Schema for Biomarker Discovery from Clinical Cancer Data

Intermediate Processes

Input Layer

Randomly select egual numbers of normal and cancer samples, with multiple iterations to avoid sampling

bias.

Apply
Random
Forest to
identify
important
biomarkers.

Random Forest for

Feature Selection

Aggregation of Important Biomarkers

Aggregate feature importance across iterations.

Filtering through Descriptive Statistics

1. Uniquely High Levels: The biomarker level must be uniquely high in the particular cancer type.

2. Higher Side Filtering: If not unique, the biomarker's level should still be relatively high, with its Q3 value in the top 2 among cancer types.

Hypothesis testing

Perform Yuen-Welch's test (a refinement of t-test) to verify that the biomarker shows a statistically significant difference between the particular cancer type and other cancer types or normal samples.

Output Layer

Finalize the set of cancer-specific biomarkers that meet all the filtering criteria.

Perform Random
Forest with only the selected biomarkers, to get accuracy scores.

Consult biological description of the biomarkers to further understanding.