Nile University

School of Information Technology and Computer Science

Program of Biomedical Informatics

AI-Assisted Prognosis Prediction using Multi-omics

Senior Project II

Submitted in Partial Fulfilment of the Requirements

For the Bachelor's Degree in Information Technology and Computer Science

Biomedical Informatics

Submitted by

Eyad El-Hawary – 202001250

Merna Medhat – 202002266

Reem Amin – 202001495

Toka Radwan – 202002280

Youmna Tarek – 202001016

Supervised by

Dr. Mai Said

Giza – Egypt

Spring 2024

Project Abstract

In this Study, we offer an integrated method thorough the analysis of RNA, miRNA, and DNA methylation data from The Cancer Genome Atlas (TCGA) utilizing Multi-Omics Factor Analysis (MOFA). Our approach includes thorough omics data preparation and normalization to guarantee precision and consistency across various molecular datasets. To gain more understanding into stomach adenocarcinoma, the MOFA framework is used to find hidden patterns and interactions within these multi-omics datasets. Additionally, to estimate patient prognosis with high accuracy, predictive models such XGBoost, Random Forest, Support Vector Machine (SVM), and Logistic Regression are developed. The test set accuracy of 0.934 indicates that SVM is the best-performing model, according to the results. These results highlight the effectiveness of combining innovative machine learning methods with multi-omics data to improve comprehension and prediction in cancer research.

Keywords: Multi-Omics Data, Multi-Omics Factor Analysis (MOFA), Stomach adenocarcinoma, Data Preprocessing, Predictive models, Prognosis Prediction.

Table of Contents

List of **Figures**

List of **Tables**

1. Chapter 1: Introduction

1.1. Background:

Gastric cancer (GC) is the fifth most common type of cancer diagnosed worldwide and the fourth leading cause of death from cancer. Despite remarkable advances in diagnosis and therapy techniques, as well as major gains in patient survival, the low malignancy stage is relatively asymptomatic, and many GC patients are discovered at advanced stages, resulting in an unfavorable prognosis and a high recurrence rate.[3] Stomach cancer is still a major cancer worldwide, accounting for over 1,000,000 new cases in 2018 and an estimated 783,000 deaths (equating to one in every 12 deaths globally) [4]. In recent years, various applications of artificial intelligence (AI) have evolved in the stomach cancer sector, owing to its efficient processing capacity, and learning capabilities, such as image-based diagnosis and prognosis prediction [7]. So, to fully comprehend SC's complex biology, multiple omics layers should be examined, such as epigenomics, and transcriptomics. Genomic alterations in stomach adenocarcinoma encompass a spectrum of mutations, copy number variations, and chromosomal rearrangements. These genomic changes contribute to dysregulated signaling pathways involved in cell proliferation, survival, and metastasis. The development and progression of stomach adenocarcinoma is significantly influenced by epigenetic dysregulation,

including alterations in DNA methylation patterns. Among tumor suppressor genes and oncogenes, abnormal DNA methylation disturbs gene expression profiles, promoting tumorigenesis and resistance to therapeutic treatments [2]. Growth in omics technologies, such as epigenomics and transcriptomics, have begun to enable personalized therapy at a previously unknown molecular level. Individually, these technologies have produced medical advances that have begun in clinical settings. However, each technique alone cannot capture the whole biological complexity of most human disorders. Integration of several technologies has evolved as a method for providing a more comprehensive understanding of biology and illness [5]. Several notable 'omics' platforms used in microbial systems biology include transcriptomics, which analyses mRNA transcript levels, and proteomics, which quantifies protein abundance.

Even so, no single 'omics' analysis can adequately explain the complexity of underlying microbial biology. To obtain a detailed image of living microorganisms, numerous layers of information must be integrated, a process known as the multi-omics approach. Despite the challenging nature of this task, recent efforts to integrate heterogeneous 'omics' datasets in various microbial systems have shown that the multi-'omics' approach is a powerful tool for understanding the functional principles and dynamics of total cellular systems [6].

1.2. Motivation:

Gastric cancer remains a serious clinical challenge due to its advanced stage at diagnosis, poor prognosis, and limited treatment choices, despite a drop in prevalence in some regions of the world. H. pylori infection, combined with dietary variables such as a high intake of nitrites and salted foods, is a major contributor to stomach cancer [9]. Improvements in food preservation techniques and H. pylori infection therapy have helped to reduce stomach cancer impact overall. However, there has been an increase in proximal gastric tumors, which often have a worse prognosis. Eastern Asian countries, which account for more than half of all stomach cancer diagnoses, benefit from effective screening programmers that promote early identification and improved outcomes. In contrast, Western countries confront obstacles due to the comparatively low prevalence of stomach cancer, making comprehensive screening economically impractical. As a result, stomach tumors in these areas are frequently detected at an advanced stage [8].

Additionally, gastric cancer has a significant societal impact, both in terms of healthcare expenses and the quality of patients' lives. As a result, there is an urgent need for improved therapy tactics and more accurate prognosis prediction tools. This emphasizes the necessity of developing novel approaches, such as combining AI and multi-omics data, to improve early detection and personalized treatment of

stomach cancer, thereby improving patient outcomes and lowering societal costs. So, we are trying to contribute to solving this issue by providing an ai system that prognose and predicts this type of cancer.

1.3. Objectives:

Stomach adenocarcinoma faces significant challenges in prognosis prediction and treatment tailoring due to its complex molecular landscape [2]. Due to their reliance on clinical characteristics, conventional prognostic models are restricted in their ability to capture the complex interplay of molecular elements that drive the course of disease and the response to treatment. Our work intends to use advanced analytical tools, such as Multi-Omics Factor Analysis (MOFA) model, in conjunction with integrated multi-omics data to solve these constraints, improve patient outcomes, and deepen our understanding of stomach cancer. Multi-Omics Factor Analysis (MOFA) model, a powerful approach for integrating heterogeneous omics data, to do this. By considering the interdependencies between various omics datasets, MOFA makes it possible to analyze them all at once and captures the complex interactions that underlie the development of cancer and its response to therapy [7]. In this study, Multi-Omics Factor Analysis (MOFA) model was constructed to integrate multi-omics data of patients from The Cancer Genome Atlas (TCGA) cohort with stomach adenocarcinoma to identify the risk and predict the prognosis, including RNA expression, miRNA expression,

and DNA-methylation. Briefly, our research aims to convert these discoveries into practical understandings for personalized medicine methods in cancer therapy, enabling the creation of customized treatment plans predicated on the distinct genetic traits of stomach adenocarcinoma cancer.

1.4. Scope:

Our significant goal is to develop an artificial intelligence (AI) model that leverages multi-omics data to predict the prognosis of stomach cancer patients. Our study intends to address the limits of existing predictive models for stomach adenocarcinoma by combining multi-omics data, which frequently rely on a set of clinical variables. The molecular landscape of this carcinoma is extremely complicated, needing a more comprehensive approach than existing models provide. To accomplish this, we intend to integrate multiple omics data types, including transcriptomics (RNA expression data) and epigenomics (DNA methylation patterns).

Furthermore, to improve the practical applicability of our research, created a website that automates the Multi-Omics Factor Analysis (MOFA) process and predicts prognosis for stomach cancer patients. MOFA is an effective technique for integrating multi-omics data, allowing the identification of common elements that drive biological variability across several types of data [7]. By automating this

procedure, our website will make it easier to analyze RNA, miRNA (microRNA), and DNA methylation data. This will give researchers and physicians a user-friendly platform for entering multi-omics data and receiving prognostic predictions.

Our holistic strategy considers the characteristics of several omics data sources to gain a better understanding of stomach cancer. Transcriptomics analyzes gene expression levels to determine which genes are elevated or downregulated in stomach cancer. Epigenomics studies DNA methylation patterns, which can either silence or trigger gene expression, providing insights into the regulatory mechanisms involved in cancer progression. Furthermore, miRNA data aids research into the role of microRNAs, which are tiny non-coding RNAs involved in gene regulation, in cancer development and prognosis.

By combining all these data sources using MOFA, our AI model will uncover latent factors that reflect the underlying structures and variances shared by the various omics layers. This integration enables a more comprehensive understanding of the biological processes underlying cancer growth, resulting in more accurate and personalized prognostic predictions. The website we created automates the MOFA pipeline, making advanced multi-omics analysis available to a wider audience, including researchers and doctors who may lack specialized computing competence. This automation will improve the analysis process by

lowering the time and effort required to extract useful insights from multi-omics data.

Finally, Our AI model will use integrated multi-omics data to estimate patient prognosis, improving treatment programs, patient outcomes, and potentially guiding targeted medicines. This technique will enhance the accessibility and usefulness of multi-omics analysis, leading to more personalized cancer treatments.

1.5. Significance of the Study:

The study aims to improve oncology, particularly in the prognosis and treatment of stomach adenocarcinoma, by integrating multi-omics data such as transcriptomics, miRNA, and DNA methylation. The AI model aims to provide a comprehensive understanding of the molecular foundations of stomach cancer, addressing constraints of current models that focus on a few clinical variables. The research also pioneers the use of Multi-Omics Factor Analysis (MOFA) in clinical settings, allowing for the identification of novel biomarkers and therapeutic targets. This could lead to more effective treatments, better patient outcomes, and a lower societal burden of stomach cancer. The study also created a website to automate the MOFA pipeline for prognostic prediction, making it more accessible to academics and clinicians. The AI model can help oncologists generate personalized

treatment strategies, enhance patient care, and optimize resource allocation within healthcare systems. The project's ability to combine diverse data sources fosters cooperation across fields like genetics, bioinformatics, and clinical research, potentially leading to advancements in cancer biology and treatment.

1.6. Outline the structure of the report:

This study is divided into six major chapters, each with a specific purpose in presenting research on AI-assisted prognosis prediction of stomach adenocarcinoma cancer using multi-omics data. Here's an overview of the report structure:

Chapter 1: Introduction

This chapter establishes the basic context and structure for the research. It has the following sections:

- 1.1. Background: An overview of stomach adenocarcinoma, its clinical significance, and current challenges in prognosis prediction.
- 1.2. Motivation: The driving reasons behind this research, emphasizing the need for improved prognostic models.
- 1.3. Objectives: Clear goals that the research aims to achieve.
- 1.4. Scope: The extent and limitations of the study.

- 1.5. Significance of the Study: The potential impact and contributions of the research to the field of oncology.
- 1.6. Outline the Structure of the Report: A brief overview of the report's organization (this section).

Chapter 2: Related Work

This chapter examines the available literature and past research relevant to the inquiry. It's separated into:

- 2.1. Introduction to Literature Review: An overview of the literature review's importance and scope.
- 2.2. Historical Perspective: A look at the historical development of stomach cancer prognosis and treatment.
- 2.3. Theoretical Framework: The theoretical underpinnings that support the research.
- 2.4. Previous Research and Studies: A summary of key studies and findings in the field.
- 2.5. Current State of the Field: An analysis of the most recent developments and ongoing research.

Chapter 3: Materials and Methods

This chapter describes the methodological technique used in the investigation. This includes:

3. Chapter 3: Materials and Methods

This chapter describes the methodological technique used in the investigation. This includes:

3.1. Preprocessing : The preprocessing section of our omics data

- 3.1.1. RNA Data Processing and Normalization
- 3.1.2. miRNA Data Processing and Normalization
- 3.1.3. DNA Methylation Data Preprocessing and Normalization

3.2. MOFA Analysis : Our integrating platform used on our omics data

- 3.2.1. Our pipeline of Running MOFA: Summary of our pipeline
- 3.2.2. MOFA Model: The structure of MOFA and how does it work
- 3.2.3. Regularization of MOFA: How does MOFA being handled
- 3.2.4. Model Training and Selection: Brief of training MOFA and features selection of the omics data

3.2.5. Downstream Analysis: Illustration of MOFA resulted model visualization

3.3. Models: Our trained used models

3.4. Website : Our User-friendly website (TRYME)

Chapter 4: Implementation and Results

This chapter includes the study's execution specifics as well as its findings. This

includes:

- 4.1. Programming Languages and Tools: The technologies used for implementation.
- 4.2. Code Structure: The organization of the codebase.
- 4.3. Data Structures and Databases: The data handling and storage mechanisms.
    - 4.3.1. Data Structures
    - 4.3.2. Databases
- 4.4. Quantitative Results: Numerical and statistical outcomes of the study.
- 4.5. Qualitative Results: Descriptive and interpretive findings.

Chapter 5: Discussion

This chapter interprets the findings and compares them to previous research. This includes:

- 5.1. Interpretation of Results: Analysis and implications of the findings.
- 5.2. Comparison with Previous Studies: How the results align or differ from existing research.
- 5.3. Limitations: The constraints and limitations encountered during the study.

Chapter 6: Conclusion and Future Work

The final chapter summarizes the research findings and offers suggestions for future research. This includes:

- 6.1. Conclusion: A recap of the main findings.
- 6.2. Future Work: Potential areas for further investigation.

References:

A comprehensive listing of all sources cited in the report.

2. Chapter 2: Related Work

2.1. Introduction to Literature Review

Due to the rapid increase in high-throughput genomic techniques, obtaining numerous omics biomarkers has become possible and is effective in terms of cost. Thus, computational methods and tools have been developed to analyze and interpret this data for the prognosis prediction of cancer. By utilizing different types of multi-omics data such as genomics, transcriptomics, proteomics, they aim to improve on the current traditional clinical or pathological approaches used in

making prognosis predictions. By utilizing these advances, personalized medicine will become more accurate and provide better tailored treatments and interventions for cancer patients with greater accuracy and efficiency.

2.2. Historical Perspective

Early cancer prognosis predictions were mostly based on clinical and pathological data. As genomic technologies advanced, the focus shifted towards integrating molecular data, such as mRNA expression, methylation, and copy number variation, to enhance prediction accuracy. Traditional statistical models like the Cox proportional hazards model were commonly used but had limitations in variable selection and handling high-dimensional data.

2.3. Theoretical Framework

The integration of multi-omics data for prognosis prediction is based on the Cox proportional hazards model, which is used to calculate the risk scores associated with patient survival. Recent breakthroughs in this area include the SKI-Cox and wLASSO-Cox models [11], which improve on classic models by better selecting significant variables from high-dimensional genomic data.

2.4. Previous Research and Studies

Several innovative models and frameworks were developed as recent advances in computational approaches for cancer prognostic prediction. When incorporating data from The Cancer Genome Atlas (TCGA), the SKI-Cox and wLASSO-Cox models [11], which use mRNA expression profiles to forecast risk scores, outperformed the traditional LASSO-Cox model by significantly improving prediction accuracy for overall survival in glioblastoma multiforme and lung adenocarcinoma patients. Deep learning techniques [12], such as autoencoders, have been used to extract essential features from multi-omics data, overcoming problems such as duplicated variables and small sample sizes, though they are still influenced by data noise. The DeepProg framework [13], which integrates deep-learning and machine-learning algorithms, has outperformed previous models in predicting patient survival subgroups across tumours, identifying similar genomic markers associated with poor survival. Franco et al. [14] performed a comparison study of various deep learning autoencoders for cancer subtype detection across four cancer types using TCGA datasets, matching with other genomic research to predict patient subgroups and survival characteristics. Furthermore, a robust autoencoder-based model for stomach adenocarcinoma [15] has a concordance index (C-index) of 0.714 and a Brier score of 0.184. Finally, a mitochondrial-related risk prognostic model [16] involving genes such as NOX4, ALDH3A2,

FKBP10, and MAOA revealed differences in immune cell infiltration and proposed that an immunosuppressive tumor microenvironment contributes to poor prognosis in high-risk groups, pointing to potential therapeutic agents for these groups.

2.5. Current State of the Field

The field is currently focusing on improving the accuracy of cancer prognosis predictions through the integration of different omics data types. Despite progress, dealing with redundant variables and limited sample numbers continues to be challenging. Innovative techniques like SKI-Cox, wLASSO-Cox [11], and deep learning frameworks such as DeepProg [13] show promise in improving prediction accuracy and uncovering valuable insights into cancer biology. Continuous development and validation of these methods are crucial for advancing precision cancer treatments and prognosis prediction.

3. Chapter 3: Materials and Methods

Figure 1. This figure shows our pipline through the whole project

3.1. Preprocessing

3.1.1. RNA Data Processing and Normalization

Data Collection:

- **Description**: RNA sequencing (RNA-Seq) data is obtained from The Cancer Genome Atlas (TCGA). This data consists of raw read counts representing the number of times each RNA molecule is sequenced.
- **Process**: Raw read counts are extracted from TCGA files and organized into a a single data frame where rows represent genes (identified by Ensemble gene IDs) and columns represent individual samples (patients).

Gene Symbol Annotation:

- **Description**: Ensemble gene IDs are unique identifiers for genes but can be less intuitive for biological interpretation.
- **Process**: The Ensemble gene IDs are translated into gene symbols (e.g., from ENSEMBL IDs to HGNC gene symbols) to enhance interpretability. Any duplicate gene symbols resulting from this translation are resolved by aggregating the counts, usually through average or summing.

Normalization Techniques:

- **Size Factor Estimation**:
    - **Description**: Technical variations and sequencing depth can generate biases into raw read counts.
    - **Process**: DESeq2 calculates size factors for each sample to normalize these differences, ensuring that counts are comparable across samples.
- **Variance Stabilizing Transformation (VST)**:
    - **Description**: The VST helps to stabilize the variance across expression levels, making downstream statistical analyses more reliable.
    - **Process**: VST is applied to the normalized count data, transforming it in a way that reduces variability.
- **Log2 Transformation**:
    - **Description**: RNA-Seq data often shows skewness, which may have an impact on statistical analyses.
    - **Process**: A log2 transformation is performed on the data to reduce skewness, making the distribution more symmetrical.
- **DESeq2 Normalization**:
    - **Description**: DESeq2 provides a robust framework for normalizing RNA-Seq data and performing differential

expression analysis.

- **Process**: DESeq2 normalization improves sequencing depth differences and other technical biases, leading to more accurate detection of differentially expressed genes.

3.1.2. miRNA Data Processing and Normalization

Data Collection:

- **Description**: MicroRNA (miRNA) expression data from TCGA consists of read counts for small non-coding RNA molecules involved in gene regulation.
- **Process**: miRNA expression data files are imported, and relevant columns are extracted. This data is organized into a single data frame, with rows representing miRNAs and columns representing samples.

Duplicate Handling:

- **Description**: Duplicate miRNA IDs can occur due to multiple probes or annotations targeting the same miRNA.
- **Process**: Duplicate miRNA IDs are checked and handled by aggregating duplicated items, usually by calculating the mean expression value for each miRNA.

Sample Metadata Integration:

- **Description**: Metadata provides crucial information about each sample, such as clinical data and experimental conditions.
- **Process**: Sample metadata from TCGA is integrated with the miRNA expression data, matching miRNA data to the corresponding sample identifiers for accurate analysis.

Data Cleaning:

- **Samples and Genes with Missing Values**:
    - **Description**: High levels of missing data can affect the quality of the dataset.
    - **Process**: Samples and genes (miRNAs) with more than 80% missing values are removed to ensure data quality.
- **Low-Variance miRNAs**:
    - **Description**: miRNAs with low variance across samples provide little information for differential expression analysis.
    - **Process**: Low-variance miRNAs are filtered out, leaving the top 25% most variable miRNAs for analysis.

Normalization Techniques:

- **Log2 Transformation**:
    - **Description**: As with RNA-Seq data, miRNA data often exhibit

skewness.

- **Process**: Data is converted to log2 scale to reduce skewness.

- **DESeq2 Normalization**:
  - **Description**: Like RNA-Seq, normalization is crucial for miRNA data to account for sequencing depth differences.
  - **Process**: DESeq2 normalization is applied to generate a normalized dataset suitable for downstream analysis.

3.1.3. DNA Methylation Data Processing and Normalization

Data Collection:

- **Description**: DNA methylation data measures the methylation levels at specific CpG sites across the genome, important for gene regulation studies.
- **Process**: Methylation data files are imported from TCGA, and columns are extracted to aggregate the data into a single data frame with rows representing CpG sites and columns representing samples.

Sample and Gene Filtering:

- **Description**: Missing values in DNA methylation data can significantly affect analysis quality.
- **Process**: Samples and genes with more than 20% missing values are eliminated to ensure data quality and reliability.

Normalization:

- **Description**: B2M normalization corrects for technical variations and biases in DNA methylation data.
- **Process**: B2M normalization is applied to adjust for technical artifacts, ensuring that methylation levels are comparable across samples.

Conversion from Beta Values to M Values

Beta Values:

- **Description**: Beta values represent the ratio of methylated probe intensity to the overall intensity (methylated + unmethylated). These values range from 0 (no methylation) to 1 (full methylation).
- **Use**: Beta values are intuitive and commonly used but can be problematic for statistical analyses at extreme values (near 0 and 1).

M Values:

- **Description**: M values provide a log2 ratio of the intensities of methylated probe to unmethylated probe. They offer a more statistically valid approach for differential methylation analysis.

- **Conversion Formula**:

$$M = \log 2 \left(\frac{\beta}{1-\beta\beta} \right)$$

- This conversion helps in stabilizing the variance of methylation levels, especially at extreme beta values (near 0 and 1).

Common Samples:

- In this analysis, we ensure that common samples (patients) are selected across different types of molecular data, resulting in each data frame containing 316 samples. This consistency allows for integrative analyses across RNA, miRNA, and DNA methylation data, providing a comprehensive view of the molecular changes in the same set of patients.

3.2. MOFA Analysis

3.2.1. Our pipeline of Running MOFA:

MOFA is an unsupervised statistical framework and is considered as a flexible, statistically sound extension of principal component analysis (PCA) for multi-omics data. In terms of a few latent components that (ideally) reflect the essential signal in the input data, MOFA infers an interpretable low-dimensional representation given several data matrices with measurements of multiple -omics data types on the same or on overlapping sets of samples. By distinguishing the factors that cause variability in a single data modality from the factors that are common across several data modalities, MOFA effectively disentangles the

sources of variation in the data [10].

To accomplish this, the RNA expression, miRNA expression, and DNA methylation datasets were among the multi-omics data that were integrated and analyzed using the Multi-Omics Factor Analysis (MOFA) pipeline. These datasets were first read from CSV files, transformed into matrices, and then assembled into a single list that represented the many categories of data. Every dataset's dimensionality was verified.

So, to connect these various datasets and provide a thorough overview and visualization of the data structure, a MOFA object was created. The model's number of components was set to 15, the maximum number of training iterations was set to 20,000, and the convergence mode was set to "slow" to ensure comprehensive model training. The default data, model, and training options were retrieved and modified accordingly.

After that, the model was created using the given training options, model, and data and saved as an HDF5 file. After making the necessary corrections to match column names and format the data, the metadata in addition to clinical dataset was integrated into the MOFA model and the trained model was loaded for further analysis.

Following evaluating and visualizing the variation described by the model from multiple perspectives and factors, significant factors and their contributions to

the data's variability were identified. Plotting of factor correlations revealed linkages between the factors. Significant correlations were found and displayed as log p-values to investigate the relationships between the variables and clinical covariates. Using various charting approaches, factors were also shown in relation to clinical variables, such as vital status and treatment response.

Additionally, in-depth analysis involved the use of scatter plots, heatmaps, and violin plots to visualize the relationships and contributions of variables within the datasets. Steps for scaling data and identifying the most important contributing features for each omics view were also included in the pipeline.
Moreover, A deeper knowledge of the underlying biological processes was made possible by this entire approach, which offered a strong foundation for integrating and evaluating multi-omics data, finding important components and their connections with clinical outcomes.

3.2.2. MOFA Model:

MOFA breaks down the data matrices Y1,..., YM of dimensions N × Dm, where N represents the number of samples and Dm the number of features in data matrix m.

Equation 1. $Y^m = ZW^{mT} + \in^m \; m = 1, \dots, M$

In this equation (2), Wm stands for the weight matrices for each data matrix

m (also known as view m), and Z stands for the factor matrix, which is common to all data matrices. The residual noise term specific to a view is denoted by ϵ^m, and its form is determined by the details of the data type. The model is formulated within a probabilistic Bayesian framework, in which we assign prior distributions to all unobserved variables (such as., factors Z, weight matrices W^m, and parameters of the residual noise term). Specifically, we utilize sparsity priors for the weight matrices and a typical normal prior for the variables Z [7].

3.2.3. Regularization of MOFA:

The weight matrices must be appropriately regularized for the model to be able to separate variation between different data sets and provide factors that can be comprehended. MOFA employs a two-stage regularization strategy. The first level promotes sparsity both view- and factor-wise, making it possible to determine which factor is active in each view. A small number of features with active weights are usually the outcome of the second level's encouragement of feature-wise sparsity. Briefly, using Gaussian priors on the factors and loadings, MOFA employs L2 regularization. Sparsity-inducing priors can be used to encourage sparsity in the loadings. These regularization methods aid in the creation of an interpretable and applicable model [7].

3.2.4. Model Training and Selection:

The number of components in this MOFA2 analysis must be determined to train the model. The model regularization procedure, which is described in full in the MOFA2 framework, includes an Automatic Relevance Determination (ARD) before dynamically deactivating factors during regularization. A threshold for the minimal fraction of variance explained is used to selectively prune components during the training process. As an alternative, we can skip the minimum variance requirement by pre-specifying the number of factors. We started the model training process in our application with K = 15 factors. This methodology guarantees a balance between interpretability and model complexity, which is essential for drawing significant conclusions from integrated omics data. Setting a maximum number of iterations at 20,000 was a crucial component of our training technique.

3.2.5. Downstream Analysis:

Many crucial processes are involved in the downstream analysis of the MOFA model, which aims to visualize and analyze multi-omics data to identify important patterns and associations.

Firstly, an overview of the data is produced to condense the distribution and features of the input datasets. To confirm the data's integrity and gain an overview of its structure, this stage is vital. The next step looks at how well each factor

explains the variance in the data to determine how well the model captures variability in the data. This aids in evaluating how well the model recognizes significant patterns in the multi-omics data.

After a model has been trained, the first step is to determine the variance explained by each factor k in each sample group g and data modality m:

Equation (3):

$$R^2{}_{gmk} = 1 - \left(\sum_{n,d}(Y_{gm} - W_m Z_g)\right)^2 / \left(\sum_{n,d} Y_{gm}\right)^2$$

R^2_{gmk} : The rate of variance for group g that can be accounted for by the k-th factor in view m.

$\sum_{n,d}$: : Total sum of characteristics (d) and samples (n).

Y_{gm} : The group g and view m data matrix.

W_m : : The view m weight matrix. The contribution of each latent factor to the data in view m is shown by these weights.

Z_g: The matrix of latent factors for group g.

$(Y_{gm} - W_m Z_g)^2$: The difference between the observed and reconstructed data is

represented by the squared residuals.

$\left(\sum_{n,d}^{Y_{gm}} - W_m Z_g\right)^2$: The squared residuals total sum.

$\sum_{n,d}$ Y_{gm}^2 : The observed data's total sum of squares.

Overall, the equation measures how well a given factor (together with the associated weights) represents the variability in the data from a given view and group. Measurement helps in determining which components are most crucial for interpreting the data and in comprehending how each component contributes to the overall model.

A correlation analysis is then used to investigate the link between various factors. This shows possible relationships and interactions between the factors, providing information on the connections between different biological processes or measures. The link between these characteristics and clinical or demographic data is evaluated to have a deeper understanding of their significance. This includes looking at the relationships between the variables and significant confounders such age, gender, treatment status, and survival rates. Additionally, to connect the latent components obtained from the MOFA model to important clinical traits and outcomes, this step is essential. To show how the factors are distributed in relation

to variables, visual representations are made. These visuals give a clear picture of how various clinical groups are represented in the variables by highlighting patterns or discrepancies across subgroups, such as treatment versus patients who were not treated.

The analysis also explores how significant individual characteristics are for each factor (e.g., genes or miRNAs). This stage aids in evaluating the biological significance of the components and identifies important biomarkers that may be of interest by determining the most significant aspects.

Moreover, the correlations between variables, features, and covariates are visualized using scatter plots and heatmaps. Heatmaps display the expression levels of the key characteristics and highlight trends in the data, while scatter plots demonstrate the distribution of data points and their relationship to clinical outcomes.

Finally, via the MOFA model, all these downstream analyses offer a thorough comprehension of the multi-omics data. They uncover important characteristics and patterns that could have biological or clinical importance, explain the fundamental structure of the data, and make links to clinical and biological variables.

3.3. Models

The top weighted features from the multi-omics data were selected using Multi-Omics Factor Analysis (MOFA). We selected 100 characteristics from DNA, 70 from RNA, and 30 from miRNA for further research. The selected features were merged into a single data frame to create a comprehensive dataset for model training. Subsequently, four different classifiers were employed: Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. Hyperparameter tuning was performed on each model to optimize its performance.

1. **Logistic Regression:**

- Data Splitting: The dataset was split into training (80%), validation (10%), and test sets (10%).
- Hyperparameter Tuning: GridSearchCV was used to optimize the hyperparameters C, penalty, and solver. The grid search spanned C values of [0.01, 0.1, 1, 10, 100], penalty options [l1, l2], and solver choices [liblinear, saga], with a 15-fold cross-validation.

2. **Support Vector Machine (SVM):**

- Data Splitting: The dataset was split into training (80%) and test sets (20%).
- Hyperparameter Tuning: GridSearchCV was used to optimize C, gamma,

and kernel parameters over a 5-fold cross-validation, with C values [0.1, 1, 10, 100], gamma values [1, 0.1, 0.01, 0.001], and kernels [rbf, linear].

3. **Random Forest:**

- Data Splitting: The dataset was split into training (70%) and test sets (30%).
- Hyperparameter Tuning: GridSearchCV was used to optimize parameters such as n_estimators, max_depth, min_samples_split, min_samples_leaf, and bootstrap. The grid search included n_estimators [19], max_depth [None, 10, 20, 30, 40], min_samples_split [2, 5, 10], min_samples_leaf [1, 2, 4], and bootstrap options [True, False], with a 35-fold cross-validation.

4. **XGBoost:**

- Data Splitting: The dataset was split into training (80%), validation (10%), and test sets (10%).
- Hyperparameter Tuning: GridSearchCV was used to optimize parameters such as n_estimators, max_depth, learning_rate, subsample, and colsample_bytree. The grid search included n_estimators [100, 200, 300], max_depth [3, 4, 5, 6], learning_rate [0.01, 0.1, 0.2], subsample [0.8, 0.9, 1.0], and colsample_bytree [0.8, 0.9, 1.0], with a 5-fold cross-validation.

To improve the robustness and accuracy of STAD (Stomach Adenocarcinoma)

prognosis prediction, we used a Voting Classifier to integrate the strengths of multiple classifiers such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. Each base model was individually optimized through GridSearchCV to ensure peak performance. The ensemble employed a hard voting strategy, wherein each classifier voted for a class label, and the final prediction was the class with the majority vote.

3.4. Website

The Python module called Streamlit[18] seeks to handle numerous problems at once. With Streamlit, developers can construct web-based front ends for Python projects by using an extensive collection of interactive components. Anywhere that a Python web program resides, the resultant website can be hosted. The best part is that all it takes to obtain good results is for the developer to create Python code that makes use of Streamlit's classes and methods; they don't need to know HTML, JavaScript, or CSS [17].

Our website includes several user-friendly interfaces that are specifically customized to different aspects of data analysis and prediction to promote thorough study on cancer. The Home, Preprocess Data, MOFA Analysis, Predict Prognosis, and About Us sections are the primary parts of the website. DNA methylation, miRNA, and RNA datasets are among the simplified data preprocessing options

available in the dashboard interface. Users can ensure clean and dependable inputs for downstream analyses by filtering data based on variance criteria. Additionally, it has the MOFA Analysis interface, which lets users specify how many factors to include in a thorough data integration to find hidden links and patterns in multi-omics datasets. Furthermore, the Prognosis Prediction tool makes use of innovative machine learning models to predict the prognosis of patients, supporting personalized care initiatives and decision-making.

Researchers can normalize their datasets for further study by preprocessing omics data from RNA, miRNA, and DNA-Methylation across different types of cancer, including stomach cancer, using the Preprocess Data interface. These preprocessed omics files are accepted by the MOFA (Multi-Omics Factor Analysis) interface, which uses a strong analytical workflow to produce models in HDF5 format. This model can be downloaded by users for display, and they can then conduct downstream studies to gather important insights about the cancer type that is being researched. Ultimately, these insights are used by our prediction model interface to evaluate the prognosis of stomach adenocarcinoma and determine if a patient has a high or low chance of survival.

4. Chapter 4: Implementation and Results

4.1. Programming Languages and Tools

Our project made use of multiple platforms and programming languages to accomplish its objectives. To be more precise, Python was used to create our models, which was very helpful in creating our ensemble model. We employed the Multi-Omics Factor Analysis (MOFA) tool for multi-omics integration. Using a collection of hidden variables that account for both biological and technological sources of variability, MOFA is a method that finds the main determinants of variance in multi-omics data sets. It separates axes of heterogeneity that are exclusive to one data modality from those that are common to several modalities [7]. This makes use of R's robust data integration capabilities to enable a variety of downstream analysis, including the detection of outlier samples, data imputation, and sample subgroup identification.

Furthermore, we built an AI-powered website using the Python-based web framework Streamlit. This website creates a tool for adenocarcinoma prediction by preprocessing the three omics data and using the MOFA framework.

4.2. Code Structure

The codebase involves prognosis prediction, multi-omics factor analysis, data

preprocessing, and an intuitive web interface to facilitate an all-encompassing workflow for cancer research. Using information from the TCGA cancer database, the codebase aims to give researchers and physicians a strong tool for preprocessing omics data, integrating it with the MOFA platform, and predicting patient prognosis for stomach adenocarcinoma cancer (STAD)[19].

R Language for Preprocessing: R is used to perform preprocessing on omics data, such as RNA, miRNA, and DNA methylation. The datasets must be cleaned and standardized in this step to enable them to be prepared for analysis later.

MOFA Platform for Omics Integration: The preprocessed omics data is integrated using the Multi-Omics Factor Analysis (MOFA). By revealing hidden patterns and linkages in the multi-omics datasets, this stage offers significant new insight into the data.

Python for Prognosis Prediction: Using information from the TCGA cancer database, an ensemble model in Python is utilized to predict the prognosis of patients with stomach adenocarcinoma cancer (STAD). Using the integrated multi-omics data and the machine learning models to do our ensemble model, this model can accurately forecast which patients will have a high or low chance of surviving.

Streamlit-based Website The complete procedure is contained in a, which makes it easily accessible and manageable for researchers and physicians. By using a smooth and engaging interface, users of this website can preprocess data, carry out MOFA analysis for all cancer types, and predict prognosis for stomach adenocarcinoma cancer.

4.3. Data Structures and Databases

4.3.1. Data Structures:

RNA Data Processing and Normalization:

Raw RNA sequencing (RNA-Seq) data is collected from The Cancer Genome Atlas (TCGA) and then processed and normalized to identify gene expression levels within patient samples using raw read counts. These raw read counts are extracted from TCGA files and organized into a comprehensive data frame format, where each row represents a gene identified by Ensemble gene IDs, and each column represents a patient sample. To enhance interpretability, Ensemble gene IDs are converted into more recognizable gene symbols, such as those from the Human Genome Navigation Consortium (HGNC). Any duplicates are then resolved through aggregation of counts, typically by averaging or summing. To ensure robustness and comparability across samples, normalization techniques are applied. We employ DESeq2 for size factors' calculation to

normalize technical variations and differences in sequencing depth between samples. Besides, we also apply Variance Stabilizing Transformation (VST) to ensure that variance is the same across expression levels of different genes leading to more reliable statistical analyses afterwards as well as removing the bias caused by skewedness of RNA-Seq data distribution through applying Log2 transform, which makes them look like normal ones.

miRNA Data Processing and Normalization:

The first step in miRNA Data Processing and Normalization is to gather DNA sequence data on microRNA (miRNA) from The Cancer Genome Atlas (TCGA) [19]. This provides counts for messenger RNA, which are important for controlling gene activities by protein messages in cells and supporting cell growth. These miRNA data files are loaded and formatted into tabular form with rows for each microRNA and columns for each patient sample.

With the aim of maintaining data integrity, it is important to handle duplicate miRNA IDs from multiple probes; and this can be done by combining duplicates through the mean expression values which helps to ensure that across samples miRNA expression is well represented. Precise sample identification requires integrating clinical and experimental metadata from TCGA [19] into miRNA expression data, thus increasing credibility and contextual understanding of the analyses done.

Then, Data cleaning procedures are then implemented to ensure high-quality data for subsequent analyses. Samples and miRNAs with more than 80% missing data are excluded to mitigate potential biases and ensure robustness in downstream analyses. Additionally, to focus on the most informative miRNAs, only the top 25% most variable miRNAs across samples are retained, filtering out low-variance miRNAs that contribute less to differential expression analyses. Normalization techniques are used to standardize and prepare the miRNA data for statistical analysis. Just like with processing of RNA-Sequencing datasets, Log2 transformation is used here to correct the skewness in gene expression levels hence making such distributions more appropriate for statistical analysis purposes.

Finally, the miRNA expression data was adjusted accurately and harmoniously by correcting for the sequencing depth differences among samples using the DESeq2 normalization methods, which made it possible to undertake further extensive analysis. It is with these steps that it can be affirmed with certainty that each stringent stage in the miRNA processing pathway as related to stomach cancer is exploited to glean significant information on the roles thereof so regulated through use TCGA datasets.

DNA-Methylation Data Processing and Normalization:

In the processing and normalization of DNA methylation data, it starts with obtaining DNA methylation data from The Cancer Genome Atlas (TCGA)[19],

whose goal is to record the amount of methylation at CpG sites in the genome. These original methylation files are uploaded and arranged into a structured format of data frame, with rows as single CpG sites and columns as various patient samples. Samples and CpG sites with more than 20% missing values are removed from the dataset to guarantee the accuracy of subsequent studies.

Standardizing procedures are crucial for proper interpretation and analysis of methylation data DNA. B2M is used to standardize the DNA methylation data to eliminate technical biases allowing for comparison of samples with minimum interference from mistakes which might give wrong results. Furthermore, a conversion from Beta Values to M Values is applied to improve the data's appropriateness for statistical analyses. M Values, which are log2-transformed ratios of methylation to unmethylated probe intensities, are created from beta values, which are the ratio of methylated probe intensity to total probe intensity and range from 0 to 1. This improvement is especially useful for discovering differential methylation patterns linked to stomach cancer, as it stabilizes variance and enables more rigorous statistical analysis.

Common patient samples are chosen from RNA, miRNA, and DNA methylation datasets to ensure uniformity among molecular datasets. By utilizing an integrative approach, extensive studies that investigate molecular interactions and correlations among various omics data types can be conducted, leading to a

clearer understanding of the molecular mechanisms behind the growth of stomach cancer. This pipeline makes sure that the DNA methylation data from TCGA [19] is prepared appropriately for the discovery of biological insights and possible biomarkers relevant to cancer research and personalized treatment methods by carefully processing and normalizing the data.

Integrated Multi-Omics Data: (MOFA)

Data on DNA methylation, miRNA, and RNA are first imported from CSV files and converted into matrices for processing. Then, these datasets are combined into a single MOFA object, which makes it possible to analyze many molecular levels simultaneously. Factor analysis is then carried out using MOFA, which creates an HDF5 file to effectively store the data. Visualizations and statistical evaluations are used in the analysis to further investigate the factors found and how they relate to clinical variables. To efficiently manage large-scale omics information, this method leverages the HDF5 format to unearth complex molecular features linked to stomach cancer.

4.3.2. Databases:

TCGA Database:

The study involved 363 patients from The Cancer Genome Atlas (TCGA) cohort who had stomach adenocarcinoma [19]. To combine the methylation, miRNA, and RNA sequencing data, an MOFA was built. For thousands of tumor samples from more than 20 different forms of cancer, the Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga/) has been producing multimodal genomes, epigenomics, and proteomics data. There are three layers to TCGA data. Level 3 data are available to the public and contain high-level summaries such as expression quantifications of genes, but many level 1 and level 2 data contain protected information such as individual germline variations and raw DNA-sequencing data. For thousands of tumor samples from more than 20 different forms of cancer, the Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga/) has been producing multimodal genomes, epigenomics, and proteomics data. There are three layers to TCGA data. Level 3 data are available to the public and contain high-level summaries such as expression quantifications of genes, but many level 1 and level 2 data contain protected information such as individual germline variations and raw DNA-sequencing data [19].

Local Storage:

Data is temporarily kept in organized directories on local file systems for preparation and analysis. Every stage of the procedure, such as the preprocessed, raw, and MOFA outputs, has its own subdirectory. R is used to run the MOFA

pipeline, which guarantees effective integration and analysis of multi-omics data. Python is used for building websites and visualization, which allows for user-friendly access and interactive exploration of the analysis results.

HDF5 Files:

The integrated multi-omics data obtained from the MOFA analysis are stored as HDF5 files. Large datasets may be efficiently stored and retrieved with the help of this format, which is essential for managing complicated multi-omics data. The system guarantees fast access to and processing of multi-dimensional arrays of scientific data by utilizing HDF5 files, hence enabling comprehensive and reliable analysis. This methodology promotes the scalability of the analytical procedures involved in multi-omics research while also improving the efficiency of data management.

Data Access and Security:

Robust access control methods are developed to guarantee that only authorized users can access the website and underlying data, given the sensitivity of patient data. To maintain patient data privacy and security, the online application offers a secure interface for data submission, preprocessing, analysis, and prediction. User-uploaded data is processed securely; all preprocessing and analysis is done server-side to guard against illegal access. Through the web

interface, users can safely download processed results and predictive models. To guarantee accurate and dependable information, the TCGA cancer database was the original source of the data.

4.4. Quantitative Results

4.4.1. Model Training and Evaluation Results

The Logistic Regression model was trained and evaluated using the parameters C: 100, penalty: l1, and solver: liblinear, which were identified as the best based on cross-validation scores. This best model was then trained on the entire training dataset. Upon evaluation, the model achieved a test set accuracy of 0.806, indicating a robust performance.

The Support Vector Machine (SVM) model was trained and evaluated with the best model parameters identified as C: 1, gamma: 1, and kernel: rbf. These optimal parameters were determined through an exhaustive grid search. The model was then trained on the entire training dataset. Evaluation results showed a test set accuracy of 0.934, demonstrating a high level of performance.

The Random Forest model was trained and evaluated using the best model parameters identified as bootstrap: False, max_depth: None, min_samples_leaf: 1, min_samples_split: 5, and n_estimators: 19. These parameters were selected based on the best cross-validation scores. The optimal model was then trained on the

entire training dataset. Upon evaluation, the model achieved a test set accuracy of 0.825, indicating strong performance.

The XGBoost model was trained and evaluated using the best model parameters identified as colsample_bytree: 1.0, learning_rate: 0.1, max_depth: 6, n_estimators: 100, and subsample: 1.0. These optimal parameters were selected based on the best cross-validation scores. The model was subsequently trained on the entire training dataset. Upon evaluation, the model achieved a test set accuracy of 0.816, demonstrating strong performance.

To improve STAD (Stomach Adenocarcinoma) prognosis prediction accuracy, we implemented a Voting Classifier, which combined the strengths of Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. Each base model was individually optimised using GridSearchCV to achieve top performance. The ensemble used a hard voting technique in which each classifier voted for a class label, and the final prediction was based on the majority vote. The ensemble model was evaluated on the test set and outperformed the individual models, with a testing accuracy of 0.886. Detailed performance metrics further highlighted the model's efficacy. The accuracy, precision, recall, and F1-score, along with a confusion matrix, demonstrated the robustness and balanced predictive capabilities of the ensemble model. This, which utilized multi-omics data, greatly enhanced the predictive accuracy and reliability of STAD

prognosis, demonstrating the efficacy of ensemble learning in integrating varied molecular data types for comprehensive cancer prognostic prediction.

4.5. Qualitative Results

Our work analyzed the molecular features of stomach cancer using data from the TCGA database, with a focus on transcriptomics (RNA and miRNA) and DNA methylation data. To guarantee the accuracy and dependability of the findings, a systematic pipeline was used for both the preprocessing and analysis of the data.

Data Preprocessing: Uploading raw RNA and miRNA data from the TCGA database was the first stage. Using the DESeq2 method for dna methylation, data overall were preprocessed to eliminate low-quality genes and samples from all three omics datasets and to normalize the read counts. The preprocessing procedures made sure that only data of the highest quality were used in the analysis that followed.

The distribution of miRNA, DNA-methylation (from beta-value to M-value), and RNA data following DESeq2 normalization are shown in Figures (2), (3), and (4), respectively. These histograms show how well the preparation procedures worked to produce normalized data distributions.

Integration and Visualization using MOFA:

Following preprocessing, the Multi-Omics Factor Analysis (MOFA) framework was used to integrate the data. The factors that capture the most significant variations across the combined omics datasets were found using MOFA. The goal of this integration was to identify the molecular characteristics that most significantly influence stomach cancer.

Following preprocessing, the Multi-Omics Factor Analysis (MOFA) framework was used to integrate the data. The factors that capture the most significant variations across the combined omics datasets were found using MOFA. The goal of this integration was to identify the molecular characteristics that most significantly influence stomach cancer.

Figure (5) provides information about the contribution of each factor of data to the total variation by displaying the variance explained by each factor for each omic. The link between the discovered factors and clinical results is highlighted in Figure 6, which shows the correlation between clinical data and MOFA factors. Figure 7 illustrates the distribution of explained variance among the various data types by presenting the overall variance explained per omic.

We used MOFA to find the key factors, and then we retrieved the top 200 genes for additional analysis from these factors; top 30 miRNA genes, top 70 RNA and top 100 DNA Methylation. Several prediction models, including XGBoost,

Random Forest, Logistic Regression, and Support Vector Machine (SVM), were trained using these features. To get the best accuracies, hyperparameter tuning was applied.

miRNA	RNA	DNA Methylation
hsa-mir-143	CCDC26	cg08366446
hsa-mir-192	MYT1L	cg22346581
hsa-mir-145	ANKRD20A8P	cg06714284
hsa-mir-194-2	DDX4	cg11820517
hsa-mir-194-1	ZAN	cg14465900
hsa-mir-100	SNTG1	cg23477406
hsa-mir-215	ADAMTS20	cg03636215
hsa-let-7c	LOC124903236	cg13334650
hsa-mir-199a-2	MUC19	cg23992449
hsa-mir-23b	LINC02909	cg27547954
hsa-mir-199a-1	PANTR1	cg20062650
hsa-mir-99a	PIWIL3	cg18324126
hsa-mir-210	CRX	cg17325959
hsa-mir-223	SLC1A6	cg01755467
hsa-mir-125b-2	DPPA4	cg16620382
hsa-mir-125b-1	OR4K2	cg00767642
hsa-mir-217	PWRN1	cg04407853
hsa-mir-146a	NLRP13	cg03527422
hsa-mir-1-2	CATSPERD	cg12615137

hsa-mir-1-1	ADGB	cg11034245
hsa-mir-133a-1	LINC01205	cg20401551
hsa-mir-133a-2	SCN1A-AS1	cg13357482
hsa-mir-490	NLRP5	cg01697732
hsa-mir-133b	LINC01102	cg04480903
hsa-mir-214	PCDHB1	cg15592945
hsa-mir-218-1	MYADML	cg02455397
hsa-mir-218-2	LINC00635	cg07923233
hsa-mir-147b	OR2AT4	cg05953927
hsa-mir-216a	LINC02054	cg01107006
hsa-mir-3926-2	SCN1A	cg09241381
	LYPD4	cg12539796
	LINC00836	cg27363327
	ARGFX	cg08371659
	LOC124902062	cg19089337
	LOC102546299	cg04794832
	LINC00837	cg27118761
	GALNTL5	cg01394819
	LINC01837	cg22863209
	LINC00507	cg06700935
	LOC101929473	cg05871997
	TRIM75	cg10192047
	LINC00320	cg20560075
	SHCBP1L	cg06316315
	LINC01405	cg27175294

	OR7D4	cg06685464
	EFCAB3	cg23139473
	OR5A1	cg08718398
	LINC00971	cg03952331
	SYCP1	cg10347759
	CABP5	cg19716902
	LOHAN2	cg19249708
	OR10G3	cg11687406
	OR1A1	cg16658099
	D21S2088E	cg26985666
	OR4D9	cg26495865
	LINC02681	cg06328855
	LOC101927575	cg10086212
	LOC101929653	cg02230854
	LINC01435	cg04593571
	LINC01551	cg12266953
	NLRP8	cg26281453
	LOC102724934	cg26857911
	LINC02064	cg16954280
	LINC02436	cg07139330
	LINC00550	cg05445632
	LINC01247	cg25797055
	TOPAZ1	cg09892426
	LINC01490	cg15344220
	ADAD1	cg21210994

	OR6C75	cg21303803
		cg13463683
		cg15243570
		cg17299935
		cg18871020
		cg12970542
		cg00109503
		cg11491381
		cg07724977
		cg26084005
		cg06367154
		cg05218346
		cg20467168
		cg15248835
		cg21293611
		cg08619651
		cg12510028
		cg26366107
		cg24199112
		cg26297547
		cg21881034
		cg13732589
		cg24126187
		cg02957576
		cg01692340

	cg12379948
	cg22889755
	cg20753954
	cg26870584
	cg20893717
	cg21555798

Table 1. This table shoes the top 200 feature from the highly explained variance factor in each omic

The top 10 weights for RNA, miRNA, and DNA methylation are shown in Figures (8), (9), and (10) correspondingly, following MOFA. These plots offer a graphic depiction of the most significant characteristics found by MOFA.

We created an ensemble model that integrates the results of the separate predictive models to improve forecast accuracy. The purpose of the ensemble model was to predict patients' chances of surviving stomach cancer. The confusion matrix for our ensemble model, which displays the model's performance in differentiating between high- and low-risk patients, is displayed in Figure (11) following prediction on the 30% testing data.

We created a website to make our analysis easily readable and approachable. The program has a dashboard with the following features:

The homepage of the website provides usage instructions (Figure 12). Users can preprocess DNA methylation, miRNA, and RNA data for any form of cancer on

the preprocessing page (Figure 13).

MOFA Interface: This interface enables users to create factors that emphasize the most influential features on cancer by integrating preprocessed data using MOFA.

Ensemble Model Interface: To predict the risk of patient survival, users can upload a CSV file containing the top features from MOFA.

This website makes sure that users may efficiently use integrated multi-omics data and sophisticated predictive modeling approaches to preprocess, analyze, and predict stomach adenocarcinoma cancer risks.

Our study offers a reliable method for patient outcome prediction and significant insights into the molecular pathways causing stomach cancer by combining and evaluating the multi-omics data through this structured pipeline. The figures (17,18,19) are essential for illustrating the different phases and outcomes of our investigation.

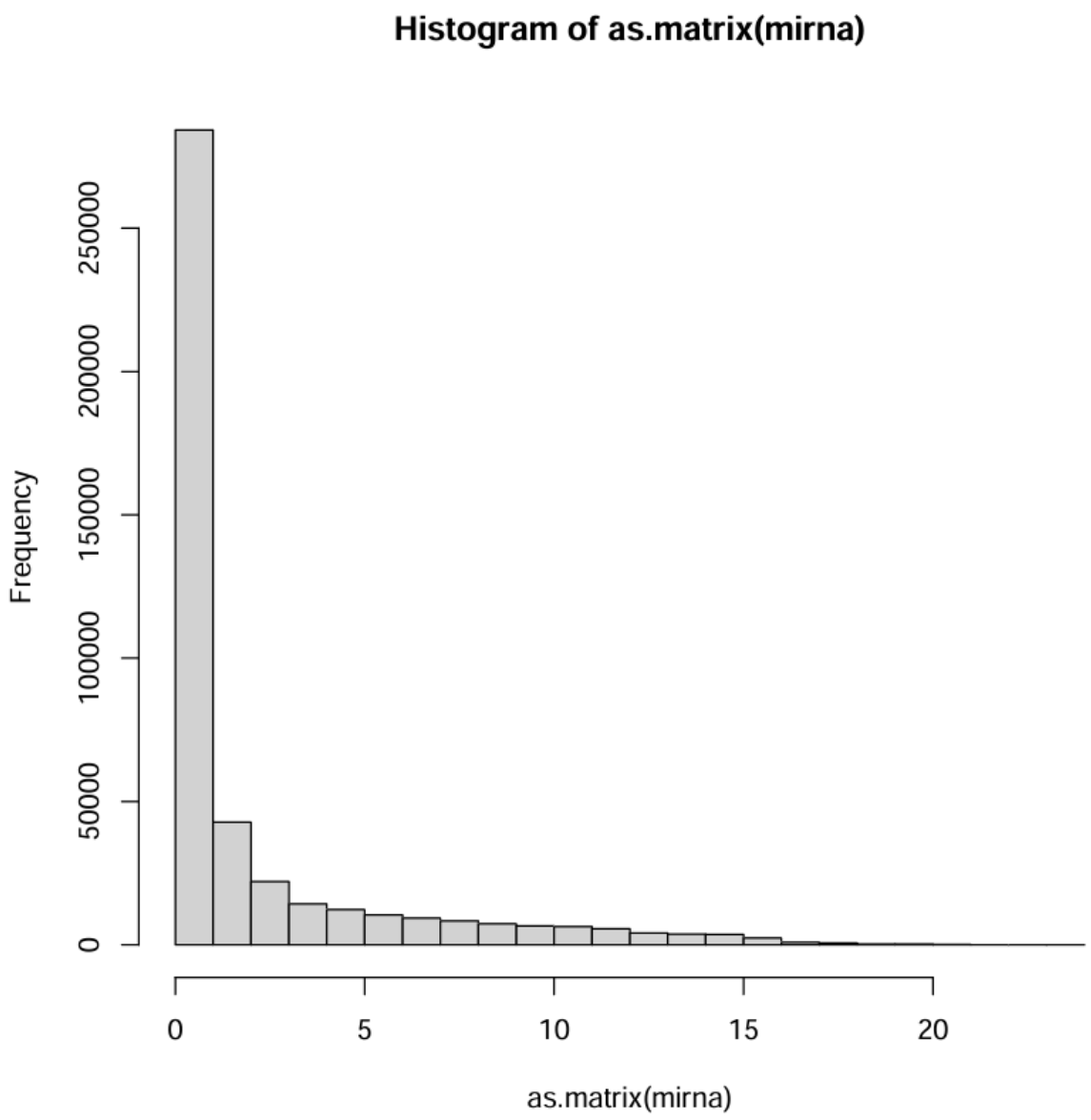

Figure 2. The histogram of miRNA after DESEQ2 normalization

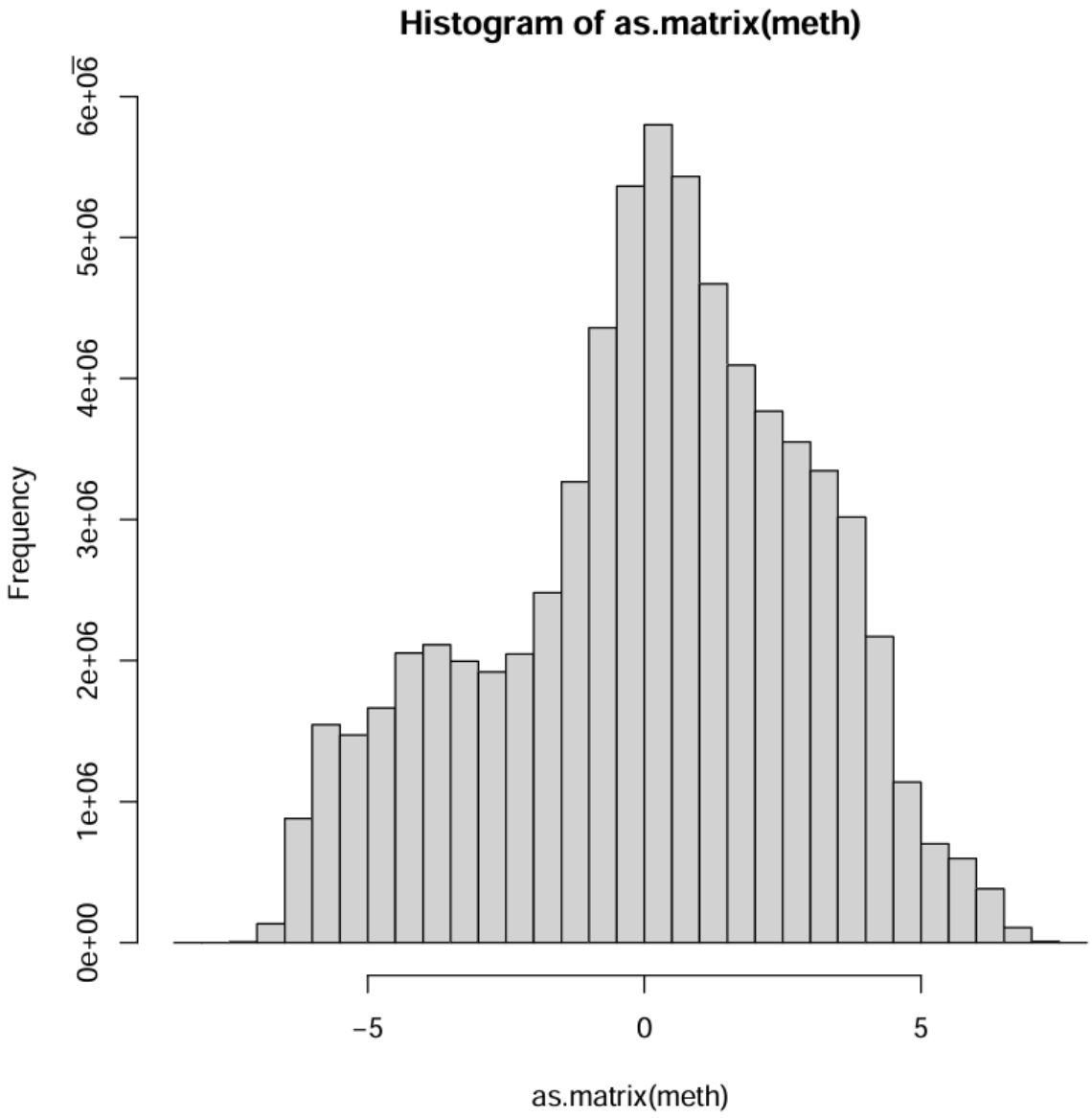

Figure 3. The histogram of DNA Methylation after converting Beta-value to M-value

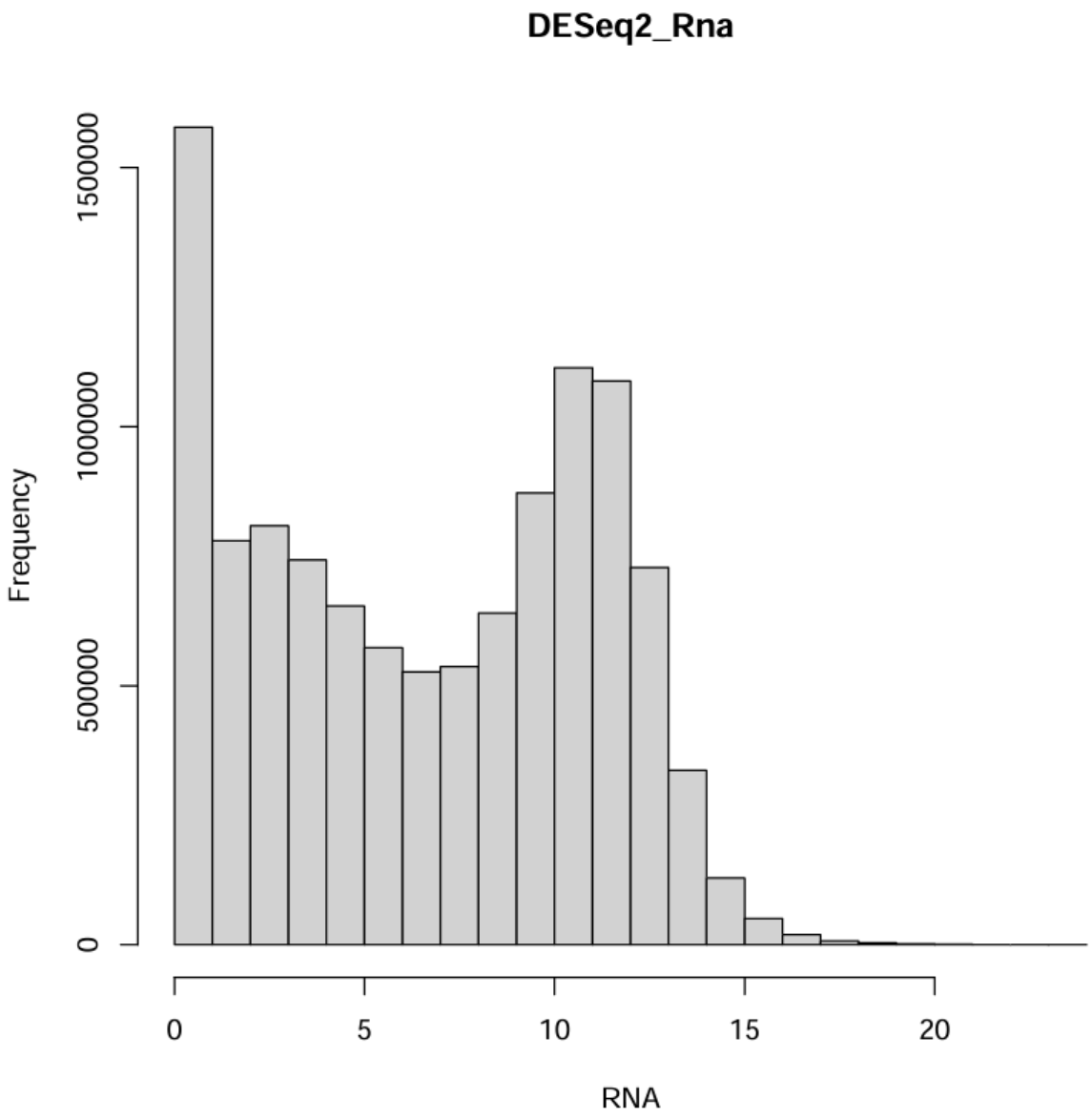

Figure 4. The histogram of RNA after DESEQ2 normalization

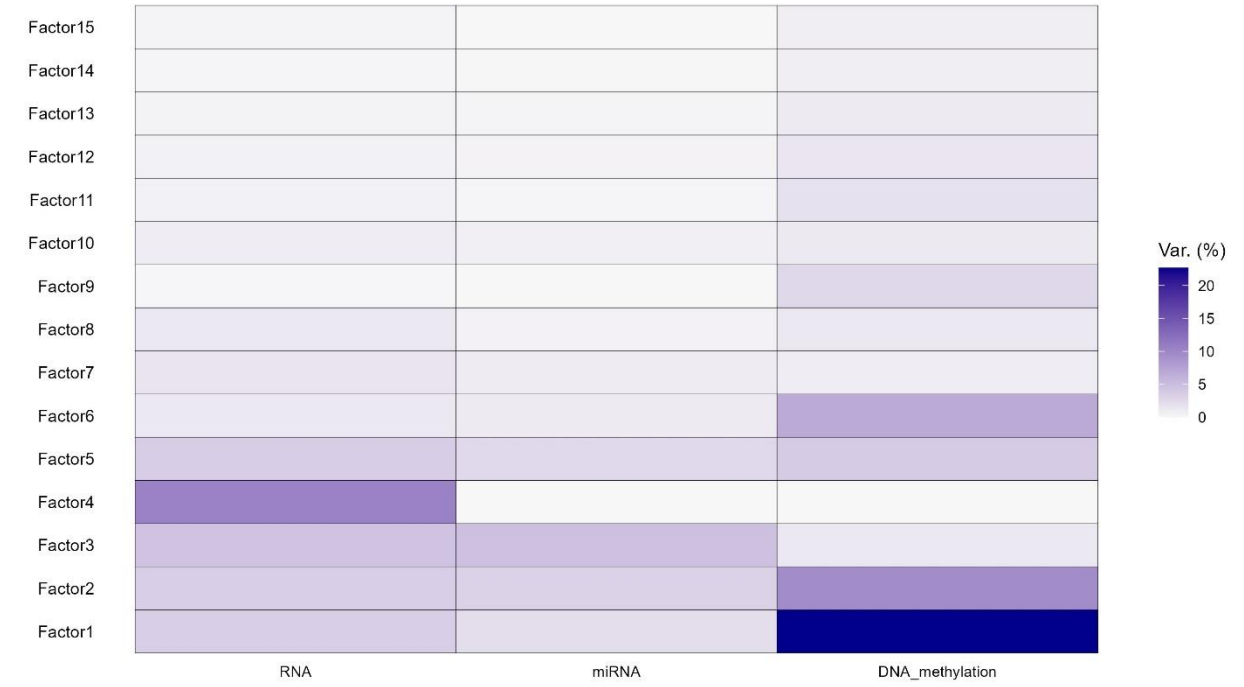

Figure 5. The variance explained in each factor for each omic

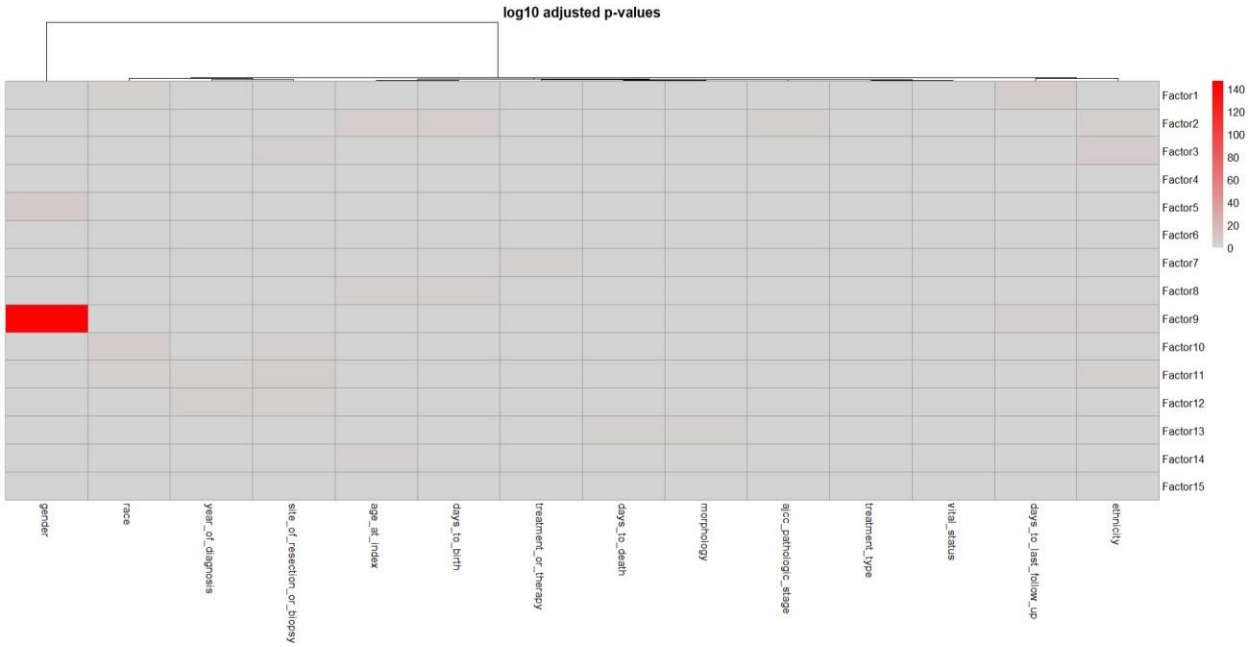

Figure 6. This plot shows the correlation between the clinical data and MOFA Factors

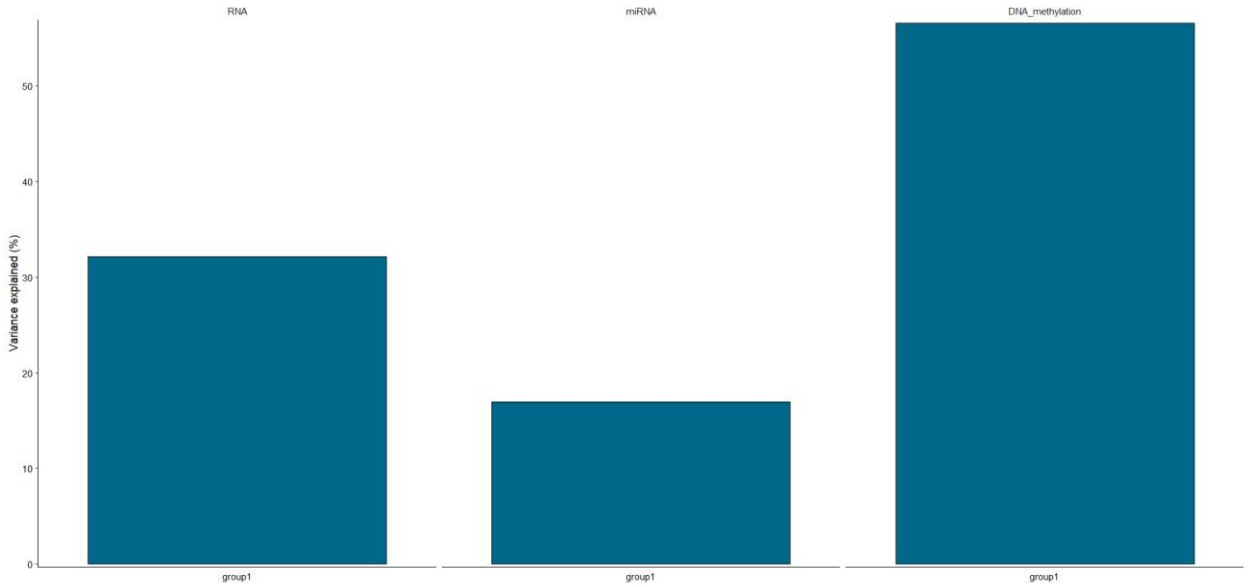

Figure 7. This plot shows total variance explained per omic

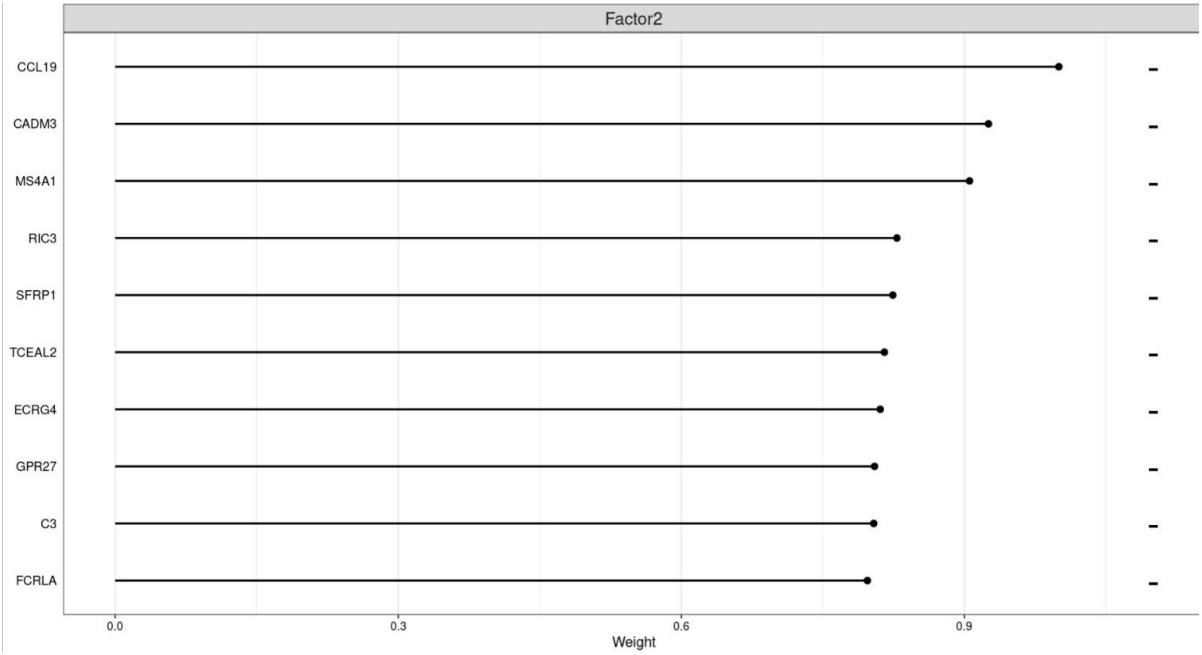

Figure 8. This plot shows the top 10 RNA weights after running MOFA

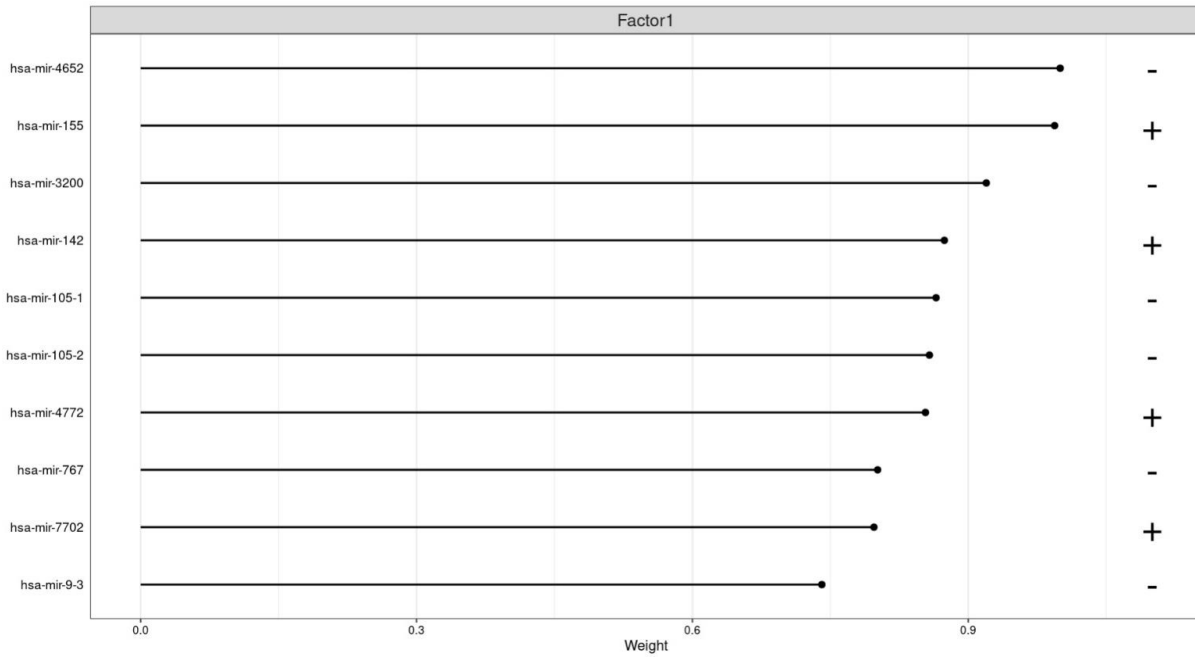

Figure 9. This plot shows the top 10 miRNA weights after running MOFA

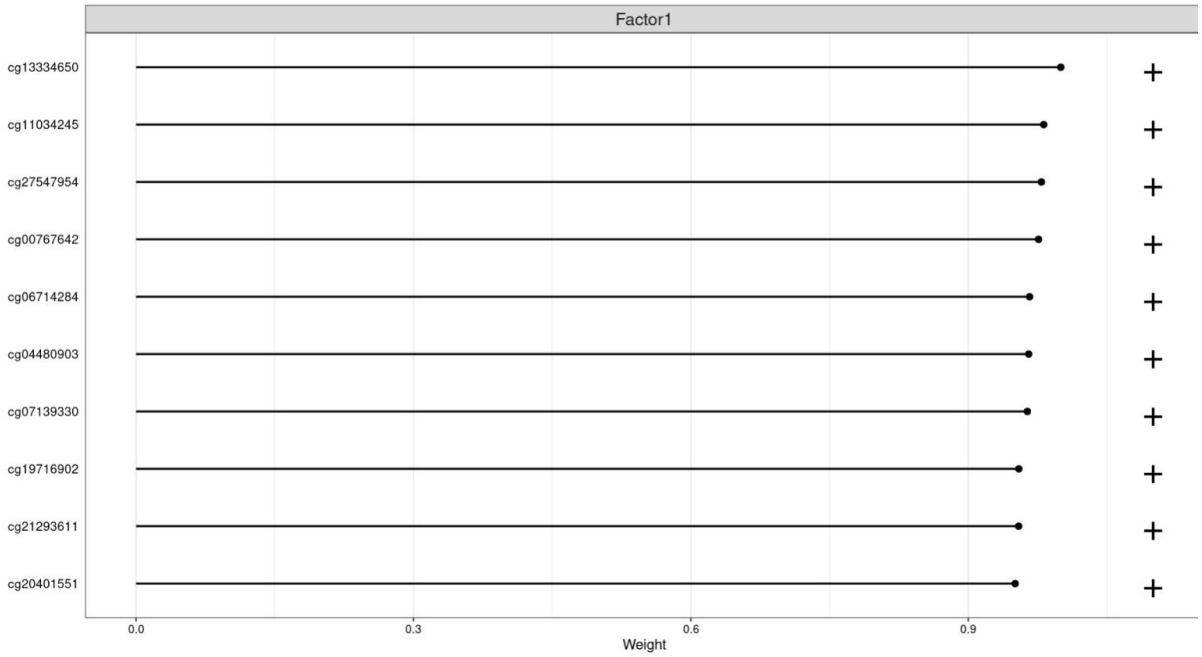

Figure 10. This plot shows the top 10 DNA Methylation weights after running MOFA

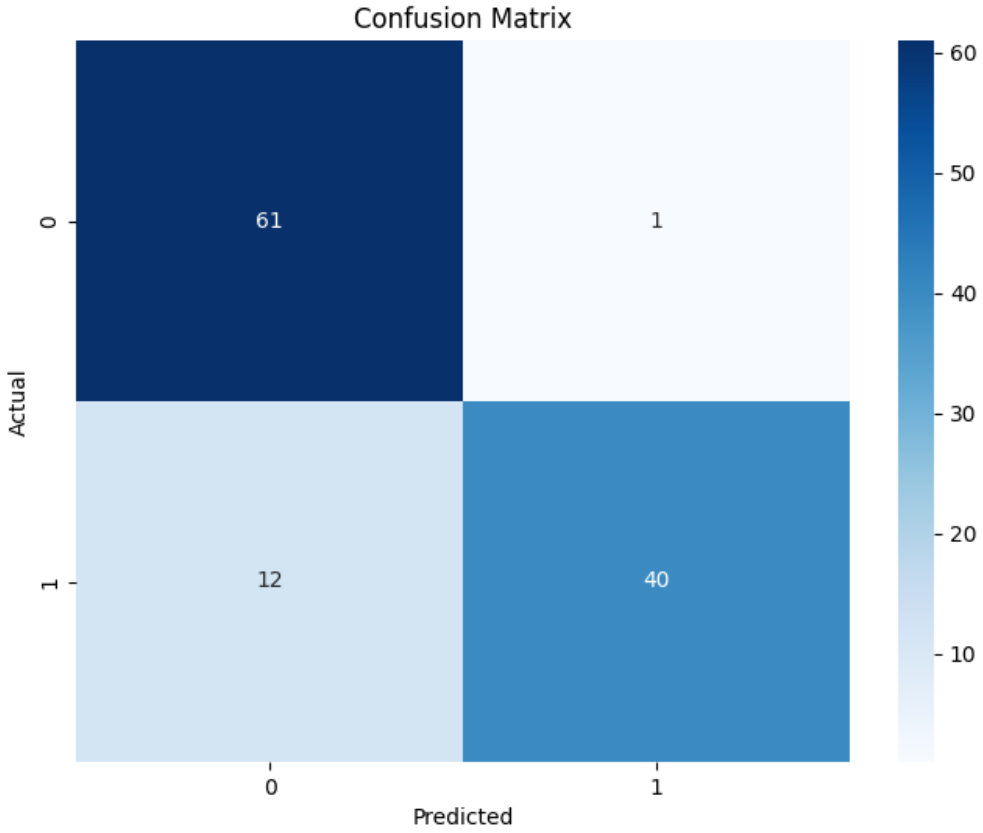

Figure 11. This plot shows the confusion matrix for our ensemble mode after predicting on out 30% testing data

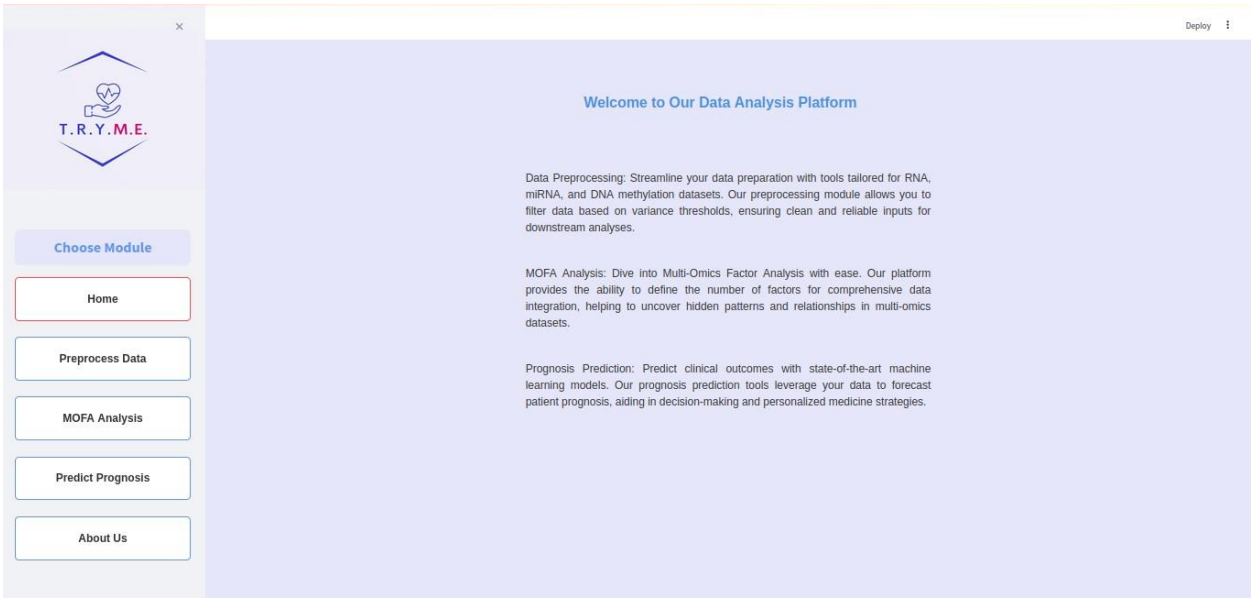

Figure 12. This figure shows the Home page of our website

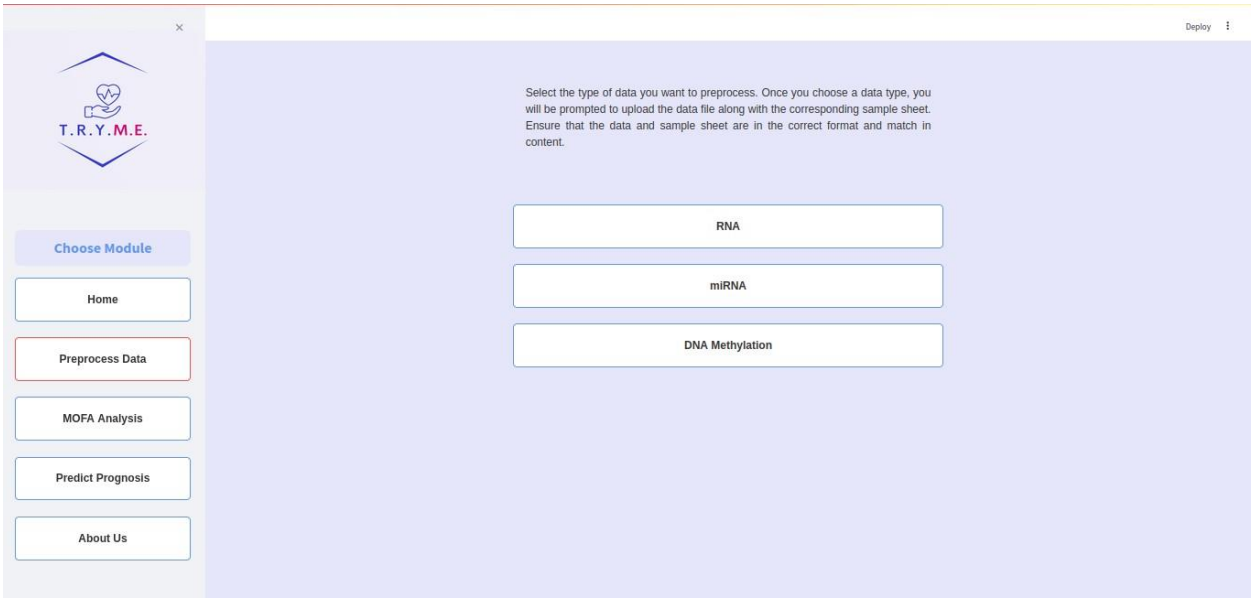

Figure 13. This figure shows the Preprocessing page of our website

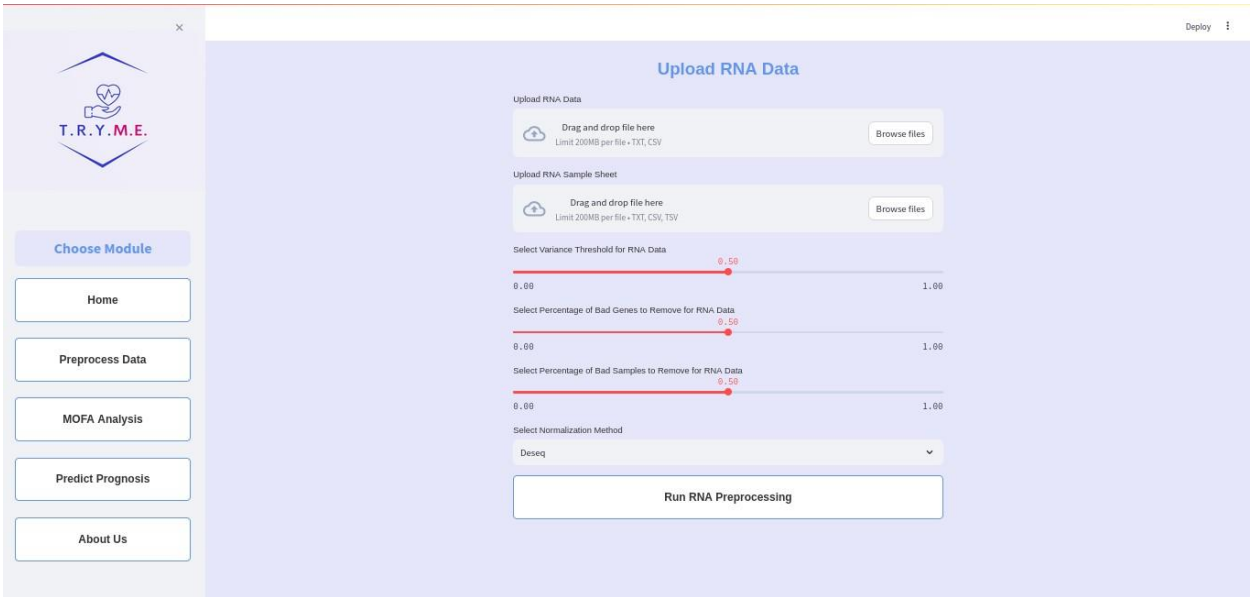

Figure 14. This figure shows the Preprocessing page of RNA

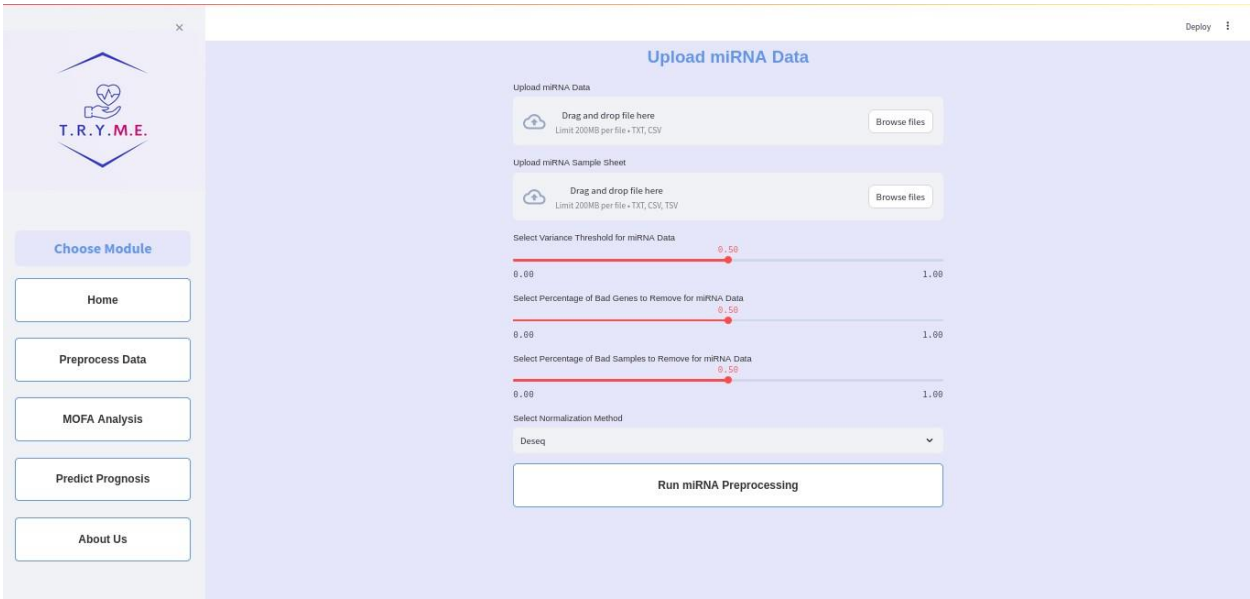

Figure 15. This figure shows the Preprocessing page of miRNA

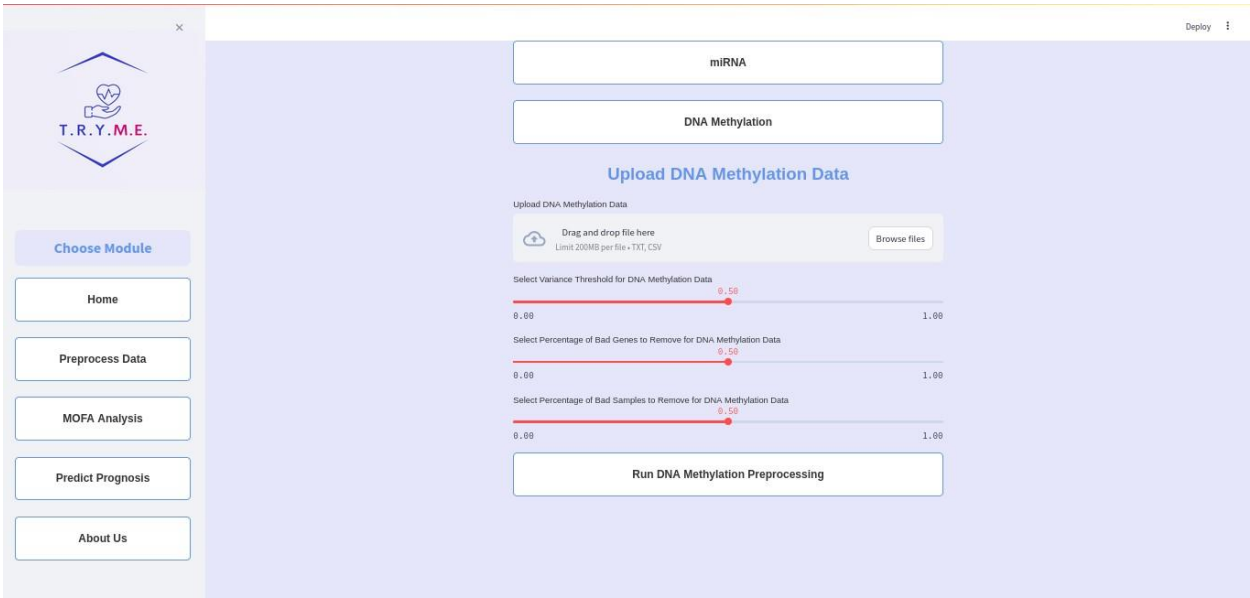

Figure 16. This figure shows the Preprocessing page of DNA-Methylation

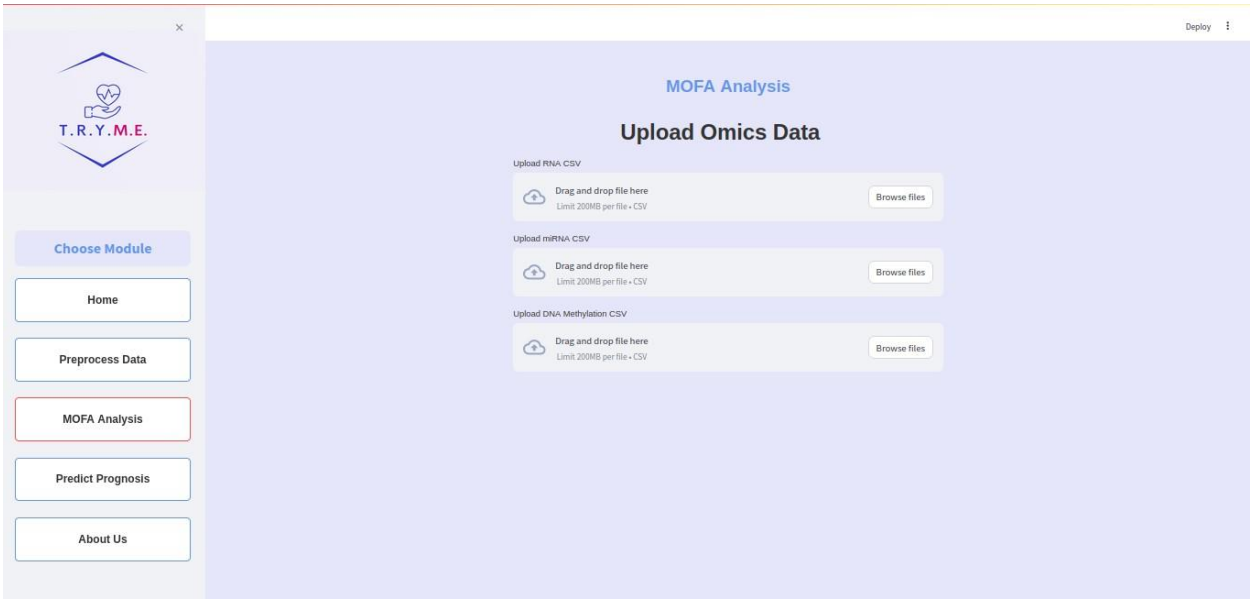

Figure 17.This figure shows the Preprocessing page of MOFA uploading omics

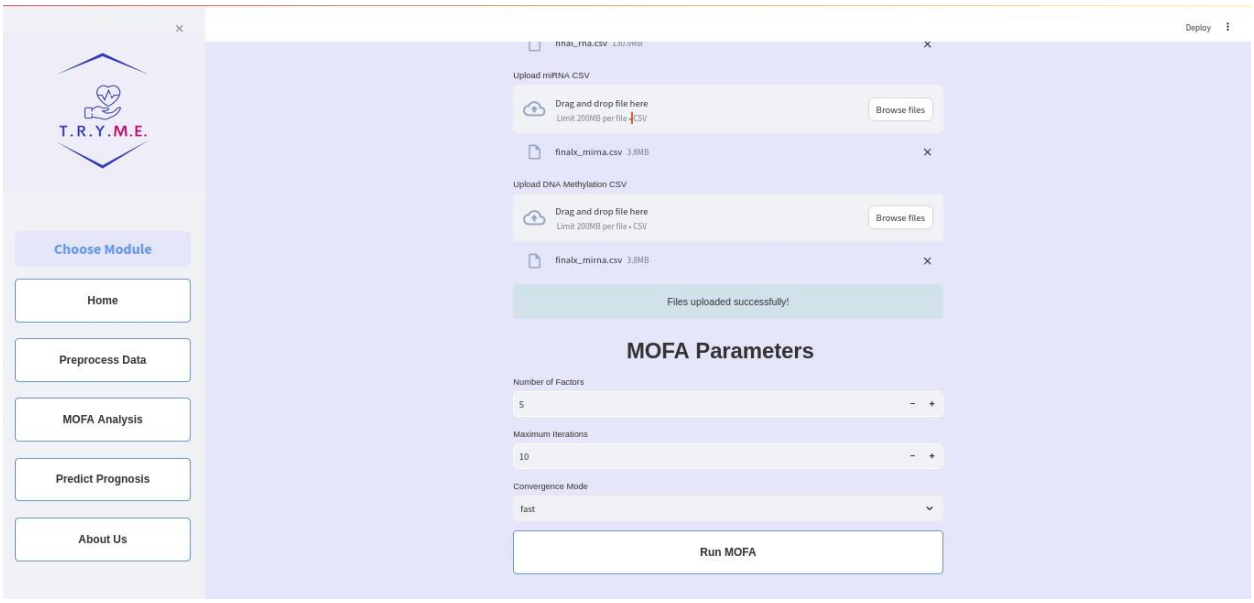

Figure 18. This figure shows the Preprocessing page of MOFA choosing parameters to run the model

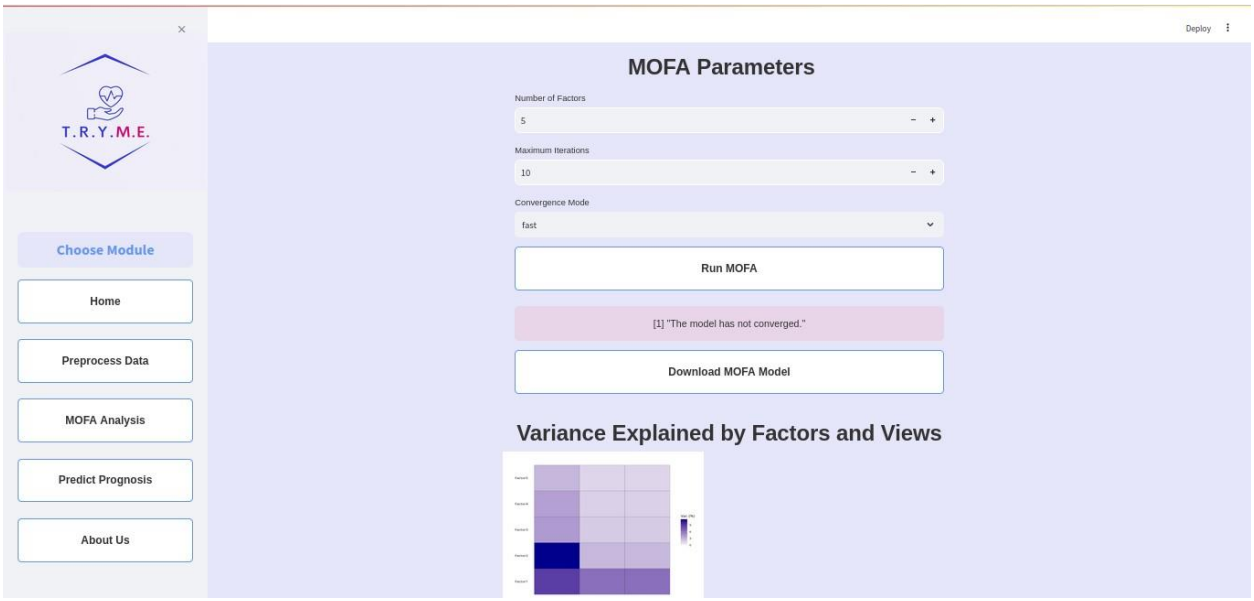

Figure 19. This figure shows the model has been converged or not on mock data

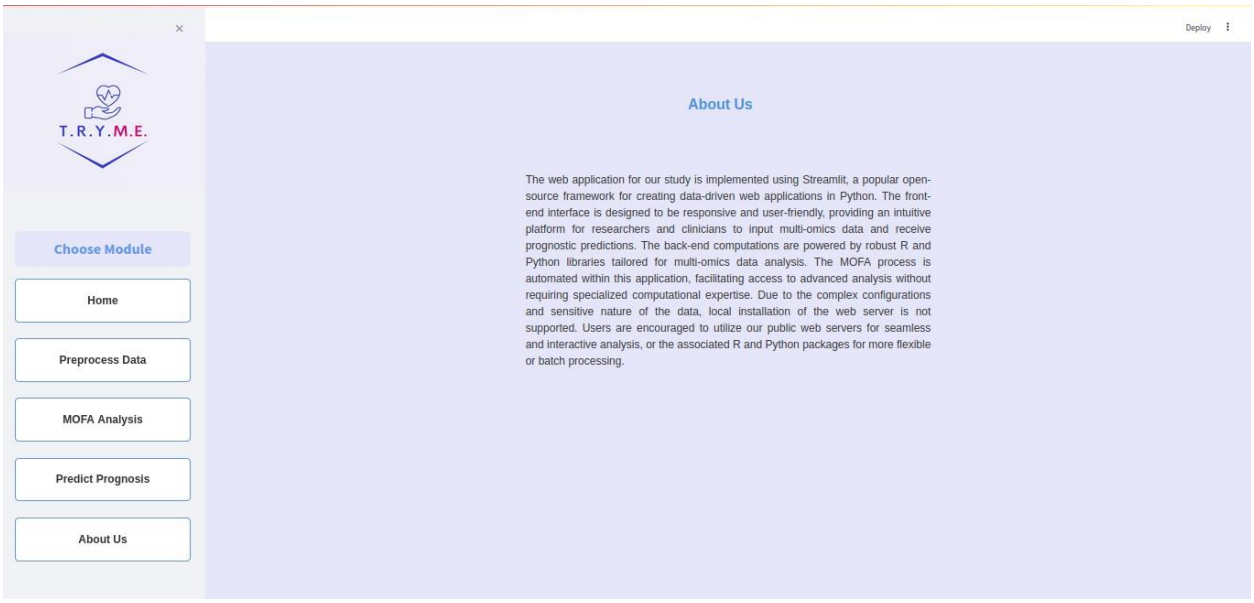

Figure 20. This figure shows the Page of About Us

5. Chapter 5: Discussion

5.1. Interpretation of Results

In our comprehensive approach to predicting the prognosis of stomach adenocarcinoma (STAD), we leveraged multi-omics data, including RNA sequencing (RNA-Seq), microRNA (miRNA), and DNA methylation, to enhance predictive accuracy. The data was gathered from The Cancer Genome Atlas (TCGA) and went through extensive preprocessing and normalization to ensure high-quality inputs. For RNA-Seq data, we started by organizing raw read counts into a single data frame, where genes were represented by rows and samples by columns. We converted Ensemble gene IDs into more interpretable HGNC gene symbols and resolved duplicates using aggregation. Normalization techniques included DESeq2 for size factor estimation, variance stabilizing transformation (VST) to stabilize variance, and log2 transformation to reduce skewness, concluding in robust DESeq2 normalization to address sequencing depth differences and technical bias.

Similarly, miRNA data underwent thorough preprocessing. We organized read counts into a single data frame, handled duplicate miRNA IDs by averaging their expression values, and integrated essential sample metadata. Data cleaning involved removing samples and miRNAs with high missing values and filtering

out low-variance miRNAs, leaving just the top 25% most variable miRNAs. To ensure data comparability, normalization involved log2 transformation and DESeq2. For DNA methylation data, we merged data into a data frame, filtered out samples and genes with significant missing values, applied B2M normalization to correct technical variations, and converted Beta values to M values to stabilize variance, especially at extreme values.

To identify the most informative features, we relied on Multi-Omics Factor Analysis (MOFA), selecting 100 features from DNA methylation, 70 from RNA, and 30 from miRNA. These features formed the basis for training various machine learning models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. Each model was trained and evaluated using optimal parameters identified through a significant grid search. The SVM model demonstrated the highest performance with an accuracy of 0.934, followed by Random Forest at 0.825, XGBoost at 0.816, and Logistic Regression at 0.806. To further enhance predictive accuracy, we implemented a Voting Classifier ensemble model, combining the strengths of the individual models. An ensemble learning, using a hard voting strategy, yielded an accuracy of 0.886, performing better than individual models and demonstrating the efficiency of integrating a variety of molecular data types.

Streamlit was used for the development of a user-friendly web interface, which

allows seamless interaction with our model and data. The website includes sections for data preprocessing, MOFA analysis, and prognosis prediction, enabling researchers to preprocess omics data, perform integrative analyses, and predict patient prognosis within a single platform. The Preprocess Data interface allows for the normalization of RNA, miRNA, and DNA methylation data, ensuring clean and dependable inputs for further analyses. The MOFA Analysis section enables users to specify the number of factors for data integration, uncovering hidden links and patterns in multi-omics datasets. By utilizing advanced machine learning models, the Prognosis Prediction tool supports the development of personalized care plans.

5.2. Comparison with Previous Studies

Our pipeline for predicting STAD (Stomach Adenocarcinoma) prognosis represents a significant advancement over previous studies by integrating multi-omics data using an ensemble learning approach. Previous studies typically focused on single types of molecular data, such as RNA-Seq, miRNA, or DNA methylation independently, which limited the predictive power and comprehensiveness of their models. In contrast, our method combines RNA-Seq, miRNA, and DNA methylation data, providing a more comprehensive perspective of the molecular changes in STAD. Furthermore, many previous studies used

traditional statistical approaches or single machine learning models, which may not have captured the complicated connections between different types of omics data. Our usage of a Voting Classifier to combine the strengths of Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost models is novel in this context. Each model was independently optimised using GridSearchCV to ensure optimal performance before being added to the ensemble. This approach resulted in a test set accuracy of 0.886, which is significantly higher than the accuracies reported in earlier studies that focused on specific data types or models.

5.3. Limitations

One of the major limitations of our STAD (Stomach Adenocarcinoma) prognostic prediction system is the sample size. The limited number of samples available for training and validation may not reflect the entire heterogeneity of STAD, resulting in model overfitting and reduced generalizability to large patient populations. Furthermore, the computational complexity of advanced methods, such as deep learning models and multi-omics integration techniques, is a significant challenge.

6. Chapter 6: Conclusion and Future Work

6.1. Conclusion:

To sum up, this work is a major contribution to cancer research, especially in stomach adenocarcinoma prognosis prediction. The research has successfully integrated RNA, miRNA, and DNA methylation data using the advanced Multi-Omics Factor Analysis (MOFA) framework by utilizing the power of multi-omics integration and machine learning. By taking this approach, stomach cancer can be better understood in terms of the intricate relationships and molecular fingerprints that underlie the disease, and prognosis models can be more accurately predicted.

By integrating RNA, miRNA, and DNA methylation data using MOFA, it was possible to uncover hidden patterns and connections between various molecular levels, which led to new discoveries about the biological processes underlying the development of stomach cancer. Researchers were able to pinpoint important biological characteristics and biomarkers that influence illness prognosis and treatment response by merging these various omics datasets.

With an impressive test set accuracy of 93.4%, the Support Vector Machine (SVM) was the best performing machine learning model among those examined models in this study. These findings demonstrate the integrated multi-omics approach's resilience in making highly reliable and accurate predictions about

patient outcomes.

Moreover, the creation of an intuitive web interface with Streamlit improves the research findings' usability and accessibility. Researchers and medical professionals can do MOFA analysis, effectively preprocess data, and forecast patient prognosis for stomach cancer with this interface. The web platform enables well-informed decision-making and individualized treatment plans based on individual molecular profiles by offering interactive tools for data exploration and visualization.

In summary, this study shows the revolutionary potential of combining multi-omics data and advanced machine learning approaches in clinical oncology. It also increases our understanding of stomach adenocarcinoma at the molecular level. The SVM, Random Forest, XGBoost, and Logistic Regression models all showed high accuracies, demonstrating how well this method works to convert complex omics data into useful insights for customized medicine. These results open the door to more investigation and practical applications with the goal of enhancing patient outcomes by using individualized treatment plans that are specific to each patient's molecular features as they relate to stomach adenocarcinoma.

6.2. Future Work:

Future work on our STAD (Stomach Adenocarcinoma) prognosis prediction

pipeline could significantly enhance its capabilities by incorporating additional omics data, such as proteomics, metabolomics, and epigenomics, to provide a more comprehensive molecular characterization and identify new biomarkers. Applying advanced deep learning models, specifically Multi-Omics Graph Convolutional Networks (MOGCN), can capture complex non-linear correlations within the data more effectively than traditional statistical methods like MOFA, potentially improving prediction performance. Moreover, extending the pipeline to provide patient-specific treatment recommendations by integrating treatment response data and optimizing models for personalized therapeutic strategies would bring us closer to precision medicine. These advancements could improve survival rates and quality of life for STAD patients.

References

[1] Piazuelo, M. B., & Correa, P. (n.d.). *Gastric cancer: Overview*. http://www.scielo.org.co/scielo.php?pid=S1657-95342013000300011&script=sci_arttext

[2] Chen, S., Zang, Y., Xu, B., Lu, B., Ma, R., Miao, P., & Chen, B. (2022). Multi-Omics Integration for Identifying Novel Molecular Subtypes of Clear Cell Renal Cell Carcinoma. *Computational and Mathematical Methods in Medicine*, 2022, 5844846. https://doi.org/10.1155/2022/5844846

[3] Lei ZN, Teng QX, Tian Q, Chen W, Xie Y, Wu K, Zeng Q, Zeng L, Pan Y, Chen ZS, He Y. Signaling pathways and therapeutic interventions in gastric cancer. Signal Transduct Target Ther. 2022 Oct 8;7(1):358. doi: 10.1038/s41392-022-01190-w. PMID: 36209270; PMCID: PMC9547882.

[4] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492.

[5] Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018 May;19(5):299-310. doi: 10.1038/nrg.2018.4. Epub 2018 Feb 26. PMID: 29479082; PMCID: PMC5990367.Zhang W, Li F, Nie L. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. Microbiology (Reading). 2010

Feb;156(Pt 2):287-301. doi: 10.1099/mic.0.034793-0. Epub 2009 Nov 12. PMID: 19910409.

[6] Niu, P. H., Zhao, L. L., Wu, H. L., Zhao, D. B., & Chen, Y. T. (2020). Artificial intelligence in gastric cancer: Application and future perspectives. *World Journal of Gastroenterology*, *26*(36), 5408-5419. https://dx.doi.org/10.3748/wjg.v26.i36.5408

[7] Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018 Jun 20;14(6):e8124. doi: 10.15252/msb.20178124. PMID: 29925568; PMCID: PMC6010767.

[8] Johnston, F. M., & Beckman, M. (2019). Updates on Management of Gastric Cancer. *Current Oncology Reports*, *21*(8). https://doi.org/10.1007/s11912-019-0820-4

[9] World Health Organization Classification of Tumours. (1999). IARCPress International Agency for Research on Cancer (IARC) 69372 Lyon, France. http://ndl.ethernet.edu.et/bitstream/123456789/36532/1/125

[10] *Tutorials*. (n.d.). Multi-Omics Factor Analysis. https://biofam.github.io/MOFA2/tutorials.html

[11] Liu, C., Wang, X., Genchev, G. Z., & Lu, H. (2017). Multi-omics facilitated variable selection in Cox-regression model for cancer prognosis prediction. *Methods*, *124*, 100–107.

https://doi.org/10.1016/j.ymeth.2017.06.010

[12] Chai, H., Zhou, X., Zhang, Z., Rao, J., Zhao, H., & Yang, Y. (2021). Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Computers in Biology and Medicine*, *134*, 104481. https://doi.org/10.1016/j.compbiomed.2021.104481

[13] Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S., & Garmire, L. X. (2021). DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Medicine*, *13*(1). https://doi.org/10.1186/s13073-021-00930-x

[14] Franco, E. F., Rana, P., Cruz, A., Calderón, V. V., Azevedo, V., Ramos, R. T. J., & Ghosh, P. (2021). Performance Comparison of Deep Learning Autoencoders for Cancer Subtype Detection Using Multi-Omics Data. Cancers, 13(9), 2013. https://doi.org/10.3390/cancers13092013

[15] Chen, S., Zang, Y., Xu, B., Lu, B., Ma, R., Miao, P., & Chen, B. (2022). An Unsupervised Deep Learning-Based Model Using Multiomics Data to Predict Prognosis of Patients with Stomach Adenocarcinoma. Computational and Mathematical Methods in Medicine, 2022, 1–20. https://doi.org/10.1155/2022/5844846

[16] Chang, J., Wu, H., Wu, J., Liu, M., Zhang, W., Hu, Y., Zhang, X., Xu, J., Li, L., Yu, P., & Zhu, J. (2023). Constructing a novel mitochondrial-related gene signature for evaluating the tumor immune microenvironment and predicting survival in stomach adenocarcinoma.

Journal of Translational Medicine, 21(1). https://doi.org/10.1186/s12967-023-04033-6

[17] Yegulalp, S. (2024, April 17). *Intro to Streamlit: Web-based Python data apps made easy*. InfoWorld. https://www.infoworld.com/article/3715100/intro-to-streamlit-web-based-python-data-apps-made-easy.html

[18] *Streamlit • A faster way to build and share data apps*. (n.d.). https://streamlit.io/

[19] https://portal.gdc.cancer.gov/projects/TCGA-STAD