Understanding the influence of DNA fragment lengths in detecting cancer

Detection of cancer using blood

Monica-Alexandra Paun[1]

Supervisor(s): Marcel Reinders[1]**, Bram Pronk**[1]**, Daan Hazelaar**[2]**, Stavros Makrodimitris**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**
[2] **Erasmus University Medical Center**

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: Monica-Alexandra Paun
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Bram Pronk, Daan Hazelaar, Stavros Makrodimitris, Johan Pouwelse

Abstract

Detecting cancer at an initial stage could change the course of the disease's development. A non-invasive examination consists of the liquid biopsy of blood, revealing biomarkers that could provide information about the existence of a tumour or not in the organism. The research touches upon the relevance of DNA fragments, precisely the length of fragments, in the detection of cancer. An in-depth interpretation of the fragment length distribution for predicting the state of a patient as being healthy or sick with cancer was approached. The distribution was explored from four perspectives: the complete fragment length distribution, the size range from 90 to 150 bp, important lengths selected by the feature extraction methods and the Fourier Transform of the initial data. These were input in three machine learning models. Using the fragment lengths between 93 and 98 produced accuracy and AUC scores of over 0.85 for all supervised classification models. Processing the data with the Fourier Transform and using the amplitude of spectrums as features in the Random Forest model resulted in an AUC of 0.99.

1 Introduction

Cancer is defined in [1] as a disease that influences the uncontrollable multiplication of certain cells, which leads to their spreading into other organs. Cancer cells are the result of mutations in cells and may proliferate without control. Therefore they tend not to die when it would be the case, causing the development of the tumor. DNA fragments originated from cancerous cells and tumours end up in the bloodstream. The cancerous genetic material is referenced as circulating tumour DNA (ctDNA). The genetic changes found in cell-free DNA (cfDNA), characteristics of ctDNA, could have a significant implication in the detection of cancer [2].

An early detection of cancer could be a vital step in determining an effective treatment, according to [3]. An accessible method for detecting cancer would be the analysis of blood measurements. The patient is subject to a non-invasive investigation, a blood test, that can detect biomarkers helping in the diagnosis of cancer [4]. The relevant biomarkers collected from blood that can be observed are the DNA fragments. The study of the DNA fragment's characteristics contributes to the detection of cancer. The length of fragments is one of the features that give an insight into the classification of a blood sample from a healthy person or a patient with cancer.

The research intends to delve deeper into the understanding of fragmentomics features for helping detect cancer by classifying the patients into healthy or not with supervised and unsupervised learning models. In [5], the analysis of fragmentomics features suggests a difference in fragmentation patterns of cfDNA in the case of healthy persons and patients with cancer. Moreover, it is mentioned that the cancer's source tissue could be determined through the fragmentation profile. An unsupervised approach for detecting cancer is described in [6]. The inspection of fragment length patterns in cfDNA is realised using a non-negative matrix factorization (NMF) way. The understanding of the implication of fragmentomics features in classification is still lacking.

The analysis of the fragment length distribution in the classification of individuals as being healthy or as having cancer is tackled in this paper. This can be visualised in Figure 1 for a breast cancer sample and a control sample. The goal of this work is to compare various tumour detection approaches based on the fragment length distribution of cell-free DNA molecules. To this end, we aim to determine which features can we extract from the given distribution, and whether a simple binary rule could achieve good classification performance. Furthermore, we will also investigate what machine learning models can be used in detection, and for which type of cancer the optimal approach performs better.

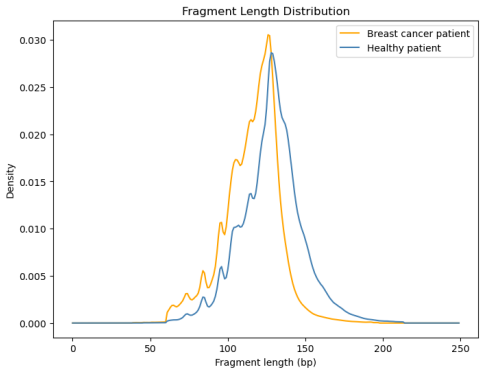

Figure 1: The fragment length distribution for a breast cancer sample and a control sample.

The paper emphasises on the study of fragment length distribution in the classification of patients. The research work includes a processing of the data into a format that could reflect the distribution. An exploration of the interpretation of the distribution is establish. For concluding the best approach of extracting the information, a comparison of the performance of models taking different inputs is presented.

Section 2 presents a description of the design decisions alongside the selected algorithms. An analysis of the results is expressed in Section 3. Further discussions are reported in Section 4. Section 5 raises the ethical aspects of the study and the issues that can be encountered when reproducing the experiment. Finally, Section 6 summarizes the work presented in the paper, and provides indications for future work.

2 Methodology

Research Approach

To address the research question, an analysis of the influence of the fragment length distribution in classification is conducted. Four approaches that determine the information to extract from the distribution are studied for understanding the effect of fragments' lengths in detecting cancer. The first idea was to assess the entire fragment length distribution. Next, the size range from 90 to 150 bp was considered, because of its relevance for ctDNA [7]. Further, an exploration of the most important features' estimated effectiveness was addressed. Lastly, the presence of oscillatory patterns in the distribution [8] motivated the implication of the Fourier Transform in the processing of the initial data. The complete distribution of each sample was input into the fft function from the NumPy library, resulting in a one-dimensional discrete Fourier Transform computed with the Fast Fourier Transform algorithm [9]. The absolute values of the transformations, representing the amplitude of the spectrums, were determined as features for the classification models.

To evaluate the significance of the four approaches in the detection of cancer, supervised and unsupervised learning models are compared. Starting from the supported literature, the Random Forest and NMF methods were considered. Additionally, the evaluation was conducted by adopting a naive classification decision and an SVM model. The performance was determined by the accuracy and the area under the receiver operating characteristic curve (ROC-AUC).

Dataset

The dataset provided for this research is the same as the one handled in [5]. The data consists of samples of plasma from healthy patients and patients with cancer. The types of cancer present in the data are breast, lung and colorectal. The 252 samples include 104 healthy patients, 49 breast cancer patients, 22 colorectal cancer patients and 77 lung cancer patients.

Each representative in the data was in the form of a Binary Alignment Map (.bam) file. A bam file represents the sequence alignment in a binary format. A processing step of these files was necessary so that the information could be human-readable. The files were transformed into a text file, consisting of the insert size metrics and the histogram data that is represented by the insert size and its corresponding number of read pairs. The conversion was achieved with the CollectInsertSizeMetrics tool from Picard [10]. Then, the information from the text files was collected in a comma-separated value (CSV) file. For gathering the data, the minimum and maximum length of all the samples was established. In the CSV file, each sample has a column with its name, its label (1 for cancer and 0 for healthy) and columns defining all the lengths in the range from the minimum to the maximum length. The lengths columns have as value the density, which was computed as the number of read pairs for a size over the total number of read pairs, or 0 if the length is missing for the sample. The size of the data features is 250 since the minimum length determined was 38 and the maximum 287.

The dataset was split into 67% training data and 33% testing data. The decision to make the split only into train and test sets was supported by the action of tuning the hyperparameters with RandomizedSearchCV from the scikit-learn library. The RandomizedSearchCV method explores different sets of parameters through cross-validation that could optimize the learning models' scoring. AUC score was chosen to be maximised for this research.

Feature Importance

Selecting a specific set of features could provide more insights into the classification of plasma samples into healthy or cancerous. Three selection methods were proposed for identifying the lengths from the complete distribution, and the frequencies from the amplitude of the spectrum obtained with Fourier. The selected features could be informative for the classification problem. The techniques were Recursive feature elimination with cross-validation (RFECV), SelectKBest and *feature_importances_* function from Random Forest model.

One of the methods presented by the scikit-learn library for selecting the important features is RFECV. Features with a small impact on the classification are recursively eliminated, and using cross-validation, the optimal collection of features is decided. A Random Forest model with the default hyperparameters and random state settled was initialized for being the estimator instance of RFECV.

SelectKBest is a feature selection approach from the scikit-learn library applied in supervised models. The method picks the k features with the highest score retrieved from an univariate statistical test. The number of dominant features, k, was chosen to be equal to six as that was the set's size output by RFECV. Mutual information was preferred as a test score for the relationship between the lengths. The problem that is solved in the research is a classification problem, treating sparse data, so the two test options were *chi-squared* and *mutual information*. *Chi-squared* was discarded because of its suitability to categorical data, a property not covered by the dataset used in the research.

The Random Forest classifier provides a method called *feature_importances_* that determines the significance of a feature in the classification process by computing the Gini importance. The model has to be trained so that the method can be accessible. As for the RFECV, the Random Forest model had set the default hyperparameters and random state.

Baseline Model

Two simple binary classification rules, one focused on the entire distribution and the other for only a specific set of lengths, were implemented as a baseline. The first step of the model for the complete distribution is to compute the mean fragment length for each data sample, which is then normalized. After the train-test split of the data, the

normalized mean lengths of the samples from the train set are transformed in the range (0, 1) using the sigmoid function, $1 - \frac{1}{1+e^{-x}}$ where x represents the mean. The threshold is established from the ROC curve with Youden's J statistic [11]. The probability of the samples from the testing set is calculated with the aforementioned formula. If the probability is greater or equal to the threshold then the plasma sample is noted to be from a patient with cancer.

A second classification rule was formulated for the size range 90-150 bp, the lengths with the most importance and the Fourier Transform features. From the training data, the mean density of cancer and healthy patients was computed, setting the threshold of the classification in the middle of the two. A sample is classified as cancerous if the mean density of the lengths is greater than the threshold.

Support Vector Machine Model

To separate different classes, SVM produces a hyper-plane or a set of hyper-planes that aim to maximize the margin. In the experiments, the C-Support Vector Classification (SVC) class from the scikit-learn library was used. The tuning of the hyperparameters was realized with RandomizedSearchCV and having as parameters setting the following values:

- Regularization parameter - $uniform(0.01, 100)$
- Kernel type - $\{'linear', 'poly', 'rbf', 'sigmoid'\}$
- Degree of the polynomial kernel function - $randint(2, 5)$
- Kernel coefficient - $\{'scale', 'auto'\}$
- coef0 - $uniform(0, 1)$

Random Forest Model

Random Forest improves the predictive accuracy by fitting multiple decision trees over subsets of data. The set of possible parameters of the RandomForestClassifier from which the model tuning selected the optimal values was as follows:

- Number of trees in the forest - $randint(100, 2000)$
- Function that measures the quality of a split - $\{'gini', 'entropy', 'log\_loss'\}$
- Maximum depth of a tree - $randint(10, 110)$
- Minimum number of samples for splitting an internal node - $randint(2, 10)$
- Minimum number of samples needed to be at a leaf node - $randint(1, 5)$
- Number of features to consider for the best split - $\{'sqrt', 'log2'\}$
- Bootstrap sample - $\{True, False\}$

Non-Negative Matrix Factorization

An unsupervised approach was implemented to have a sense of the performance of the two ways. The classification method had the focus on NMF as described in [6]. Since a binary classification problem needs to be solved, we have decided to use NMF with two components. In this way, two weighted vectors and two signatures were resulted. The reasoning for choosing two components is that we can now associate the signatures with the two labels: cancerous, or healthy. Given the assumption that the signature with a lower mean fragment length is cancer-related, the two weights of the weight matrix were compared. Assuming that the weight corresponding to the cancer signature is greater, the sample is classified as originating from a patient with cancer.

Furthermore, the classification was accomplished based on a threshold chosen from the ROC curve as well. After the two weighted vectors and the two signatures were returned, the probability of samples being cancerous was computed. Having the probabilities, the threshold was picked from the ROC curve using Youden's J index [11]. The sample was characterized as cancerous if the probability was greater or equal to the threshold.

3 Results

Data Characteristics

The data employed in the research was analyzed to identify distinctive attributes. Computing the mean of the fragment length distribution for the two classes facilitated the understanding of the data trends. Figure 2 presents the mean fragment length distribution captured for cancer and healthy samples. The average length of data originating from the patients with cancer is 165 bp, smaller than for the healthy data which has a value of 167 bp. Furthermore, in Figure 2 the fragments with size between 80 and 160 bp seem to be more common in cancer data as opposed to fragments ranging from 180 to 230 bp. The healthy data exposes a mean density of 0.0039 in the range from 80 to 160 bp, and as for the other class, the mean density is 0.0046. The values of the mean density between 180 and 230 bp for the cancer and healthy samples are 0.0034, and 0.0043 respectively. According to [5], the short fragments have lengths from 100 to 150 bp while the long fragments have lengths from 150 to 220 bp. In light of this categorization, the range size of short fragments is prevalent in cancer samples.

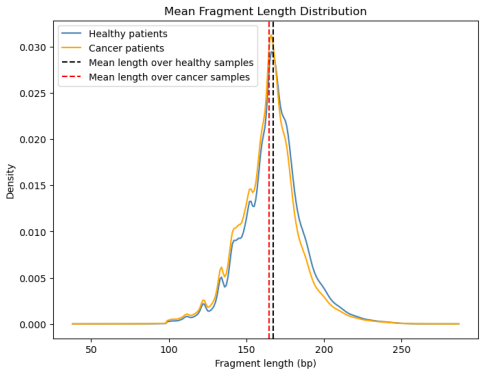

Figure 2: The mean fragment length distribution over the cancer and healthy samples.

Feature Importance Results

The results of the three feature selection methods described in Section 2 for the initial complete fragment length distribution are outlined next. RFECV selected Length 93-98 as being valuable features in classification. The second approach, SelectKBest, highlighted Length 92-98. The *feature_importances_* attribute of the Random Forest classifier returned the list of lengths in decreasing order based on the Gini importance. The first 10 features with the highest score resulted from the *feature_importances_* method were plotted, Figure 3. The features derived from the three approaches have a set of common lengths that was decided to be used as the third possibility of features for the detection of cancer. As noticed the frequent set of lengths is Length 93-98.

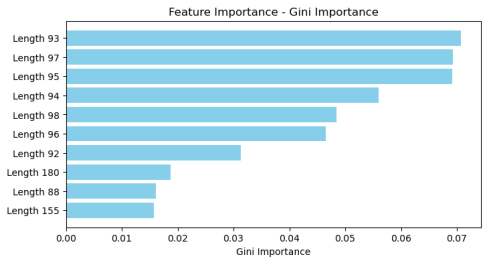

Figure 3: The first 10 most important features extracted by the *feature_importances_* model from Random Forest.

The determination of these specific lengths could have been influenced by their distinct characteristic for the two classes. Figure 4 illustrates through a boxplot a comprehension of data collected from cancer and healthy patients. The boxplot depicts the distribution and any outliers of cancer and healthy data for the lengths chosen in the feature importance step. It can be observed that there is a clear separation between the cancer and healthy patients data. There is a consistently higher median and interquartile range for the cancer samples.

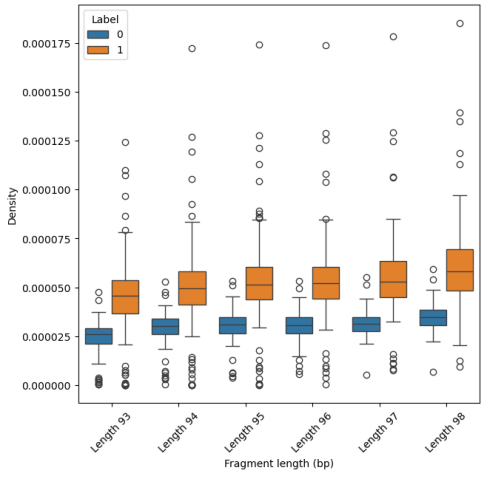

Figure 4: Comparison of cancer (label 1) data and healthy (label 0) data for the lengths selected through feature importance methods.

The feature selection techniques were applied to the amplitude of spectrums from the Fourier Transform likewise. The RFECV established a set of 136 frequencies to be informative in the classification, more than half of the feature set's size. A larger set of Fourier Transform features imply to capture variations in data that support the classification problem. Considering the first six features with the highest Gini importance derived from the *feature_importances_* and the SelectKBest methods, the same collection of frequencies was revealed: 107, 108, 115, 135, 142 and 143.

Evaluation

To understand the importance of the fragment lengths distribution and the specific characteristics they present for detecting cancer, an evaluation of the four approaches listed in Section 2 was conducted. The sets of features representing all lengths, the lengths in size 90 to 150 bp, the important lengths selected and the DFT amplitude spectrums were analysed against the baseline model, the SVM model and the Random Forest model. The NMF method made use of the whole distribution. The performance was measured with regards to two global measures of diagnostic accuracy [12], accuracy and AUC. The two were taken into consideration, because of their different goals. The accuracy metric outputs the percentage of correct predictions, while AUC provides insights into measuring the model sensitivity and specificity.

Baseline Classification Model

The performance of the benchmark predictive model was compared between the four types of information taken from the data distribution and the results can be viewed in Table 1. A significant difference was noticed. The lengths selected by the feature importance methods (lengths 93-98) performed better for this model in comparison to the set of amplitude spectrums. The results of the performance for the complete distribution and the set of lengths in the range from 90 to 150 bp seem to be relatively close. The ROC curves for the four setups are visible in Figure 5.

	Accuracy	AUC
Complete Distribution	0.75	0.795
Range 90 - 150 bp	0.702	0.767
Important Lengths	0.857	0.910
Amplitude Spectrums	0.666	0.683

Table 1: Results obtained after performing the classification with the baseline model.

Support Vector Machine Model

The four approaches for identifying valuable information from the distribution were input in the SVM model, results being present in Table 2. The classifier had a similar performance when using unprocessed data from the initial dataset. Applying the Fourier Transform to the data collection could have made it harder for the model to differentiate between the two classes. As observed in Figure 6 the set of amplitude spectrums presents an AUC score lower than the others with 0.1.

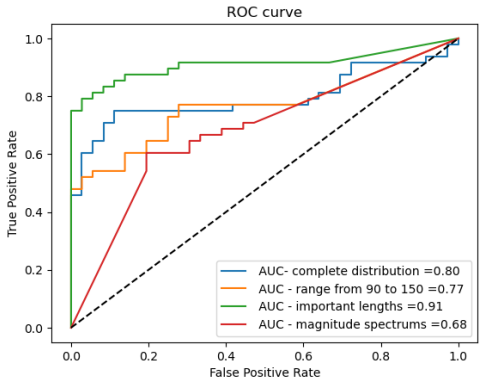

Figure 5: ROC curve of baseline model having different information from the data.

	Accuracy	AUC
Complete Distribution	0.892	0.965
Range 90 - 150 bp	0.869	0.962
Important Lengths	0.892	0.968
Amplitude Spectrums	0.809	0.872

Table 2: Results obtained after performing the classification with the SVM model.

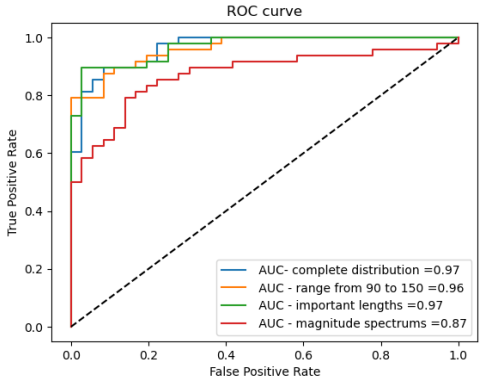

Figure 6: ROC curve of SVM model having different information from the data.

Random Forest Model

The sets of possible features were analysed with the Random Forest model as well, Table 3. The best performance in terms of both accuracy and AUC score was achieved by the set of features represented by the amplitude spectrums. The other three approaches have an accuracy above 0.9 and an AUC score above 0.95, relatively close to each other. Overall, the model tends to capture more informative data from the set of amplitude spectrums. Figure 7 illustrates the sensitivity and specificity of the model in the report with the four sets of features.

Non-Negative Matrix Factorization

A difference in AUC score was noticed when implementing the NMF model as stated in [6] (AUC = 0.742) for the set of

	Accuracy	AUC
Complete Distribution	0.916	0.989
Range 90 - 150 bp	0.916	0.985
Important Lengths	0.916	0.966
Amplitude Spectrums	0.940	0.986

Table 3: Results obtained after performing the classification with the Random Forest model.

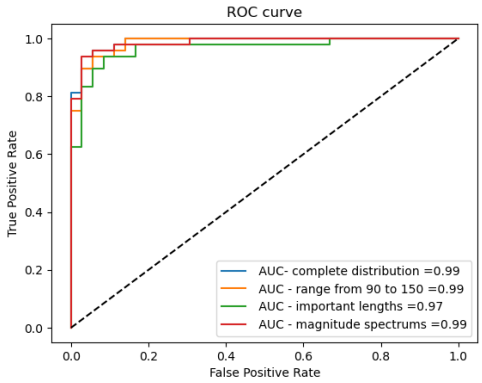

Figure 7: ROC curve of Random Forest model having different information from the data.

data used in this research. The same data samples were engaged, however with a lack of five types of cancer for the set of this research. That could lead to the two dissimilar scores. The evaluation of the NMF model with a change in choosing the threshold from the ROC curve for the predictive task was done. In Table 4 a distinction in accuracy can be observed. The value of the threshold after computing the ROC between the true labels and probabilities of samples from the train data to be cancerous was 0.47, Figure 8.

	Accuracy	AUC
NMF from [6]	0.761	0.808
NMF using threshold from ROC curve	0.821	0.81

Table 4: Results obtained after performing the classification with the NMF.

Evaluation of the Optimal Setting for Each Type of Cancer

The Random Forest model with the amplitude of spectrums as features for classification was the setting with the optimal performance. Further analysis of its behaviour was done for the three types of cancer available in the dataset, Table 5. The model accurately predicted for each group whether the sample belongs to a healthy or a cancer patient as seen in Figure 9. No additional conclusion can be taken since the size of data for each cancer type and for healthy is not balanced.

4 Discussion

The comparisons investigated in Section 3 could lead to an understanding of the fragment lengths influence in the

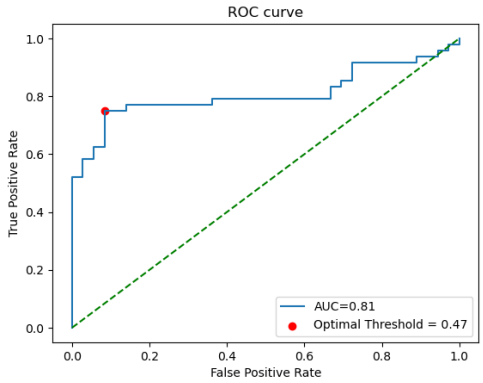

Figure 8: ROC curve for the NMF that is using the threshold from the curve.

	Accuracy	AUC
Breast Cancer	0.941	0.991
Colorectal Cancer	0.928	0.915
Lung Cancer Cancer	0.950	0.980

Table 5: Results obtained after performing the classification with the Random Forest model and amplitude of spectrums as data features. The classification of samples into healthy or cancerous was performed on each type of cancer.

detection of cancer in the blood samples. Using the set of lengths resulting from the feature selection methods seemed to have a notable performance improvement over models that use the complete distribution. This implementation manages to achieve an accuracy and AUC score above 0.85. However, the set of data features represented by the amplitude of spectrums obtained from the Fourier Transform had a higher result when the predictive task was executed by the Random Forest model. It can be reasoned that using a Fourier Transform for the preprocessing step and the Random Forest model is the most favorable for the classification of plasma samples into healthy or cancer.

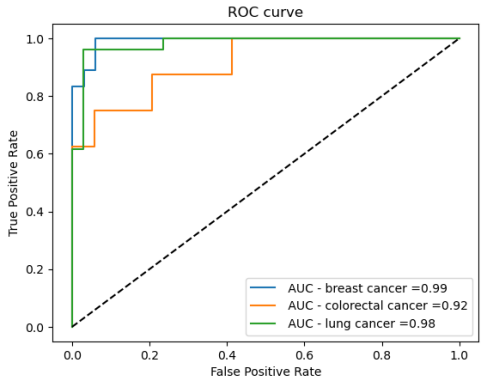

Figure 9: ROC curve of Random Forest model and amplitude of spectrums as data features. The classification of samples into healthy or cancerous was performed on each type of cancer.

In a previous work [8] the prediction of tumor content using statistical learning was employed with Fourier and wavelet transforms for a particular range of fragment lengths. The main distinction between the existing work and the one done in this research is the use of the Fourier transform in processing the distribution. It was considered that a study of the amplitude of spectrums captured from the Fourier Transform for the complete distribution would be relevant. This approach could bring more insights than only limiting to a smaller range. The use of complete distribution provides an understanding of the data's pattern for the two classes. Observation referring to the different characteristics of cancer and healthy data noticed in Figure 2 supports the decision of analysing the entire range of lengths.

The work proposed in [7] concluded that an enhancement in fragments with lengths from 90 to 150 bp is distinguishable for the ctDNA. The Random Forest model performed better compared to the other models for this range. Furthermore, an interesting finding was that the lengths resulting from the feature selection methods lie between 90 and 150 bp. The significant lengths for the classification were from 93 to 98 bp, which demonstrated a generally favorable performance when input into the learning models considered. The selection of sizes from 93 to 98 bp could be due to the altered genes representative of cancer patients. The genomic regions bound by the regulatory proteins can be determined by the fragments with lengths smaller than 100 bp [13]. An important factor between cancer and healthy patients is the presence of Tumor Protein 53 (TP53), mutated gene p53 characteristic in human cancer [14]. p53 is a regulatory protein that controls the cell cycle and suppresses the tumor.

5 Responsible Research

The ethical aspects of the research and the reproducibility of the methods should be brought to attention to ensure a responsible research process. The description of the experiment in the previous sections should give the reader guidance for reproducing the steps of the research. The conclusions stated in the paper could have an impact on the diagnosis of patients, and thus a thorough discussion regarding the ethical implications has to be mentioned. A critical reflection on the ethical considerations and methodological reproducibility ensures the integrity and transparency of the research.

The dataset used for performing the experiments is the one that was employed in [5]. The data was stored in the database of Genotypes and Phenotypes, from where it was retrieved for the research of cancer detection. In [5] is specified that the samples were collected under Institutional Review Board protocols. Furthermore, all the participants gave their consent to have their blood samples taken for research purposes. The samples were provided anonymized and no correlation with the donor can be realized. No other extra assessments of the data quality were concluded during the research.

The detailed description of the methods is presented in Sec-

tion 2, ensuring the reproducibility of those. However, slight differences can be noticed in results, since machine learning experiments are naturally stochastic. It is important to maintain the hyperparameters set along the research the same when reproducing the experiment, to obtain similar results. A processing of the data into a format that could be inserted into the learning models was necessary, a report of the approach realized to accomplish the transformation is found in Section 2.

6 Conclusions and Future Work

The research aimed to fill the knowledge gap regarding the implication of fragmentomics features in the detection of cancer. The work proposed in this paper highlighted the understanding of the complete fragment length distribution's influence in predicting the origin of liquid biopsy samples. Considerations of information derived from the distribution were evaluated against a baseline classification, SVM, Random Forest, and NMF models. The complete fragment length distribution, the size range from 90 to 150 bp, the set of lengths resulting from the feature extraction, and the Fourier Transform's amplitude of spectrums were compared.

After performing the experiments, it was concluded that using the lengths extracted by the feature selection methods gave a considerable performance boost for all three supervised machine learning models. The cancer and healthy data presents representative characteristics for this specific set of lengths ranging from 93 to 98 bp. These differences between the two classes impact the classification task. Nonetheless, the Random Forest classifier with the amplitude of spectrums had the best performance with an accuracy of 0.94 and an AUC score of 0.99.

The optimal setting was analysed against each cancer type available. The classification of blood samples into healthy or cancer was fairly precise. However, because of the imbalance between the size of healthy data and the size of each type of cancer, a detailed conclusion cannot be drawn. The absence of more data samples for the already available types of cancer was considered to be one of the study's limitations. Additionally, a broad dataset with more various types could give a more accurate interpretation of the models' behaviour. Finally, an in-depth analysis of the implication of the Fourier Transform in the prediction of blood samples would be recommended for future research.

References

[1] "What is cancer?." https://www.cancer.gov/about-cancer/understanding/what-is-cancer, 2021. 08 May 2024.

[2] M. Cisneros-Villanueva, L. Hidalgo-Pérez, M. Rios-Romero, A. Cedro-Tanda, C. Ruiz-Villavicencio, K. Page, R. Hastings, D. Fernandez-Garcia, R. Allsopp, M. Fonseca-Montaño, *et al.*, "Cell-free dna analysis in current cancer clinical trials: a review," *British journal of cancer*, vol. 126, no. 3, pp. 391–400, 2022.

[3] W. H. Organization *et al.*, *Guide to early cancer diagnosis*. World Health Organization, 2017.

[4] S. Das, M. K. Dey, R. Devireddy, and M. R. Gartia, "Biomarkers in cancer detection, diagnosis, and prognosis," *Sensors*, vol. 24, no. 1, p. 37, 2023.

[5] S. Cristiano, A. Leal, J. Phallen, J. Fiksel, V. Adleff, D. C. Bruhm, S. Ø. Jensen, J. E. Medina, C. Hruban, J. R. White, *et al.*, "Genome-wide cell-free dna fragmentation in patients with cancer," *Nature*, vol. 570, no. 7761, pp. 385–389, 2019.

[6] G. Renaud, M. Nørgaard, J. Lindberg, H. Grönberg, B. De Laere, J. B. Jensen, M. Borre, C. L. Andersen, K. D. Sørensen, L. Maretty, *et al.*, "Unsupervised detection of fragment length signatures of circulating tumor dna using non-negative matrix factorization," *Elife*, vol. 11, p. e71569, 2022.

[7] F. Mouliere, D. Chandrananda, A. M. Piskorz, E. K. Moore, J. Morris, L. B. Ahlborn, R. Mair, T. Goranova, F. Marass, K. Heider, *et al.*, "Enhanced detection of circulating tumor dna by fragment size analysis," *Science translational medicine*, vol. 10, no. 466, p. eaat4921, 2018.

[8] M. Cardner, F. Marass, E. Gedvilaite, J. L. Yang, D. W. Tsui, and N. Beerenwinkel, "Predicting tumour content of liquid biopsies from cell-free dna," *BMC bioinformatics*, vol. 24, no. 1, p. 368, 2023.

[9] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of computation*, vol. 19, no. 90, pp. 297–301, 1965.

[10] "Picard toolkit." https://broadinstitute.github.io/picard/, 2019.

[11] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[12] A.-M. Šimundić, "Measures of diagnostic accuracy: basic definitions," *ejifcc*, vol. 19, no. 4, p. 203, 2009.

[13] I. Hudecova, C. G. Smith, R. Hänsel-Hertsch, C. S. Chilamakuri, J. A. Morris, A. Vijayaraghavan, K. Heider, D. Chandrananda, W. N. Cooper, D. Gale, *et al.*, "Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free dna," *Genome research*, vol. 32, no. 2, pp. 215–227, 2022.

[14] S. Tsai and T. C. Gamblin, "Molecular characteristics of biliary tract and primary liver tumors," *Surg. Oncol. Clin. N. Am*, vol. 28, no. 4, pp. 685–693, 2019.