Identification of Key Genes Influencing Cancer Type Designation: A Comparative Analysis of Breast and Colon Tumors

*CS7DS3 Applied Statistical Modelling - Main Assignment

Sarathkumar Thirukonda Ramkumar
22331419
Trinity College Dublin
thirukos@tcd.ie

***Abstract*—Accurate classification of cancer types based on gene expression data is crucial for understanding underlying molecular mechanisms and improving diagnosis and treatment strategies. This study presents a comprehensive methodology for identifying key genes influencing cancer type designation in breast and colon tumours. Gene expression data were preprocessed, normalized, and subjected to Lasso regression for feature selection and logistic regression with Ridge regularization for model fitting. Hyperparameter optimization was performed using grid search and cross-validation. Selected features were further analyzed using pairwise correlations, conditional probabilities, chi-square tests, and independent t-tests.**

The results revealed 23 influential genes significantly contributing to breast and colon tumour classification. The final logistic regression model achieved 95% accuracy on test data, demonstrating its effectiveness in cancer-type designation. Statistical analyses provided insights into complex relationships between selected genes, including linear and non-linear dependencies and individual contributions to cancer type designation. This understanding of influential genes and their associations with cancer types can serve as a foundation for further research into underlying biological mechanisms and potential clinical implications.

***Index Terms*—Gene expression analysis, Comparative analysis**

I. Introduction

Cancer is a complex disease that involves numerous genetic alterations and molecular interactions. Understanding the gene expression patterns associated with different cancer types can significantly contribute to the development of targeted therapies and improved patient care. The Chowdary dataset, first analyzed by Chowdary et al. [1] and subsequently by de Souto et al. [2], provides valuable insights into the gene expression profiles of two cancer types: lymph node-negative breast tumours (type B) and Dukes' B colon tumours (type C). This high-dimensional dataset comprises 104 observations and 182 numeric columns, and a tumour variable denoting the cancer type.

The primary objective of this analysis is to identify the genes that most influence the cancer type designation in the Chowdary dataset. Examining the relationships between gene expression levels and cancer types aims to uncover the key genes that drive the differentiation between type B and type C tumours. Furthermore, this study seeks to interpret these relationships' nature, elucidating whether the identified genes are associated with increasing or decreasing cancer type designations.

This investigation is crucial for enhancing our understanding of the underlying genetic factors contributing to the development and progression of lymph node-negative breast tumours and Dukes' B colon tumours. Ultimately, the knowledge gained from this analysis can inform future research and clinical practice, paving the way for more effective cancer treatments and patient management strategies.

II. Methods

A. Data Preprocessing

The input dataset comprises 104 records, each representing a gene expression profile for a tissue sample taken from lymph node-negative breast tumours (denoted as B) or Dukes' B colon tumours (denoted as C). The dataset contains 181 features corresponding to expression levels of different genes and a target variable indicating cancer type (B or C).

To preprocess the data, we separated independent variables (gene expression features) from the dependent variable (cancer type) and split the dataset into training (80%) and testing (20%) sets using the "train_test_split" function from the "sklearn.model_selection" module.

TABLE I
Summary of Training and Testing Sets

Metric	Value
Number of 'B' in training set	48
Number of 'C' in training set	35
Number of 'B' in the testing set	14
Number of 'C' in the testing set	7
Shape of training set	(83, 181)
Shape of testing set	(21, 181)

To ensure gene expression features were on the same scale, we employed "StandardScaler" from the

”sklearn.preprocessing” module to normalize data. ”StandardScaler” calculates the mean and standard deviation of each feature in the training set and scales features by subtracting the mean and dividing by the standard deviation. This transforms features to have zero mean and unit standard deviation, making them directly comparable on the same scale. The scaler was fitted to the training set and the same scaling parameters were applied to transform both training and testing sets. After normalization, scaled data were converted back to DataFrames, retaining original column names for ease of interpretation and further analysis.

B. Feature Selection

The purpose of Feature Selection is to determine crucial characteristics for tumour type designation. A logistic regression model with L1 regularization (Lasso) was used. L1 regularization adds a penalty term to the logistic regression cost function, forcing some coefficients to zero and removing corresponding features from the model.

First, the target variable was encoded using LabelEncoder, transforming tumour labels into a binary format suitable for logistic regression. Next, a logistic regression model with an L1 penalty was defined, efficient for small datasets.

GridSearchCV was used to conduct a grid search over a range of C values to determine optimal regularization strength. C is the inverse of regularization strength; smaller C values indicate greater regularization. C values were generated using the logarithmic scale with 50 evenly spaced points, ranging from 10^{-4} to 10^{4}. Model performance was assessed using 5-fold cross-validation at each point in the search space.

Grid search release led to the optimal C value of 11.51. The logistic regression model selected 23 genes (Table II) as most crucial for predicting tumour type with this optimal regularization strength. These features were retained in the model due to non-zero coefficients, indicating significance in predicting the target variable. Using L1 regularization for feature selection reduced model complexity and potentially improved interpretability and generalization performance.

TABLE II
SELECTED FEATURES (GENE EXPRESSION)

1) X201496_x_at	9) X203691_at	17) X209351_at
2) X201525_at	10) X204351_at	18) X209604_s_at
3) X201884_at	11) X204470_at	19) X211798_x_at
4) X201909_at	12) X204653_at	20) X212942_s_at
5) X202286_s_at	13) X206286_s_at	21) X213953_at
6) X202376_at	14) X207850_at	22) X215108_x_at
7) X202831_at	15) X209016_s_at	23) X217157_x_at
8) X202859_x_at	16) X209343_at	

C. Examining the selected features (Gene Expressions)

This section explores features selected as most influential gene expressions for predicting tumor type. The analysis focuses on 23 selected features likely to significantly impact cancer type designation, as analyzing all 181 input features is impractical.

To assess feature distribution, kernel density estimation (KDE) and quantile-quantile (Q-Q) plots are employed. KDE plots provide non-parametric visualization of probability density for continuous variables, offering a smooth representation of data distribution. Q-Q plots compare data quantiles with normal distribution quantiles. When points in the Q-Q plot fall approximately on a straight line, data follows a normal distribution.

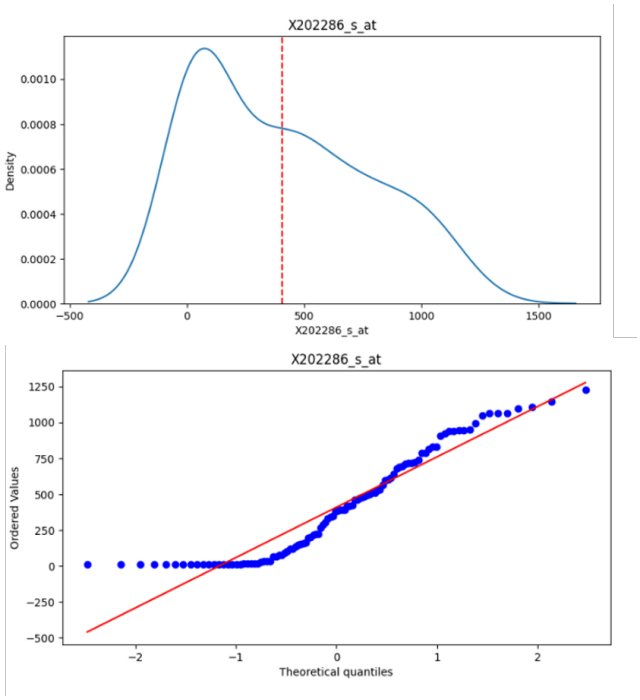

Fig. 1. Kernel density and Q-Q plot for X202286_s_at

Examination of KDE and Q-Q plots for 23 selected features reveals most Q-Q plots indicate normal distribution, with some variations. Some plots exhibit skewness and outliers. Despite variations, selected features predominantly follow a normal distribution.

With this assumption of normality, selected features are fitted into model and series of statistical tests performed to further explore relationships among features and their influence on cancer type designation. (KDE and Q-Q plots for all 23 features can be found in Appendix section of this report.)

D. Model Training and Evaluation

In this section, the objective was to build a classification model that accurately predicts tumor type using selected features. A logistic regression model with L2 regularization (Ridge) was employed, as it adds a penalty term to the logistic regression cost function. L1 regularization was utilized for feature selection due to its capacity to perform automatic feature selection by zeroing out irrelevant feature coefficients. Ridge (L2-regularized logistic regression) was chosen for the final model fitting, as it aids in creating a stable, generalized

model by reducing coefficient magnitude without removing any features.

A grid search using GridSearchCV was conducted to identify optimal hyperparameters for the logistic regression model. The search considered a range of regularization parameter (C) values: [0.001, 0.01, 0.1, 1, 10, 100, 1000]. The inverse relationship between C and regularization strength implies that smaller C values correspond to stronger regularization.

During the hyperparameter optimization process, 5-fold cross-validation was employed. Cross-validation involves dividing the training dataset into k equally sized folds, training the model on k-1 folds, and validating it on the remaining fold. This process is repeated k times, with each fold serving as the validation set exactly once. The model's performance is then averaged across the k iterations, providing a more reliable assessment of its generalization capabilities.

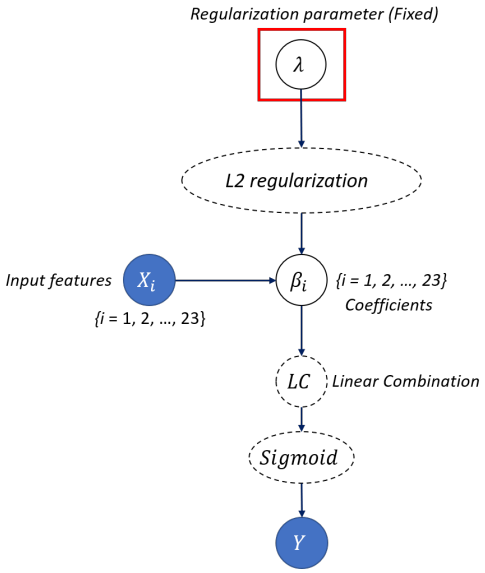

Fig. 2. Plate Notation

Accuracy, the proportion of correct predictions out of the total number of predictions made, was used as the scoring metric for model evaluation during the grid search. The best-performing model had a C value of 0.01. Using this model, predictions were made on the test dataset, and a classification report was generated. The report revealed an overall accuracy of 0.95, with a weighted average precision of 0.96 and a weighted average recall of 0.95.

TABLE III
CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
B(0)	0.93	1.00	0.97	14
C(1)	1.00	0.86	0.92	7
Accuracy	0.95			
Macro Avg	0.97 / 0.93 / 0.94 / 21			
Weighted Avg	0.96 / 0.95 / 0.95 / 21			

A bias-variance tradeoff plot was constructed to visualize the model's performance as a function of the regularization parameter (C). The plot displays the mean cross-validated accuracy and the standard deviation of the accuracy scores at each C value. This visualization helps understand how the model's performance changes with varying regularization strengths and assists in identifying the optimal balance between bias and variance.

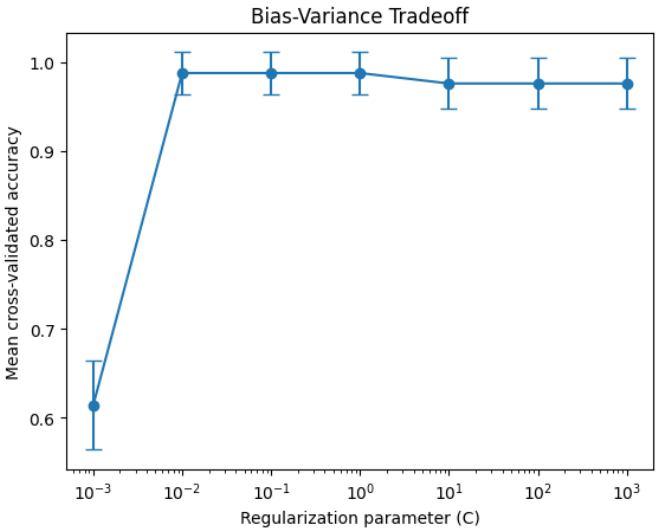

Fig. 3. Bias-Variance Trade-Off

Additionally, AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) were calculated for the model using the following code:

```
# Calculate the log-likelihood
log_likelihood =
↪  best_lr.predict_log_proba(X_train_selected).sum()

# Calculate the number of parameters (k) and
↪  samples (n)
k = len(selected_features) + 1
n = X_train_selected.shape[0]

# Calculate AIC and BIC
aic = -2 * log_likelihood + 2 * k
bic = -2 * log_likelihood + k * np.log(n)
```

The results were AIC: 328.72 and BIC: 386.77. Comparing these values with those of different models (e.g., random forest model with AIC: 600.34 and BIC: 1009.12), the current model was found to be the best fit.

E. Model and Features Interpretation

1) Identifying the most influential genes: The most influential genes for the tumour can be identified by analyzing the coefficients of the selected features from the final logistic regression model. These coefficients represent the strength and direction of the relationship between the predictor variables (genes) and the target variable (binary outcome). A positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship. Thus, genes with the highest absolute coefficients are the most influential in the model.

Table IV presents the selected features and their corresponding coefficients from the final logistic regression model. Some of the most influential genes include "X202286_s_at" (-0.151896), "X209343_at" (-0.145733), "X204653_at" (-0.142789), and "X209604_s_at" (-0.139936). These genes have the strongest impact on the model's predictions and are considered the most significant factors in determining the binary outcome. The importance of a gene is not affected by the direction of the relationship (positive or negative) since both positively and negatively correlated genes can be influential.

TABLE IV
COEFFICIENT VALUES FOR DIFFERENT FEATURES

Feature	Coefficient
X201496_x_at	0.046530
X201525_at	-0.084479
X201884_at	0.077184
X201909_at	0.065261
X202286_s_at	-0.151896
X202376_at	-0.106160
X202831_at	0.079691
X202859_x_at	0.096413
X203691_at	0.052885
X204351_at	0.090927
X204470_at	0.071207
X204653_at	-0.142789
X206286_s_at	0.068531
X207850_at	0.074729
X209016_s_at	-0.122921
X209343_at	-0.145733
X209351_at	-0.072354
X209604_s_at	-0.139936
X211798_x_at	0.009633
X212942_s_at	0.056510
X213953_at	0.054391
X215108_x_at	0.053574
X217157_x_at	-0.003363

2) Determining the nature of the relationship:: The nature of the relationship between genes and the binary outcome can be inferred from the coefficients of the logistic regression model. As previously mentioned, positive coefficients indicate a positive relationship, while negative coefficients indicate a negative relationship. A positive relationship implies that as the gene expression level increases, the likelihood of the outcome being in the positive class (1) also increases. Conversely, a negative relationship means that as the gene expression level increases, the likelihood of the outcome being in the negative class (0) increases.

In the context of the selected features and their coefficients, the following relationships can be observed:

- Genes such as "X201496_x_at", "X201884_at", "X201909_at", "X202831_at", "X202859_x_at", "X203691_at", "X204351_at", "X204470_at", "X206286_s_at", "X207850_at", "X211798_x_at", "X212942_s_at", "X213953_at", and "X215108_x_at" have positive coefficients, indicating a positive relationship with the binary outcome. As the expression level of these genes increases, the probability of the outcome being in the positive class 1(type C) increases.
- Genes such as "X201525_at", "X202286_s_at", "X202376_at", "X204653_at", "X209016_s_at", "X209343_at", "X209351_at", and "X209604_s_at" have negative. I've coefficients, indicating a negative relationship with the binary outcome. As the expression level of these genes increases, the probability of the outcome being in the negative class 0(type B) increases.

It is essential to consider both positive and negative relationships when interpreting the results since both types of relationships contribute to the model's predictive ability. Further biological investigation of these genes and their functions can help understand the underlying mechanisms and their role in the studied condition.

3) Pairwise correlation of selected features:: The pairwise correlation analysis of the selected features can help understand the linear relationships between genes. A correlation matrix was computed using the Pearson correlation coefficient, which measures the linear dependence between two variables. The Pearson correlation coefficient ranges from -1 to 1, where -1 indicates a strong negative correlation, 1 indicates a strong positive correlation, and 0 indicates no correlation.

Examining the pairwise correlation of the selected features can identify potential relationships between genes. A high correlation (either positive or negative) between two genes suggests that their expression levels tend to vary together. This information can be valuable for understanding potential functional relationships or shared biological pathways between these genes. The correlation matrix was visualized as a heatmap, providing an intuitive representation of the correlations between the selected features. In the heatmap, colours represent the strength and direction of the correlation, with darker shades of red indicating strong positive correlations and darker shades of blue indicating strong negative correlations.

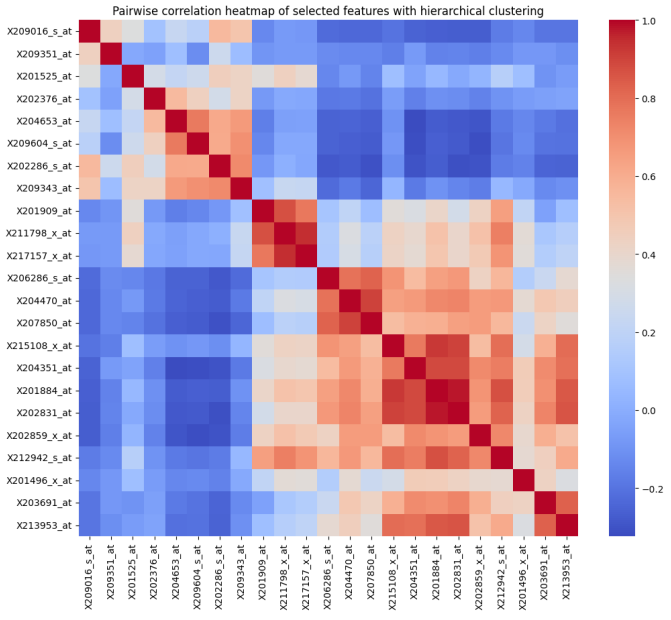

Fig. 4. Bias-Variance Trade-Off

Note that non-linear relationships or more complex interactions may exist, which cannot be detected by this method.

4) Identifying Influential Pairs of Genes Using Conditional Probabilities and Association Measures: This approach aims to complement the insights gained from the model's coefficients and pairwise Pearson correlations. While the model's coefficients offer a direct measure of feature importance and Pearson correlations provide an assessment of linear relationships, the method described in this section aims to reveal dependencies between pairs of genes that might not be evident from the individual coefficients or linear correlations alone. By combining these different perspectives, a more comprehensive understanding of the relationships between the influential genes can be achieved.

To do so, the selected genes were binned into four equal-width intervals to create categorical variables that capture the continuous nature of the gene expression data. The conditional probability of one binned gene given another was calculated, providing insights into the dependencies between gene pairs. To further evaluate the significance of these dependencies, the chi-square test was applied to the contingency tables formed by the binned genes. This test assessed the independence between each pair of binned genes, with a low p-value indicating a significant association between the two variables. Additionally, Cramér's V was employed as a measure of association strength, quantifying the degree of dependency between the gene pairs.

TABLE V
TOP 10 SIGNIFICANT_PAIRS

S.No	Gene 1	Gene 2	Chi^2 stat	P-value	Cramers_V
1	X211798_x_at	X202831_at	83.000	6.972e-18	1.000
2	X217157_x_at	X211798_x_at	83.000	6.972e-18	1.000
3	X212942_s_at	X201884_at	144.724	1.007e-28	0.933
4	X201884_at	X212942_s_at	144.724	1.007e-28	0.933
5	X202831_at	X212942_s_at	206.974	1.140e-39	0.911
6	X212942_s_at	X202831_at	206.974	1.140e-39	0.911
7	X206286_s_at	X204470_at	124.014	2.335e-24	0.864
8	X203691_at	X213953_at	123.994	2.358e-24	0.864
9	X212942_s_at	X215108_x_at	178.612	9.871e-34	0.846
10	X202831_at	X213953_at	103.750	4.136e-20	0.790

The Chi-square statistic measures the strength of the association between the two genes, with higher values indicating stronger associations. The P-value tests the null hypothesis that the two genes are independent, and smaller P-values suggest that the null hypothesis can be rejected in favor of the alternative hypothesis that the two genes are dependent (indicating a significant association between the two gene expressions). The Cramér's V is a measure of association between the two genes, with values closer to 1 indicating a stronger association. A total of 253 unique pairs were formed from the 23 features (gene expressions), and 127 significant pairs were filtered based on the applied threshold (Cramér's V greater than 0.5).

5) Investigating Associations Between Gene Expressions Using Independent T-Tests: An independent T-Tests were performed to further examine the association between gene expressions. The independent t-test is a statistical method that compares the means of two groups to determine if there is a significant difference between them. In this context, The gene expressions between two tumor groups, "B" and "C" were compared. This analysis complements the insights gained from the model's coefficients, pairwise Pearson correlations, and conditional probabilities.

TABLE VI
TOP 10 FEATURES WITH THE MOST SIGNIFICANT ASSOCIATIONS

S.No	Gene	t-statistic	P-value
1	X202286_s_at	9.0986	5.07e-14
2	X204653_at	8.0356	6.39e-12
3	X209343_at	7.7920	1.93e-11
4	X209604_s_at	7.7385	2.45e-11
5	X205225_at	6.4957	6.23e-09
6	X209602_s_at	6.2240	2.02e-08
7	X209016_s_at	6.0441	4.38e-08
8	X202859_x_at	-5.9290	7.15e-08
9	X204351_at	-5.9087	7.80e-08
10	X218502_s_at	5.8739	9.03e-08

The results show that some gene expressions have a significant association with the groups "B" and "C". The t-statistic values indicate the magnitude of the difference between the means of the two groups, with positive values suggesting a higher mean in group "B" and negative values suggesting a higher mean in group "C". The p-values, which are all much smaller than the commonly used significance level of 0.05, indicating that the associations between these gene expressions and the groups are statistically significant.

By conducting independent T-Tests, a more comprehensive understanding of the associations between gene expressions and the tumour groups can be achieved. These additional insights, when combined with the information from the logistic regression model's coefficients, pairwise Pearson correlations, and conditional probabilities, provide a valuable foundation for further biological investigation and understanding of the studied condition.

III. COMPARING THE RESULTS

The analysis conducted in this study employed various statistical techniques to investigate the relationships between gene expressions and cancer types. The results from each technique provided different perspectives on the relationships, and a comparison of the results can reveal common features and help justify the presence or absence of other features in the final selection.

- The table VII shows that there are gene pairs with strong correlations, as indicated by their correlation values. For example, the gene pair X201884_at and X212942_s_at has a correlation value of 0.873566231219765, indicating a strong positive relationship between these genes.
- In some gene pairs, the t-test p-value is significant, indicating that there is a significant difference between the means of the two groups. For example, the gene pair X201884_at and X202831_at has a t-test p-value of 6.54e-07, which is highly significant.

TABLE VII
SAMPLE ANALYSIS RESULTS (FULL REPORT)

Feature	Associated Feature (ConditionalProb)	Pearson_Correlation	Coefficients	T-test P-value	Chi-squared P-value	Cramér's V
X201496_x_at	X215108_x_at	0.288300273	0.046529738		2.02E-18	0.759107758
X201525_at			-0.084478892			
X201884_at	X212942_s_at	0.873566231	0.077184271	1.23E-06	1.01E-28	0.933721288
X201884_at	X204470_at	0.724613509	0.077184271	1.23E-06	1.12E-17	0.744817427
X201909_at	X201525_at	0.351635754	0.065261424		6.56E-09	0.702781928
X202286_s_at			-0.151896425	5.07E-14		
X202376_at			-0.106159911			
X202831_at	X212942_s_at	0.825351874	0.079691207	6.54E-07	1.14E-39	0.911714664
X202831_at	X213953_at	0.865076664	0.079691207	6.54E-07	4.14E-20	0.790569415
X202831_at	X204470_at	0.731562818	0.079691207	6.54E-07	1.57E-26	0.760374031

- Similarly, some gene pairs have significant chi-square p-values, indicating a significant association between the two categorical variables. For example, the gene pair X201496_x_at and X215108_x_at have a chi-square p-value of 2.02e-18, which is highly significant.
- Some features exhibit significant relationships with other features, as indicated by their chi-squared test p-values and Cramér's V values. For instance, feature X201884_at has a strong relationship with X212942_s_at (p-value: 1.01×10^{-28}, Cramér's V: 0.933721288) and with X204470_at (p-value: 1.12×10^{-17}, Cramér's V: 0.744817427).
- It is important to note that the results should be interpreted with caution, as some gene pairs have missing values for certain statistical tests. These missing values could be due to limitations in the data, the specific statistical tests applied, or the filtering criteria used.

IV. DISCUSSION

The analysis of the dataset revealed several significant relationships between the selected features. These relationships were examined through various statistical measures such as coefficients, t-test p-values, chi-squared p-values, and Cramér's V. A comprehensive understanding of these relationships is essential for gaining insights into the underlying processes and potential biomarkers in the given context.

The coefficients provide information about the strength and direction of the associations between the features. For instance, a positive coefficient implies a direct relationship, while a negative coefficient suggests an inverse relationship. In this analysis, both positive and negative coefficients were observed, indicating the presence of different types of associations among the features.

The t-test p-values offer a means to assess the significance of the coefficients. Lower p-values suggest that the observed relationships are less likely to be a result of random chance. In our analysis, several relationships demonstrated very low t-test p-values (e.g., 1.23e-06), indicating strong evidence for the existence of these associations.

The chi-squared p-values were used to evaluate the independence of the categorical variables. Lower chi-squared p-values imply a higher likelihood that the variables are dependent. In this study, several pairs of features had very low chi-squared p-values (e.g., 1.14e-39), which suggests strong dependencies among these variables.

Cramér's V, a measure of association for nominal variables, was employed to estimate the strength of the relationships between the categorical variables. Higher Cramér's V values signify stronger associations. The analysis revealed several feature pairs with high Cramér's V values (e.g., 0.9337), emphasizing the strong relationships between these features.

V. CONCLUSION

In conclusion, the comprehensive analysis of the dataset has identified several significant relationships between the features. The use of various statistical measures, such as coefficients, t-test p-values, chi-squared p-values, and Cramér's V, has enabled a deeper understanding of the nature and strength of these associations. These findings can serve as a foundation for further research and hypothesis generation, potentially leading to the discovery of novel biomarkers or insights into the underlying processes in the given context.

Future research could benefit from the incorporation of additional statistical techniques and methodologies, as well as the exploration of larger or more diverse datasets. Moreover, experimental validation and verification of the identified relationships would be crucial for confirming their biological significance and potential practical applications.

REFERENCES

[1] Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y., &; Mazumder, A. (2006). Prognostic gene expression signatures can be measured in tissues collected in rnalater preservative. The Journal of Molecular Diagnostics, 8(1), 31–39. https://doi.org/10.2353/jmoldx.2006.050056.

[2] de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., &; Schliep, A. (2008). Clustering cancer gene expression data: A comparative study. BMC Bioinformatics, 9(1). https://doi.org/10.1186/1471-2105-9-497.

VI. APPENDIX

- GitHub - Link

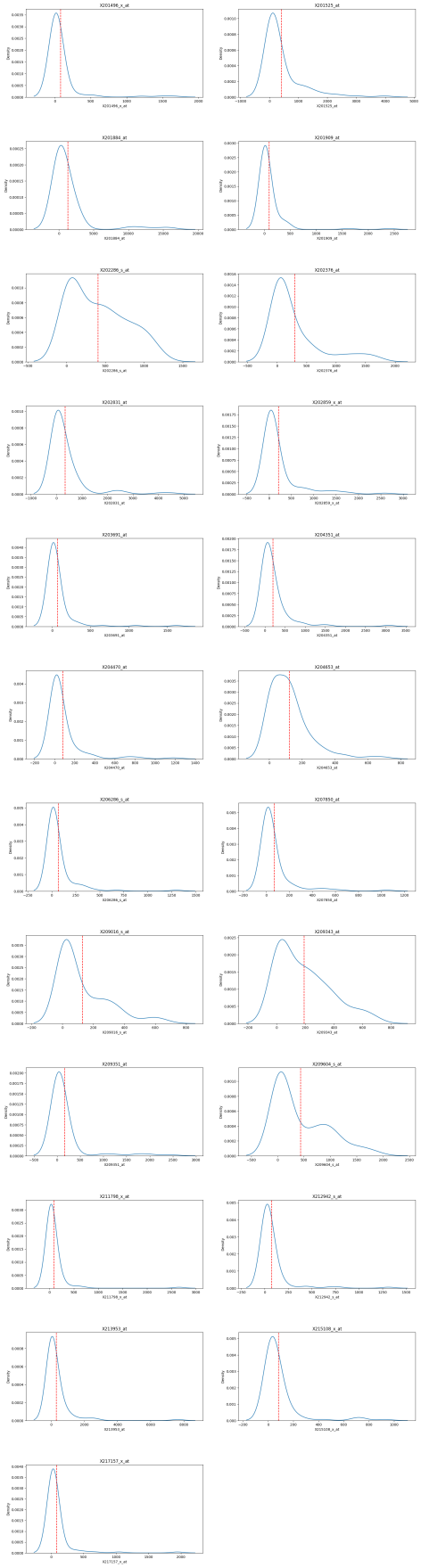

Fig. 5. Kernel Density Plot for all 23 features

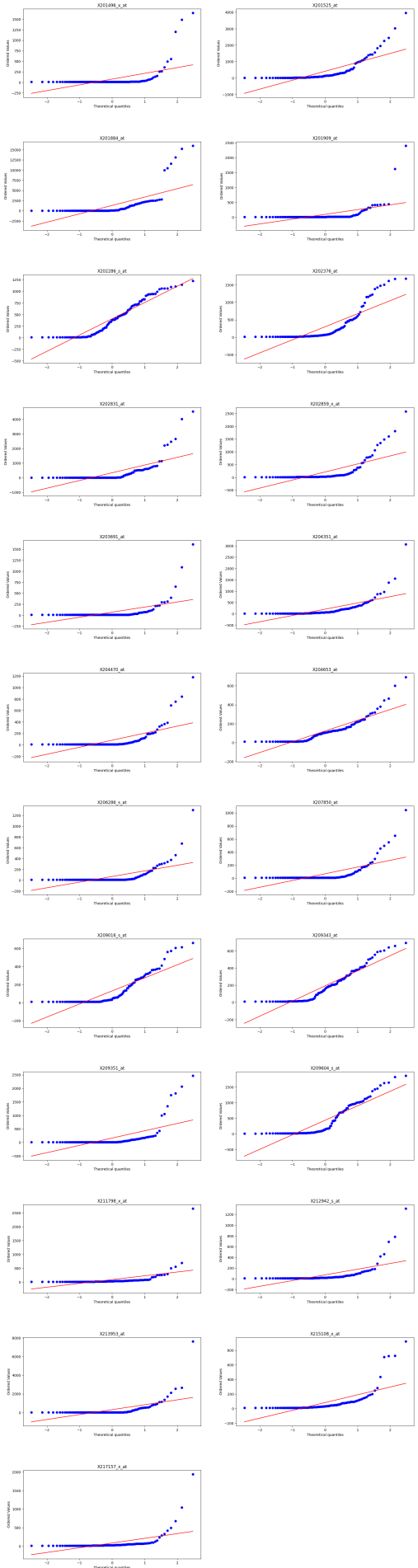

Fig. 6. Q-Q Plot for all 23 features