OMICS DATA ANALYSIS FOR SUBPOPULATION DISCOVERY AND FEATURE SELECTION IN BLADDER CANCER

485711: OMICS FOR NON-BIOLOGISTS

Shunyu Wu, Jiahui Yu

Friday 19th January, 2024

1 Dataset Description

1.1 Proteomics Dataset

proteomics.txt: This dataset contains proteomic profiles of 140 patients with bladder cancer. The proteomic data is the result of high-throughput mass spectrometry, which quantitatively measures the abundance of proteins in biological samples.

Rows: Each row represents a different protein.

Columns: The first column is the Protein ID, and each of the following columns represents a patient (labeled BC.1, BC.2, ..., BC.140).

Entries: The numerical values are the expression levels or abundances of each protein in each patient's sample.

1.2 Metadata Dataset

metadata.csv: This file contains additional information about the 140 patients whose proteomic data you have. Common types of metadata include demographic information (age, gender, ethnicity), clinical data (tumor size, histological grade, lymph node involvement, metastasis, treatment and response, survival status), and lifestyle factors (smoking history, family history of cancer). This data can provide context to the proteomic data and may reveal important correlations.

2 Subpopulation Identification

The objective of the Subpopulation Identification phase is to discern inherent subgroups within a dataset of 140 bladder cancer patients, each presenting varying stages of the disease. By deploying sophisticated clustering algorithms, we aim to unravel the complex stratifications within the Omics data, unveiling subpopulations that might respond differently to treatment or exhibit unique disease progression patterns.

2.1 Pipeline for Clustering Patients

A systematic approach to clustering is pivotal for revealing meaningful subpopulations. We detail our pipeline for clustering patients as follows:

2.1.1 Preprocessing

- Transpose the Omics dataset to align with clustering algorithms' expectations, where each row represents a unique patient and columns correspond to various proteomic features.
- Ensure data integrity by confirming the absence of missing values, thereby eliminating the need for imputation strategies.
- Normalize the data using StandardScaler to mitigate the influence of disparate scales and to facilitate a fair comparison across proteomic features.

2.1.2 Dimensionality Reduction

Dimensionality reduction is a precursor to clustering to manage the curse of dimensionality inherent in high-throughput Omics data.

- Implement Sparse PCA to reduce the dataset to its most informative axes while preserving sparsity, aiding interpretability.
- Employ t-SNE for a more nuanced visualization of clusters in a reduced dimensional space, providing an intuitive understanding of the data structure.
- Apply UMAP to maintain the global structure of the dataset, leveraging its ability to balance local versus global data features efficiently.

2.1.3 Clustering

We explore several clustering algorithms, acknowledging that each brings a unique perspective to the identification of patient subpopulations.

- Begin with K-Means clustering as a baseline to gauge the broad structure of the dataset.
- Advance to Hierarchical Clustering to investigate potential nested relationships among patients.
- Use DBSCAN to capture clusters of arbitrary shapes and to manage noise within the dataset.
- Finalize with Gaussian Mixture Models for a probabilistic model of cluster assignments, offering a flexible alternative to K-Means.

2.1.4 Determining the Number of Clusters

The determination of the optimal number of clusters is critical and requires a combination of techniques to validate the clustering outcome.

- The Elbow Method provides an initial estimate for the number of clusters, though its efficacy may be limited in high-dimensional spaces.

- The Silhouette Score offers a quantitative measure of cluster quality, indicating how similar an object is to its own cluster compared to others.
- The Calinski-Harabasz Index serves as an additional criterion, evaluating the dispersion between and within clusters.
- Assess the stability of clusters by comparing solutions across random subsamples of the data, ensuring that our findings are not artifacts of particular data partitions.

2.1.5 Iterative Approach

An iterative refinement process is employed, reevaluating the choice of dimensionality reduction techniques and clustering algorithms to optimize the subpopulation discovery.

2.1.6 Visualization and Interpretation

Visualization tools are leveraged to interpret the clustering results, ensuring that the identified subpopulations are both statistically and clinically meaningful.

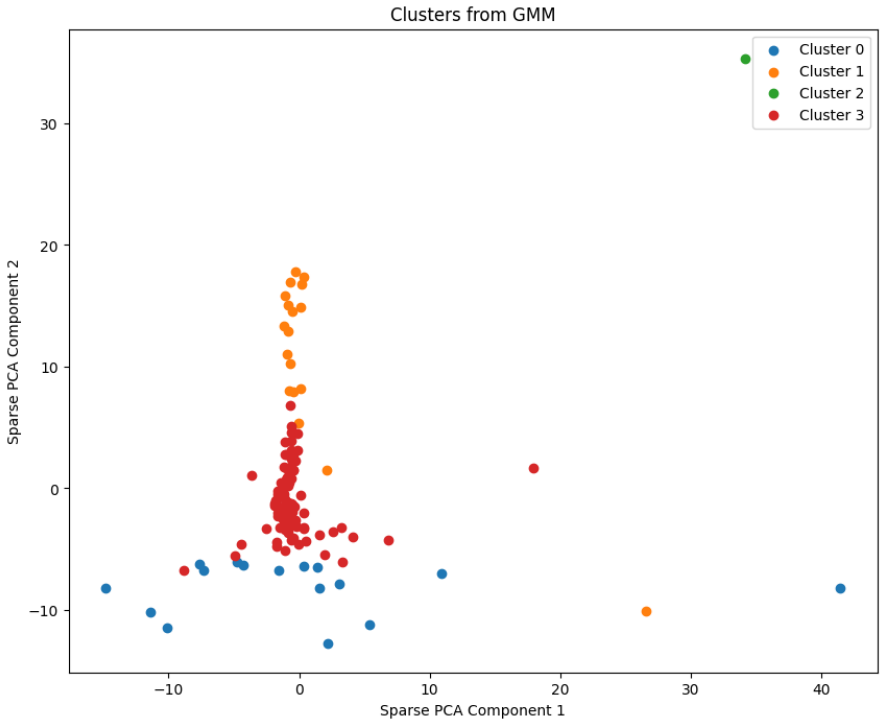


Figure 1: Visualization of patient clusters using Gaussian Mixture Model. The green dot cluster is identified as a potential outlier.

2.2 Identifying the Correct Number of Patient Subpopulations

The GMM clustering suggests the existence of four patient subpopulations. However, upon closer inspection and considering the biological relevance, the green dot cluster is more likely an

outlier. Hence, we postulate that the true number of patient subpopulations is three.

2.3 Pipeline for Feature Selection

The pipeline for selecting distinguishing features is as follows:

2.3.1 Remove Outliers

- Prior to feature selection, we remove data points that pertain to the outlier cluster to preclude skewed interpretations of feature relevance.

2.3.2 Feature Importance Methods

- Utilize ANOVA F-tests to compare the means of the features across clusters, identifying those that exhibit statistically significant differences.
- Implement Mutual Information to assess the dependency between features and cluster labels, favoring features with high mutual information scores.
- Leverage the intrinsic feature importance rankings provided by tree-based models like Random Forests and Gradient Boosting Machines, which are indicative of the features' predictive power.

2.3.3 Determining the Correct Number of Features

- Retain features that consistently rank high across all feature importance methods, ensuring a robust selection process.

2.3.4 Validation

- Reapply clustering with the selected features to confirm enhanced cluster definition and to validate the discriminative power of the features.
- Employ classification models to predict cluster membership, using the feature set as input, to further assess the features' effectiveness.

2.4 Identifying the Correct Number of Features

The reevaluation of clusters using the selected 18 features demonstrates a 74% congruence with the original cluster assignments. A combination of visualization techniques and quantitative measures such as the Adjusted Rand Index confirms the robustness of the 18-feature set in capturing the essence of the patient subpopulations.

Selected features include ['NINJ1', 'CTBP2', 'DOCK1', 'FAHD2A', 'IGLV5-45', 'STAT5A', 'MCM3', 'SELENOF', 'MANF', 'FKBP5', 'SEC24A', 'CEACAM7', 'ITGAM', 'ANKRD13A', 'S100A12', 'HBG2', 'PODN', 'FCN3']

Adjusted Rand Index comparing the non-outlier groups: 0.7444328423606514

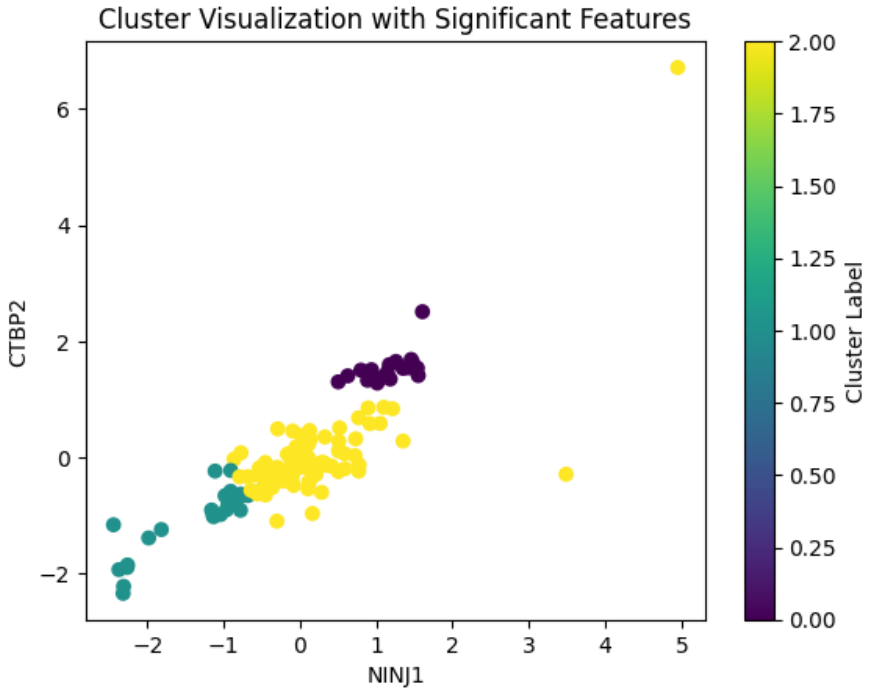


Figure 2: The refined clustering output post-feature selection, demonstrating more pronounced cluster separation.

3 Metadata Analysis

In this part, we interlinks patient metadata with the clusters previously identified from proteomic data, and with selected features that potentially influence clinical outcomes. This multifaceted analysis aims to decode the intricate relationships between clinical attributes and molecular signatures within the patient cohorts.

3.1 Cluster Analysis with Metadata

We explores the relationship between patient metadata features and predefined patient clusters in this section. The clusters were identified in previous analyses based on proteomic data.

3.1.1 Data Integration

We merged the patient metadata with the cluster assignments previously derived from proteomic analysis. This integration was pivotal in constructing a unified dataset, allowing for a holistic examination of the patients' clinical and biological characteristics in relation to their respective clusters.

3.1.2 Statistical Examination

A rigorous statistical examination was performed to discern patterns and disparities in the distribution of metadata features across the identified clusters. For continuous variables such

as age and tumor size, we employed Analysis of Variance (ANOVA) to identify statistically significant differences between groups. For categorical variables like gender and smoking history, Chi-square tests were utilized to evaluate the independence of these attributes across different clusters.

3.1.3 Cluster Characterization

We employed descriptive statistical methods to characterize each cluster, delving into the central tendencies such as the mean and median, as well as measures of spread including variance and standard deviation. This characterization provided nuanced insights into the distinct features defining each patient subgroup, enabling us to understand the demographic and clinical contours that demarcate the clusters.

3.1.4 Result

The Figure 3 provides a multi-dimensional analysis of patient clusters based on metadata features. From the box plots, we observe a significant variation in age and tumor size among the clusters, indicating potential differences in disease progression or stage. Notably, Cluster 3 exhibits a broader age range and follow-up duration, suggesting it may consist of both younger patients with longer follow-up times and older patients. The count plots reveal imbalances in gender distribution across clusters, with Cluster 3 showing a striking predominance. Smoking history, lymph node involvement, metastasis, and family history of cancer are distributed unevenly, suggesting these factors may play an important role in the clustering mechanism. Treatment approaches and survival status also vary, with Cluster 3 again demonstrating distinct patterns, possibly reflecting different treatment efficacies or disease severities. Ethnicity count plots underscore demographic diversity within clusters, which could have implications for personalized medicine and epidemiological studies.

3.2 Feature Analysis with Metadata

We investigated the relationship between selected proteomic features and patient metadata in this section.

3.2.1 Statistical Analysis

We conducted a comprehensive correlation analysis to uncover significant associations between the patients' metadata and their proteomics data. For continuous variables, we analyzed patterns and strengths of linear relationships, whereas for categorical features, we encoded the variables to quantify their correlations. The resulted heat map is shown in figure 4 and figure 5.

3.2.2 Machine Learning for Clinical Predictions

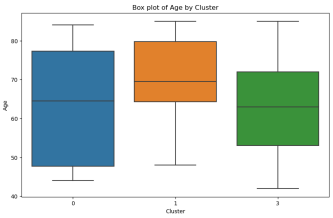
To further our understanding, we employed a Random Forest algorithm that provided insights into the importance of various features in predicting key clinical outcomes, notably tumor size and decrease survival status. The feature importance is shown in figure 6 and figure 7.

3.2.3 Result

The heatmaps reveal a modest correlation between proteomic features and tumor size, hinting at a potential link to tumor growth, whereas follow-up duration shows no clear pattern, suggesting

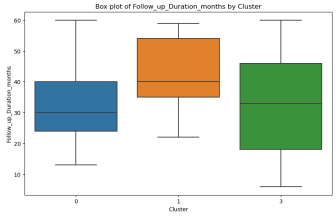
complex, non-linear relationships. The categorical metadata heatmap shows generally low correlations, with some features negatively associated with specific treatments and positively with others, such as partial remission, indicating a need for further investigation into their clinical relevance.

Feature importance analysis highlights biomarkers like MANF and ITGAM as significant predictors of tumor size and survival status, with the models showing a moderate degree of predictive accuracy. The Mean Squared Error (289.55) and R-squared values (0.55) for tumor size, along with the classification report metrics for survival status, suggest that while the models perform adequately, there is room for enhancement. These insights point towards the potential utility of proteomic features in clinical prognostics and the necessity for advanced analytical methods to fully unravel the intricate relationships in the data.

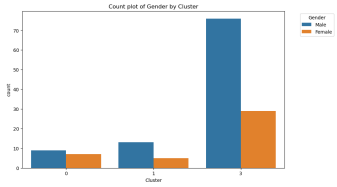


(a) Box plot of Age by Cluster

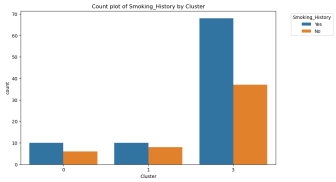
(b) Box plot of Tumor Size mm by Cluster



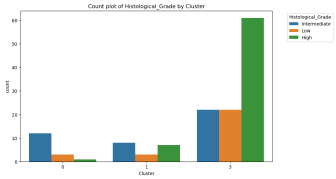
(c) Box plot of Follow up Duration months by Cluster



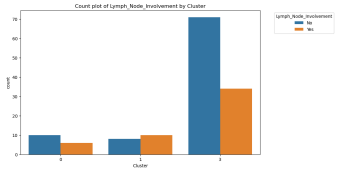
(d) Count plot of Gender by Cluster by Cluster



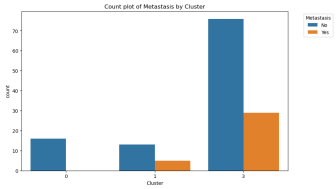
(e) Count plot of Smoking History



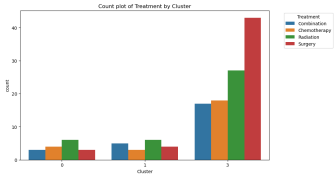
(f) Count plot of Histological Grade by Cluster



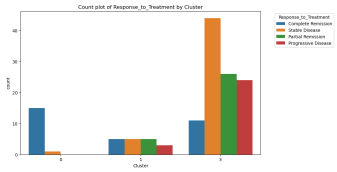
(g) Count plot of Lymph Node Involvement by Cluster



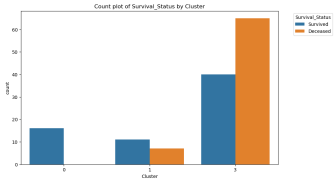
(h) Count plot of Metastasis by Cluster



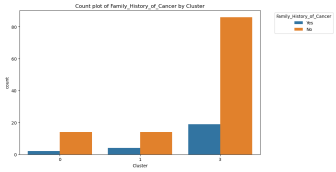
(i) Count plot of Treatment by Cluster



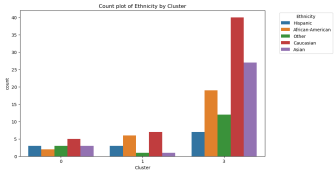
(j) Count plot of Response to Treatment by Cluster



(k) Count plot of Survival Status by Cluster



(l) Count plot of Family History of Cancer by Cluster



(m) Count plot of Ethnicity by Cluster

Figure 3: Metadata analysis with clusters

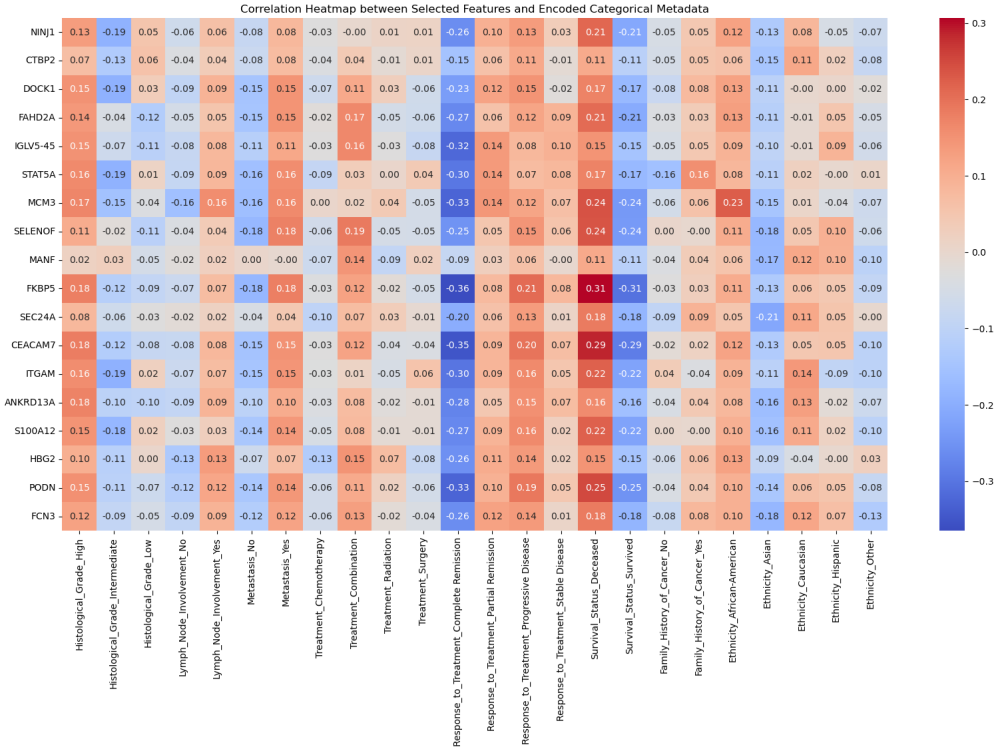


Figure 4: Correlation heatmap of selected features and encoded categorical metadata.

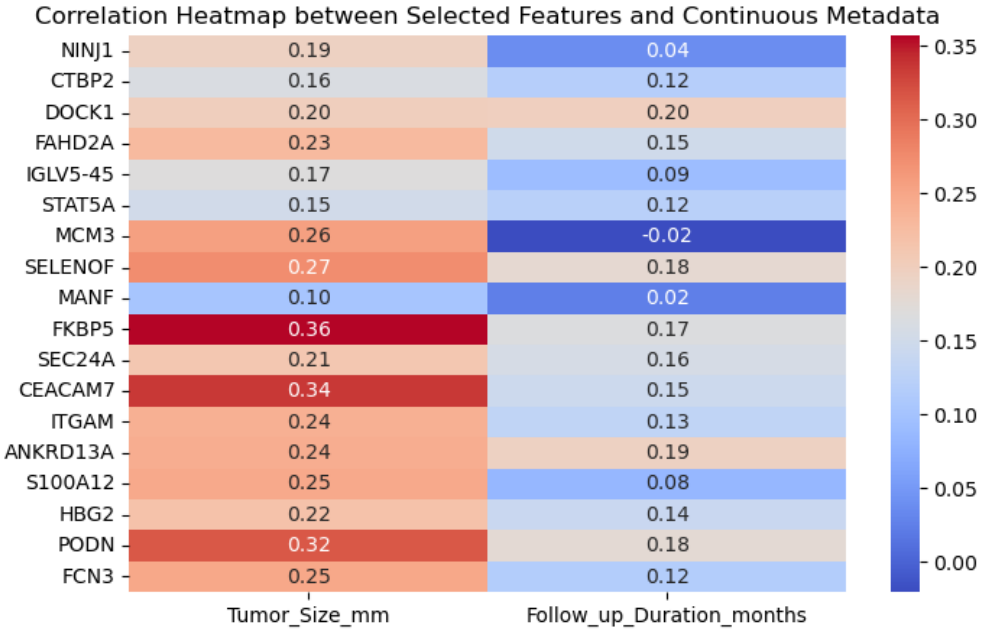


Figure 5: Correlation heatmap between selected features and continuous metadata.

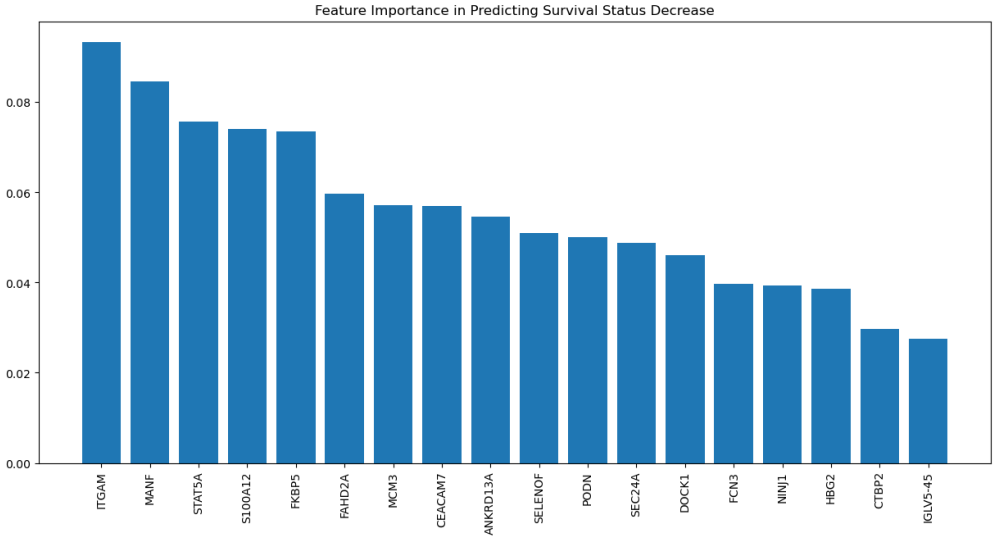


Figure 6: Feature importance in predicting survival status decrease.

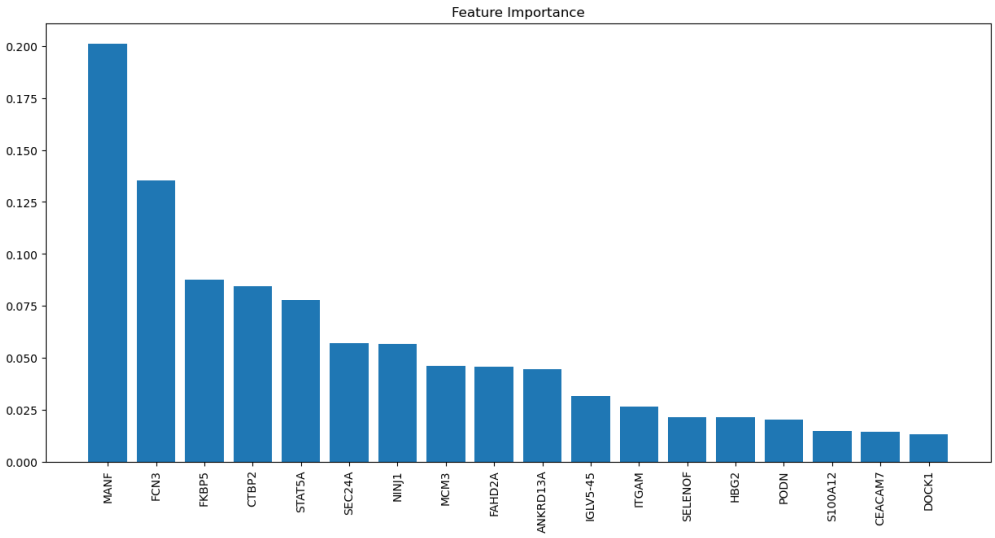


Figure 7: Feature importance in predicting tumor sizes.