Oral Squamous Cell Carcinoma

A comprehensive proteomics and metabolomics analysis utilizing machine learning algorithms to identify robust molecular classifiers for the early detection of oral squamous cell carcinoma

Project Report: DH 307

Soumitra Darshan Nayak
22B0984
Guide: Prof Sanjeeva Srivastava
Guide: Avinash Singh

December 19, 2024

Abstract

This report outlines the work undertaken during the Oral Squamous Cell Carcinoma (OSCC) Research and Development (R&D) project, focusing on the identification of potential biomarkers for OSCC using proteomics and metabolomics data. OSCC is a prevalent and aggressive form of cancer with low survival rates due to late-stage diagnosis, underscoring the critical need for reliable biomarkers to enable early detection and targeted treatment strategies.

The study employed a systematic pipeline that began with extensive data preprocessing to handle missing values, normalize datasets, and ensure consistency. This was followed by robust feature selection processes to reduce the high-dimensional data and retain only the most relevant features. Statistical methods such as Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) and Partial Least Squares Discriminant Analysis (PLS-DA) were utilized to identify the relevant features from datasets, after which machine learning models were applied to refine the selection and validate the identified biomarkers.

Key machine learning methods included Logistic Regression with Elastic Net regularization, Random Forest, and Support Vector Machines (SVM) with Recursive Feature Elimination (RFE) & Lasso Regression. Each model was evaluated using cross-validation techniques to ensure the reliability and generalizability of the results. Furthermore, a comparative analysis of the selected features was performed to identify common proteins and metabolites that could serve as robust biomarkers for OSCC.

The integration of proteomics and metabolomics data, along with machine learning, enabled the identification of the most discriminative biomarkers, which can potentially enhance early diagnosis, improve prognosis, and guide personalized treatment strategies for OSCC. The findings underscore the importance of combining high-dimensional biological data with advanced computational methods to unravel insights into complex diseases like OSCC, paving the way for improved clinical outcomes.

Contents

1 Oral Squamous Cell Carcinoma (OSCC)

Oral Squamous Cell Carcinoma (OSCC) is one of the most common malignancies of the head and neck, representing a significant global health burden. It typically arises from the epithelial cells of the oral cavity and can spread to other parts of the body if not detected and treated early. Despite advancements in surgical techniques and therapies, the overall survival rate for OSCC remains low due to late-stage diagnosis, often when the disease has already progressed to an advanced stage.

1.1 Importance of Early Detection and Biomarkers

Early detection of OSCC is critical for improving patient prognosis, as it significantly increases the chances of successful treatment and recovery. Traditional diagnostic methods, such as physical examination and biopsy, often detect the disease at later stages, where treatment options are limited. Therefore, identifying reliable biomarkers for OSCC is essential to enable early diagnosis, allowing for more effective interventions and personalized treatments.

Biomarkers, such as specific proteins, metabolites, or genetic changes, can serve as valuable indicators for the presence of OSCC at its earliest stages. Through advanced techniques like metabolomics and proteomics, we can identify and validate these biomarkers, facilitating the development of non-invasive diagnostic tests that can detect OSCC in its asymptomatic phase. This approach holds the potential to revolutionize early detection, making it possible to initiate treatment before the disease spreads, thereby improving survival rates and quality of life for patients. In this project, our goal is to identify potential biomarkers, including proteins and metabolites, that can accurately assess the status of a patient, enabling early diagnosis and personalized treatment strategies.

2 Proteomics Analysis

2.1 Introduction

In the first part of the project, we analyze biomarker proteins for classifying Tumor versus Normal Adjacent Tissue using various machine-learning models.

Dataset:- The dataset comprises protein expression data from 84 samples collected from 42 patients (Tumor and Normal from each patient), encompassing 3611 proteins. We applied several models, including Logistic Regression with Elastic Net, Support Vector Machine (SVM) with Recursive Feature Elimination (RFE) & Lasso Regression, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes, to identify significant biomarkers.

2.2 Methodology

2.2.1 Using MetaboAnalyst for Feature Reduction

We utilized MetaboAnalyst 6.0 (Visit MetaboAnalyst Module View) to remove unnecessary features from our dataset. The Partial Least Squares Discriminant Analysis (PLS-DA) method was employed, resulting in eight components. Each component provided Variable Importance in Projection (VIP) scores for each protein. We selected only those proteins with $avg\_vip\_score > 1$, leading to the identification of 894 significant proteins. After that, we reduced our original dataset to retain information on only these top **894** proteins. The formula for the VIP score is:-

$$\text{VIP} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\text{Var}(x_i|y)}{\text{Var}(x_i)} \right) \cdot \text{Corr}(x_i, y)^2$$

where n is the number of components, x_i represents the individual proteins, $\mathrm{Var}(x_i|y)$ is the variance of protein x_i conditioned on class y, and $\mathrm{Corr}(x_i, y)$ is the correlation between protein x_i and the class variable y.

Now, our objective is to identify the optimal subset of these 894 relevant features that contribute most effectively to the classification task.

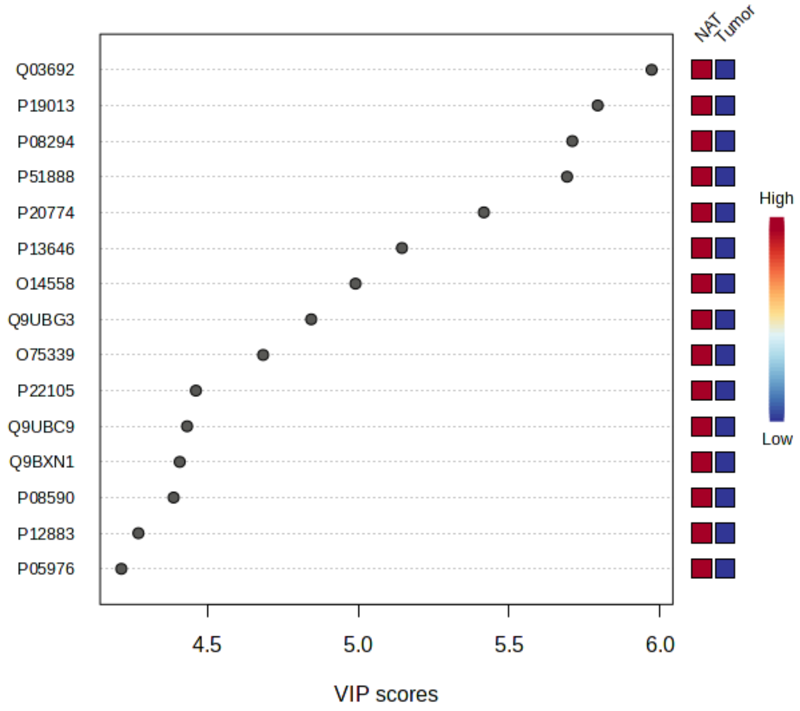


Figure 1: VIP scores of top proteins from MetaboAnalyst 6.0

2.2.2 Logistic Regression with Elastic Net Regularization

Logistic regression is a widely used statistical model for binary classification. The Elastic Net regularization combines L1 and L2 penalties, balancing the trade-off between feature selection and multicollinearity. The objective function for Elastic Net is given by:

$$L(\beta) = -\sum_{i=1}^{n} \left(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right) + \lambda \left(\alpha\|\beta\|_1 + \frac{(1-\alpha)}{2}\|\beta\|_2^2\right)$$

where p_i is the predicted probability, λ is the regularization parameter, and α controls the mixing of L1 and L2 penalties.

L1 Regularization (Lasso):- L1 regularization promotes sparsity in the model by driving some coefficients to exactly zero, thus performing implicit feature selection. This is particularly advantageous when dealing with high-dimensional data, such as in our case, where many features may be irrelevant. By keeping only the most relevant predictors, L1 regularization simplifies the model and enhances interpretability. However, it may also lead to instability in coefficient estimates, especially when features are highly correlated.

L2 Regularization (Ridge): In contrast, L2 regularization shrinks the coefficients towards zero without completely eliminating any features. This approach can stabilize the model in the presence of multicollinearity, where predictors are correlated with each other. By applying a penalty proportional to the square of the coefficient values, L2 regularization helps maintain all features while reducing their influence on the model predictions.

Elastic Net regularization combines both L1 (Lasso) and L2 (Ridge) penalties, allowing us to achieve a balance between their individual strengths. We selected the top 50 features from this model using the `SelectFromModel` method using the feature relevance.

2.2.3 Support Vector Machine (SVM) with L1 Regularization

SVM is a powerful classification technique that finds the optimal hyperplane to separate classes in the feature space. In our implementation, we used a linear kernel with L1 regularization. The decision boundary is determined by:

$$w^T x + b = 0$$

where w is the weight vector, x is the input vector, and b is the bias. The top 27 features with non-zero coefficients were selected from the fitted model using the 'SelectFromModel' approach.

We also tried a Support Vector Machine (SVM) model with Recursive Feature Elimination (RFE) for feature selection. RFE iteratively removes the least important features, identifying significant predictors for our classification task. The SVM with RFE achieved an accuracy of approximately 100% using 5-fold cross-validation.

While RFE effectively selects features, it may lead to overfitting. To address this, we applied L1 regularization, which not only performs feature selection but also penalizes large coefficients. This approach mitigates overfitting, particularly in high-dimensional spaces, by shrinking some coefficients to zero, thus simplifying the model and enhancing its generalizability. By retaining only the most relevant features, we reduce the complexity associated with irrelevant predictors.

2.2.4 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. It is robust to overfitting and works well for high-dimensional data. The feature importance is calculated as:

$$\text{Importance}(f) = \sum_{t=1}^{T} \frac{N_t}{N} \cdot \Delta Gini_t$$

where N_t is the number of samples in tree t and $\Delta Gini_t$ is the decrease in Gini impurity contributed by feature f. We selected the top 50 features based on importance scores to retrain the model and got a 5-fold cross-validation accuracy of around 97%.

2.2.5 K-Nearest Neighbors (KNN) and Naive Bayes

Both KNN and Naive Bayes classifiers were tested using the top common selected features. KNN classifies a data point based on the majority class among its k nearest neighbours, while Naive Bayes applies Bayes' theorem with the assumption of independence among features:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

We found around 94% accuracy with both KNN and Naive Bayes using only the top 5 common features from the selected features of the above models.

2.3 Results of Different Models

2.3.1 Logistic Regression with Elastic Net regularisation

The Logistic Regression model, utilizing Elastic Net regularization, successfully identified the top 50 proteins from the dataset. After selecting the top features, we reduced the train and test datasets to only contain these top 50 proteins and then retrained the logistic regression model with this modified dataset. The accuracy achieved by the Logistic Regression model was 94% on the reduced test set and mean cross-validation accuracy (train set): 0.84.

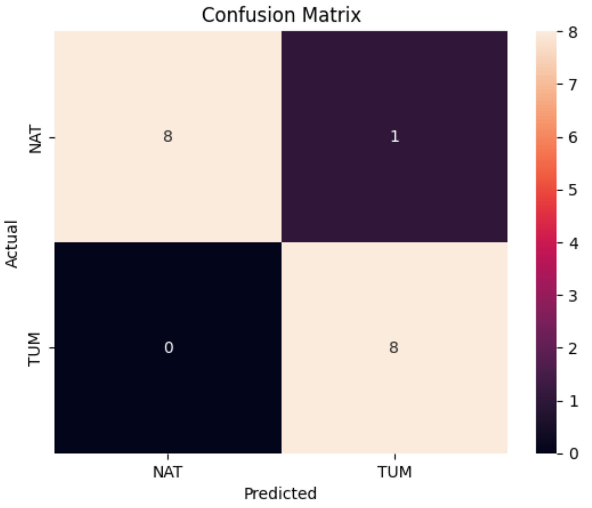


Figure 2: Confusion Matrix of logistic regression model

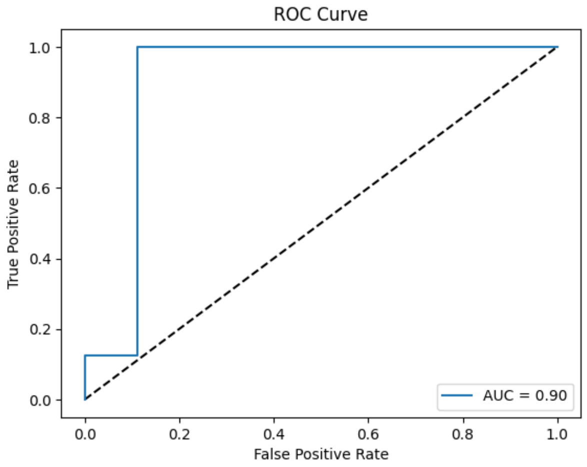


Figure 3: Roc curve of logistic regression model

The top 50 proteins identified by the model as per their UniProt IDs, are as follows:

- Q96IU4, P30838, Q9BXN1, P55957, P35613, Q13895, O75339, P23946, Q03692, Q05707
- P02461, Q8NBJ5, P15088, P52943, Q9UBG3, O75718, Q53TN4, O95865, P07099, P34913
- Q96AY3, O95302, P29992, P22352, P07305, P00738, Q92598, P13646, P19013, Q13751
- Q9UHB6, Q08AI8, P27338, Q99685, Q99735, P35749, O43795, Q8TCD5, P20774, Q6UWY5
- P02763, P13674, Q08174, P51888, Q13308, P50454, Q9NR46, P08195, P08294, P02786

2.3.2 Random Forest Model

The Random Forest model identified the top 50 proteins from the dataset. After selecting the top features, we reduced the train and test datasets to only contain these top 50 proteins and retrained the Random Forest model with the modified dataset. The accuracy achieved by the Random Forest model was 94% on the reduced test set and mean cross-validation accuracy (train set): 0.97.

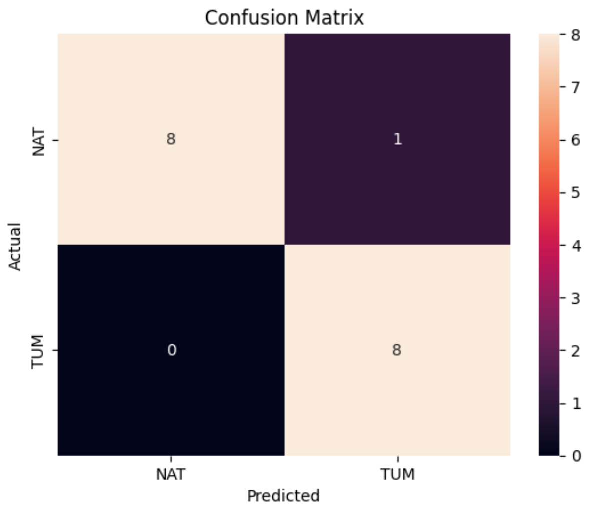


Figure 4: Confusion Matrix of Random Forest model

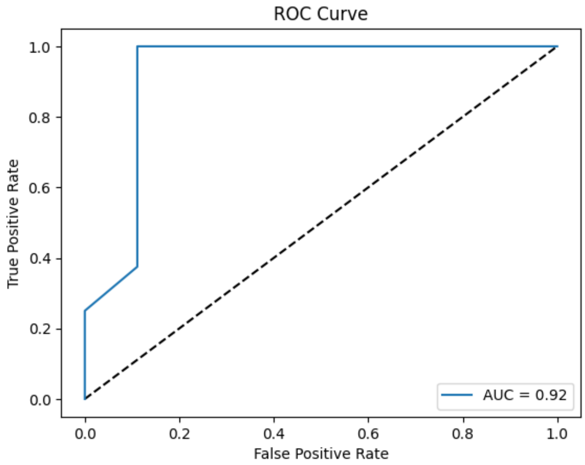


Figure 5: ROC Curve of Random Forest model

The top 50 proteins identified by the Random Forest model, according to their UniProt IDs, are as follows:

- P27338, Q13813, O43795, P08195, O95865, Q6UWY5, Q99735, P08294, Q9UN36, Q8N335
- Q16853, P00746, P29992, Q03692, P05166, Q13228, Q53TN4, Q08J23, Q92598, O95302
- P30086, P55268, P51888, Q9Y6K5, P20774, Q9UEY8, Q07507, O14933, Q96AG4, Q16762
- Q13509, P05091, P13674, Q13308, O00339, P49189, P00167, Q9HBL0, P17812, Q9H0A0
- Q969V3, P24821, Q92506, P51608, Q9BX66, P25311, P32455, P51884, P14927, Q63ZY3

2.3.3 Support Vector Machine (SVM) Model

The Support Vector Machine (SVM) model, employing Lasso regularization, selected the top 27 proteins from the dataset. After selecting these proteins, we reduced the train and test datasets to only include these features and retrained the SVM model. The SVM model achieved an accuracy of 94% on the reduced test set and mean cross-validation accuracy (train set): 0.98.

The top 27 proteins identified by the SVM model, according to their UniProt IDs, are as follows:

- P55957, O75339, Q03692, Q9NZJ6, P52943, Q03001, P34913, Q96AY3, P35269
- P00738, Q92598, Q92743, P01876, P13646, Q2M2I5, Q9NX58, P27338, Q9Y623
- Q8TCD5, Q6UWY5, P02763, Q08174, Q15063, P51888, Q9BXM0, Q9UBD6, Q66K66

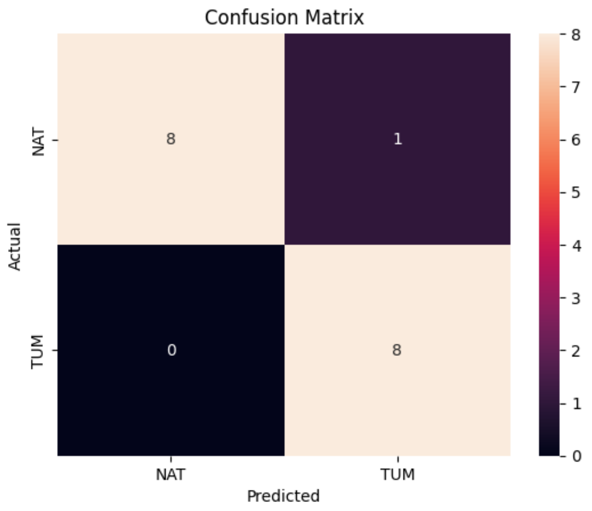


Figure 6: Confusion Matrix of SVM model

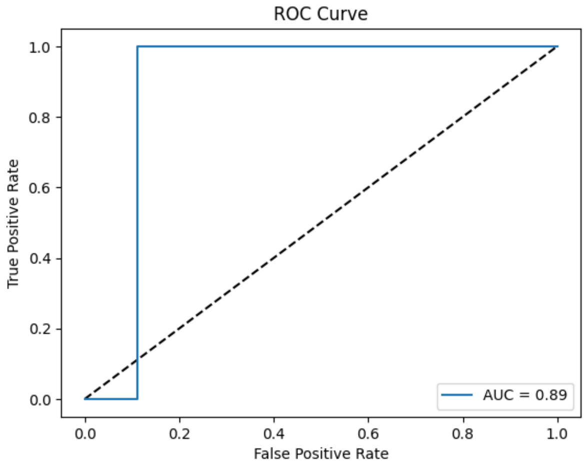


Figure 7: ROC Curve of SVM model

2.3.4 Comparison of Common Proteins Across Models

After comparing the proteins selected by all three models, we identified 5 common proteins that appeared in the feature sets of all models. These proteins, with their corresponding UniProt IDs, are:

- `P27338`
- `P51888`
- `Q03692`
- `Q6UWY5`
- `Q92598`

To evaluate the significance of these proteins, we modified the dataset to include only these five proteins. Upon retraining the models using only these features, we achieved an accuracy of 94% with the Random Forest model and 88% with both the Logistic Regression and SVM models. So, the Random Forest Model emerged as the best-performing model among those evaluated.

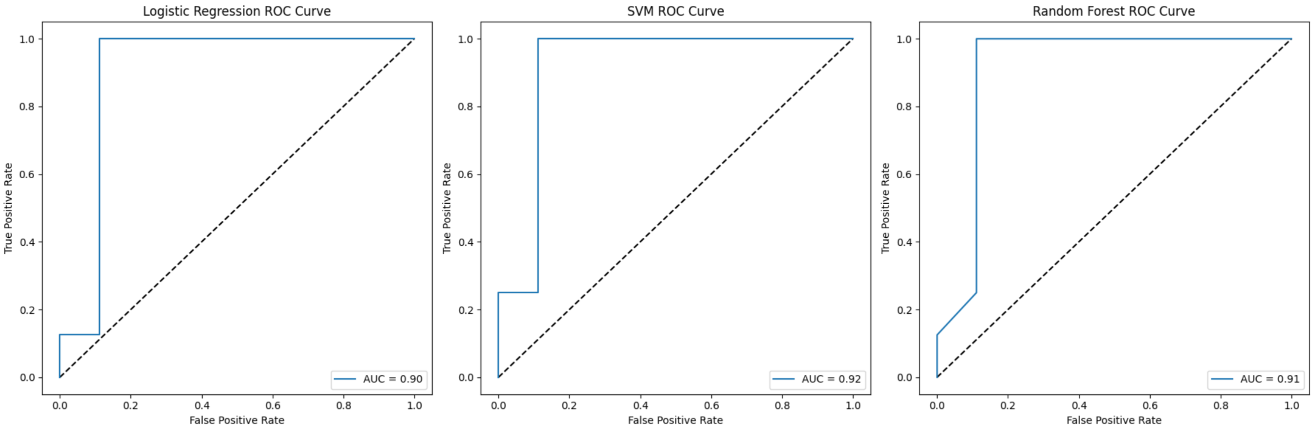


Figure 8: ROC Curves for Common Proteins Across Models

2.3.5 Observational Trends

We plotted the boxplots for the 5 common proteins, along with their corresponding t-statistics and p-values, using data from the original dataset of 42 patients. The boxplots highlight a decreasing trend for the first four proteins from NAT to Tumor and an increasing trend for the last protein from NAT to Tumor. Below are the boxplots for these 5 common proteins:

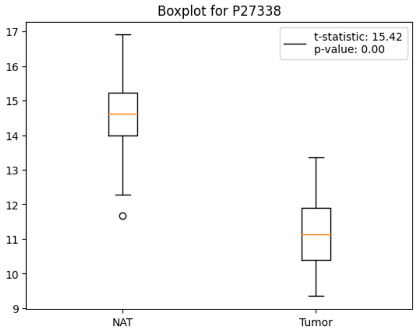


Figure 9: Boxplot for P27338

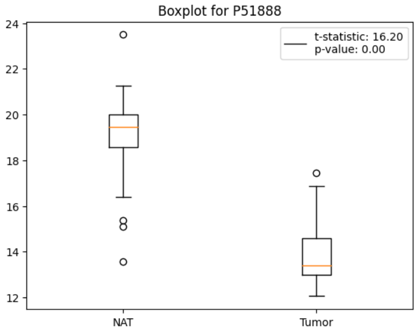


Figure 10: Boxplot for P51888

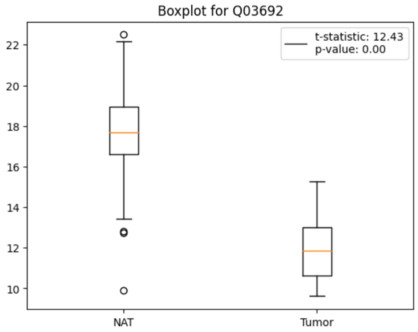


Figure 11: Boxplot for Q03692

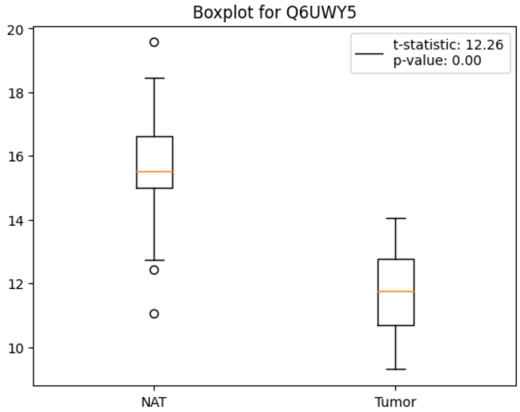


Figure 12: Boxplot for Q6UWY5

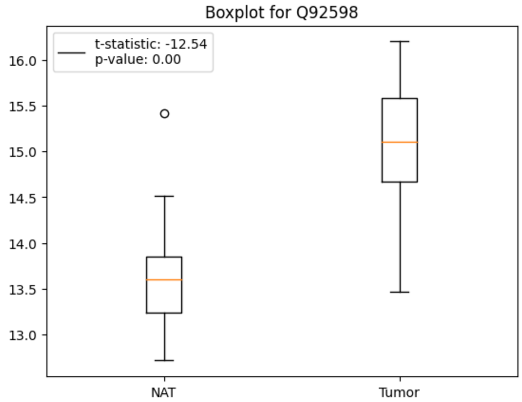


Figure 13: Boxplot for Q92598

2.4 Validation of Selected Biomarkers

CPTAC (Clinical Proteomic Tumor Analysis Consortium) is a collaborative initiative by the National Cancer Institute (NCI) aimed at advancing the understanding of cancer biology through the comprehensive characterization of tumour proteomes. By analyzing cancer samples from various tumour types, CPTAC seeks to identify molecular biomarkers for early detection, diagnosis, and treatment.

To validate the selected biomarkers, we utilized the `CPTAC_TumorVsNormal.xlsx` dataset, which comprises approximately 172 tumour and normal samples, each characterized by 9,666 distinct genes. We employed KNN imputation with a neighbourhood of 5 to address any missing values in the dataset. Next, we mapped our top-selected proteins to their corresponding genes and filtered the genes to retain only those that represent our top features. We evaluated the performance of the Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression models, achieving 5-fold cross-validation accuracies of 97.5%, 97.85%, and 95.83%, respectively, along with 100% accuracy on the test set for the SVM model and around 96% for both RF and Logistic Regression. The SVM model exhibited the best performance, followed closely by the RF model on the CPTAC dataset.

3 Metabolomics Analysis

3.1 Introduction

In the second part of the project, we have focused on the metabolomics study. While proteomics data provides insight into protein-level changes, metabolomics data highlights alterations in metabolic pathways. This study focuses on the analysis of two metabolomics datasets:

- **3Groups-OSCC_PM_Normal.xlsx**: Contains 41 columns representing mixed groups (OSCC, PM, Normal) with 1,378 metabolites per column across two sheets (Positive and Negative).
- **TumorVsNAT_Metabolomics_Tissue.xlsx**: Contains 31 columns (mixed Tumor and NAT groups) with 3,771 metabolites per column across two sheets (Positive and Negative).

We integrate these metabolomics analyses with the previously conducted proteomics study to identify the top discriminative features using statistical and machine learning techniques.

3.2 Methodology

3.2.1 Data Preprocessing

- Each dataset was preprocessed to ensure consistent handling of missing values, normalization, and scaling.
- Metabolomics data was analyzed separately for each dataset's Positive and Negative sheets.

3.2.2 MetaboAnalyst Analysis

- For the **Tumor vs NAT** dataset:
  - Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) was performed in MetaboAnalyst 6.0.
  - Features with predictive variance score greater than a certain threshold were selected, narrowing the list to approximately top 500 metabolites.
- For the **3Groups** dataset:
  - Partial Least Squares Discriminant Analysis (PLS-DA) was applied using MetaboAnalyst 6.0.
  - The average score across 8 components was calculated, and a threshold was applied to select approximately top 250 metabolites.

After this we have reduced our dataset to contain only these top metabolites, which would be further refined using different machine learning models. Below are the graphs of the vip scores of the top metabolites selected by the MetaboAnalyst 6.0.

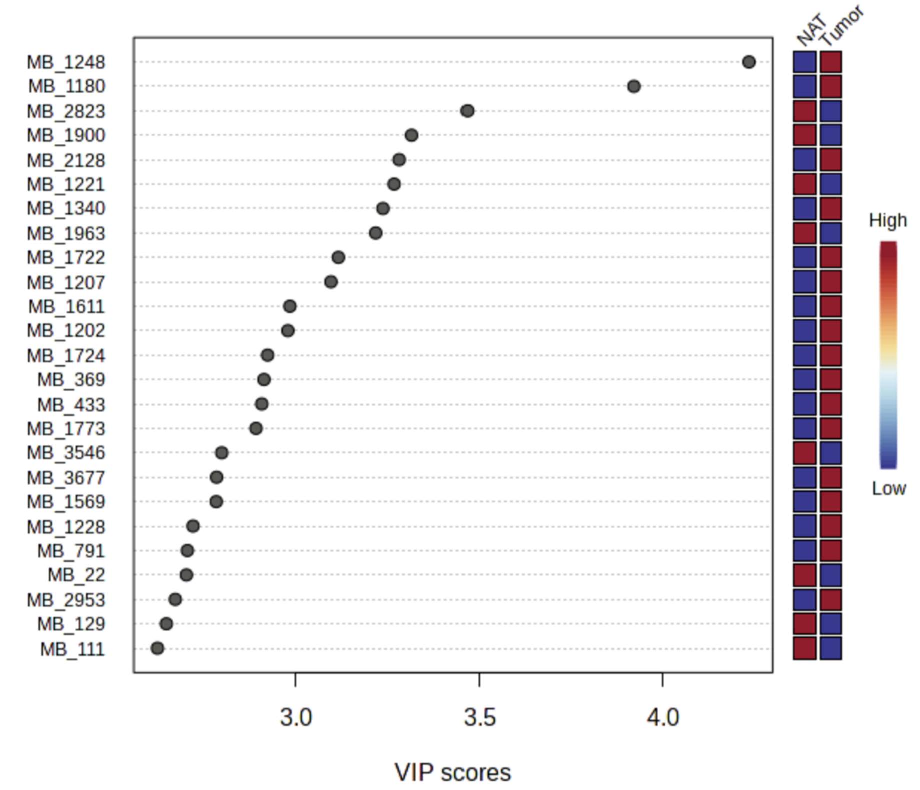


Figure 14: VIP scores of top metabolites (Positive sheet - Tum vs NAT)

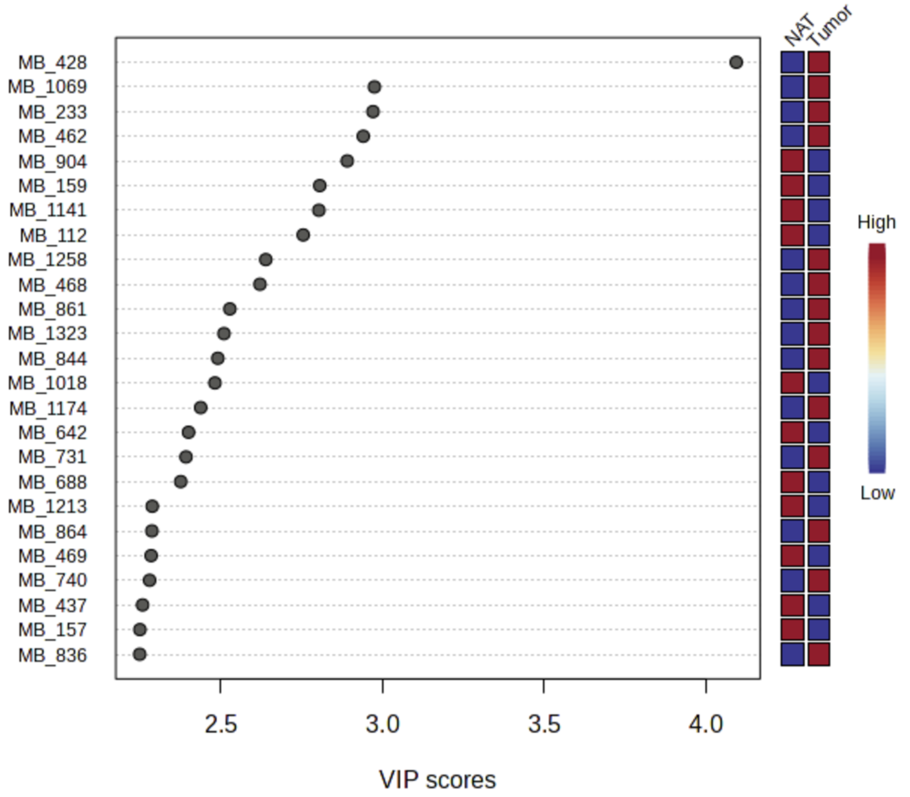


Figure 15: VIP scores of top metabolites (Negative sheet - Tum vs NAT)

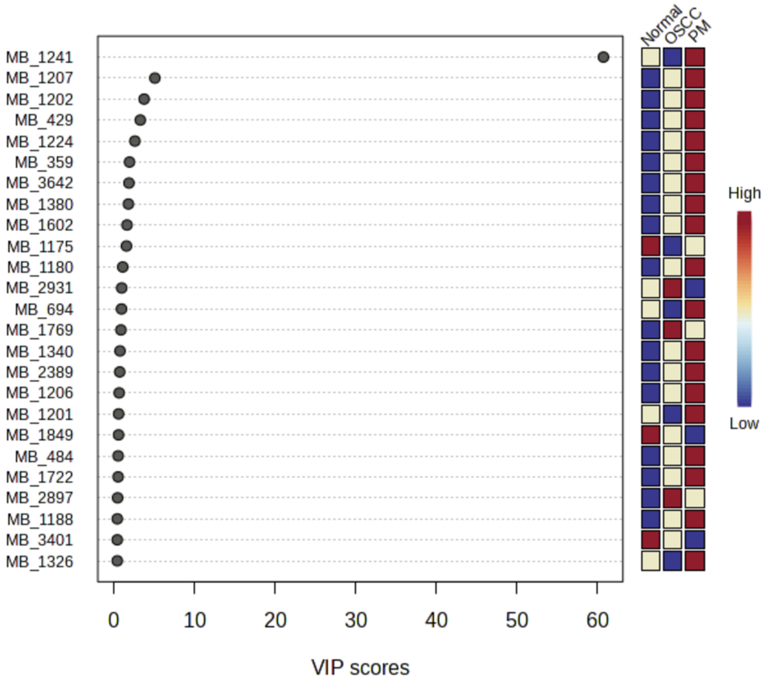


Figure 16: VIP scores of top metabolites (Positive sheet - 3Groups: OSCC, PM, Normal)

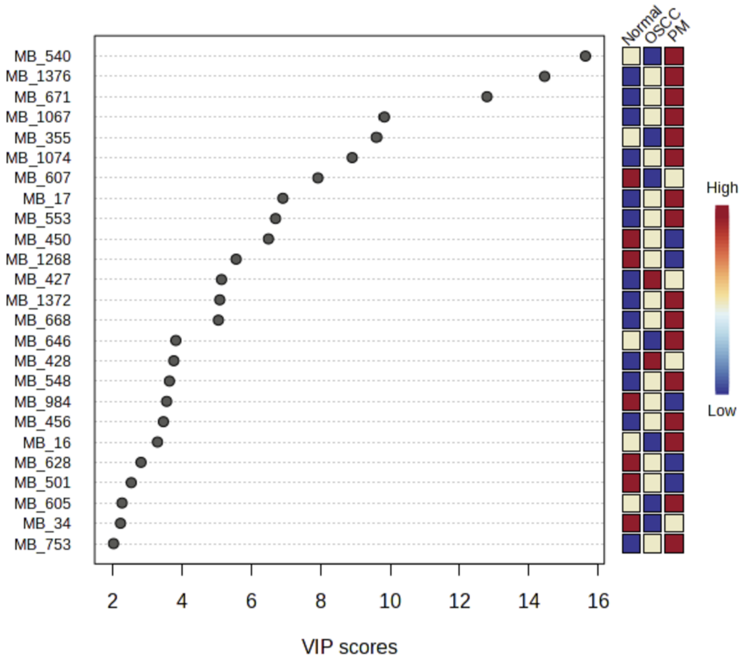


Figure 17: VIP scores of top metabolites (Negative sheet - 3Groups: OSCC, PM, Normal)

3.2.3 Machine Learning Models

To further refine the features, the following machine learning models were employed:

1. **Logistic Regression with Elastic Net Regularization**:
   The logistic regression model was implemented with Elastic Net regularization, which combines L1 (Lasso) and L2 (Ridge) penalties to balance feature selection and regularization. The model parameters were tuned using 5-fold cross-validation, ensuring robust performance evaluation. We have selected top 25 features using the `SelectFromModel` method.

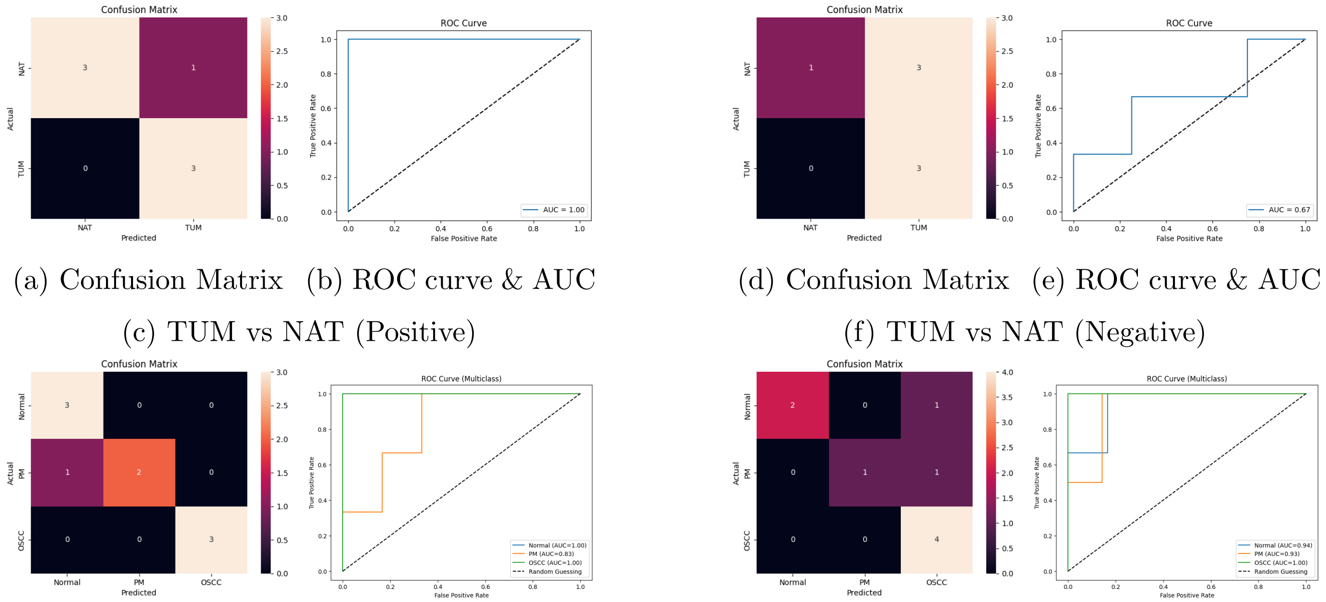
2. **Random Forest**:
   A Random Forest classifier was implemented, and hyperparameters were optimized using `GridSearchCV` with 5-fold cross-validation. The tuned parameters included: **Number of Trees** (100, 200, 500), **Max Depth** (10, 20, None), and **Minimum Samples Split** (2, 5, 10). The best hyperparameters were identified and used to train the model. Feature importance scores were extracted, and the top 25 features were selected for further analysis.

3. **SVM with L1 Regularization**:
   A Linear Support Vector Machine (SVM) with L1 regularization was implemented to perform simultaneous classification and feature selection. The configuration included: **Penalty:** $L1$, enabling sparse feature selection; **Dual Optimization:** set to `False` for compatibility with smaller datasets; and **Maximum Iterations:** increased to 10,000 for convergence. Using L1 regularization, the model selected top features based upon non-zero coefficients.

   To adapt the Linear Support Vector Machine (SVM) model with L1 regularization for multiclass classification, several modifications were made. The dataset labels were binarized using `label_binarize` to facilitate the calculation of individual ROC curves for each class. A multiclass confusion matrix was generated to assess classification performance across the categories, namely *Normal*, *PM*, and *OSCC*. Feature selection was performed using non-zero coefficients from the L1-regularized model to select the top 25 features.

3.3 Results of Different Models

3.3.1 Logistic Regression with Elastic Net



(a) Confusion Matrix (b) ROC curve & AUC

(c) TUM vs NAT (Positive)

(d) Confusion Matrix (e) ROC curve & AUC

(f) TUM vs NAT (Negative)

(g) Confusion Matrix (h) ROC curve & AUC

(i) 3groups (Positive)

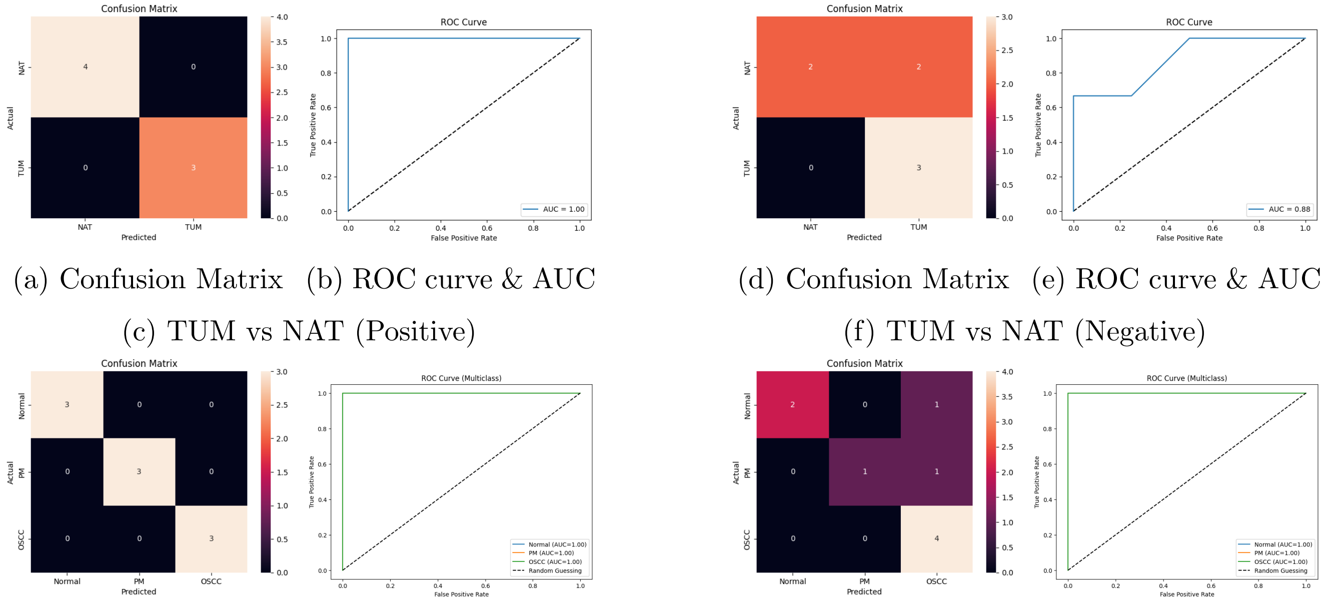
(j) Confusion Matrix (k) ROC curve & AUC

(l) 3groups (Negative)

Figure 18: Model outputs from datasets with top 25 features

The logistic regression model is not performing well as compared to the other 2 models with accuracy approximately 60% in each dataset.

3.3.2 Random Forest Model



(a) Confusion Matrix (b) ROC curve & AUC

(c) TUM vs NAT (Positive)

(d) Confusion Matrix (e) ROC curve & AUC

(f) TUM vs NAT (Negative)

(g) Confusion Matrix (h) ROC curve & AUC

(i) 3groups (Positive)

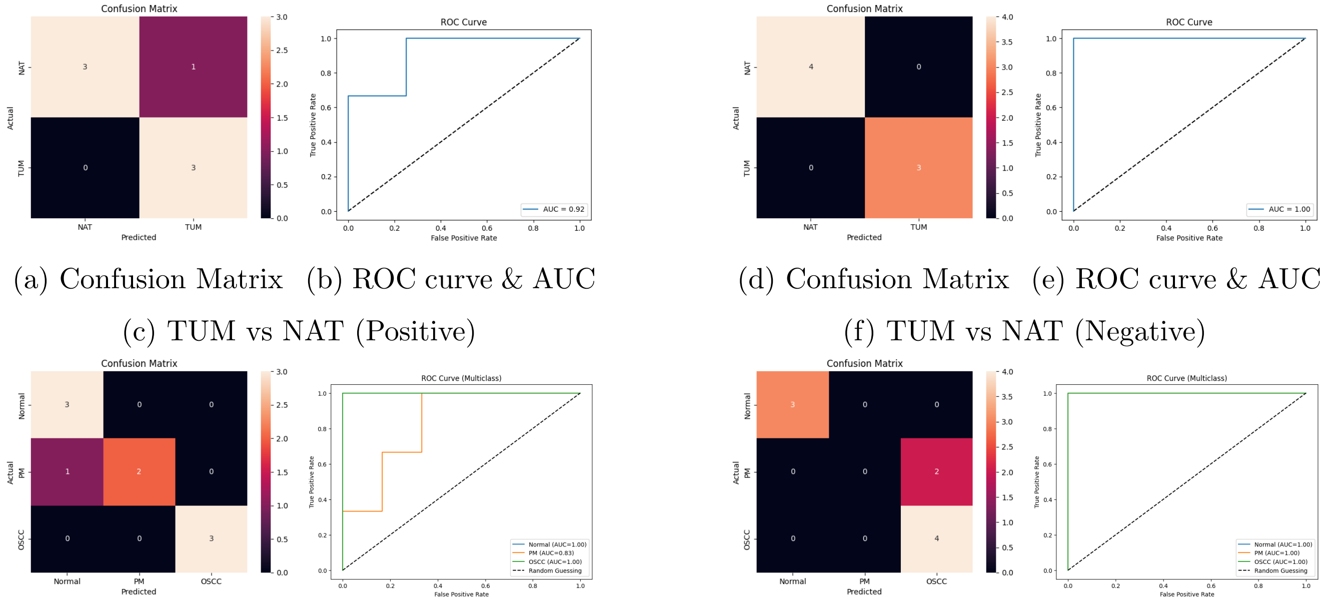
(j) Confusion Matrix (k) ROC curve & AUC

(l) 3groups (Negative)

Figure 19: Model outputs from datasets with top 25 features

The Random Forest model is outperforming the other two models in positive datasets with accuracy on the test sets 100% in each positive dataset.

3.3.3 SVM with L1 regularisation



(a) Confusion Matrix (b) ROC curve & AUC

(c) TUM vs NAT (Positive)

(d) Confusion Matrix (e) ROC curve & AUC

(f) TUM vs NAT (Negative)

(g) Confusion Matrix (h) ROC curve & AUC

(i) 3groups (Positive)

(j) Confusion Matrix (k) ROC curve & AUC

(l) 3groups (Negative)

Figure 20: Model outputs from datasets with top 25 features

The SVM model with L1 regularisation is outperfoming the other two models in the negative datasets with accuracy of 100% in the Tum Vs Nat (negative) dataset and around 80% in the

3groups (negative) dataset. The confusion matrices provide insights into the classification accuracy, while the ROC curves highlight the model's ability to distinguish between classes across datasets. We can see from the above plots that the Random Forest model consistently performs better than the SVM with L1 regularisation across different datasets.
I have put the top 25 features selected by various models from all the 4 datasets in the result section of the metabolomics directory (Results Directory). I could not find any top feature which is common in the top 25 selected features by the Random Forest model across 4 datasets.

4 Conclusion

This study demonstrated the power of integrating proteomics and metabolomics datasets with advanced machine learning techniques to identify robust biomarkers for the early detection of Oral Squamous Cell Carcinoma (OSCC). By leveraging high-dimensional datasets, we applied systematic feature selection and classification methodologies, achieving promising results in both the proteomics and metabolomics domains.

Key Findings

1. **Proteomics Analysis**:

   - Using machine learning models such as Logistic Regression with Elastic Net, Support Vector Machines (SVM) with L1 regularization, and Random Forest, we identified a subset of highly relevant proteins from a dataset of 3,611 proteins.
   - Notably, five proteins (P27338, P51888, Q03692, Q6UWY5, Q92598) were common across all models, demonstrating their potential as robust biomarkers for OSCC classification.

2. **Metabolomics Analysis**:

   - A combination of statistical techniques (e.g., OPLS-DA and PLS-DA) and machine learning models narrowed down thousands of metabolites to a concise set of top features.
   - Models such as Random Forest and SVM demonstrated superior performance, achieving accuracies up to 100% on specific datasets, highlighting the discriminative power of the selected metabolites.

3. **Model Comparison**:

   - The Random Forest model emerged as the most consistent performer across various datasets, particularly excelling in positive metabolomics datasets, achieving 100% accuracy on test sets.
   - The SVM model with L1 regularization was particularly effective in negative datasets, demonstrating its adaptability for different dataset characteristics.

Significance

By integrating the top 25 features from metabolomics with proteomics biomarkers, we developed a comprehensive biomarker panel. This integrative approach has the potential to significantly improve the early detection of OSCC, enabling personalized treatment strategies and better patient outcomes. The ability to identify common biomarkers across independent datasets further strengthens the reliability of these findings.

Limitations and Future Directions

1. **Validation**:

   - Although the models achieved high accuracy, further validation on independent datasets, such as those from CPTAC, is necessary to confirm the generalizability of the identified biomarkers.
   - External validation using larger, more diverse patient cohorts will enhance the translational potential of these biomarkers.

2. **Clinical Applicability**:

   - Developing cost-effective, non-invasive diagnostic tools based on the identified biomarkers will be a crucial next step. Techniques like liquid biopsy and targeted metabolomics can be explored for this purpose.

3. **Integration with Clinical Data**:

   - Incorporating clinical variables, such as patient demographics and medical history, alongside omics data may improve model performance and provide more comprehensive insights into OSCC progression.

In conclusion, this study highlights the utility of combining machine learning techniques with high-dimensional biological data to uncover critical insights into OSCC. The integration of proteomics and metabolomics features holds significant promise for improving early detection, guiding therapeutic interventions, and ultimately enhancing clinical outcomes. Continued efforts in validation and clinical translation are imperative to realize the full potential of these findings.

References

[1] Judith, Laura, et al. "Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment." *Frontiers in Microbiology*, vol. 12, 2021, p. 634511. https://doi.org/10.3389/fmicb.2021.634511. Accessed 1 Oct. 2024.

[2] Shi, Zhiao, et al. "Feature Selection Methods for Protein Biomarker Discovery from Proteomics or Multiomics Data." *Molecular & Cellular Proteomics*, vol. 20, 2020, p. 100083. https://doi.org/10.1016/j.mcpro.2021.100083. Accessed 1 Oct. 2024.

[3] Leclercq, Mickael, et al. "Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data." *Frontiers in Genetics*, vol. 10, 2019, p. 449967. https://doi.org/10.3389/fgene.2019.00452. Accessed 1 Oct. 2024.

[4] Al-Tashi, Qasem, et al. "Machine Learning Models for the Identification of Prognostic and Predictive Cancer Biomarkers: A Systematic Review." *International Journal of Molecular Sciences*, vol. 24, no. 9, 2023. https://doi.org/10.3390/ijms24097781. Accessed 1 Oct. 2024.

[5] Chen, Yangzi, et al. "Metabolomic Machine Learning Predictor for Diagnosis and Prognosis of Gastric Cancer." Nature Communications, vol. 15, no. 1, 2024, pp. 1-13, https://doi.org/10.1038/s41467-024-46043-y. Accessed 27 Nov. 2024.

[6] Chong, J., Wishart, D.S., Xia, J. "Using MetaboAnalyst 6.0 for Comprehensive and Integrative Metabolomics Data Analysis." Current Protocols in Bioinformatics, vol. 68, 2019, e86. https://dev.metaboanalyst.ca/ModuleView.xhtml. Accessed 1 Oct. 2024.