Machine learning detection of cancer in liquid biopsies

Pascal Bärtschi and Alina Martin
Department of Quantitative Biomedicine, University of Zurich

Introduction

Analysing **cell-free DNA** (cfDNA) in liquid biopsies is a cutting edge approach in cancer care. The analysis of cfDNA fragmentation patterns has been shown to provide additional information compared to mutation-based approaches like somatic mutation calling or copy number analysis (CNA).
cfDNA fragmentation patterns provide information about the **presence** and the **origin of the tumour**, which could improve diagnosis, targeted therapy and non-invasive disease monitoring[1].

The aim of this project is to develop a **machine learning model** to distinguish healthy from cancer samples.

Materials and Methods

Fragment length, copy number and nucleosome footprint features extracted from whole-genome cfDNA sequencing BAM files of a publicly available dataset[2] were used to classify healthy and cancer samples (Fig. 1) .

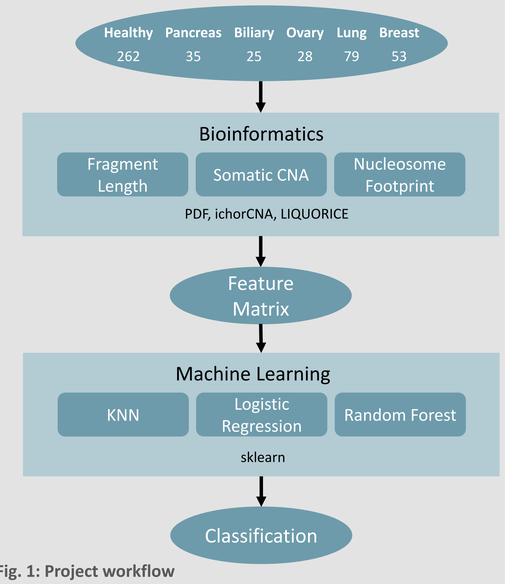

Fig. 1: Project workflow

First, the data was explored in a bioinformatic analysis involving visualizing fragment length distributions, detection of somatic CNA (ichorCNA[3]) and determination of nucleosome footprints (LIQUORICE[4]). The resulting feature matrix was then fed into different machine learning models. Cross validation (CV) was used to find the best performing model and to filter for the most important features (sklearn[5]) to solve the binary classification problem.

Results

Obtained feature matrix results from exploring characteristics of the following features: Fragment length distribution (Fig. 2), CNA and tumour fraction (Fig. 3) and nucleosome footprints (Fig. 4). Cancer samples show different characteristics.

Increased amount of shorter fragments

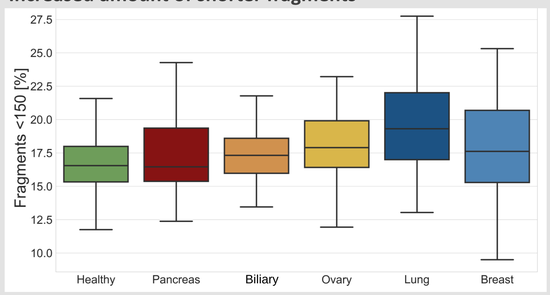

Fig. 2: Percentage of fragments <150 bp per diagnosis. cfDNA fragments of cancer samples display higher percentage of short fragments compared to healthy samples.

CNAs and higher tumour fractions

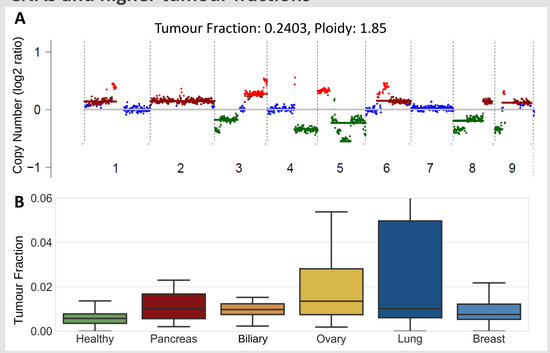

Fig. 3: Output of ichorCNA of a lung cancer sample. A: ichorCNA estimates tumour fraction and ploidy and predicts CNAs (green/ red). **B**: Most of the samples do not display increased tumour fractions compared to healthy controls.

Altered coverage profiles

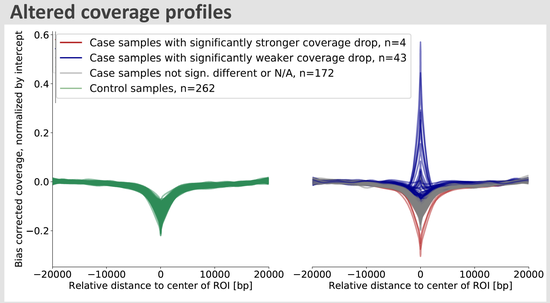

Fig. 4: LIQUORICE output. Overlay plot for hematopoietic specific region-set. Compared to healthy samples, some cancer samples display increased coverage indicating silenced gene regions, while others show decreased coverage indicating increased expression.

To find the best performing model, 5-fold cross validation on a 80/20 train-test split was used while minimizing overfitting. Selected model performs logistic regression and classifies upon fragment length features.

Important features of the fragment length distribution

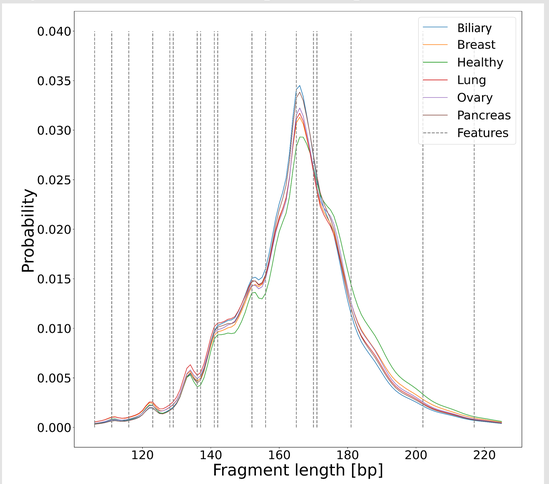

Fig. 5: Fragment length distribution. Fragment lengths features selected by Lasso penalty (C value) of logistic regression. Typically found close to local maxima and minima.

Logistic regression model yields highest F1 scores

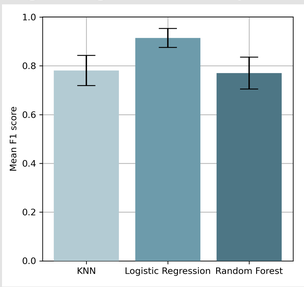

Fig. 6: Mean and standard deviation of the macro F1 scores across folds (5-fold CV) for different models. Logistic regression (0.91) outperformed Random Forest (0.79) and K-nearest neighbours (0.78).

Optimal performance with C value of 0.17

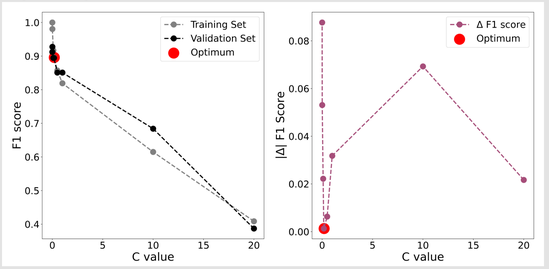

Fig. 7: F1 score values of logistic regression against C value to avoid overfitting. The F1 score drops with increasing C value. The best C being 0.17 is found by determining optimal trade-off between |Δ| performance and F1 score.

Performing logistic regression revealed that the simplest model, which only included information of fragment length, produced the highest F1 scores. Including information on tumour fraction and nucleosome footprint information did not improve the model (Fig. 8).

Final model only uses fragment length features

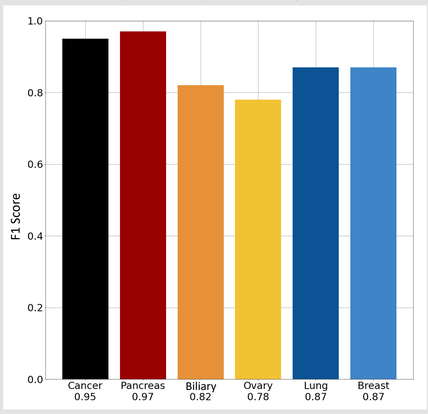

Fig. 8: F1 scores obtained by final model. Logistic regression on previously unseen test data, with post-hoc computation of performance score for each type of cancer.

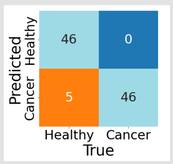

Fig. 9. Confusion Matrix of Cancer classification model. n = 97. 46/46 cancer samples identified (true positives), 5/51 healthy samples misdiagnosed (false positive).

Conclusion

The results show that **logistic regression** when focusing on the relative concentration of specific **length features** allows for better classification than using other algorithms.
Reliable performance of the weakly penalised model in detecting cancer and its subtypes shows the advantage of its **simplicity**. Nevertheless, the model should still be treated with caution, as **overfitting** is difficult to rule out, even if crossvalidation has been preformed to minimize it.

References

1. Lo, Y. M. D. et al. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372**, eaaw3616. (2021). https://doi.org/10.1126/science.aaw3616
2. Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019). https://doi.org/10.1038/s41586-019-1272-6
3. Adalsteinsson, V.A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, 1324 (2017). https://doi.org/10.1038/s41467-017-00965-y
4. Peneder, P. et al. LIQUORICE: detection of epigenetic signatures in liquid biopsies based on whole-genome sequencing data, *Bioinformatics Advances* **2** (2022) vbac017, https://doi.org/10.1093/bioadv/vbac017
5. Buitinck, L. et al. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 108–22 (2013)