# KEY FRAGMENTOMICS FEATURES FOR CANCER DETECTION

**Author**
David-Ștefan Peța
peta-1@student.tudelft.nl

**Responsible Professor**
Prof. dr. ir. Marcel Reinders

**Supervisors**
Bram Pronk
Stavros Makrodimitris
Daan Hazelaar

**References**
[1]P. Peneder, A. StÃŒtz, D. Surdez, et al., "Multimodal analysis of cell-free dna whole-genome sequencing for pediatric cancers with low mutational burden,"
[2] S. Cristiano, A. Leal, J. Phallen, et al., "Genome-wide cell-free dna fragmentation in patients with cancer,"
[3] G. Renaud, M. NÃžrgaard, J. Lindberg, et al., "Unsupervised detection of fragment length signatures of circulating tumor dna using non-negative matrix factorization,"

## 1 BACKGROUND

- Cancer poses significant challenges for patients and researchers due to its widespread prevalence and complexity
- Studying circulating DNA fragments in the bloodstream of individuals with cancer emerged as a promising path in cancer investigation - **Fragmentomics**
- Literature proved that fragmentomics features offer great insights into cancer detection, origin and treatment response [1], [2], [3]

## 2 RESEARCH QUESTION

**Which fragmentomics features are most important for cancer detection?**

## 3 OBJECTIVE

- **Apply feature importance and selection techniques to obtain the most important fragmentomics features from the available data**
- Understanding the level of importance of these features can lead to improved diagnostic tools such as cancer detection based on blood tests

## 4 METHODOLOGY

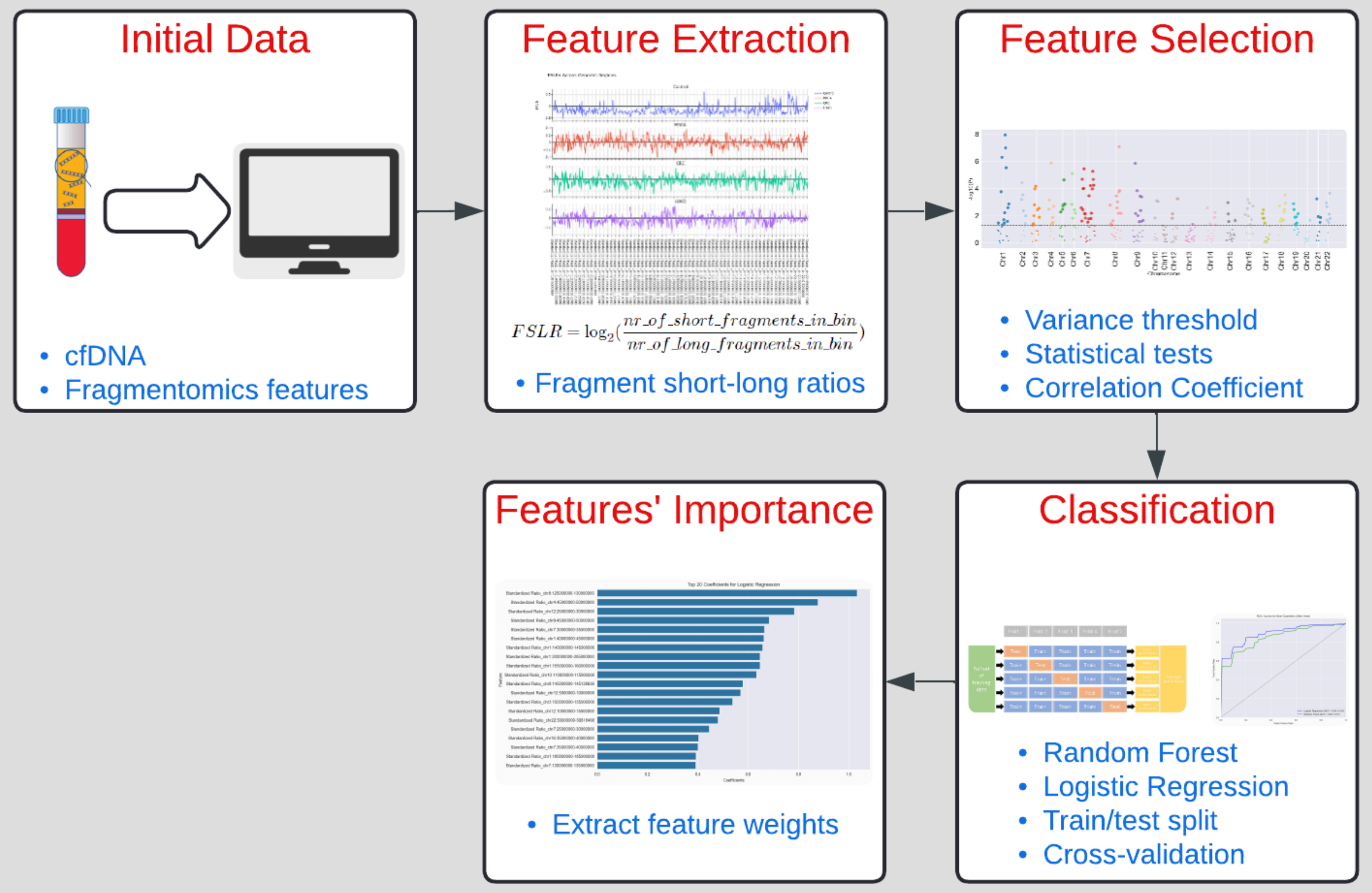

Figure 1: Pipeline describing the method used in this research

## 5 RESULTS/FINDINGS

- The logistic regression classifier outperforms the random forest classifier in all scenarios
- Chromosomes 1, 7 and 8 contain the most genomic bins that contribute the most to the classification task of both models
- The genomic bins from Table 1 are part of the 20 features that contributed the most to the classification task of both classifiers

| |
|---|
| Chr1_bin[40000000-45000000] |
| Chr1_bin[155000000-160000000] |
| Chr1_bin[160000000-165000000] |
| Chr7_bin[35000000-40000000] |
| Chr7_bin[130000000-135000000] |
| Chr8_bin[95000000-100000000] |
| Chr8_bin[125000000-130000000] |

Table 1: Bins whose values are within the 20 most important features in both classifiers. The numbers in the brackets represent the start and end positions of the genomic bin

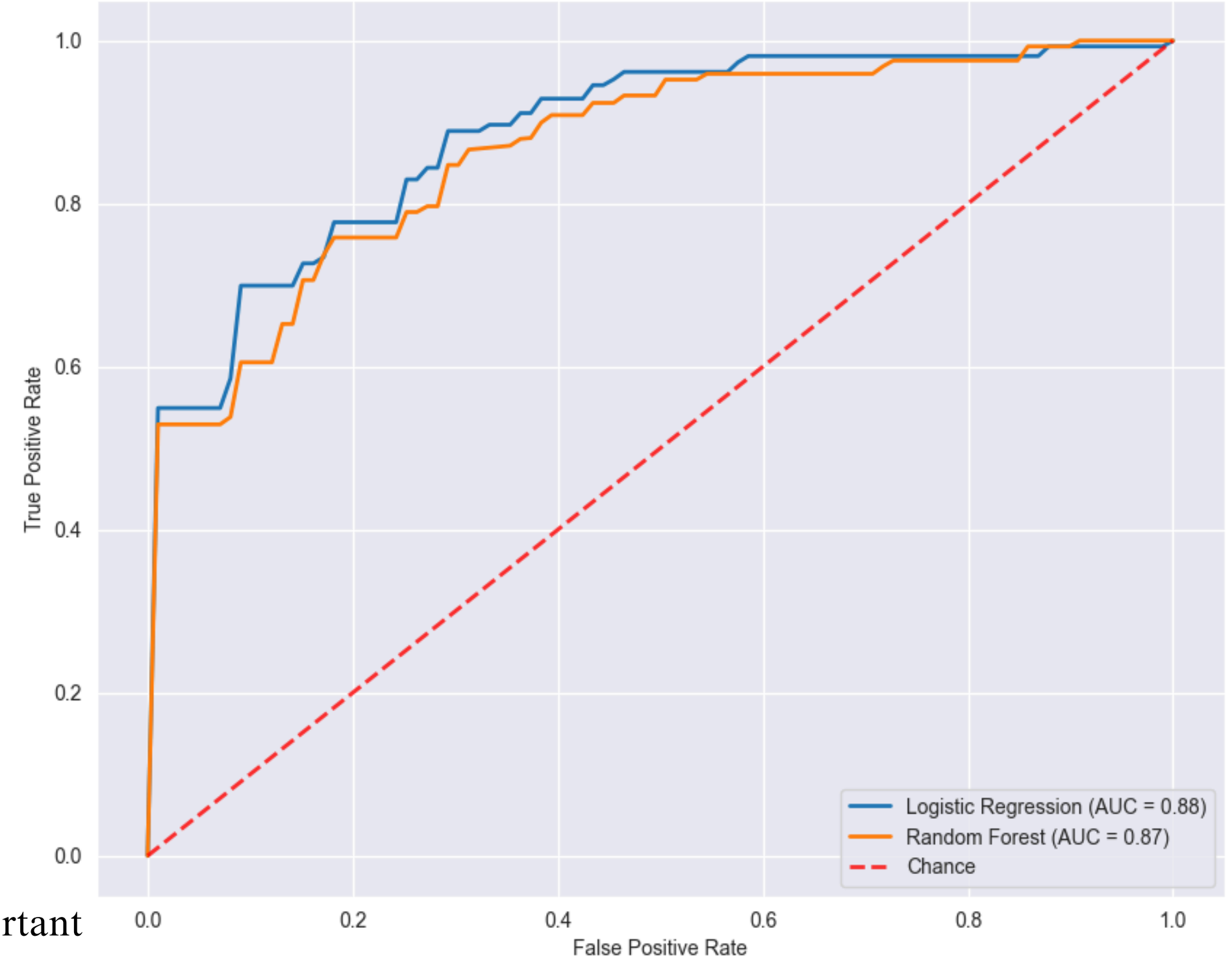

Figure 3: ROC curve illustrating the performance of the two classifiers when nested cross-validation is used for evaluation

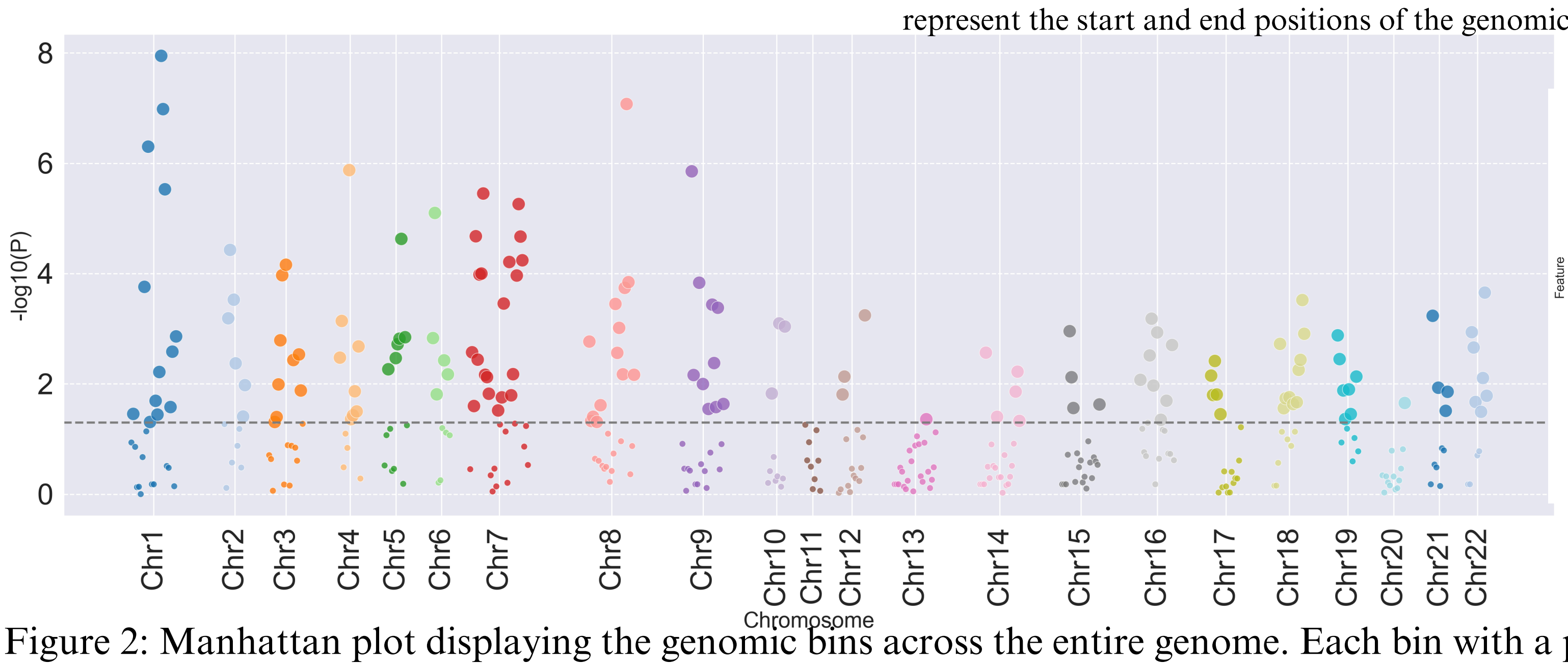

Figure 2: Manhattan plot displaying the genomic bins across the entire genome. Each bin with a p-value above the dashed line is selected during the selection procedure

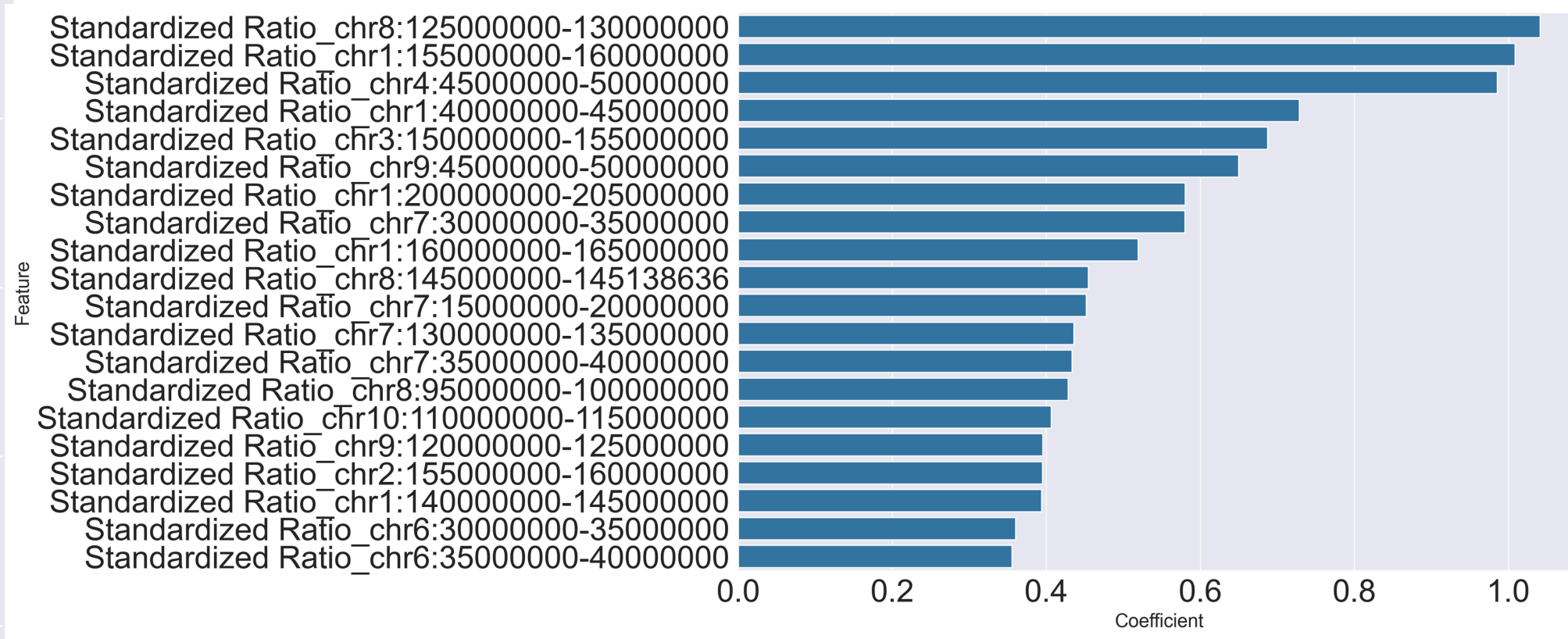

Figure 4: The coefficient value of the 20 features that contributed the most to the classification task using logistic regression. For this setting, the classifier was trained with the feature subset obtained after applying t-test and correlation-based filtering

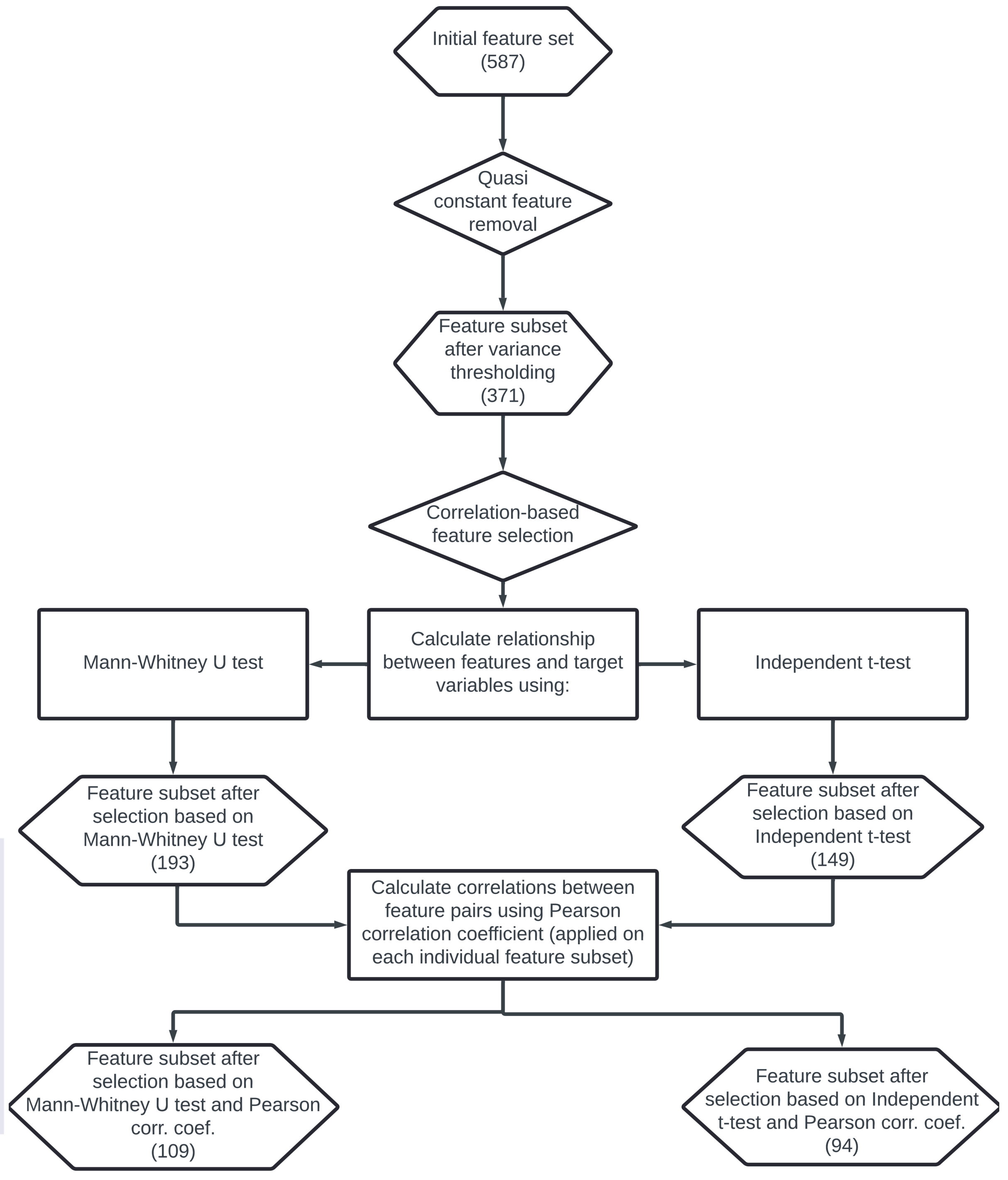
Initial feature set (587)
Quasi constant feature removal
Feature subset after variance thresholding (371)
Correlation-based feature selection
Calculate relationship between features and target variables using:
Mann-Whitney U test
Independent t-test
Feature subset after selection based on Mann-Whitney U test (193)
Feature subset after selection based on Independent t-test (149)
Calculate correlations between feature pairs using Pearson correlation coefficient (applied on each individual feature subset)
Feature subset after selection based on Mann-Whitney U test and Pearson corr. coef. (109)
Feature subset after selection based on Independent t-test and Pearson corr. coef. (94)

Figure 5: Flowchart describing the feature selection approaches used. The number at each hexagon's end represents the amount of features selected in each subset

## 6 LIMITATIONS

- The features used during the experiments are extracted based on a single extraction approach
- A conclusion about features' importance can only be derived manually by inspecting Figures 2 & 4
- These results should be validated by experts in the medical & bio-informatics field since they are obtained from a purely computer-science perspective

## 7 CONCLUSION

- Features are extracted from the available samples
- Feature selection is applied to filter out the redundant features
- The remaining ones are used for the classification task
- After training the classifiers, the weights of the features are extracted
- **The features with the highest coefficient values represent the most important fragmentomics used in cancer detection using blood**