Contribution Report – Multi-Omics Cancer Analysis

Student Weilin He
Student ID 2476216
Course Data Science Mini Project (M28)

Team Collaboration Overview

At the early stage of the project, I took responsibility for the structuring of our collaboration and technical workflow. I independently compiled more than 20 guidance and management documents, which covered preprocessing instructions, data merging pipelines, weekly progress records, naming conventions, file directory rules, and output alignment standards. These materials enabled group members to efficiently collaborate under a common framework and ensured the reproducibility of the multi-stage pipeline.

In parallel, I produced the *Understanding Data* document, which provided detailed interpretations of the 17 raw data tables provided at the beginning of the project. For each dataset, I explained field meanings, origin, biological context, and integration methods. This served as the foundational reference for all subsequent modeling and analysis, and greatly enhanced the team's shared understanding of data structure.

Technical Report and Presentation Preparation

For the final technical report, I contributed 2,764 words out of a total of 4,670, accounting for 59.2% of the full document. The sections I independently authored include the Abstract, Introduction, the entire Methodology chapter, Section 5.3 on DNA methylation co-regulation, Section 6.3 on tumor versus normal tissue analysis, and all parts of the final Conclusions. These sections form the analytical and conceptual backbone of the report.

Additionally, I led the creation of the final 24-slide presentation. Around 90% of the slide content—including layout design, code result visualisations, figure annotations, and textual summaries—was independently completed by me. The remaining slides were drafted by teammates following my structural guidance. The overall presentation flow was based on the technical logic I outlined in the main report.

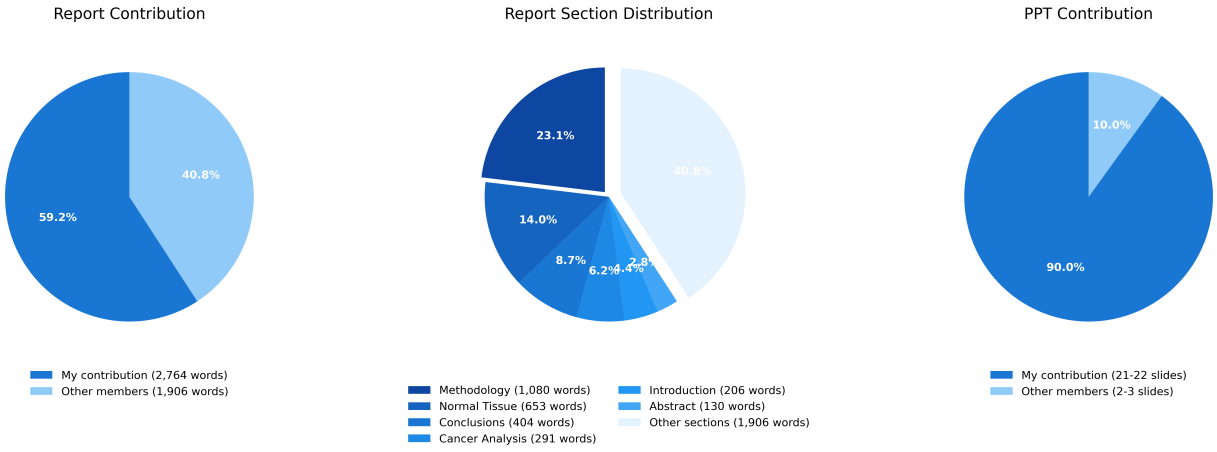

Figure 1: Writing and presentation contribution breakdown across report and slides.

Cancer Analysis and Modeling Tasks

Among the five cancer types investigated in this project—PRAD, SKCM, STAD, PAAD, and LUAD—I independently completed the full analytical pipeline for three cancers: PRAD, SKCM, and STAD. This included the entire workflow from raw data cleaning, sample harmonization, and annotation mapping to differential expression (DEG) and methylation (DMA) analyses, transcription factor (TF) network reconstruction, pathway enrichment, and result visualization. These tasks account for 60% of the cancer-type-specific analytical work.

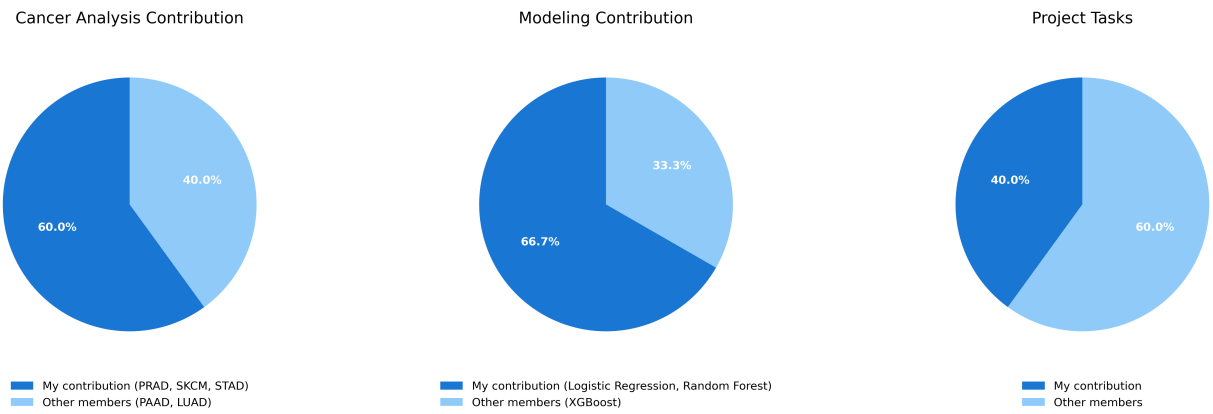

Figure 2: Distribution of contributions across cancer types and modeling tasks.

With respect to machine learning modeling, each cancer type was modeled using three approaches: Logistic Regression, Random Forest, and XGBoost. I independently implemented two of these—Logistic Regression and Random Forest—across all assigned cancer types. These implementations involved complete pipelines for data preprocessing, feature selection, model training, cross-validation, and performance interpretation, comprising 66.7% of the modeling workload.

Beyond specific cancer assignments, I also contributed substantially to the unified analytical framework. I took part in almost every core step of the multi-omics pipeline, including omics feature integration, high-dimensional variable screening, result standardization, and biological interpretation. According to GitHub activity and line-level statistics, I completed approximately 40% of the total code contribution to the project. My work ensured coherence and reusability of the analytical structure across cancers and models.

GitHub and Code Contributions

In the development phase of the project, I served as the key contributor in terms of source code implementation, maintenance, and standardization. Although my GitHub account (WeilinHe00619) shows 27 commits due to an upload issue I encountered in April (which was communicated to the TSR and handled locally), my overall code changes were by far the most substantial among all group members.

Specifically, I contributed 378,119 lines of additions and 285,029 lines of deletions, according to GitHub contributor statistics. These figures represent the highest volume of active code editing within the team, and were the result of extensive model development, result visualization, data processing pipelines, and automated output generation scripts. My commit frequency appeared lower because I followed a strategy of local staging and periodic bulk uploads to avoid data inconsistency during the repository access issue.

As shown in Table 1, I ranked first in both code additions and deletions, demonstrating my leading role in the project's implementation phase.

Beyond code development, I also authored the main `README.md` file for the GitHub repository. This document provides a full explanation of the folder structure, analytical pipeline, reproduction instructions, and dataset configuration. Moreover, the directory layout, naming conventions, and output schema I established were adopted across all group contributions to ensure organizational consistency and downstream integration. My work on GitHub formed the technical backbone of our collaborative infrastructure and version control framework.

	Additions	**Deletions**	**Commits**
WeilinHe00619	378,119	285,029	27
Rank	1	1	2

Table 1: Summary of GitHub contributions by Weilin He.