracy across five-fold cross-validation for samples with low tumor burden across increasing sequencing depths*

Figure 2.5 displays the test accuracy distribution across different sequencing depths for a control dataset consisting of non-cancer conditions: hepatitis B and liver cirrhosis. Unlike the cancer-versus-healthy classification, where accuracy improved consistently with higher read depth, this plot shows limited and relatively flat performance across all read depths. Median test accuracies range from approximately 0.45 to 0.56, with minimal improvement even at higher sequencing depths (up to 5 million reads). The modest accuracy and overlapping distributions across conditions suggest that end motif patterns are less distinctive in non-cancer liver-related diseases, supporting the hypothesis that high classification accuracy using end motifs is specific to cancer versus healthy comparisons.

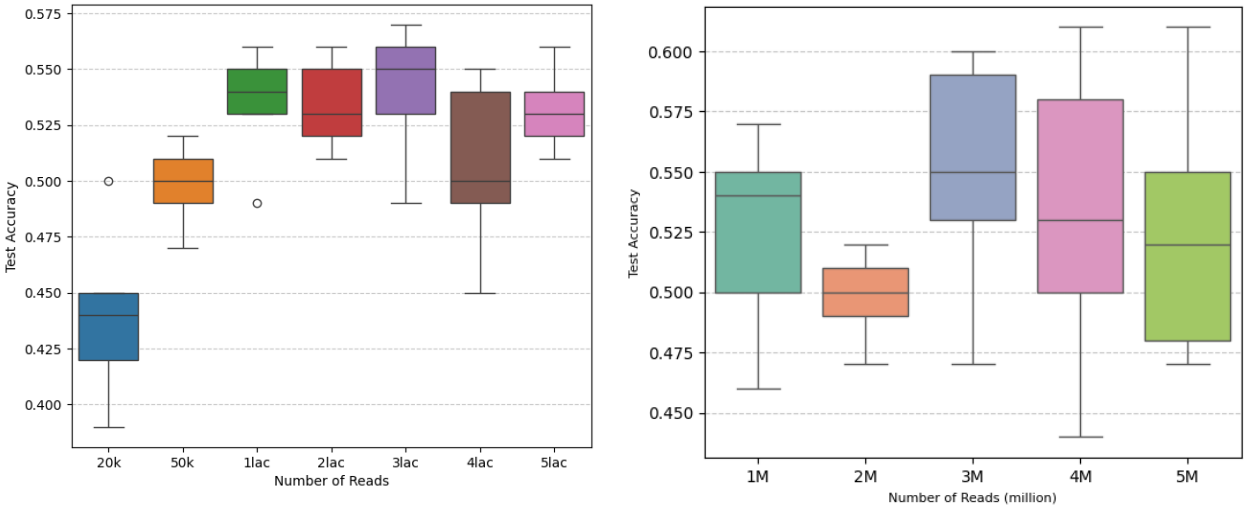


*Figure 2.5 Variation in classification accuracy of end motif classification in non-cancer control*

### 2.4.1 Evaluating end motif signatures for tissue of origin prediction

To evaluate whether end motif signatures derived from cfDNA can distinguish the tissue of origin of different cancers, a binary classification task for each cancer type was conducted. The approach involved comparing one cancer type against all others (one-vs-rest strategy). Two independent datasets with different sets of cancer types were used. Dataset 1 included bileduct, breast, colon, gastric, lung, ovarian, and pancreatic cancer types**.** Dataset 2 included prostate, lung, kidney, colon, breast, and bladder cancer types. **Five-fold cross-validation** was used to ensure robustness and to mitigate overfitting. The test accuracy from each fold was recorded to assess model stability and predictive performance. The classification results, measured by test accuracy across the 5 CV folds, are shown in the box plots below for each dataset. For the first dataset (Fig. 2.6) breast, gastric, and colon cancers showed high classification performance with median test accuracies above 0.85. Lung and pancreatic cancers showed higher variability, suggesting overlapping end motif profiles with other cancer types. Bileduct and ovarian cancers showed moderate accuracy and lower variance.

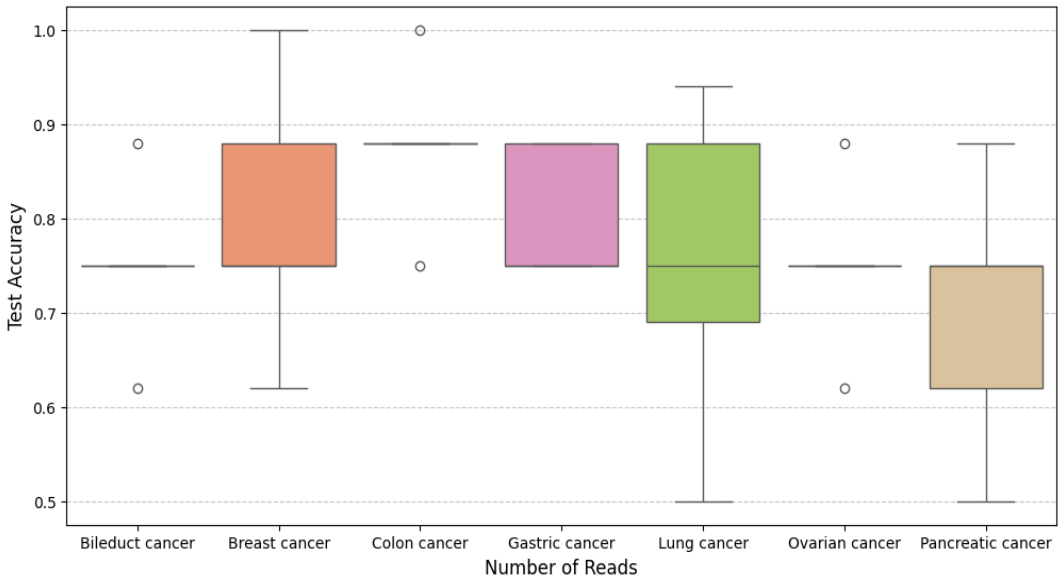


*Figure 2.6 Variation in classification accuracy for each cancer type against all others for dbGaP study ID 34536*

For the second dataset (fig. 2.7) Bladder and lung cancers demonstrated the highest test accuracies, with bladder cancer achieving >0.9 in most folds. Colon and breast cancers had relatively lower and more variable performance. Prostate and kidney cancers showed intermediate results. These results suggest that end motif features can indeed capture cancer-type-specific signals, with certain cancer types (e.g., bladder, breast, gastric) being more distinguishable than others.

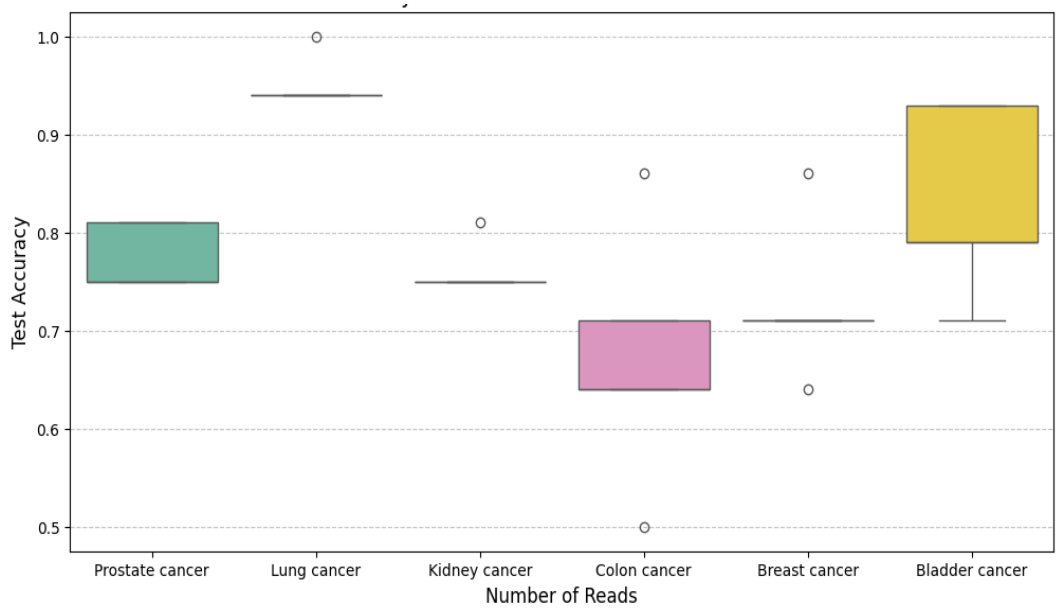


*Figure 2.7 Variation in classification accuracy for each cancer type against all others for dataset GSE202988*

## 2.5 Conclusion

The current research validates the use of cfDNA end motif analysis as an effective method for classifying cancer patients vs. healthy controls through machine learning methods. Through the use of 6-mer features derived from cfDNA fragment ends and their predictive potential on different sequencing depths, we offer strong evidence that end-motif patterns contain bioinformatically significant signals related to cancer pathology. Our results indicated that classification accuracy rose with read depth and leveled off at 4-5 million reads, suggesting that after a point, more sequencing provides diminishing returns for classification performance. In addition, our utilization of a control dataset that included hepatitis B and cirrhosis patients further supported the cancer-specific character of end motif changes.

Our results are consistent with and build upon those of Jiang et al. (2020), who showed that cfDNA end motifs not only mark the tissue of origin but are also disease-state modulated by conditions like cancer, pregnancy, and organ transplantation [34]. The biological explanations for these findings have been increasingly linked to the activity of nucleases, most notably DNASE1L3. Serpas et al. (2019) had provided definitive evidence that mice with Dnase1l3 deletion exhibit abnormal plasma DNA fragmentation. These end motifs were abnormally underrepresented in Dnase1l3-deficient mice, suggesting that DNASE1L3 is an essential enzyme in determining the landscape of plasma DNA end motifs [30].

Additionally, the present study reiterates the idea of tissue-of-origin specificity contained in cfDNA fragmentation profiles. Ding and Lo (2022) discussed how nucleosome positioning and enzymatic DNA processing yield varied fragmentation signatures, such as favored ends, jagged ends, and terminal 5' end motifs [24]. Absence of such motifs in cancer could be a result of a change in tissue contribution to the cfDNA pool, in addition to dysregulated nuclease expression.

A key feature of our approach is the employment of a random forest classifier to reduce dimensions and select features. This method enabled us to determine the most 50 discriminative k-mers, facilitating interpretable and effective modeling. Our application of ensemble-based learning is transparent and minimizes the risk of overfitting, particularly at low sequencing depth. The fact that consistent performance improvements are noted with increasing read depth implies that even with limited sequencing, such classification with significance is possible, which is reassuring for cost-conscious clinical applications.

The addition of a control set comprising non-cancerous liver diseases (cirrhosis and hepatitis B) was critical to confirming the cancer specificity of the end-motif signal. Although in each of these patients there was liver inflammation and injury, the performance of the model to differentiate them from normal samples did not move beyond chance levels, further affirming the finding that the predictive signal is specific to cancer-associated cfDNA fragmentation. Taken together, these findings support the tissue-of-origin specificity and biological relevance of cfDNA end motif patterns. They provide strong rationale for incorporating motif-based classifiers into multi-cancer early detection platforms. Furthermore, assessment of end motif dynamics longitudinally through cancer treatment may provide a minimally invasive biomarker for tracking therapeutic response or disease progression. However, technical aspects of library

preparation and sequencing chemistry may further impact the accuracy of captured end motifs, especially in low-input or low-signal scenarios. Finally, the present findings support the promise of cfDNA end motif analysis as an auspicious biomarker methodology for the detection of cancer. The agreement with previous research and the distinctively clear cancer-specific signal make this strategy both biologically sound and technically viable. Further investigation into enzymatic and structural factors of cfDNA fragmentation will certainly make our knowledge and uses of liquid biopsy in precision oncology more robust.

# Chapter 3

# Transfer Learning

## 3.1 Introduction

Transfer learning has become a key component of machine learning in recent years. It aims to use knowledge from previously acquired tasks or domains to improve performance on new but related problems. This paradigm works particularly well in scenarios where labelled data is hard to come by or quite expensive, which is prevalent in many real-world applications, ranging from tailored treatment to fraud detection. By allowing knowledge transfer from source domains to target domains, transfer learning overcomes the assumption made by traditional machine learning models that training and testing data have the same distribution. This is true even in cases where there are differences in distribution, feature spaces, or task objectives [35]. One of the most interpretable and widely used model families in supervised learning is decision trees and their ensemble extensions, such as random forests. These models are appealing for fields where model transparency is crucial because they provide the benefits of interpretability, efficiency, and resilience. However, traditionally, the advancements in support vector machines (SVMs) and deep learning have outpaced the use of transfer learning in decision tree-based models.

The foundation of transfer learning is the capacity to apply past knowledge from one or more source domains to enhance performance on a target domain. Formally, transfer learning uses knowledge from and, where and/or, to increase the prediction function in a target domain with a task and a source domain with a learning task [36]. There are various levels at which this information transfer might take place, including instance, feature, parameter, and model. Without having direct access to the initial training data, model-based transfer learning occurs through the structural or parametric adaptation of a pretrained model. This is particularly helpful in fields where data sharing is restricted or privacy is a concern [37].

Decision trees (DTs), although conceptually simple, are powerful classifiers known for their interpretability and adaptability. Decision trees can be extended to facilitate transfer learning by changing their decision thresholds or structure in response to a modest amount of target domain data. Two model-transfer methods, Structure Expansion/Reduction (SER) and Structure Transfer (STRUT), are among the first attempts in this field. By leveraging target data to expand misclassified leaves into subtrees and then removing superfluous complexity, SER alters the source tree. On the other hand, STRUT modifies node thresholds to more accurately represent the distribution of target data while maintaining the structure of the source tree [38]. These techniques are effective because they can adapt locally without necessitating a complete retraining. This makes them particularly appropriate for fields like industrial diagnostics, personalized healthcare, or adaptive interfaces where retraining is costly or data is scarce [37]. When source and target tasks are different but connected, this is known as inductive

transfer learning [39]. Tasks stay the same but domains vary in transductive transfer learning. Unsupervised transfer learning occurs when neither domain has labeled data accessible [39].

Negative transfer, which happens when knowledge from the source domain negatively impacts the performance of the target model, is a significant difficulty in transfer learning. Inappropriate model modifications or notable variations in distributions may be the cause of this. Transfer techniques based on decision trees are especially susceptible to these discrepancies, particularly when there is a class imbalance [38]. Data is extremely unbalanced in many real-world situations, particularly in fields like fraud detection or uncommon disease diagnosis. When applied to an unbalanced target dataset, a model trained on a balanced source dataset may perform poorly because it overfits the majority class. SER* and STRUT* are modifications of the SER and STRUT procedures that have been proposed to remedy this. To preserve performance in the face of imbalance, these variations incorporate sampling strategies and cost-sensitive learning. For instance, STRUT* modifies threshold selection to maintain the accuracy of the decision boundary across distributions, whereas SER* employs class-dependent pruning based on leaf error metrics [38]. For transfer cases, ensemble-based techniques such as adaptive boosting and bagging have also been modified. Methods such as hierarchical adaptation and forest fine-tuning are employed in domain-adapted random forests. The Adaptive Random Forest, for instance, which gradually retrains trees on fresh target data streams [40]. Although they are not tree-based, deep transfer learning techniques provide information about cross-domain representation learning that may help develop hybrid tree-deep models. For instance, domain-adversarial training and feature disentanglement are possible approaches to improve decision tree ensembles' flexibility [36].

Decision tree-based transfer learning has been used in a number of industries. In healthcare, modifying diagnostic models for home monitoring devices that have been trained on hospital data. In finance, modifying fraud detection algorithms for various geographical areas or clientele groups. Decision trees' modularity and transparency are advantageous to these applications because they enable domain experts to examine and evaluate the transferred models, which is essential in regulated businesses [36]. Even with encouraging advancements, a number of possibilities are still unexplored. One important area of research is extending decision tree transfer learning to diverse environments with different feature spaces [39].

### 3.2 Rationale and Motivation

Next-generation sequencing (NGS) has revolutionized genomics by enabling large-scale, high-throughput profiling of nucleic acids. Among the various NGS technologies, different platforms and protocols generate data with distinct characteristics in terms of sequencing chemistry, read error profiles, and data distribution. As a result, models trained on data from one sequencing methodology or protocol often fail to generalize well to data obtained from another, even at matched read depths. This study is motivated by the challenge of domain adaptation and limited data in the context of transfer learning. Specifically, to investigate the potential of transfer learning to bridge the performance gap between models trained on Illumina WGS data and their deployment on target domains derived from two biologically and

technically distinct sources: 5-hydroxymethylcytosine (5hmC)-seal pulldown sequencing (different protocol) and Oxford Nanopore sequencing (different sequencing). The datasets were generated by the methodology described in Chapter 2 and consisted of kmer frequency distributions of healthy and cancer samples. The rationale for this work stems from two observations: (a) **Cross-Platform Discrepancy**: Illumina WGS data, particularly when obtained from short-read platforms, exhibits different sequence composition, error models, and base modification profiles compared to Illumina 5hmC-seal and Nanopore sequencing. These domain shifts challenge the generalization ability of conventional machine learning models. (b) **Data Scarcity in Target Domains:** The models were adapted using 6 samples only.

### 3.3 Methodology

This study investigates the application of transfer learning using decision trees to adapt models trained on Illumina whole genome sequencing (WGS) data to two distinct target domains: Illumina 5-hydroxymethylcytosine (5hmC)-seal pulldown sequencing and Nanopore sequencing. We conducted experiments across five sequencing read depths (1M–5M reads), applying both original and class imbalance-adapted transfer algorithms. The methodological pipeline comprises several key stages: data preprocessing, feature selection, model training, domain adaptation using transfer algorithms, and evaluation.

#### 3.3.1 Data Preparation

Source and target datasets were generated using 6-mer frequencies at the end motif sequences as described in Chapter 2. The aim was to classify healthy samples from cancer samples.

#### 3.3.2 Feature Selection

To reduce dimensionality and focus the learning process on the most discriminative features, a random forest classifier was trained on the source dataset. Feature importances were calculated, and the top 50 features were selected based on their importance scores. These features were used consistently across both the source and target domains to ensure comparability. A decision tree classifier was trained using the selected features from the WGS dataset. This base model served as the source model for transfer learning. Four model transfer learning algorithms were employed to adapt the source model to each target domain:

- **SER (Structure Expansion/Reduction)**: Expands and prunes the source decision tree to better fit the target data.
- **SER***: An enhanced version of SER that accounts for class imbalance during pruning.
- **STRUT (Structure Transfer)**: Transfers the structure of the source model and fine-tunes decision thresholds using target data.
- **STRUT***: A refined version of STRUT that includes node pruning and adaptive propagation mechanisms.

These algorithms are implemented as described in Segev et al. (2017) and Minvielle et al. (2019). The adapted models were evaluated on the held-out test set.

### 3.4 Results

Across all read depths, the decision tree classifier trained on the source domain (Illumina) using the topmost informative features consistently achieved perfect training accuracy (1.0), confirming the model's capacity to learn the source data well. However, its performance on the unseen target test sets remained poor or moderate (ranging from 0.5 to 0.576), indicating limited generalization across sequencing platforms despite matched read depth. This pattern exemplifies the negative transfer problem, where a well-trained source model underperforms on the target domain due to domain discrepancies.

For the 5hmc-seal sequencing dataset, transfer learning results conducted for different read depths were as follows:

**One million Reads:** STRUT and STRUT* outperformed all other methods, achieving an accuracy of 0.808, F1 score of 0.800, and AUC of 0.808. SER* delivered strong recall (1.0) but suffered in precision (0.591), resulting in a slightly lower F1 score of 0.743. SER (original) and STRUT (original) either showed low accuracy (SER = 0.385) or less robustness under class imbalance (fig 3.1). These results are consistent with the risk of minority class *leaf loss* identified by Minvielle et al. (2019), where SER without class-dependent pruning can prune meaningful leaves under imbalance.

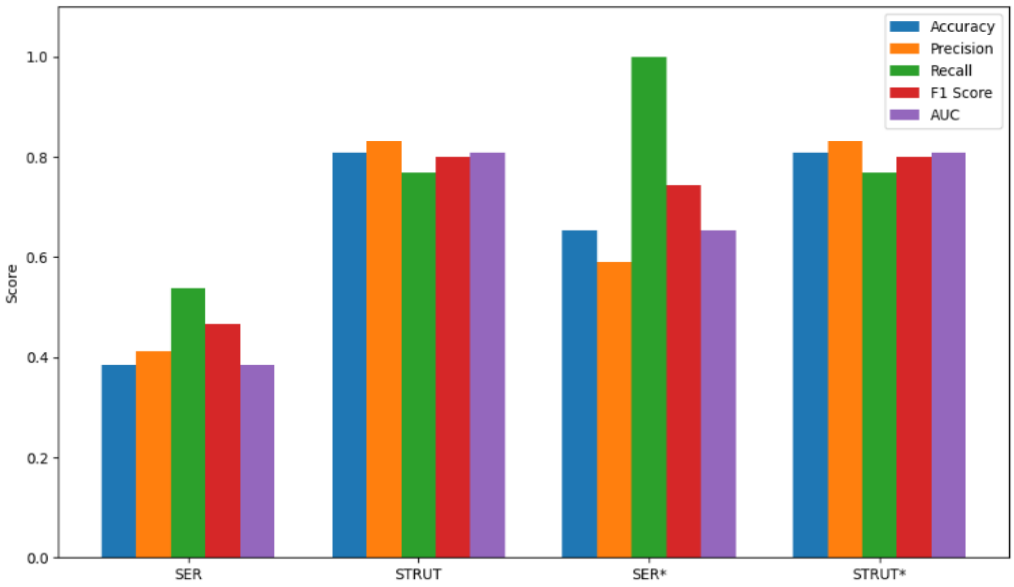


*Figure 3.1 Performance comparison of different variants of Transfer Learning on 1 million reads*

**Two million Reads:** SER* was the best-performing method (Accuracy: 0.731, F1: 0.720, AUC: 0.731), indicating that at this read depth, structure refinement paired with class-preserving pruning worked best. STRUT and STRUT* performed poorly (accuracy 0.500), indicating that divergence gain-based threshold optimization may have misaligned splits in the presence of class imbalance (fig 3.2). SER and SER* outperformed both target-only and source-only baselines, confirming their ability to adapt structure effectively.

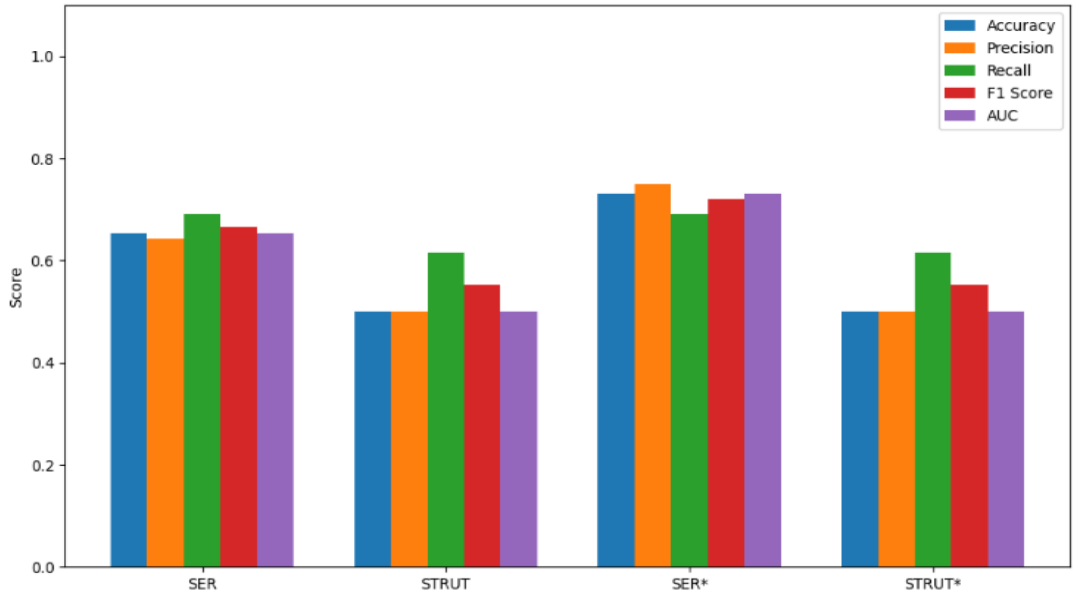


*Figure 3.2 Performance comparison of different variants of Transfer Learning on 2 million reads*

**Three million Reads:** STRUT and STRUT* again topped with accuracy and F1 scores reaching 0.808 and 0.815, respectively. SER* lagged behind (accuracy: 0.385). SER remained consistent with a strong recall (0.923), but weaker precision reduced its overall effectiveness (fig 3.3).

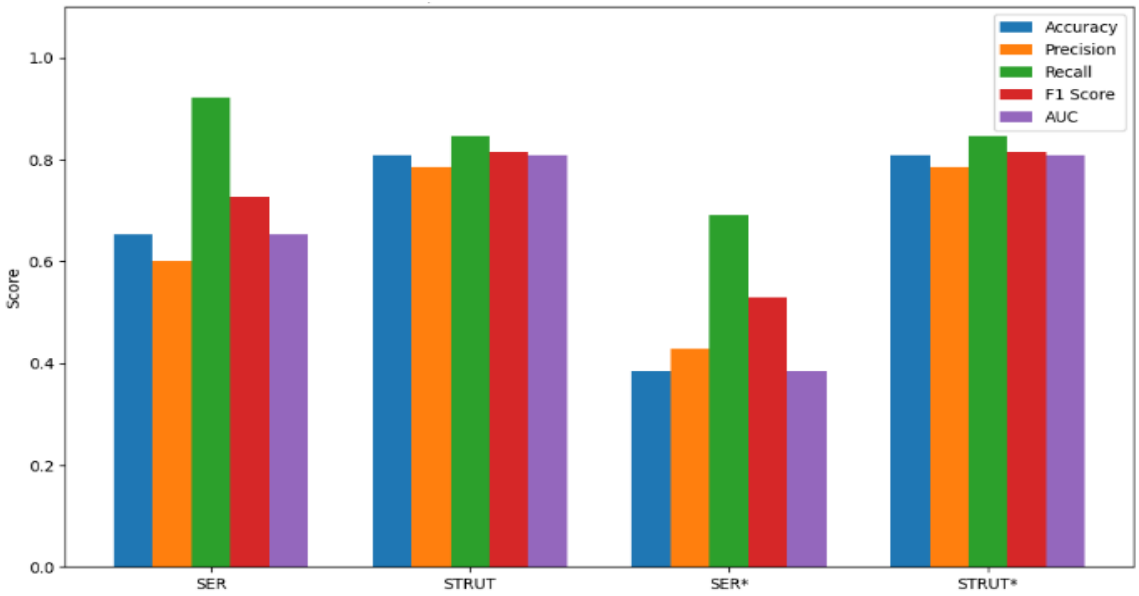


*Figure 3.3 Performance comparison of different variants of Transfer Learning on 3 million reads*

**Four million Reads:** An unexpected inversion occurred. SER outperformed STRUT methods (accuracy: 0.731, AUC: 0.731), likely due to more stable structure modifications. STRUT and STRUT* plummeted in performance (accuracy: 0.269, SER* slightly outperformed its original variant in F1 score (0.516 vs. 0.467), indicating a marginal gain from protecting minority class leaves (fig 3.4).

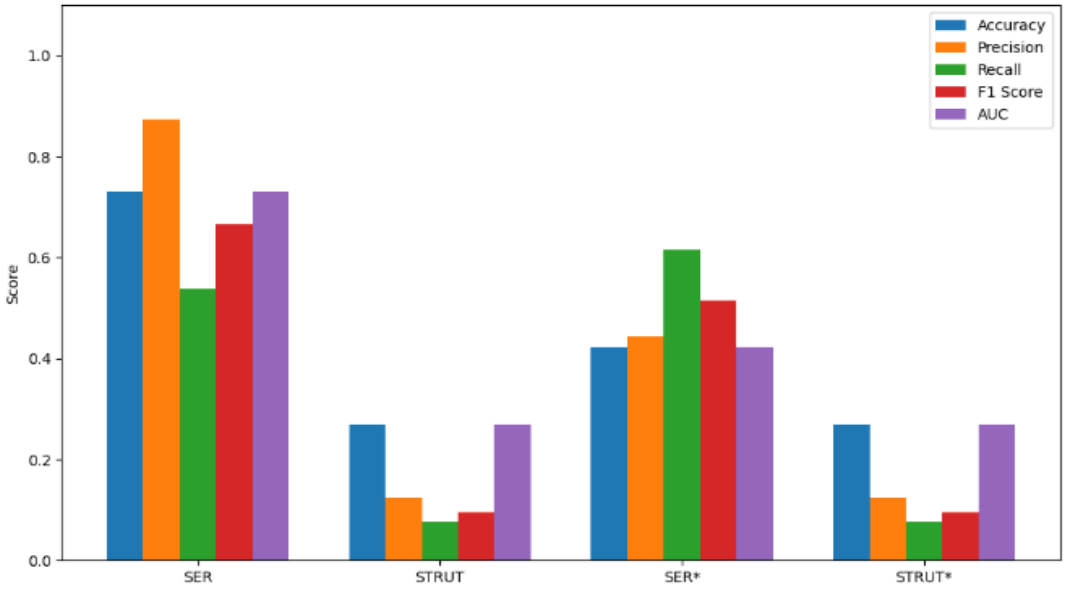


*Figure 3.4 Performance comparison of different variants of Transfer Learning on 4 million reads*

**Five million Reads:** STRUT and STRUT* regained dominance, with both achieving 0.692 across all primary metrics. SER* also performed robustly, especially in recall (1.0) and F1 score (0.743), reaffirming its ability to favor minority class recovery under moderate imbalance (fig 3.5). The improvements in STRUT* highlight the effectiveness of class-ratio-adjusted divergence gain in the threshold selection phase.

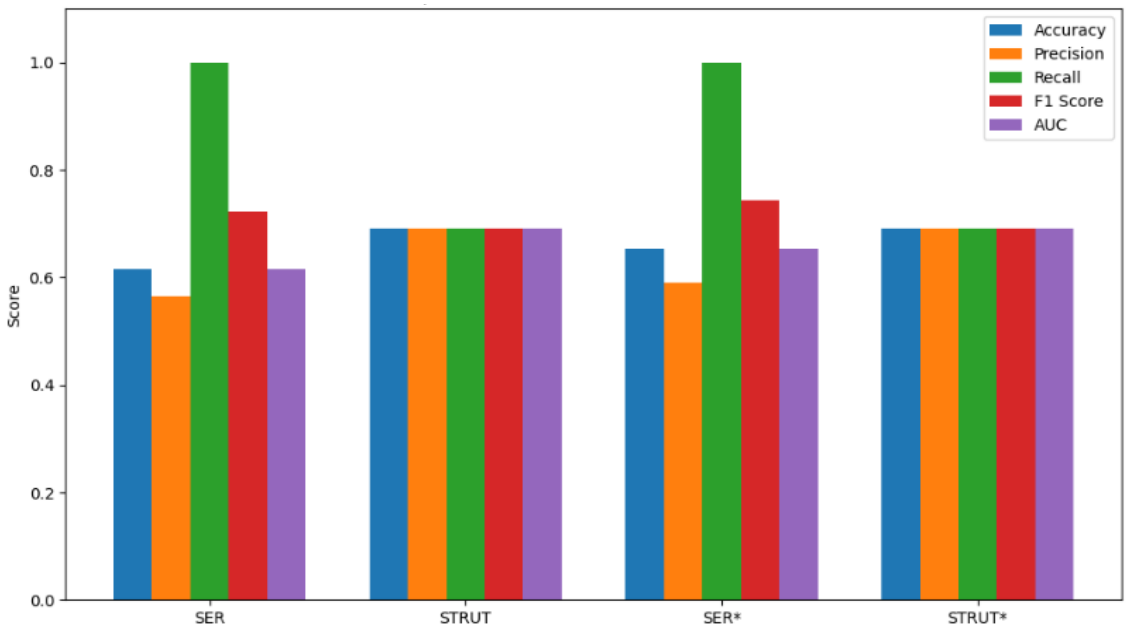


*Figure 3.5 Performance comparison of different variants of Transfer Learning on 5 million reads*

Fig. 3.6 depicts the performance comparison across read depths for the first target dataset (GSE202988).

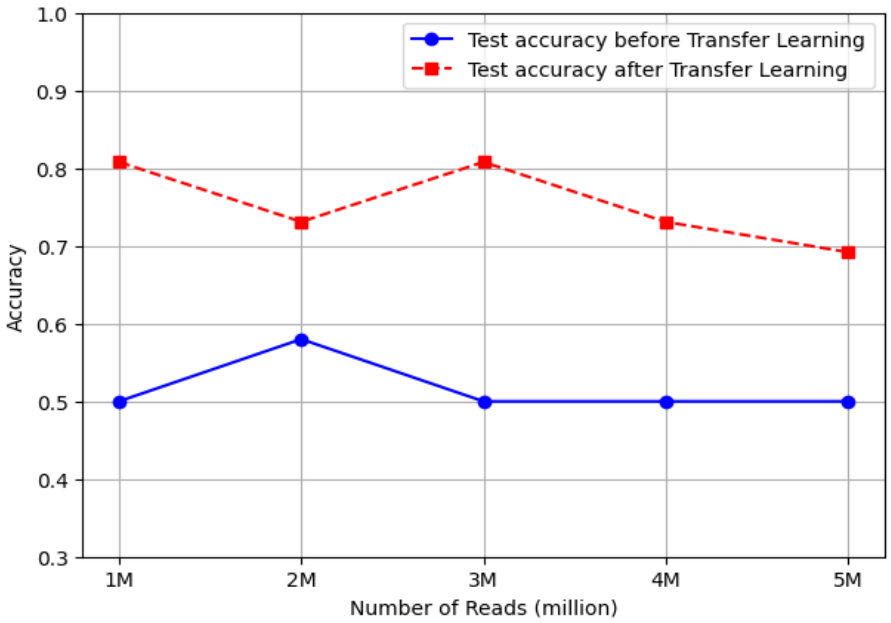


*Figure 3.6 Performance comparison across read depths for dataset GSE202988*

Figure 3.7 shows the results for the nanopore sequencing dataset (GSE185307), where most reads did not exceed one million. Due to data limitations, transfer learning was only conducted at this read depth. SER* outperformed all other methods consistently across all metrics. It achieved the highest scores in Accuracy (0.808), F1 Score (0.800), and AUC (0.808), indicating strong generalization and balanced classification capability even under potential class imbalance. STRUT* closely followed, with Recall matching SER* at 0.769 and an AUC of 0.808. This shows that STRUT*, with its class-adjusted divergence gain, can effectively recalibrate decision thresholds for target domain adaptation. STRUT (original) showed improvement over SER, especially in Recall and F1 score, demonstrating that threshold adjustment was beneficial, though it was less robust than STRUT*. SER (original) lagged behind the other methods in all metrics, achieving an accuracy of only 0.385. This confirms the theoretical concern raised by Minvielle et al. (2019) that pruning strategies in SER may result in minority leaf loss, especially when the available target samples are limited and imbalanced.

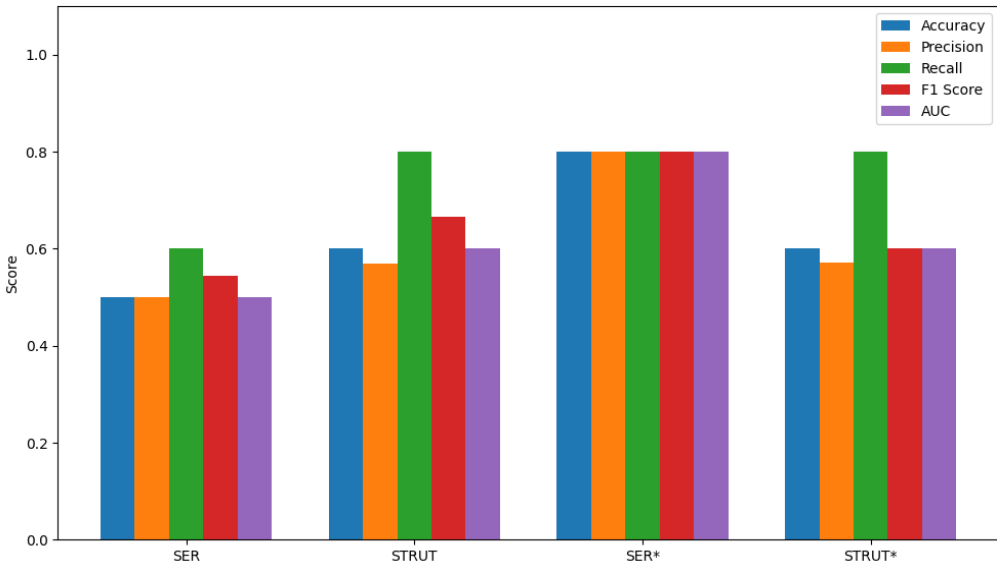


*Figure 3.7 Performance comparison of different variants of Transfer Learning for nanopore dataset*

## 3.5 Conclusion

The objective of this study was to evaluate the effectiveness of model-based transfer learning using decision trees in adapting classifiers trained on Illumina whole genome sequencing (WGS) data to target domains derived from biologically and technically distinct sequencing protocols, specifically Illumina 5hmC-seal pulldown and Nanopore sequencing at matched read depths. To achieve this, we employed four transfer learning algorithms: SER, STRUT, and their respective class imbalance-aware variants SER* and STRUT*. The findings of this study provide valuable insights into the performance and limitations of these methods under constrained, real-world conditions. Across all tested configurations and read depths, it was observed that naive direct application of a decision tree trained on the Illumina platform to target domains resulted in poor test performance, with accuracy often close to random guessing (≈0.5). This outcome clearly illustrates the classic domain shift problem and reaffirms the necessity of domain adaptation techniques in cross-platform genomic analysis. Transfer learning methods, particularly the adapted variants SER* and STRUT*, consistently improved classification metrics in comparison to the base model. This supports the argument that localized adaptation of either tree structure (SER) or decision thresholds (STRUT) allows decision trees to generalize more effectively in the presence of feature distribution discrepancies, class imbalance, and limited target data.

At **lower read depths (1–2M)**, structure refinement through SER* was most beneficial. This suggests that coarse-to-fine restructuring of decision boundaries is effective when target data is sparse and imbalanced. At **moderate to higher depths (3–5M)**, STRUT and STRUT* emerged as more robust performers. These algorithms emphasize threshold optimization rather than structural changes, which may be better suited when more target data is available to refine splits without overfitting. These results align with theoretical expectations. SER*, with its class-dependent pruning, mitigates the **minority leaf loss risk** described by Minvielle et al. (2019), which is more enhanced when labeled target data is limited [38]. By modifying

threshold splits with a class-ratio-aware divergence gain, STRUT*, on the other hand, exhibits robustness and maintains decision boundary accuracy as data availability increases. This study was limited to homogeneous feature spaces, i.e., both source and target domains shared identical features. In real-world scenarios, feature space mismatches are common (e.g., due to missing values, platform-specific markers, or epigenetic annotations), requiring extensions into **heterogeneous transfer learning** [36], [39]. Despite this, the results strongly support the deployment of transfer learning in genomics. In particular, class imbalance-aware model transfer algorithms such as SER* and STRUT* provide interpretable, robust frameworks for adapting decision tree models across domains with minimal retraining. This has important implications for clinical diagnostics, where training data is often well-structured but deployment settings are noisy, imbalanced, and rapidly evolving.

# Chapter 4

# Nucleosome occupancy

## 4.1 Introduction

Within eukaryotic cells, the genome is compacted into a dynamic structure called chromatin, whose fundamental repeating unit is the **nucleosome**. Each nucleosome comprises approximately 147 base pairs of DNA coiled around a core of histone proteins (H2A, H2B, H3, and H4), separated by variable-length stretches of linker DNA [41]. While nucleosomes serve to organize DNA spatially within the nucleus, their positioning also plays a regulatory role in many cellular functions such as gene transcription, DNA replication, and repair. A key aspect of this regulation is **nucleosome occupancy**, that is, the degree to which specific regions of the genome are covered by nucleosomes. Changes in nucleosome positioning can influence gene accessibility, impacting which genes are active or silenced in different physiological or pathological conditions.

Gene promoters, in particular, are influenced heavily by nucleosome arrangement. Promoters with low nucleosome density, often termed **nucleosome-depleted regions (NDRs)** are typically associated with active gene expression, whereas high occupancy near transcription start sites (TSSs) can inhibit transcription initiation [41]. Epigenetic elements, including DNA methylation, histone modifications, and chromatin remodeling complexes, work together to establish and maintain these nucleosome landscapes. Disruption of this balance is a common feature in cancer, where aberrant chromatin remodeling and epigenetic reprogramming lead to gene misregulation and uncontrolled cell growth.

A significant portion of cfDNA fragments corresponds to nucleosome-protected DNA segments, typically around 167 base pairs in length [42]. It is a footprint left by the histone-DNA interaction [42]. When tumors release DNA into circulation, the patterns of nucleosome protection are preserved, offering a snapshot of the chromatin state within the tumor [42].

This insight has led to the development of **nucleosome profiling** from cfDNA, particularly in the context of circulating tumor DNA (ctDNA), the tumor-derived fraction of cfDNA. Advanced analytical approaches now allow researchers to extract information about nucleosome positioning and occupancy directly from cfDNA sequencing data. These patterns correlate with the epigenetic status of tumor tissue, making cfDNA a powerful marker for studying transcriptional regulation and chromatin structure in vivo without invasive procedures [42].

In a recent study, De Sarkar and colleagues sequenced whole-genome cfDNA from mouse models of prostate cancer and identified nucleosome signatures at transcription start sites and transcription factor binding sites (TFBSs) that were specific to tumor phenotypes [42]. By applying machine learning to these features, they successfully predicted the key transcription factors and mixed subtypes from ctDNA samples with high accuracy. Their findings

demonstrate the potential of nucleosome occupancy patterns in cfDNA as a predictor for tumor gene regulation and chromatin accessibility.

Beyond classification, nucleosome positioning also has implications for genome integrity. Research has shown that regions of high nucleosome occupancy tend to accumulate more mutations over time, particularly when DNA repair pathways are intact [43]. A study by Yazdi et al. (2015) revealed a strong correlation between increased nucleosome occupancy and elevated mutation rates, including both germline and somatic mutations [43]. This is likely due to restricted access of repair enzymes to DNA wrapped tightly around histones, reducing the efficiency of repair mechanisms. This pattern's absence in systems with altered genetic repair machinery further emphasizes the significant function of nucleosomes in managing exposure to mutations. [43].

Mishra et al. (2023) introduced *scEpiSearch*, a novel tool designed to match single-cell open chromatin profiles to extensive reference datasets of single-cell transcriptomes and epigenomes. The method significantly improves the annotation and classification of single-cell data across various platforms and species. By applying scEpiSearch to leukemia samples and embryonic stem cells (ESCs), the authors identified distinct regulatory features, including heterogeneity in leukemia cell states and stress-response readiness in ESC subpopulations. This highlights the tool's potential in uncovering functional states and lineage potentials from epigenomic landscapes at single-cell resolution [44].

Chandra et al. (2023) developed *GFPredict*, a machine learning framework that predicts gene functions—both coding and non-coding—using transcription factor (TF) binding, histone marks, DNase hypersensitivity, and CAGE-tag signals at gene promoters. The model achieved high predictive accuracy (AUROC > 0.9 for many gene sets), with TF binding patterns being especially informative. By clustering functions based on shared TF predictors, they linked gene sets to major cellular processes like the cell cycle and immune response. Their predictions were validated using CRISPR screens and PubMed mining, demonstrating the method's utility in inferring functions of poorly characterized non-coding RNAs [45].

Peng et al. (2023) discovered that pioneer transcription factors involved in cell differentiation exhibit higher nucleosome-region enrichment. Hence it supported their role in driving lineage-specific gene expression programs [46]. Furthermore, they showed that pioneer activity is often cell-type-specific and can act as master regulators of chromatin accessibility [46]. Hence, they may serve as critical determinants in cancer and developmental reprogramming.

Genome-wide analysis has shown that changes in nucleosome placement take place early in the development of malignancies including lung and colorectal adenocarcinomas. Druliner et al. (2015) used a method known as MNase-TSS sequencing to show that nucleosomes were widely redistributed around TSSs in early-stage carcinomas [47]. They found that these chromatin changes were similar in all tumor types and mostly determined by their DNA sequence [47]. Hence, it raises the possibility that they could be used as early diagnostic indicators. Moreover, these structural shifts made transcription factor binding sites more accessible, implying that nucleosome alterations could precede and support the development of cancer. [47].

Researchers can determine the tumor's identification, chromatin state, transcriptional activity, and even resistance and mutational tendency by deciphering nucleosome patterns from cfDNA. This marks a substantial advancement in liquid biopsy technology, providing physicians with a more comprehensive and enlightening toolkit for therapeutic decision-making and disease monitoring. The possibility of nucleosome occupancy as a biomarker for cfDNA-based cancer detection is examined in this chapter.

### 4.2 Rationale and Motivation

Cancer is fundamentally a disease of dysregulated gene expression, with alterations in transcriptional regulation and chromatin architecture occurring at the core. Among the key drivers of these alterations are transcription factors (TFs). TFs are proteins that orchestrate gene expression by binding to specific DNA motifs and recruiting or evicting chromatin-remodeling complexes, thereby influencing whether a gene is turned on or off. Because nucleosomes (DNA wrapped around histone proteins) physically block or expose these binding motifs, the local arrangement of nucleosomes around a TF's binding site directly affects TF access and function. In cancer, changes in TF abundance or activity frequently manifest as shifts in nucleosome positioning, regions that should be open become inaccessible, and vice versa. These cancer-associated nucleosome occupancy signatures, when centered on TF binding sites, can therefore act as highly informative indicators of oncogenic disruption.

By integrating data sources that mark TF binding (e.g., motif matches, ChIP-seq peaks) and regions of open chromatin, we target the genomic loci most likely to exhibit disease-specific chromatin remodeling. For instance, a TF that drives proliferation in pancreatic cancer may leave a distinct footprint of phased nucleosomes around its binding sites in tumor cells, whereas in healthy pancreatic tissue or in other cancers, that footprint is absent or altered. Analyzing nucleosome patterns across these defined TF-centered regions enhances sensitivity to tumor-specific chromatin changes, enabling us to distinguish one cancer type from another.

However, genome-wide and even TF-centered analyses can produce thousands of potential features that can lead to nucleosome occupancy values at many genomic positions across numerous TFs. Machine learning, and specifically Random Forests, excels at handling high-dimensional data. By first training a model on all available features, then retraining on only the top 10 most informative features, we achieve two critical objectives that are noise reduction and interpretability and potential biomarkers. High-throughput chromatin assays capture a vast amount of background signal. Retraining on the top 10 features filters out noisy loci, focusing on those few TF sites whose nucleosome patterns most powerfully differentiate one cancer type from others. This elimination of irrelevant dimensions not only preserves (or often improves) classification accuracy but also yields a concise set of candidate biomarkers. The features that rise to the top of the importance ranking are almost always TF binding sites or chromatin regions that undergo the most pronounced cancer-specific remodeling.

Taken together, this multi-tiered strategy, identifying TF binding sites, measuring nucleosome occupancy around those sites, and then employing Random Forests to obtain the highest-value

features might produce a powerful framework for cancer classification and tissue of origin achievement.

## 4.3 Methodology

To carry out nucleosome occupancy-based classification of different cancer types, two different approaches was applied. The first approach was a genome-wide analysis, which is described as follows. To identify transcription factors (TFs) potentially involved in cancer-specific regulatory mechanisms, position frequency matrix (PFM) files for a wide range of human TFs from the JASPAR database were downloaded. A curated list of TFs that are highly expressed across various cancer types was compiled through an extensive literature review. These PFM files were then processed using the MOODS (Motif Occurrence Detection Suite) motif scanning suite to convert them into BED format, mapping their binding sites across the genome. To analyze the nucleosome occupancy signals around these binding sites, I used a custom C-based program. This program takes two inputs: the transcription factor bed files and bed files of patient samples from different cancer types. It calculates the average signal intensity across genomic regions flanking each TF binding site, providing a positional signal distribution that reflects nucleosome presence or absence. The tool uses a kernel-based windowing system with high-resolution binning (in this case, 5 base pairs within a 200 base pair window). This method enabled a comprehensive and comparative analysis of nucleosome patterns around transcription factor binding sites across multiple cancer genomes, contributing to the understanding of chromatin-level regulation in cancer. The second approach involved cancer specific ChIP-Seq data (ChIP Atlas) and Open Chromatin Sites (TCGA).

### 4.3.1 Frequency-Based Feature Extraction Using Fast Fourier Transform (FFT)

Following the generation of transcription factor (TF)-specific signal profiles, each patient sample was represented by a text file where each row corresponded to the smoothed signal of a single TF across genomic regions. To extract quantitative features reflective of nucleosome periodicity, a frequency-domain transformation approach was employed using the Fast Fourier Transform (FFT). An algorithm that calculates a sequence's discrete Fourier transform (DFT) or its inverse (IDFT) is called a fast Fourier transform (FFT). A signal's original domain, which is frequently time or space, can be transformed into an equivalent signal in the frequency domain and vice versa using a Fourier transform.

Figure 4.1(a) shows a TF-specific signal profile generated by aggregating nucleosome occupancy signals across genomic loci, smoothed to highlight periodic patterns. The x-axis denotes base pair positions, while the y-axis reflects the level of nucleosome occupancy. The central peak suggests high nucleosome density at the core TF binding site, with flanking regions displaying periodic oscillations that may reflect phased nucleosome arrangements.

Figure 4.1(b) displays the magnitude of frequency components derived from the FFT of the smoothed nucleosome occupancy signal. The x-axis represents the frequency components, while the y-axis denotes their corresponding magnitudes. The presence of distinct peaks, especially among the lower frequency components, indicates periodic structures within the

spatial signal, suggestive of nucleosome phasing. This transformation enables the extraction of quantitative features corresponding to nucleosome periodicity.

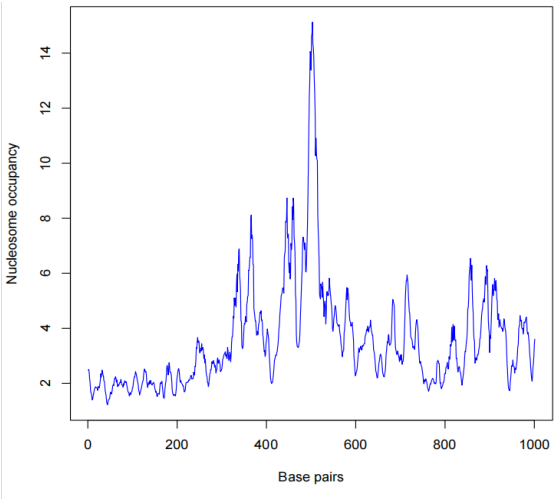


Figure 4.1 (a)

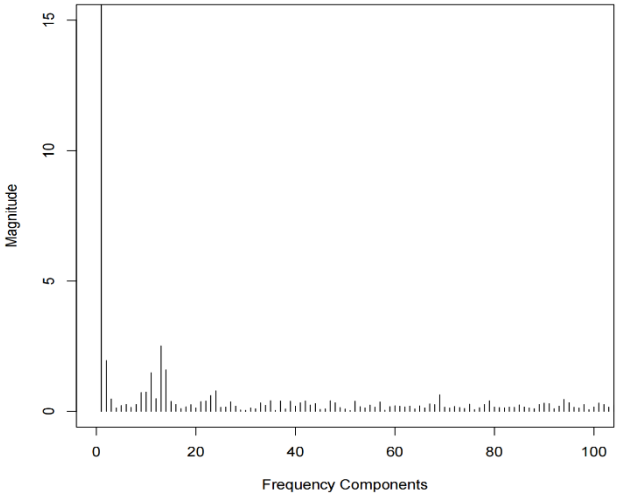


Figure 4.1 (b)

*Figure 4.1 (a) Smoothed Nucleosome Occupancy Signal Around Genomic Regions. (b) Frequency-Domain Representation of Nucleosome Signal via FFT*

For each TF signal, an FFT was applied to transform the signal from the spatial domain to the frequency domain. The magnitude of the FFT coefficients was calculated, and a nucleosome enrichment score was computed as the ratio of the mean magnitude within frequency bins where nucleosome occupancy is expected to be high divided by the background signal. These specific bins were chosen based on their relevance to nucleosome phasing patterns, as periodicities corresponding to nucleosome spacing (~147 bp) typically appear within this frequency range. This process was repeated for all transcription factors in each sample file. The enrichment scores for each TF were compiled into a matrix for each sample, and subsequently, all individual matrices were merged into a single unified matrix. Then classification using Random Forest was carried out.

After initial training, feature importance scores were extracted. The top 10 most informative features were selected, and the model was retrained using only these features. Performance metrics were recalculated to compare model efficacy with reduced feature sets.

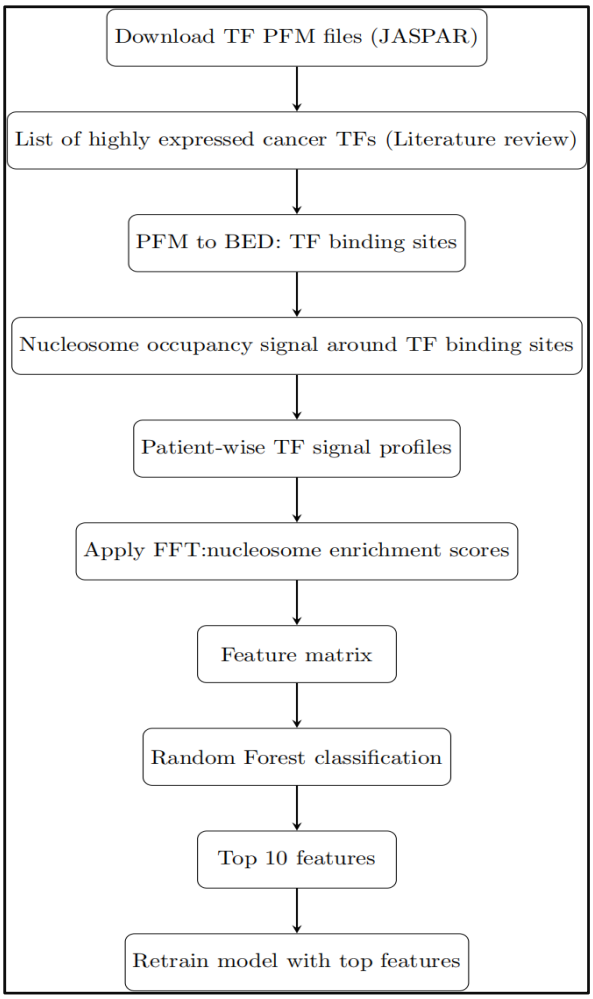


*Figure 4.2 Workflow for nucleosome occupancy-based classification*

## 4.4 Results

For each strategy, the results are reported first on the full feature set and then on the top 10 most important features (as determined by the initial model). Accuracy, precision, recall, F1-score, and AUC-ROC are computed for seven cancer types: bile duct, breast, colon, gastric, lung, ovarian, and pancreatic cancer. Figures 1 and 2 display summary bar plots of accuracy and AUC before and after retraining on the top 10 features for the genome-wide approach.

### 4.4.1 Genome-wide TF Occupancy Features

#### 4.4.1.1 By using all features

When using the full set of genome-wide nucleosome occupancy features (i.e., aggregated and smoothed signals around every TF binding site), classifier performance varied by cancer type (Table 4.1). Overall accuracies ranged from 0.67 (bile duct and breast cancer) to 0.78 (pancreatic cancer). Notably, colon and pancreatic cancers exhibited the highest AUC values (0.94 and 0.81, respectively), indicating that genome-wide TF occupancy alone captures strong phasing signals sufficient to distinguish these tumor types. In contrast, bile duct and breast

cancers had lower AUCs (0.64 and 0.78), suggesting that TF occupancy patterns alone are somewhat less distinctive for these tissues.

#### 4.4.1.2 Using the top 10 features

After ranking feature importance from the full model and selecting the top 10 TF–occupancy features per cancer, we retrained the classifier (Table 4.2). Across nearly all cancer types, there was either stable or improved performance. In summary, retraining on only the top 10 genome-wide features yielded comparable or better accuracy for six of seven cancer types. The largest gains in AUC occurred for bile duct ($\Delta = +0.11$) and ovarian ($\Delta = +0.03$) cancers. Figure 4.2 illustrates these accuracy improvements, and Figure 4.3 shows corresponding AUC gains (green bars = full feature set; orange bars = top 10 features).

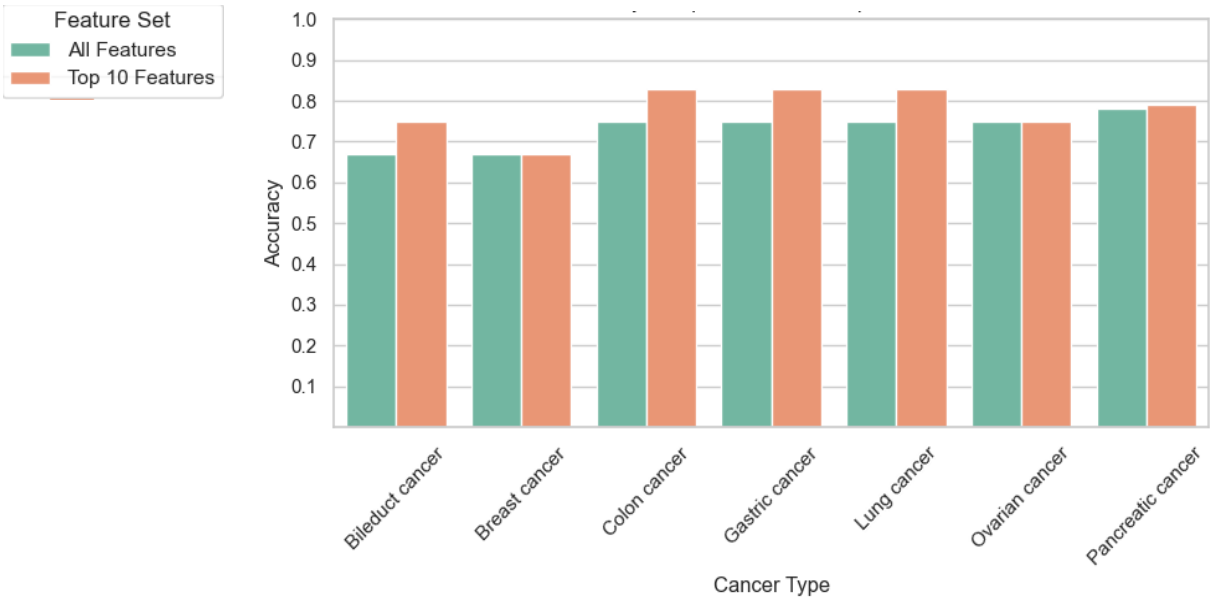


*Figure 4.3 Accuracy comparison using all features genome wide versus top 10 features across different cancer types*

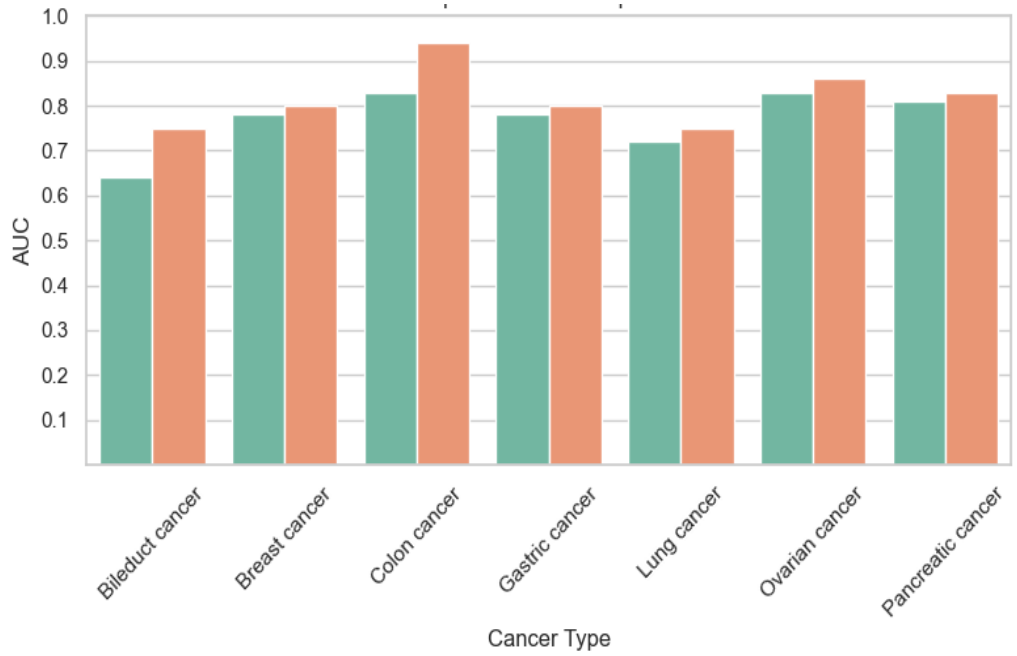


*Figure 4.4 AUC comparison using all features genome wide versus top 10 features across different cancer types*

***Table 4.1 Classification metrics using all features (Genome wide)***

| Cancer type | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| Bile duct cancer | 0.67 | 0.69 | 0.67 | 0.66 | 0.64 |
| Breast cancer | 0.67 | 0.67 | 0.67 | 0.67 | 0.78 |
| Colon cancer | 0.75 | 0.76 | 0.75 | 0.75 | 0.94 |
| Gastric cancer | 0.75 | 0.83 | 0.75 | 0.73 | 0.78 |
| Lung cancer | 0.75 | 0.76 | 0.75 | 0.75 | 0.72 |
| Ovarian cancer | 0.75 | 0.83 | 0.75 | 0.73 | 0.83 |
| Pancreatic cancer | 0.78 | 0.79 | 0.78 | 0.77 | 0.81 |

***Table 4.2 Classification metrics using top 10 features (Genome wide)***

| Cancer type | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| Bile duct cancer | 0.75 | 0.76 | 0.75 | 0.75 | 0.75 |
| Breast cancer | 0.67 | 0.67 | 0.67 | 0.67 | 0.80 |
| Colon cancer | 0.83 | 0.83 | 0.83 | 0.83 | 0.83 |
| Gastric cancer | 0.83 | 0.87 | 0.83 | 0.83 | 0.80 |
| Lung cancer | 0.83 | 0.83 | 0.83 | 0.83 | 0.75 |
| Ovarian cancer | 0.75 | 0.83 | 0.75 | 0.73 | 0.86 |
| Pancreatic cancer | 0.78 | 0.79 | 0.78 | 0.77 | 0.84 |

### 4.4.2 Using combined features genome-wide, ChIP-seq and Open Chromatin sites

Next, we incorporated ChIP-seq TF-binding peaks (from ChIP Atlas) and open chromatin sites (from TCGA) alongside genome-wide TF occupancy. We again evaluated performance on all features and then on the top 10 combined features.

#### 4.4.2.1 Using All Features

Using genome-wide occupancy plus ChIP-seq and open chromatin signals, the classifier's performance (Table 4.3) was as follows: Compared to genome-wide alone, adding ChIP-seq and open chromatin modestly decreased accuracy for most cancers (e.g., bile duct: 0.67→0.62, breast: 0.67→0.62, colon: 0.75→0.69), although AUC improved for bile duct (0.64→0.80) and breast (0.78→0.73 remained similar). This suggests that although combined features introduce greater biological context, they also introduce additional noise or complexity, which as a full set, does not uniformly translate to higher classification accuracy.

#### 4.4.2.2 Using Top 10 Features

We then extracted the top 10 most important features for each cancer type, retrained the model, and measured performance (Table 4.4). Retraining on the top 10 combined features yielded substantial gains over the full combined set: In particular, bile duct and breast cancers saw the largest accuracy increases (Δ = +0.22 and +0.15, respectively) when restricting to the top 10 combined features. Figure 4.4 displays the combined approach accuracy and figure 4.5 diaplays the AUC before (all features) and after (top 10 features).

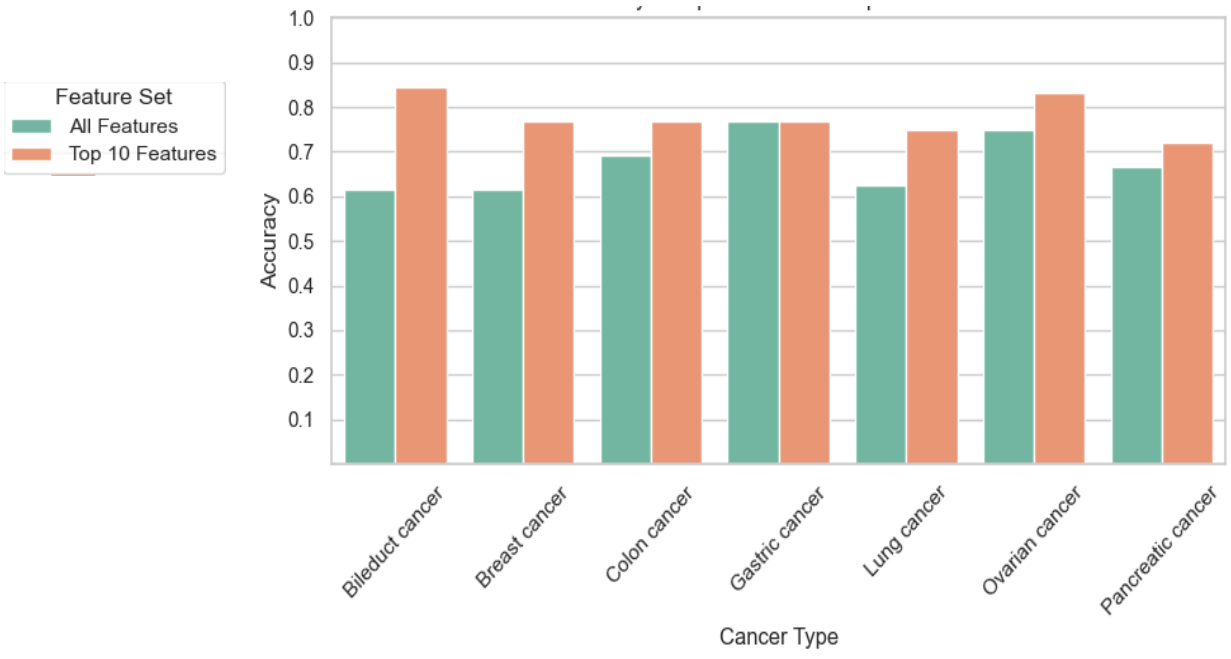


*Figure 4.5 Accuracy comparison using all features (combined) versus top 10 features across different cancers*

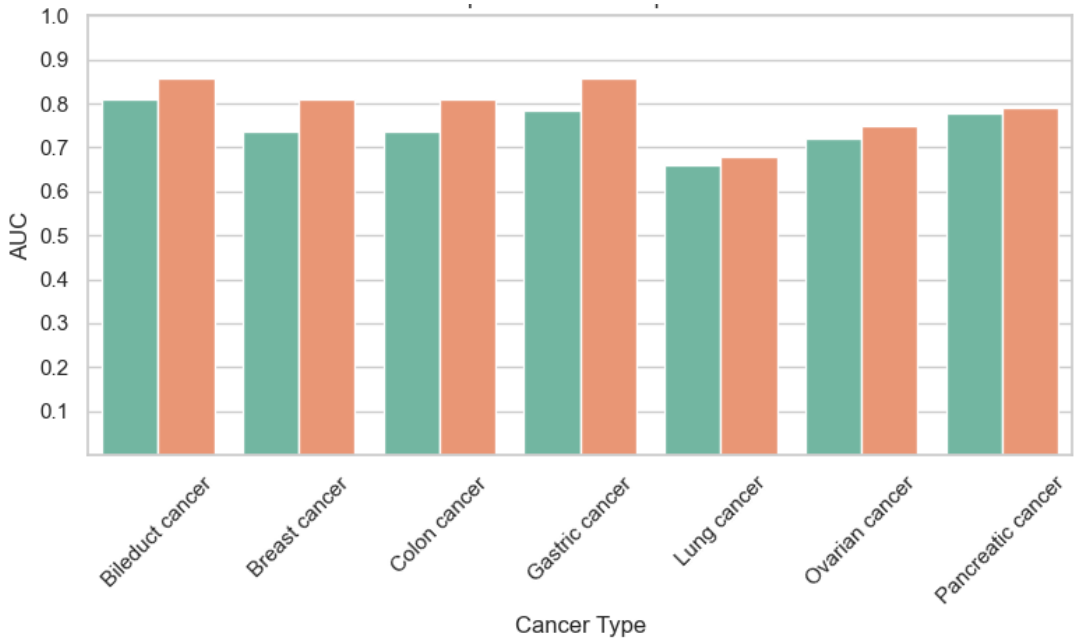


*Figure 4.6 AUC comparison using all features (combined) versus top 10 features across different cancers*

***Table 4.3 Classification metrics using all features (combined)***

| Cancer type | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| Bile duct cancer | 0.62 | 0.65 | 0.63 | 0.60 | 0.80 |
| Breast cancer | 0.62 | 0.62 | 0.62 | 0.61 | 0.73 |
| Colon cancer | 0.69 | 0.80 | 0.71 | 0.67 | 0.73 |
| Gastric cancer | 0.77 | 0.77 | 0.76 | 0.76 | 0.78 |
| Lung cancer | 0.62 | 0.63 | 0.62 | 0.62 | 0.66 |
| Ovarian cancer | 0.75 | 0.76 | 0.75 | 0.75 | 0.72 |
| Pancreatic cancer | 0.67 | 0.67 | 0.67 | 0.66 | 0.78 |

***Table 4.4 Classification metrics using top 10 features (combined)***

| Cancer type | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| Bile duct cancer | 0.84 | 0.84 | 0.84 | 0.84 | 0.86 |
| Breast cancer | 0.77 | 0.77 | 0.77 | 0.77 | 0.81 |
| Colon cancer | 0.77 | 0.77 | 0.77 | 0.77 | 0.81 |
| Gastric cancer | 0.77 | 0.77 | 0.77 | 0.77 | 0.86 |
| Lung cancer | 0.75 | 0.83 | 0.75 | 0.73 | 0.68 |
| Ovarian cancer | 0.83 | 0.83 | 0.83 | 0.83 | 0.75 |
| Pancreatic cancer | 0.72 | 0.72 | 0.72 | 0.72 | 0.79 |

Figure 4.6 shows the pairwise classification performance heatmap for distinguishing six cancer types based on nucleosome occupancy patterns. The diagonal values represent self comparisons. Off-diagonal values indicate the model's accuracy when distinguishing each pair of cancer types (after feature selection).

Overall, high pairwise performance is observed for certain comparisons, such as Breast cancer vs. Lung cancer (0.91) and Prostate cancer vs. Colon cancer (0.84), suggesting distinct nucleosome accessibility profiles for these tissue types. Comparisons involving Bladder cancer tend to show relatively lower discrimination, with values closer to 0.75, indicating more similar chromatin footprints among some tissue pairs.

These results support the hypothesis that nucleosome occupancy signatures around transcription factor binding sites capture biologically meaningful chromatin context that can differentiate tissues of origin. This reinforces the potential of fragmentomic features, beyond sequence motifs, to contribute to tissue-specific cancer diagnostics using cfDNA.

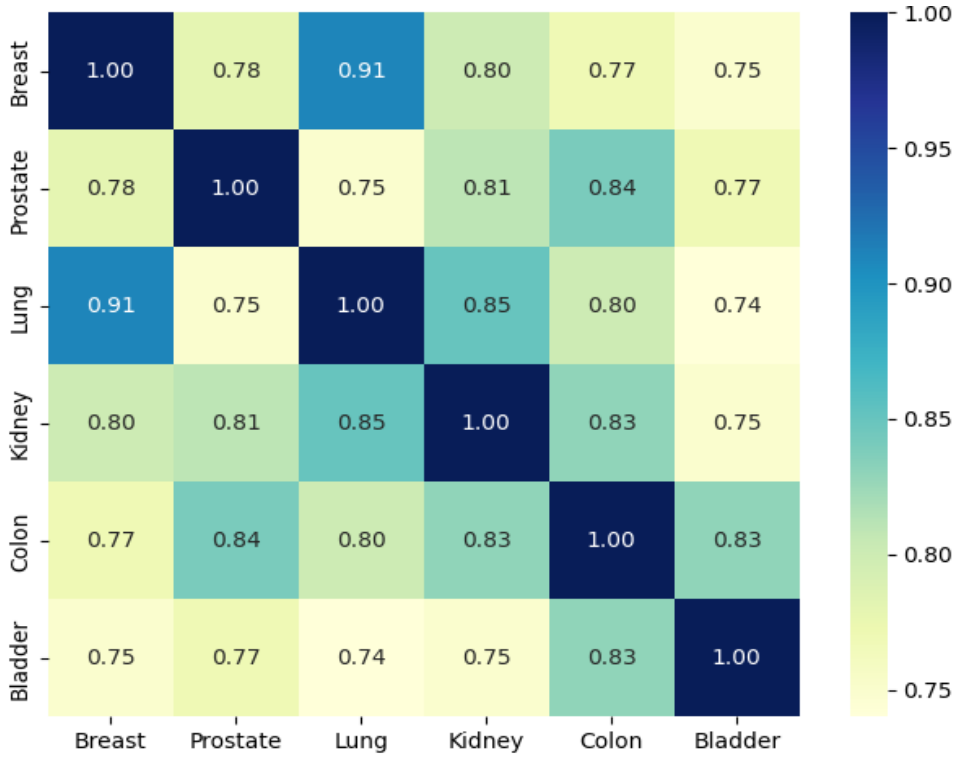


*Figure 4.7 Heatmap of pairwise classification accuracies for six cancer types*

### 4.5 Key Insights

**Feature Reduction Boosts Performance and Interpretability**: Across both genome-wide and combined strategies, retraining on the top 10 features consistently improved or matched accuracy, F1-score, and AUC for nearly all cancer types. The most dramatic improvements were seen when using combined features: bile duct (0.62→0.84) and breast (0.62→0.77). This underscores that many raw combined features add noise; selecting the highest-importance loci yields a cleaner, more discriminative signal. When using all features, the genome-wide approach slightly outperformed the combined features for some cancers (e.g., colon: Accuracy 0.75 vs. 0.69; ovarian: 0.75 vs. 0.75), largely because adding ChIP-seq/open chromatin introduced extra complexity. However, after feature selection, combined features often matched or surpassed genome-wide performance, especially in AUC (e.g., gastric: 0.80→0.86 combined vs. 0.78→0.80 genome-wide). This suggests that although integrating multiple data types can dilute raw model performance, careful selection of key TF/ChIP/OC loci yields powerful biomarkers. Across seven tumor types, top 10 feature models consistently produced high F1-scores ($\geq 0.73$ for all cancers) and AUCs ($\geq 0.68$). Best overall performance (accuracy $\geq 0.80$, AUC $\geq 0.80$) was achieved for colon, gastric, ovarian, and breast cancers under at least one scenario, highlighting that nucleosome occupancy around selected TF sites can serve as reliable tissue-of-origin markers in cfDNA.

### 4.6 Conclusion

In this chapter, we demonstrated a multi-tiered strategy for leveraging nucleosome occupancy around transcription factor (TF) binding sites to distinguish among seven cancer types and infer tissue of origin from cfDNA data. The findings underscore two key points: (1) Interpretability through Feature Selection, i.e., isolating the top 10 TF sites or regulatory loci not only removes noise inherent in thousands of genome-wide measurements but also highlights a concise set of candidate biomarkers. Many of these top features correspond to TFs or chromatin regions previously implicated in tumorigenesis, suggesting that the model is effectively capturing meaningful signals. (2) Although genome-wide nucleosome occupancy alone provided strong predictive signals, especially for colon and pancreatic cancers, adding ChIP-seq and open chromatin data enabled the identification of combined feature sets that further refine classification for cancers (e.g., bile duct and breast) that were more challenging to separate. This implies that multi-layered epigenomic integration can enhance sensitivity to subtle, tissue-specific chromatin remodeling events.

In conclusion, our pipeline, from motif scanning and nucleosome profiling to FFT-based feature extraction and Random Forest classification offers a robust framework for classifying cancer tissue of origin from cfDNA. The increased performance gains achieved by retraining on only the top 10 most informative features point toward a small but powerful panel of TF-centered loci that may serve as future diagnostic biomarkers. As sequencing datasets grow and additional epigenomic layers (such as DNA methylation or histone modifications) become available, this modular approach can be extended to further improve accuracy and uncover the molecular foundations of cancer-specific chromatin landscapes.

# Chapter 5

## Discussion and future scope

Collectively, the findings presented in this thesis contribute meaningfully to the advancement of non-invasive cancer diagnostics through cell-free DNA (cfDNA) analysis. The demonstrated utility of end-motif and nucleosome occupancy classifiers, based on shallow sequencing suggests that high diagnostic sensitivity and specificity can be achieved without the need for extensive sequencing depth or mutation profiling. This is particularly advantageous in clinical scenarios where cost, time, and sample input are constrained. The ability to extract biologically informative signals from fragmentomic patterns, rather than relying on somatic mutations, broadens the applicability of cfDNA testing to cancers with low mutational burdens or limited actionable variants.

Importantly, the integration of multiple cfDNA features such as end motifs and nucleosome footprints has shown to enhance classification performance across both cancer detection and tissue-of-origin inference. This multi-parametric approach reflects a growing trend in precision oncology, wherein composite signals derived from chromatin structure and fragmentation dynamics are leveraged to increase diagnostic resolution. In practice, this could translate to improved early detection of asymptomatic cancers, more accurate disease monitoring, and better stratification for therapeutic interventions.

Our computational framework supports the development of cost-effective, interpretable, and scalable assays. Using low-coverage whole-genome sequencing combined with targeted motif or occupancy analysis that are suitable for deployment in diverse clinical environments.

Moreover, by emphasizing interpretability and biological grounding (e.g., TF motif involvement, nuclease activity), this work facilitates the identification of clinically actionable biomarkers. These features not only improve model transparency but also enable useful insights into cancer-associated chromatin remodeling, which can inform future therapeutic targets. Transfer learning makes it easier to apply models to new situations. It allows models trained on well-understood groups to be adjusted for use with different sequencing tools or populations, even when only a small amount of new data is available.

In sum, the computational strategies developed in this thesis position cfDNA fragmentomics as a practical and powerful tool in the liquid biopsy landscape. With further clinical validation, these methods have the potential to transform cancer diagnostics enabling early detection, subtyping, and longitudinal monitoring through minimally invasive, blood-based assays.

## REFERENCES


[1] L. N. Phuong, S. Salas, and S. Benzekry, "Computational modeling approaches for circulating cell-free DNA in oncology." Feb. 28, 2024. Accessed: May 22, 2025. [Online]. Available: https://hal.science/hal-04481689

[2] A.-L. Volckmar *et al.*, "A field guide for cancer diagnostics using cell-free DNA: From principles to practice and clinical applications," *Genes. Chromosomes Cancer*, vol. 57, no. 3, pp. 123–139, 2018, doi: 10.1002/gcc.22517.

[3] P. Muluhngwi, R. Valdes Jr, R. Fernandez-Botran, E. Burton, B. Williams, and M. W. Linder, "Cell-Free DNA Diagnostics: Current and Emerging Applications in Oncology," *Pharmacogenomics*, vol. 20, no. 5, pp. 357–380, Apr. 2019, doi: 10.2217/pgs-2018-0174.

[4] Q. Gao *et al.*, "Circulating cell-free DNA for cancer early detection," *The Innovation*, vol. 3, no. 4, Jul. 2022, doi: 10.1016/j.xinn.2022.100259.

[5] X. Peng, H.-D. Li, F.-X. Wu, and J. Wang, "Identifying the tissues-of-origin of circulating cell-free DNAs is a promising way in noninvasive diagnostics," *Brief. Bioinform.*, vol. 22, no. 3, p. bbaa060, May 2021, doi: 10.1093/bib/bbaa060.

[6] P. Song, "Limitations and opportunities of technologies for the analysis of cell-free DNA in cancer diagnostics," *Nat. Biomed. Eng.*, vol. 6, 2022.

[7] C. F. Wright and H. Burton, "The use of cell-free fetal nucleic acids in maternal blood for non-invasive prenatal diagnosis," *Hum. Reprod. Update*, vol. 15, no. 1, pp. 139–151, Oct. 2008, doi: 10.1093/humupd/dmn047.

[8] K. G. Weerakoon and D. P. McManus, "Cell-Free DNA as a Diagnostic Tool for Human Parasitic Infections," *Trends Parasitol.*, vol. 32, no. 5, pp. 378–391, May 2016, doi: 10.1016/j.pt.2016.01.006.

[9] Y. Long *et al.*, "Diagnosis of Sepsis with Cell-free DNA by Next-Generation Sequencing Technology in ICU Patients," *Arch. Med. Res.*, vol. 47, no. 5, pp. 365–371, Jul. 2016, doi: 10.1016/j.arcmed.2016.08.004.

[10] I. De Vlaminck *et al.*, "Noninvasive monitoring of infection and rejection after lung transplantation," *Proc. Natl. Acad. Sci.*, vol. 112, no. 43, pp. 13336–13341, Oct. 2015, doi: 10.1073/pnas.1517494112.

[11] J. C. M. Wan *et al.*, "Liquid biopsies come of age: towards implementation of circulating tumour DNA," *Nat. Rev. Cancer*, vol. 17, no. 4, pp. 223–238, Apr. 2017, doi: 10.1038/nrc.2017.7.

[12] M. Sharma, R. K. Verma, S. Kumar, and V. Kumar, "Computational challenges in detection of cancer using cell-free DNA methylation," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 26–39, Dec. 2021, doi: 10.1016/j.csbj.2021.12.001.

[13] C. Bedin, M. V. Enzo, P. Del Bianco, S. Pucciarelli, D. Nitti, and M. Agostini, "Diagnostic and prognostic role of cell-free DNA testing for colorectal cancer patients," *Int. J. Cancer*, vol. 140, no. 8, pp. 1888–1898, 2017, doi: 10.1002/ijc.30565.

[14] N. Wan *et al.*, "Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA," *BMC Cancer*, vol. 19, no. 1, p. 832, Aug. 2019, doi: 10.1186/s12885-019-6003-8.

[15] P. Peneder *et al.*, "Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden," *Nat. Commun.*, vol. 12, no. 1, p. 3230, May 2021, doi: 10.1038/s41467-021-23445-w.

[16] Y. Hou, X.-Y. Meng, and X. Zhou, "Systematically Evaluating Cell-Free DNA Fragmentation Patterns for Cancer Diagnosis and Enhanced Cancer Detection via Integrating Multiple Fragmentation Patterns," *Adv. Sci.*, vol. 11, no. 30, p. 2308243, 2024, doi: 10.1002/advs.202308243.

[17] H. Shen, M. Yang, J. Liu, K. Chen, and X. Li, "Development of a deep learning model for cancer diagnosis by inspecting cell-free DNA end-motifs," *Npj Precis. Oncol.*, vol. 8, no. 1, pp. 1–12, Jul. 2024, doi: 10.1038/s41698-024-00635-5.
[18] B.-W. Han *et al.*, "A Deep-Learning Pipeline for TSS Coverage Imputation From Shallow Cell-Free DNA Sequencing," *Front. Med.*, vol. 8, p. 684238, 2021, doi: 10.3389/fmed.2021.684238.
[19] E. Katsman *et al.*, "Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing," *Genome Biol.*, vol. 23, no. 1, p. 158, Dec. 2022, doi: 10.1186/s13059-022-02710-1.
[20] W. Li *et al.*, "CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data," *Nucleic Acids Res.*, vol. 46, no. 15, p. e89, Sep. 2018, doi: 10.1093/nar/gky423.
[21] M. E. Norton *et al.*, "Cell-free DNA Analysis for Noninvasive Examination of Trisomy," *N. Engl. J. Med.*, vol. 372, no. 17, pp. 1589–1597, Apr. 2015, doi: 10.1056/NEJMoa1407349.
[22] M. Oellerich *et al.*, "Donor-derived cell-free DNA as a diagnostic tool in transplantation," *Front. Genet.*, vol. 13, Oct. 2022, doi: 10.3389/fgene.2022.1031894.
[23] I. De Vlaminck *et al.*, "Circulating Cell-Free DNA Enables Noninvasive Diagnosis of Heart Transplant Rejection," *Sci. Transl. Med.*, vol. 6, no. 241, pp. 241ra77-241ra77, Jun. 2014, doi: 10.1126/scitranslmed.3007803.
[24] S. C. Ding and Y. M. D. Lo, "Cell-Free DNA Fragmentomics in Liquid Biopsy," *Diagnostics*, vol. 12, no. 4, p. 978, Apr. 2022, doi: 10.3390/diagnostics12040978.
[25] K. A. Chan *et al.*, "Cancer Genome Scanning in Plasma: Detection of Tumor-Associated Copy Number Aberrations, Single-Nucleotide Variants, and Tumoral Heterogeneity by Massively Parallel Sequencing," *Clin. Chem.*, vol. 59, no. 1, pp. 211–224, Jan. 2013, doi: 10.1373/clinchem.2012.196014.
[26] M. Ivanov, A. Baranova, T. Butler, P. Spellman, and V. Mileyko, "Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation," *BMC Genomics*, vol. 16 Suppl 13, no. Suppl 13, p. S1, 2015, doi: 10.1186/1471-2164-16-S13-S1.
[27] P. Jiang *et al.*, "Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients," *Proc. Natl. Acad. Sci.*, vol. 112, no. 11, pp. E1317–E1325, Mar. 2015, doi: 10.1073/pnas.1500076112.
[28] M. E. Norton, "Noninvasive prenatal testing to analyze the fetal genome," *Proc. Natl. Acad. Sci.*, vol. 113, no. 50, pp. 14173–14175, Dec. 2016, doi: 10.1073/pnas.1617112113.
[29] D. S. C. Han *et al.*, "The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB," *Am. J. Hum. Genet.*, vol. 106, no. 2, pp. 202–214, Feb. 2020, doi: 10.1016/j.ajhg.2020.01.008.
[30] L. Serpas *et al.*, "*Dnase1l3* deletion causes aberrations in length and end-motif frequencies in plasma DNA," *Proc. Natl. Acad. Sci.*, vol. 116, no. 2, pp. 641–649, Jan. 2019, doi: 10.1073/pnas.1815031116.
[31] Y. M. Lo *et al.*, "Presence of fetal DNA in maternal plasma and serum," *Lancet Lond. Engl.*, vol. 350, no. 9076, pp. 485–487, Aug. 1997, doi: 10.1016/S0140-6736(97)02174-0.
[32] T.-R. Lee *et al.*, "Integrating Plasma Cell-Free DNA Fragment End Motif and Size with Genomic Features Enables Lung Cancer Detection," *Cancer Res.*, vol. 85, no. 9, pp. 1696–1707, May 2025, doi: 10.1158/0008-5472.CAN-24-1517.
[33] Z. Zhou *et al.*, "Fragmentation landscape of cell-free DNA revealed by deconvolutional analysis of end motifs," *Proc. Natl. Acad. Sci.*, vol. 120, no. 17, p. e2220982120, Apr. 2023, doi: 10.1073/pnas.2220982120.

[34] P. Jiang *et al.*, "Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation," *Cancer Discov.*, vol. 10, no. 5, pp. 664–673, May 2020, doi: 10.1158/2159-8290.CD-19-0622.
[35] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, p. 9, Dec. 2016, doi: 10.1186/s40537-016-0043-6.
[36] F. Zhuang *et al.*, "A Comprehensive Survey on Transfer Learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.
[37] N. Segev, M. Harel, S. Mannor, K. Crammer, and R. El-Yaniv, "Learn on Source, Refine on Target: A Model Transfer Learning Framework with Random Forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1811–1824, Sep. 2017, doi: 10.1109/TPAMI.2016.2618118.
[38] L. Minvielle, M. Atiq, S. Peignier, and M. Mougeot, "Transfer Learning on Decision Tree with Class Imbalance," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, Portland, OR, USA: IEEE, Nov. 2019, pp. 1003–1010. doi: 10.1109/ICTAI.2019.00141.
[39] S. Niu, Y. Liu, J. Wang, and H. Song, "A Decade Survey of Transfer Learning (2010–2020)," *IEEE Trans. Artif. Intell.*, vol. 1, no. 2, pp. 151–166, Oct. 2020, doi: 10.1109/TAI.2021.3054609.
[40] H. M. Gomes *et al.*, "Adaptive random forests for evolving data stream classification," *Mach. Learn.*, vol. 106, no. 9, pp. 1469–1495, Oct. 2017, doi: 10.1007/s10994-017-5642-8.
[41] C. V. Andreu-Vieyra and G. Liang, "Nucleosome occupancy and gene regulation during tumorigenesis," *Adv. Exp. Med. Biol.*, vol. 754, pp. 109–134, 2013, doi: 10.1007/978-1-4419-9967-2_5.
[42] N. De Sarkar *et al.*, "Nucleosome Patterns in Circulating Tumor DNA Reveal Transcriptional Regulation of Advanced Prostate Cancer Phenotypes," *Cancer Discov.*, vol. 13, no. 3, pp. 632–653, Mar. 2023, doi: 10.1158/2159-8290.CD-22-0692.
[43] P. G. Yazdi *et al.*, "Increasing Nucleosome Occupancy Is Correlated with an Increasing Mutation Rate so Long as DNA Repair Machinery Is Intact," *PLOS ONE*, vol. 10, no. 8, p. e0136574, Aug. 2015, doi: 10.1371/journal.pone.0136574.
[44] S. Mishra *et al.*, "Matching queried single-cell open-chromatin profiles to large pools of single-cell transcriptomes and epigenomes for reference supported analysis," *Genome Res.*, vol. 33, no. 2, pp. 218–231, Feb. 2023, doi: 10.1101/gr.277015.122.
[45] O. Chandra *et al.*, "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes," *Comput. Struct. Biotechnol. J.*, vol. 21, pp. 3590–3603, 2023, doi: 10.1016/j.csbj.2023.07.014.
[46] Y. Peng, W. Song, V. B. Teif, I. Ovcharenko, D. Landsman, and A. R. Panchenko, "Detection of new pioneer transcription factors as cell-type specific nucleosome binders," *eLife*, vol. 12, Dec. 2023, doi: 10.7554/eLife.88936.3.
[47] B. R. Druliner *et al.*, "Comprehensive nucleosome mapping of the human genome in cancer progression," *Oncotarget*, vol. 7, no. 12, pp. 13429–13445, Dec. 2015, doi: 10.18632/oncotarget.6811.