Schema for Biomarker Discovery from Clinical Cancer Data

Intermediate Processes

Output Layer

Input Layer

Feature Selection using Random Forest

Randomly select equal numbers of normal and cancer samples, with multiple iterations to avoid sampling bias.

Apply *Random Forest* to identify important biomarkers.

Aggregation of Important Biomarkers

Aggregate feature importance across iterations.

Filtering Selected Features through Descriptive Statistics

1. Uniquely High Levels: The biomarker level must be uniquely high in the particular cancer type. We establish uniqueness using a *Coefficient of Variation* check and *Median Absolute Deviation* based outlier detection.

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.

Shared Nature of the Biomarkers through Yuen-Welch's test

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. If that is not the case, we take note of the shared nature.

1. Finalize the set of cancer-specific biomarkers that meet all the filtering criteria, while considering their shared nature.

2. Perform *Random Forest* for the cancer types with biomarkers that show uniquely high levels. Note the **accuracy scores**.

3. Consider **biological mechanisms** of the biomarkers to further understanding.