UNDERSTANDING THE INFLUENCE OF DNA FRAGMENT LENGTHS IN DETECTING CANCER

DETECTION OF CANCER USING BLOOD

AUTHOR

Monica-Alexandra Paun <m.a.paun@student.tudelft.nl>

RESPONSIBLE PROFESSOR & SUPERVISORS

Prof. Dr. Ir. Marcel Reinders
Bram Pronk
Daan Hazelaar
Stavros Makrodimitris

TU Delft Delft University of Technology

INTRODUCTION

- An early detection of cancer could be a vital step in determining an effective treatment.
- An accessible method for detecting cancer would be the analysis of liquid biopsy.
- Study of the DNA fragments characteristics (fragmentomics) contribute in detection of cancer.

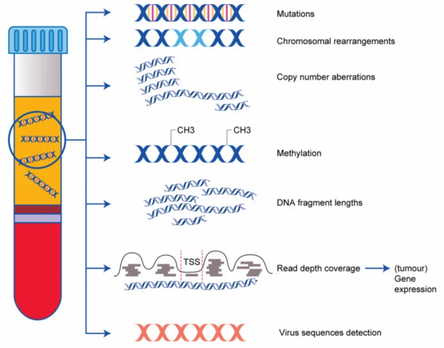


Figure 1: ctDNA features and their relevant clinical importance [1]

OBJECTIVE

Understanding of the fragment length distribution in detection of cancer.

- Compare various tumour detection approaches based on the fragment length distribution of cell-free DNA molecules.
- Determine which features can be extracted from the given distribution, and whether a simple binary rule could achieve good classification performance.
- Investigate what machine learning models can be used in detection, and for which type of cancer the optimal approach performs better.

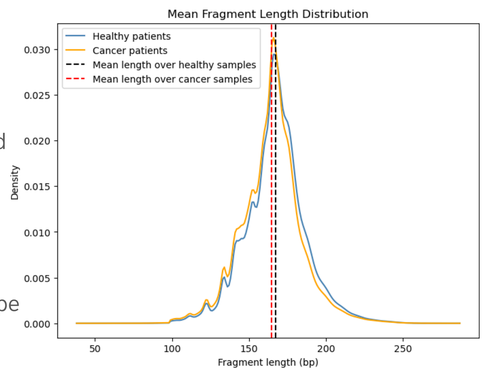
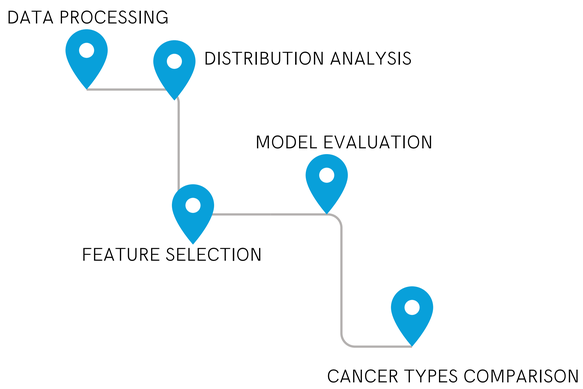


Figure 2: The mean fragment length distribution over the cancer and healthy samples

METHODOLOGY



Experimental setup:

- Processing of the initial data (104 control & 148 cancer samples) that was in the form of a Binary Alignment Map (.bam) file into a human-readable format.
- The distribution was explored from four perspectives: the complete fragment length distribution, the size range from 90 to 150 bp, the set of important lengths and the amplitude of spectrums obtained from the Fourier Transform .
- Selecting the set of features that could provide more insights for the classification task.
- The four analysis of distribution were evaluated against a naive model, the SVM model and the Random Forest model. An NMF model proposed in [2] was compared with the others as well.

RESULTS

Feature importance:

- The features derived from the three approaches have a set of common lengths (Length 93 - 98) that was decided to be used as the third approach for the detection of cancer. A clear separation between the cancer and healthy patients data for this range can be noticed in Figure 3.
- The RFECV established a set of 136 frequencies to be informative in the classification, more than half of the feature set's size.

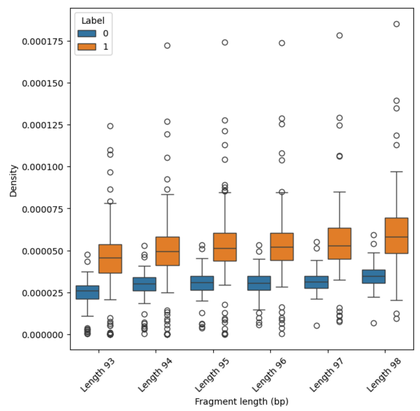


Figure 3: Comparison of cancer (label 1) data and healthy (label 0) data for the lengths selected through feature importance methods

	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

	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

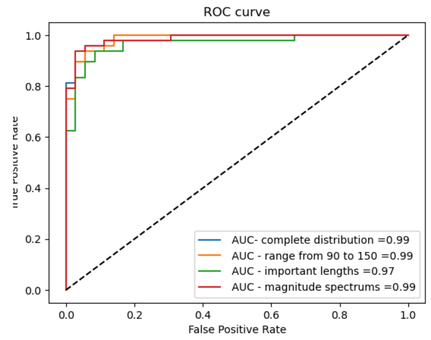


Figure 4: ROC curve of Random Forest model having different information from the data

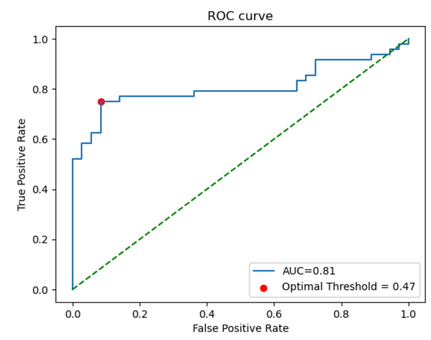


Figure 5: ROC curve for the NMF that is using the threshold from the curve

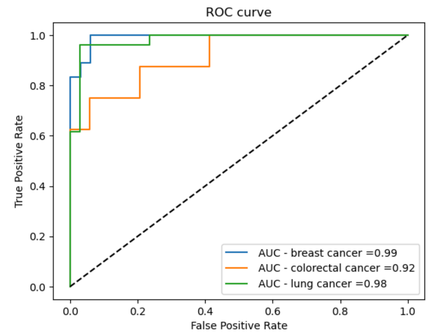


Figure 6: ROC curve of Random Forest model and amplitude of spectrums as data features. The classification of samples into healthy or cancerous was performed on each type of cancer

DISCUSSION

- 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 with an accuracy and AUC score above 0.85.
- 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.
- An interesting finding was that the lengths resulting from the feature selection methods lie between 90 and 150 bp, size range specific for ctDNA [3]. This selection could be due to the altered genes representative for cancer patients.

CONCLUSION

- The amplitude of spectrums resulted after applying Fourier Transform to the distribution had an outstanding result when input into Random Forest.
- A broad dataset could give a more accurate interpretation of the models' behaviour.
- An in-depth analysis of the implication of the Fourier Transform in the prediction of blood samples would be recommended for future research.

REFRENCES

[1] Keller, L., Belloum, Y., Wikman, H. et al. "Clinical relevance of blood-based ctDNA analysis: mutation detection and beyond." Br J Cancer 124, 345–358 (2021). https://doi.org/10.1038/s41416-020-01047-5.

[2] G. Renaud, M. Nørgaard, J. Lindberg, H. Gronberg, ¨ 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.

[3] 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.

[4] 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.