Analysis of cell deconvolution methods
A comparison of reference-based and reference-free cell deconvolution

Stanisław Howard

Responsible Professor: Prof. Marcel Reinders
Supervisors: PhD. Stavros Makrodimitris,
Bram Pronk, Daan Hazelaar

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Stanisław Howard
Final project course: CSE3000 Research Project
Thesis committee: **Prof. Marcel Reinders**, **PhD. Stavros Makrodimitris, Bram Pronk, Daan Hazelaar**, **Prof. Johan Pouwelse**

Abstract

In recent years, a new way of cancer diagnostics has emerged, the analysis of DNA fragments circulating in the blood of cancer patients known as fragmentomics. This DNA, known as cell-free DNA (cfDNA), is an easily available biomarker for cell types. Deducing the tissue origin of cfDNA can reveal anomalies in cell death caused by diseases. That holds great potential in cancer detection and monitoring. The process of establishing the cell composition of a blood sample is called **cell deconvolution**. This research paper focuses on the comparison of two methods of cell deconvolution. The first one **UXM**, solves this problem by employing a reference-based technique using a methylation atlas. The second one reference-free **cfSort**, utilizes a Deep Learning Neural Network to perform the sample analysis. In the paper, however, a simpler architecture was trained due to the difficulties in reproduction. Experiments have been conducted to assess the sensitivity of both methods. Experiments consisted of 5 major cell types together mixed with white blood cell DNA fragments to assess the sensitivity of each method. Furthermore, different metrics such as Pearson's correlation coefficient have been used to determine the accuracy of both methods. In the end, UXM outperformed cfSort in most metrics, including Pearson's correlation coefficient, indicating its superior accuracy in deconvolution tasks. However, cfSort showed potential for higher prediction accuracy with further development and better documentation. The findings highlight the strengths and limitations of both methods. This study suggests that while UXM is currently more reliable, future improvements in cfSort could make it a viable alternative. Continued research is recommended to enhance the accuracy and transparency of these methods, ensuring their effectiveness in real-world healthcare applications.

1 Introduction

In the field of cancer treatment, one of the most important challenges is the timely and accurate diagnosis of the disease. Early detection is crucial for successful treatment outcomes, however, achieving this goal is very difficult. Cancer diagnosis typically involves invasive procedures and can be both a physical and emotional burden for patients [4]. Usually, the standard approach to detect cancer is to perform a biopsy. A biopsy is a medical procedure that involves the extraction of a small sample of tissue from the body for diagnostic examination [1]. This tissue sample is typically analyzed under a microscope to detect the presence of cancer, helping in the determination of the nature, extent, and stage of the disease. Depending on the type of biopsy, it can be a very invasive procedure where the patient has to undergo an extraction under anaesthesia.

Due to the invasiveness and associated difficulties that the extraction might pose, there is a great demand for non-invasive, equally sensitive and accurate diagnostic approaches. In recent years, a new way of cancer diagnostics has emerged. That is a form of liquid biopsy [6], specifically the analysis of circulating DNA fragments in the blood. This field is known as fragmentomics. The idea behind fragmentomics is that every cell sheds fragments of DNA into the bloodstream when it dies. Thanks to that we can obtain useful information from these fragments while at the same time ensuring that the sample extraction process is very short, noninvasive and cost-effective. Several studies have demonstrated the utility of fragmentomics in detecting the presence of cancer, for instance by using a machine learning model to detect anomalies in the fragmentation profile of white blood cells [3] or by using the length distribution of DNA fragments from cancer patients to infer tumour load and types [5]. Despite these advancements, there remains a large knowledge gap in our understanding of how this information can be used in cancer detection. The gap stems from the lack of understanding of the complexity of the human organism and all the intricate processes that are occurring on a cellular level.

The main focus of this research paper is to compare two methods of extracting cell origin information from the circulating DNA fragments. Cell-free DNA (cfDNA) found in blood plasma comes mainly from different types of blood cells such as white blood cells, along with cells lining blood vessels and tissues such as the liver and kidneys. In cancer patient samples, cfDNA can include DNA from tumours. This mix of DNA in plasma reflects how cells naturally break down and release their genetic material into the bloodstream.

A real-world application would be to detect any anomalies in these fragments. One way to do so is to estimate the percentage proportion of each origin, that is being able to tell how many fragments come from which cell type. This process lies at the very foundation of anomaly detection and is known as **cell deconvolution**.

The information enclosed in the cell type proportions can allow us to analyse the patient's health. This is all possible thanks to the property of cancerous cells. Namely, a cancerous cell's main priority is to grow and spread rapidly. When it detects irregularities the organism, attempts to fight the disease and kill the cancer cells. This way as a result we can observe a higher level of deaths of a particular cell type through the DNA fragment levels.

Using cell deconvolution techniques we can compare healthy and diseased patients, and establish levels at which we can assume irregularities. We can then guide the patient to perform more checks. However, this all relies on the deconvolution methods that are in place. An inaccurate deconvolution algorithm can produce many false positives which can lead to DNA fragment levels being diagnosed as normal. Thus, it is extremely important to use a reliable and robust method to accurately predict the cell proportions.

1.1 Cell Deconvolution

To perform the deconvolution, we use the DNA fragments circulating in the blood to draw information from them. DNA is the instruction manual for the body. It is made up of building blocks called nucleotides. The order of them specifies the instructions for the organism to build necessary proteins and compounds. DNA methylation is a process where a methyl

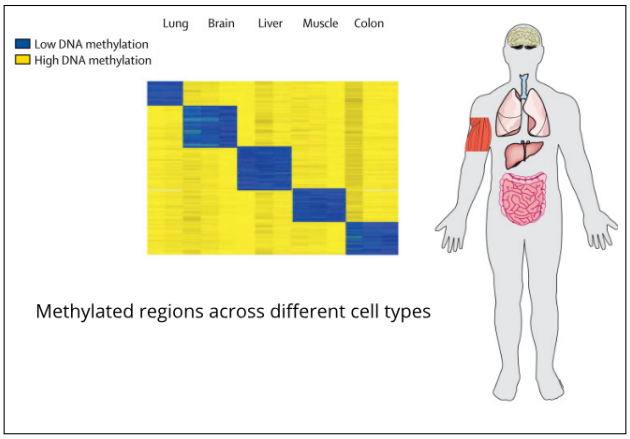

Figure 1: Methylation across different cell types [4]

group is added to specific DNA molecules, mostly cytosine. This can alter gene expression without changing the underlying DNA sequence. These methylation patterns vary significantly between cell types as seen in Fig 1. By analyzing the methylation profile of a bulk tissue sample, containing a mixture of different cell types, deconvolution algorithms can identify cell type-specific methylation signatures. These signatures are then used to estimate the relative abundance of each cell type within the sample. Deconvolution methods can be broadly divided into two categories described below.

1.2 Reference-based Deconvolution

Reference-based deconvolution requires a pre-defined set of methylation profiles for specific cell types. These profiles act as a reference, allowing the algorithm to identify cell type-specific methylation signatures within the bulk tissue sample. A general idea is described in Fig. 2 [1]. By comparing the sample's methylation profile to the reference profiles, the algorithm estimates the relative composition of each cell type contributing to the sample.

This approach can offer a high accuracy when a reliable reference set is available, but its limitation lies in the requirement for pre-characterized cell types. If a specific cell type is absent from the reference, the deconvolution may miss its contribution or misinterpret it as another similar cell type. The particular method that was used in this work was the **UXM deconvolution tool** [8]

1.3 Reference-free Deconvolution

Reference-free deconvolution, on the other hand, doesn't rely on pre-defined cell type profiles that much. It requires them

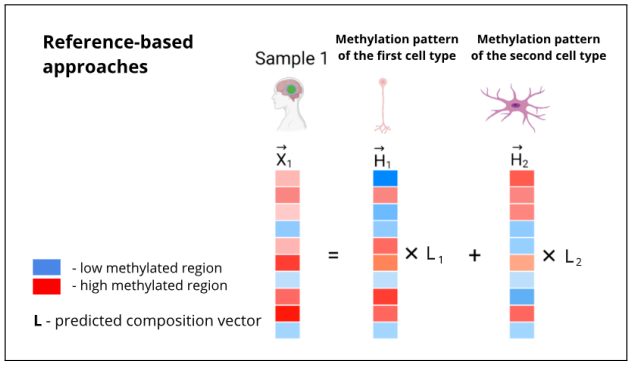

Figure 2: General idea of Reference-based deconvolution

to train the model. However, during the deconvolution, the model only requires the input mixture. The model analyzes the inherent properties of the methylation data itself. These methods often employ sophisticated model architectures to identify underlying patterns within the bulk sample data. This can be observed in Fig. 3 [2]

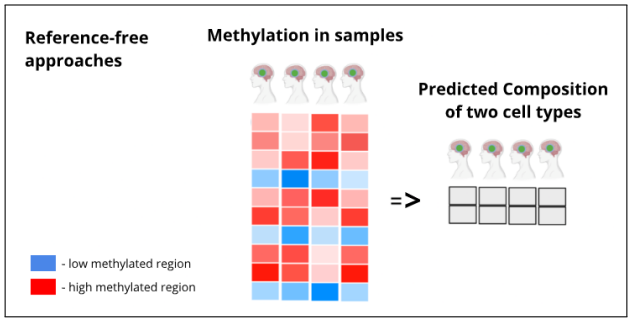

Figure 3: General idea behind reference-free deconvolution

Reference-free methods offer greater flexibility as they do not require reference patterns. Until recently reference-based methods were more popular. However, thanks to the fast development of Deep Learning Models and the overall progress of genetics, a new reference-free tool, namely **cfSort** [7] was developed and according to the author, it performs better than the state-of-the-art methods.

1.4 Motivation

This paper will focus on comparing these two methods. Comparing the specific reference-free and reference-based deconvolution methods is crucial for several reasons. It allows for a neutral evaluation of their performance across a unified dataset, ensuring the generalizability of findings beyond specific experimental conditions. Furthermore, such comparisons provide insights into the strengths and weaknesses of each approach, helping researchers select the most suitable approach based on the specific characteristics of their data and research goals.

[1] Image taken and modified from https://en.wikipedia.org/wiki/Cellular_deconvolution#/media/File:ReffreevsBased.png

[2] https://en.wikipedia.org/wiki/Cellular_deconvolution#/media/File:ReffreevsBased.png

1.5 Problem Decomposition

To effectively address the main research question, we decomposed it into the following sub-problems:

- **Which method is better in terms of performance?** This question involves measuring the accuracy of the models using Pearson's correlation coefficient. Additionally, we will perform a short qualitative analysis of each approach in terms of their codebase reproducibility.
- **How do the two models compare in a sensitivity detection test?** We investigate their performance in detecting low percentages of a secondary cell type in a mixture with white blood cells (WBCs). The reason for this experiment is to establish how well the two methods detect cell types if the corresponding proportion of DNA fragments is low. This is important as it affects the way we draw conclusions from results. results of each method.

2 Methodology

In this chapter, we describe the approach to answering the research question. We present the experiment setup and the method modifications.

2.1 Algorithm Selection and Design

For this research, we selected the **UXM** deconvolution tool for reference-based deconvolution and **cfSort** for reference-free deconvolution.

UXM

A well-established method developed by Loyfer et al. [8]. **UXM**[3] is a computational reference-based deconvolution algorithm for DNA methylation sequencing data. It constructs a reference atlas where the percentages of unmethylated fragments are computed for every marker in each cell type. A non-negative least squares (NNLS) algorithm [8] is then used to fit an input sample and estimate its relative contributions. It covers the 39 most common human cell types.

cfSort

A newly released method in August 2023 by Li et al. [7]. **cfSort** [4] is a Deep Learning-based approach for sensitive and accurate tissue deconvolution in cfDNA. It is built upon tissue markers covering 29 major human tissue types.

2.2 Experimental Setup

Data Collection: We used a dataset of 521 genomic DNA samples, including 464 non-WBC tissue samples from the GTEx project [2] and 57 WBC samples from UCLA hospitals taken from the cfSort dataset [7]. We choose 5 types of cells to evaluate the methods. Specifically, breast tissue with 15 samples, colon tissue with 29 samples, lung tissue with 16 samples, prostate tissue with 13 samples and kidney tissue with 13 samples. These types were chosen due to their occurrence in cancer studies and the availability of samples.

[3]https://github.com/nloyfer/UXM_deconv

[4]https://github.com/jasminezhoulab/cfSort

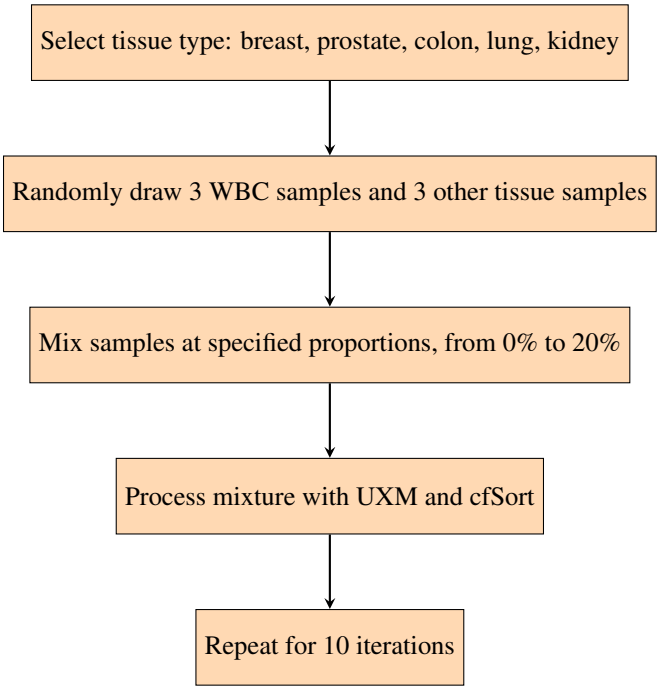

Figure 4: Flow of the experiment

Experimental Design: To evaluate the deconvolution methods, we created mixtures of WBCs with five other tissue types: breast, prostate, colon, lung, and kidney tissue. We created mixtures with the following concentration of a tissue type in a white blood cell mixture: 0%, 0.5%, 1%, 2%, 3%, 5%, 10%, 15%, and 20%. For each ratio, we randomly drew three WBC files and three files from another tissue type, generating mixtures that reflect varying levels of the secondary cell type. Each mixture was processed using both the UXM and cfSort deconvolution methods. This process was repeated 10 times to ensure the reliability of the results. A general overview is seen in Flowchart 4

2.3 Evaluation Metrics

To assess the performance of the deconvolution methods, we calculated two key metrics:

- **Minimal Composition Detection**: For each iteration, we took the minimal composition of the secondary cell type that could be reliably detected. By reliable detection, we assumed that the detected proportion of the type is within 50% of the original proportion. That is, in the case of 10% of the composition of another cell type, between 5% and 15% would be considered detected
- **Pearson Correlation Coefficient** [5]: We computed the average Pearson's correlation coefficient for each cell type across all iterations. This metric helps quantify the accuracy of the deconvolution by comparing the predicted cell type proportions to the actual proportions.

[5]https://www.britannica.com/topic/Pearsons-correlation-coefficient

The Pearson correlation coefficient r between two variables X and Y is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where:

- X_i and Y_i are the individual sample points.
- $\overline{X}$ and $\overline{Y}$ are the means of the sample points X and Y, respectively.
- n is the number of sample points.

2.4 Implementation Details

The implementation involved several critical steps:

- **Code availability:** We developed code (available at the TU Delft **GitLab Repository**[6]) to generate DNA fragment mixtures. The pipeline includes scripts for data preprocessing, mixture generation, and running the deconvolution methods.
- **File sampling:** Each sample was in the form of a `.pat` file, a `.pat.csi` file and a `.bed` file. A `.pat` file contains fragments of DNA methylation patterns, including the position of each methylated cytosine and its context within the genome. A `.pat.csi` is an index of a corresponding `.pat` file, allowing for efficient querying and retrieval of methylation data based on genomic coordinates. To create a mixture we drew a certain percentage of DNA fragments from each of the 3 files of a particular cell type. The rest was filled with fragments from the WBC files.
- **Method Modification**: Both UXM and cfSort were modified to accept `.pat` files as input. This involved adapting the input formats and ensuring compatibility with our data.

  UXM modification:

  The UXM deconvolution tool was well-written and documented and only minor adjustments were necessary. The primary change involved mapping the cell types in our dataset to those used by UXM, which has a higher resolution with 39 cell types. For example, UXM's multiple white blood cell types were mapped to our dataset's general "WBC" category. The exact mapping is in the Appendix 7. This higher resolution of UXM is primarily due to the more diverse dataset used for training and extensive experiments conducted by the UXM creators, rather than the method itself. Additionally, we modified UXM to use the human genome version 19, as cfSort's markers were created based on that version.

  cfSort modification: Due to the poor maintenance of the cfSort codebase, its reliance on Python 2 and the lack of proper comments or documentation, we encountered major difficulties in using it directly. As a result, we decided to train our model and reproduce the exact steps of the cfSort method but with a simpler model architecture. We used the methylation markers provided with the original codebase and followed these steps.

  For each marker, we noted its chromosome position, alpha value, and cluster-ID. By marker, we define a location on the DNA sequence in which methylation differs across different tissue types. Alpha-value is defined as the fraction of methylated positions out of all pairs on a DNA fragment. This fragment-level measurement has been utilized in several studies to identify cancer-specific methylation markers [9]. The cluster-ID was a marker clustering strategy designed to merge individual tissue markers into a marker cluster that is robust against the impact of nucleosome positioning. Constrained K-means clustering was performed on the individual markers based on their methylation profiles across training samples, allowing four to seven individual markers in a cluster. All this information was already provided in the marker files of cfSort.

  We simulated 400 mixtures as described in the cfSort paper. A detailed description can be found in the Appendix 8. For each marker region, we counted how many reads mapped to it, and in how many of these fragments the fraction of methylated cytosines was less than the alpha-value of that marker. We then added together these two numbers for all markers with the same cluster-ID to get, for each cluster, the fraction of fragments that were below the alpha value, which is the cfSort feature.
  We followed the preprocessing steps, including log-transform and min-max scaling, as described in the supplement of cfSort to arrive at the features. Finally, we trained a linear regression model using the features from the mixtures we generated. We added a softmax layer at the output to ensure the predicted composition results were always positive.

3 Results

3.1 Specification of types for UXM

To compare the performance of cfSort and UXM deconvolution tools, we evaluated the accuracy of UXM both with and without specifying the exact cell types present in the sample. When running UXM without specifying the cell types, it was unable to reliably detect the presence of underlying cell types. The fraction of actual cell types in the sample mixture that was correctly detected was notably low, with average values ranging from 0.608 to 0.712 across different tissue types. That is, if a sample was a mixture of WBC and colon DNA fragments only 71,2% of all fragments were detected to be of these specific types. The results are visible in Tab. 1.

Due to these low detection rates, we decided to run the UXM with specified cell types to test its accuracy in the following tests. When specifying the cell types, UXM

[6] https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Reinders_Pronk_Hazelaar/showard-Detection-of-cancer-using-blood.git

Table 1: Fraction of each cell type sample mixture correctly detected by UXM without specifying cell types

Tissue Type	Average	Standard Deviation
Breast Tissue	0.614	0.074
Colon Tissue	0.712	0.068
Kidney Tissue	0.635	0.028
Prostate Tissue	0.685	0.026
Lung Tissue	0.608	0.062

Table 2: Fraction of each cell type sample mixture correctly detected by UXM with specifying cell types

Tissue Type	Average	Standard Deviation
Breast Tissue	0.999	0.000271
Colon Tissue	0.999	0.000171
Kidney Tissue	0.999	0.000179
Prostate Tissue	0.999	0.000190
Lung Tissue	0.999	0.000241

showed a significantly higher accuracy with average values close to 1.0 and very low standard deviations. Only a 0.1% fraction was classified as different from the provided types. This pattern was consistent across all tissue types, as shown in Table 2.

As for the cfSort, it was trained in a way to predict all the underlying types and the results of this are shown in Table 3.

Table 3: Fraction of each cell type sample mixture correctly detected by cfSort

Tissue Type	Average	Standard Deviation
Breast Tissue	0.921	0.0415
Colon Tissue	0.912	0.0844
Kidney Tissue	0.904	0.0477
Prostate Tissue	0.917	0.0517
Lung Tissue	0.958	0.0243

3.2 Pearson's Correlation Coefficient Experiment

In this experiment, we evaluated the performance of the UXM and cfSort deconvolution tools by calculating Pearson's correlation coefficients for different tissue types. The results indicate a better performance of the UXM with an average correlation of **0.920** against cfSort's **0.826** correlation across all samples and types. The data also shows that the standard deviation of Pearson's coefficients is generally higher for cfSort compared to UXM. For example, in the case of colon tissue, UXM has a standard deviation of 0.0193, whereas cfSort's standard deviation is 0.0687. This suggests that the cfSort method produces more variable results than UXM. The box plot in Fig. 5 visually demonstrates the difference in variability and all the results are stored in table 4. One thing worth indicating is the two outliers in the colon tissue visible both

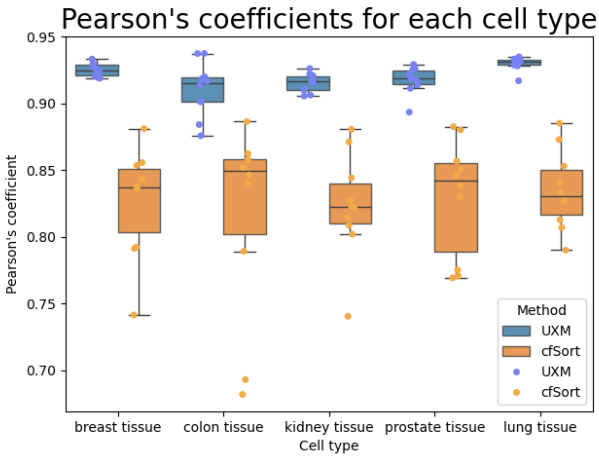

Figure 5: Pearson's correlation results of both methods

in the UXM and cfSort correlation results.

Table 4: Mean and standard deviation of Pearson's correlation coefficients for UXM and cfSort across different tissue types

Tissue Type	UXM		cfSort	
	Mean	Std	Mean	Std
Breast Tissue	0.925	0.0046	0.827	0.0385
Colon Tissue	0.911	0.0193	0.817	0.0687
Kidney Tissue	0.915	0.0066	0.823	0.0369
Prostate Tissue	0.918	0.0096	0.830	0.0412
Lung Tissue	0.930	0.0047	0.835	0.0278

3.3 Sensitivity Detection Experiment

In our earliest detection test, we compared the performance of the UXM and cfSort deconvolution tools across various tissue types. The results are summarized in Table 5. As mentioned earlier we defined detection when the predicted level is within 50% range of the actual level. Notably, UXM consistently detected the other types at the lowest compositions all types apart from colon and prostate. The mean value of those types was 10 times larger than others. CfSort's results on the other hand were not that spread out, ranging from the earliest detection at 0.037 of colon tissue to 0.09 of breast tissue. Overall, UXM had an average mean value of 0.0207, while cfSort had a higher average mean value of 0.0595. This suggests that cfSort generally detected the presence of tissue types at higher mean values compared to UXM.

4 Discussion

In this section, we analyse our experimental results, examining the sensitivity detection experiment and the Pearson's correlation coefficient experiment. We also explore future research directions and underline the limitations of our study.

Table 5: Earliest detection mean values and number of samples for UXM and cfSort across different tissue types

Tissue Type	UXM Mean	cfSort Mean
Breast Tissue	0.0055	0.09
Colon Tissue	0.04	0.037
Kidney Tissue	0.0050	0.07
Prostate Tissue	0.0479	0.0375
Lung Tissue	0.0050	0.063
Overall	0.0207	0.0595

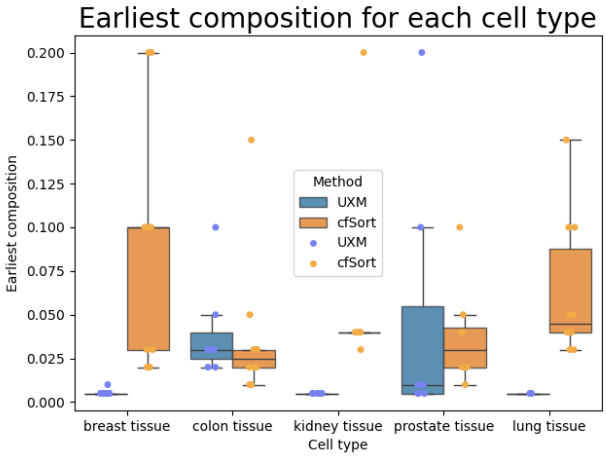

Figure 6: Sensitivity test of the UXM for both methods

4.1 Pearson's Correlation Coefficient Experiment Analysis

The Pearson's correlation coefficient experiment revealed a trend in the performance of the UXM and cfSort deconvolution tools, with UXM outperforming cfSort in terms of average correlation and lower standard deviation across different tissue types. However, an interesting aspect of the results is the presence of outliers in the tissue correlation results for both UXM and cfSort methods. This is particularly visible in the colon tissue. The presence of outliers in both methods suggests that the variability might not be due to the deconvolution methods themselves but could be caused by other factors related to the samples.

To further investigate this anomaly, a deeper analysis of the coverage for the files used in each sample is recommended. Coverage, defined as the number of reads for each region of each marker in the .pat file, could potentially explain the lower correlation coefficients observed in these outliers. It is possible that these samples had significantly lower coverage, resulting in fewer markers being sampled for the colon tissue type, which could impact the results of the experiment.

To validate this hypothesis, an experiment should be conducted to analyze the correlation between the coverage of samples and their corresponding Pearson correlation coefficients. To ensure the reliability of the results the number of samples and iterations should be increased. Such an analysis could provide valuable insights into these outliers.

4.2 Sensitivity Detection Experiment Analysis

In our sensitivity detection experiment comparing UXM and cfSort across various tissue types (Table 5), we observed that cfSort consistently achieved higher mean detection values compared to UXM. It exhibited significantly larger mean values for colon and prostate tissues, approximately 10 times higher than those for other tissue types. This suggests that the differences in detection performance between the two methods may be influenced by the specific composition of markers used for these tissue types. Further investigation into the markers utilised by each method could provide insights into optimising their respective detection accuracies. A simple experiment to count the number of discovered markers per each type could help understand these results.

4.3 Limitations

One limitation to mention is that we selected lung, kidney, colon, breast, and prostate tissues for our study. It is important to note that breast and prostate tissues are associated with gender-specific cancers, as breast cancer affects mostly women and prostate cancer exclusively affects men. This gender specificity should be considered when interpreting the results, as it may influence the tissue type proportions and their implications for health and disease detection. Additionally, the current study did not cover the entire dataset of 29 cell types, which includes a broader range of tissues and potential markers.

4.4 Future Research Directions

Future research could explore and refine the designed architecture of cfSort. As mentioned in the analysis of the experiments the results should be further investigated. To ensure a more comprehensive evaluation of the deconvolution methods, future research should test UXM and cfSort across all 29 cell types. This expanded analysis would provide a more detailed understanding of the methods' performance and reliability in detecting various tissues.

4.5 General Remarks About Both Methods

A good codebase is crucial for the practical application of research ideas. While the concepts presented in the cfSort paper may be innovative, their value is significantly diminished if the code is not reproducible or well-maintained. Poorly documented and outdated code can lead to substantial time wasted trying to understand and implement the proposed methods. This issue was evident in our experience with cfSort which was intended to be a Deep Neural Network (DNN) for cell deconvolution but due to time constraints had to be changed to a simpler model. Despite its potential for better prediction accuracy, cfSort is essentially a black box model, offering no transparency in its decision-making process. This lack of explainability requires extensive reasoning and testing before such a model can be trusted in real-world scenarios, particularly in the field of healthcare

where the decisions can have great consequences.

In contrast, the reference-based UXM approach, which utilizes methylation patterns derived from tissues, offers a more straightforward and mathematically anchored solution. Although the need to provide the types for UXM can be a limitation, its transparency and explainability make it a more reliable choice, especially when making decisions related to human health. The ability to understand and justify how conclusions are reached in medical applications is often more important than the high accuracy a model can reach.

5 Responsible Research

5.1 Data Collection

The data diversity in our study is likely limited due to the origins of the samples. The samples are primarily sourced from various institutions in the USA, including the National Disease Research Interchange, Roswell Park Cancer Institute, Science Care, Inc., and the ELSI study at Virginia Commonwealth University. The GTEx project [2] is led by the Broad Institute of MIT and Harvard. This geographical concentration suggests that the data may not be very diversified, showing a predominantly US-based population. This limitation is difficult to address, as expanding the sample base internationally would require extensive cooperation and resources. Despite this, the data still provides valuable insights, but it's important to recognize the potential bias introduced by its limited diversity.

5.2 Use of Generative AI and Large Language Models

In our work, we used large language models primarily to help understand the existing code, especially the cfSort algorithm. The cfSort code lacked useful comments and documentation, making it hard to figure out what the functions did and how the algorithm functioned. By using large language models, we were able to explain the functions and get a clear picture of the algorithm's structure. We made sure not to feed any sensitive data into the model during this process.

5.3 Reproducibility

To ensure the reproducibility of our experiments:

- **Software and Tools**: All software and tools used are publicly available, and custom scripts are provided in the **GitLab Repository**[7].
- **Data Availability**: Datasets used in this study are publicly available [2].
- **Documentation**: Detailed documentation of procedures, including data processing pipelines and parameter settings, is included in the repository. The code is well-commented and the architecture is designed to be easily understandable.

[7]https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Reinders_Pronk_Hazelaar/showard-Detection-of-cancer-using-blood.git

6 Conclusion

This study set out to compare two methods of cell deconvolution, namely UXM and cfSort, in the context of detecting cancer using blood samples. Both UXM and cfSort have their strengths and weaknesses, with UXM performing better or the same as cfSort across all scenarios. Specifically, UXM consistently showed higher accuracy and lower variability in identifying cell types within mixed tissue samples. CfSort, while demonstrating potential in some cases, was hindered by its simple architecture. Future research should focus on addressing the main limitations identified in this study. These are, providing more extensive testing using more cell types and evaluating cfSort using its designed architecture. Our findings underscore the critical importance of both transparency and explainability in choosing deconvolution methods, especially in medical applications where accuracy and the ability to understand decision-making processes are crucial.

References

[1] Catherine Alix-Panabières, Dario Marchetti, and Julie E. Lang. Liquid biopsy: from concept to clinical application. *Scientific Reports*, 13(1):21685, 2023.

[2] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585, June 2013.

[3] S. Cristiano. Genome-wide cell-free dna fragmentation in patients with cancer. *Nature*, 570(7761):385–389, May 2019.

[4] David Crosby, Sangeeta Bhatia, Kevin M. Brindle, Lisa M. Coussens, Caroline Dive, Mark Emberton, Sadik Esener, Rebecca C. Fitzgerald, Sanjiv S. Gambhir, Peter Kuhn, Timothy R. Rebbeck, and Shankar Balasubramanian. Early detection of cancer. *Science*, 375(6586):eaay9040, 2022.

[5] Renaud G. et al. Unsupervised detection of fragment length signatures of circulating tumor dna using non-negative matrix factorization. *eLife*, 11, Jul. 2022.

[6] T. Hirahata, R. Ul Quraish, A. U. Quraish, S. Ul Quraish, M. Naz, and M. A. Razzaq. Liquid biopsy: A distinctive approach to the diagnosis and prognosis of cancer. *Cancer Informatics*, 21:11769351221076062, Feb 2022.

[7] Zeng W et al. Li S. Comprehensive tissue deconvolution of cell-free dna by deep learning for disease diagnosis and monitoring. *Proceedings of the National Academy of Sciences*, 120(28), Jul. 2023.

[8] Loyfer N., Magenheim J., and Peretz A. et al. A dna methylation atlas of normal human cell types. *Nature*, 613:355–364, 2023.

[9] Mary L. et al. Stackpole. Cost-effective methylome sequencing of cell-free dna for accurately detecting and locating cancer. *Nature Communications*, 13(1):5566, 2022.

A Appendix

A.1 Dataset types to UXM types mapping

```
dataset_to_uxm_map = {
 'WBC':
 ['Blood-B', 'Blood-Granul', 'Blood-Mono+
    Macro', 'Blood-NK', 'Blood-T'],
 'breast tissue':
 ['Breast-Basal-Ep', 'Breast-Luminal-Ep'],
 'colon tissue':
 ['Colon-Ep', 'Colon-Fibro'],
 'kidney tissue':
 ['Kidney-Ep'],
 'prostate tissue':
 ['Prostate-Ep'],
 'lung tissue':
 ['Lung-Ep-Alveo', 'Lung-Ep-Bron']
}
```

Figure 7: Cell types mapping. Each key of the dictionary is a type in the dataset mapped to a list of types that the UXM was trained to distinguish.

A.2 Mixture generation for the training of linear regression

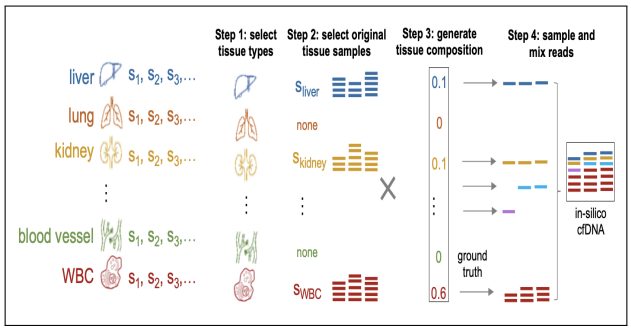

Figure 8: Mixture generation for training

In four steps, we generated a simulated sample. In Step 1, we first selected the tissue types that contributed positive fractions to the simulated sample. WBC always contributed positively to the final mixture. In Step 2, we chose an original tissue sample at random for each selected tissue type and WBC. In Step 3, we created a random tissue composition for the simulated sample. We set the tissue fraction to zero if a tissue type was not chosen in Step 1, and we required WBC to always have the highest tissue fraction. In Step 4, we sampled sequencing reads at random from the selected samples (from Step 2) based on tissue composition (generated in Step 3).