Vrije Universiteit Amsterdam

BSc Artificial Intelligence Thesis

Robust Multi-Omics Integration: Comparative Analysis of Feature Engineering and Intermediate Fusion Techniques for Cancer Prediction

Krzysztof Piotr Nowak

VU Amsterdam Supervisor: prof. dr. Annette ten Teije

OUH Supervisors: PhD Student Alberto López, PhD student Katarina Willoch

Bachelor of Science in *Artificial Intelligence*

July 10, 2025

Abstract

The integration of multi-omics data is crucial for advancing precision oncology, yet its clinical translation is impeded by a lack of systematic benchmarks for computational pipelines. This research addresses this gap by conducting a large-scale comparative analysis to identify robust and efficient strategies for cancer prediction, with a focus on intermediate fusion methods and their resilience to missing data. By developing and deploying a reusable benchmarking framework, this study systematically evaluated 10,206 unique configurations across nine cancer datasets from The Cancer Genome Atlas (TCGA). The pipeline combined fourteen feature engineering algorithms (supervised and unsupervised), eight intermediate fusion strategies, and six machine learning models. The robustness of each configuration was rigorously tested under simulated missing data in modalities (Gene Expression, miRNA Expression and DNA Methylation) scenarios of 0%, 20%, and 50%. The results demonstrate that supervised feature extraction methods, such as Linear Discriminant Analysis (LDA) and Partial Least Squares (PLS), are critical for high performance, significantly outperforming unsupervised approaches. Furthermore, adaptive fusion techniques, particularly Multiple Kernel Learning (MKL), proved superior in both predictive accuracy and robustness to data incompleteness. Notably, the top-performing combinations are also the most computationally efficient. This work concludes that the optimal strategy for multi-omics integration involves combining supervised feature extraction with adaptive fusion. It provides a foundational benchmark that supports the development of clinically viable predictive models that are accurate, robust, and efficient.

Keywords: Multi-Omics Integration, Cancer Prediction, Machine Learning, Feature Engineering, Benchmarking, Missing Data

Contents

1. Introduction

1.1 Background and Motivation

Machine learning has emerged as a powerful tool in oncology, enabling the development of predictive models that can distil complex molecular data into clinically actionable insights. However, the full potential of these models is often unrealised because many current approaches do not adequately integrate multi-omics data, which combines genomic, transcriptomic, and other biological layers for a more holistic view of disease mechanisms. This gap is addressed by focusing on the optimisation of intermediate fusion, which merges data after initial feature processing to balance modality-specific signals with cross-modal interactions. By systematically evaluating and identifying robust analytical tools for this fundamental bottleneck in multi-omics analysis, this thesis aims to contribute to significant advancements in cancer detection. It is conducted as part of a project for a research group with expertise in translating complex molecular data into predictive models for personalised therapy at the Oslo University Hospital, Institute for Cancer Research. (Ous-research.no (2024))

A central goal of this larger effort is to create maximally predictive models from complex, multi-view patient data to diagnose cancer more accurately and rapidly. However, the clinical translation of these methods is hindered by significant technical challenges, including high data dimensionality, heterogeneity, and the frequent occurrence of missing values. While many studies apply feature engineering to individual data types before analysis, a critical and unresolved question is how to best combine these processed features in an intermediate step to achieve optimal predictive performance. This thesis directly confronts this challenge by conducting a rigorous, comparative analysis of intermediate fusion techniques, aiming to provide the foundational, empirical guidance needed to advance the development of next-generation predictive models in oncology.

1.2 Problem Statement and Research Gaps

The central problem addressed by this thesis is the lack of a standardised framework for benchmarking multi-omics integration pipelines. This lack of a rigorous comparative process hinders both the identification of optimal configurations and the development of clinically viable predictive models. The literature reveals two primary gaps that this research aims to fill. First, there is a notable absence of systematic evaluations focused specifically on *intermediate* fusion methods, strategies that merge data after initial feature processing to balance the preservation of modality-specific signals with the learning of cross-modal interactions. Second, the robustness of these computational pipelines to missing data, a common and persistent challenge in real-world clinical settings, is rarely assessed in a controlled manner, thereby limiting the practical utility and reliability of many proposed techniques.

1.3 Research Questions and Objectives

To address the identified gaps, the primary research question guiding this thesis is:

> How do different intermediate fusion strategies and feature engineering methods compare in terms of predictive performance, computational efficiency, and robustness to missing data for multi-omics cancer prediction?

This study is structured around two primary, interconnected objectives, designed to answer this question comprehensively:

1. To develop and present a modular, extensible framework for the systematic benchmarking of multi-omics integration pipelines.

2. To deploy this framework in a large-scale comparative analysis aimed at identifying the most robust, performant, and efficient pipeline configurations for cancer prediction. This primary objective is achieved through the following goals:

   (a) To benchmark eight intermediate fusion strategies across nine distinct cancer datasets, utilising Matthew's Correlation Coefficient (MCC) for classification tasks and the Coefficient of Determination (R^2) for regression tasks as the primary performance metrics.

   (b) To evaluate how the application of various combinations of feature selection and extraction algorithms, fusion techniques, and models influences the predictive performance and computational cost of the overall pipeline.

   (c) To systematically assess the robustness of each configuration under clinically relevant scenarios of simulated missing data in modalities (0%, 20%, and 50%).

2. Literature Review

2.1 Strategies for Multi-Omics Data Integration

The integration of multi-omics data has been shown to capture the complex biological interplay underlying cancer more effectively than single-omics approaches, leading to more accurate and clinically relevant predictive models (Sammut et al. (2022); Hernández-Lemus and Ochoa (2024)). Machine learning-based data fusion strategies are typically categorised into three main paradigms: early, intermediate, and late integration (Picard et al. (2021)). Early integration involves the simple concatenation of features before model training, while late integration aggregates the outputs of models trained independently on each data modality (Yang et al. (2024)).

This research concentrates on intermediate fusion to address a critical gap: the lack of systematic benchmarks for this specific integration strategy. Theoretically, this paradigm offers an optimal balance by enabling tailored, modality-specific feature processing while still permitting the model to learn the complex cross-modal interactions fundamental to cancer biology. This architectural choice provides a strategic juncture to manage practical challenges, such as building robustness against missing data, which is a core objective of this work. Notable intermediate fusion methods include Multiple Kernel Learning (MKL), which integrates modalities through an optimised combination of kernels (Gönen and Alpaydın (2011)), and attention-based mechanisms that learn to weight data sources based on their predictive relevance (Cai et al. (2022)). By focusing on this pivotal stage, the study provides guidance for constructing robust and clinically viable predictive models.

2.2 State of Benchmarking in Multi-Omics Research

Several key studies have benchmarked multi-omics integration methods, providing valuable insights into the landscape of available algorithms. For instance, extensive comparisons of clustering methods revealed that performance varies significantly across different cancer types, highlighting the need for context-specific evaluation (Rappoport and Shamir (2018)). Similarly, assessments of integration methods for survival prediction have shown that adding more omics layers does not universally improve performance, underscoring the importance of employing adaptive fusion strategies that can selectively utilise data (Duan et al. (2021)).

Other studies have focused on specific stages of the integration pipeline, such as joint dimensionality reduction, confirming the utility of methods like Principal Component Analysis (PCA) for managing data complexity (Cantini et al. (2021)). More recently, benchmarks of deep learning models have found that no single fusion architecture is consistently superior across all datasets (Leng et al. (2022)). A recurring theme throughout this body of work is the clear and persistent need for systematic, reproducible, and comparative analyses across a wide range of conditions, a need which this thesis directly addresses.

3. Data

The multi-omics datasets used in this research are derived primarily from The Cancer Genome Atlas (TCGA), a publicly accessible resource that offers comprehensive genomic datasets for multiple cancer types (Rappoport and Shamir (2018)). TCGA provides robust, standardised data facilitating systematic comparative studies, making it particularly suitable for evaluating multi-omics fusion methodologies. For this thesis, selected datasets cover different cancer types, including Acute Myeloid Leukemia (AML), Breast, Sarcoma, Colon, Kidney, Liver, Lung, Melanoma, and Ovarian. (Table 3.1) Each dataset includes three omics modalities: gene expression (RNA-seq), microRNA expression (miRNA-seq), and DNA methylation. (Appendix A)

Table 3.1: Overview of all datasets including Target explanation and class imbalance.

Dataset	Type	Target	Target Description	Target samples	Target balance
AML	Regression	lab_procedure_bone _marrow_blast_cell _outcome _percent_value	The percentage of immature blood cells found in a patient's bone marrow	200	**0%–20%:** 89; **20%–40%:** 21; **40%–60%:** 36; **60%–80%:** 24; **80%–100%:** 30
Sarcoma	Regression	pathologic_ tumor_length	The measured length (cm) of a tumour	271	**1cm–8cm:** 70; **8cm–16cm:** 99; **16cm–24cm:** 44; **24cm–32cm:** 17; **32cm–40cm:** 4
Breast	Classification	pathologic_T	Refers to the tumour size and extent component of the TNM cancer-staging system	1247	**T1:** 45; **T1a:** 2; **T1b:** 18; **T1c:** 253; **T2:** 720; **T2a:** 1; **T2b:** 2; **T3:** 150; **T3a:** 1; **T4:** 9; **T4b:** 34; **T4d:** 4; **TX:** 3
Colon	Classification	pathologic_T	Refers to the tumour size and extent component of the TNM cancer-staging system	551	**T1:** 11; **T2:** 90; **T3:** 377; **T4:** 36; **T4a:** 20; **T4b:** 11; **Tis:** 1
Kidney	Classification	pathologic_T	Refers to the tumour size and extent component of the TNM cancer-staging system	985	**T1:** 37; **T1a:** 250; **T1b:** 205; **T2:** 101; **T2a:** 15; **T2b:** 8; **T3:** 7; **T3a:** 234; **T3b:** 102; **T3c:** 4; **T4:** 22
Liver	Classification	pathologic_T	Refers to the tumour size and extent component of the TNM cancer-staging system	438	**T1:** 210; **T2:** 109; **T2a:** 1; **T2b:** 2; **T3:** 56; **T3a:** 33; **T3b:** 9; **T4:** 16; **TX:** 1
Lung	Classification	pathologic_T	Refers to the tumour size and extent component of the TNM cancer-staging system	626	**T1:** 61; **T1a:** 28; **T1b:** 51; **T2:** 218; **T2a:** 112; **T2b:** 42; **T3:** 81; **T4:** 30
Melanoma	Classification	pathologic_T	Refers to the tumour size and extent component of the TNM cancer-staging system	481	**T0:** 23; **T1:** 10; **T1a:** 22; **T1b:** 10; **T2:** 32; **T2a:** 32; **T2b:** 15; **T3:** 15; **T3a:** 9; **T3b:** 39; **T4:** 16; **T4a:** 26; **T4b:** 14; **TX:** 48
Ovarian	Classification	clinical_stage	Initial stage of cancer determined before treatment (closest available variable to pathologic_T)	630	Stage **IA:** 4; **IB:** 3; **IC:** 11; **IIA:** 4; **IIB:** 5; **IIC:** 24; **IIIA:** 8; **IIIB:** 25; **IIIC:** 429; **IV:** 89

3.1 Omics Modalities

The omics modalities analysed in this research include (Appendix A):

- **Gene Expression (RNA-seq)**: Quantifies the expression levels of thousands of genes simultaneously, providing insights into cellular function and disease processes.
- **miRNA Expression**: Reflects short RNA molecules that regulate gene expression post-transcriptionally, influencing cancer progression and prognosis.
- **DNA Methylation**: Measures epigenetic modifications which can silence or activate gene expression, playing critical roles in tumorigenesis and cancer progression.

3.2 Clinical Targets and Labels

Each dataset includes clinical targets tailored for prediction tasks using both regression and classification models. Specifically:

- **Regression Targets**: AML and Sarcoma datasets are used for regression tasks, as they contain continuous tumour size measurements expressed as length or percentage. (Table 3.1)
- **Classification Targets**: Breast, Colon, Kidney, Liver, Lung, Melanoma, and Ovarian datasets are used for classification tasks, based on tumour size categories (e.g., T1, T2, T3, T4). (Table 3.1)

Survival time was intentionally excluded as a target variable because such data is commonly censored. It means the event of interest (such as death or hospital discharge) has not been observed for all patients by the end of the study period. For instance, if a patient's survival time is recorded as 150 days, it's unknown whether they passed away on day 151 or were discharged and remained healthy. The analysis of censored data requires specialised statistical methods and evaluation metrics, which are distinct from standard regression or classification. Therefore, it was omitted to maintain a focused analytical approach.

3.3 Data Challenges

Several data-related challenges inherent to multi-omics studies were encountered:

- **Missing Data**: Due to technical limitations and incomplete sampling, missing data is a significant issue. Modalities may lack certain measurements for subsets of patients, presenting challenges in consistent model training. (Appendix A)
- **High Dimensionality**: Omics data often include thousands of features, significantly complicating feature selection and extraction. (Appendix A)
- **Data Imbalance**: Clinical outcomes may show class imbalances, necessitating special handling during analysis to avoid biased models. (Table 3.1)

4. Methodology

The research is centred on a modular software architecture designed to systematically assess combinations of feature engineering, intermediate fusion, and predictive modelling, under varying conditions of data completeness. In total, 10,206 unique model configurations were trained and evaluated across nine cancer datasets:

- **Regression Domain (2 Datasets)**: 3 Missing Data scenarios → 6 Extraction and 5 Selection Algorithms for Regression → 8 Intermediate Fusion Techniques → 3 Regression Models
- **Classification Domain (7 Datasets)**: 3 Missing Data scenarios → 6 Extraction and 5 Selection Algorithms for Classification → 8 Intermediate Fusion Techniques → 3 Classification Models

Source code, configuration files, and reproducibility scripts are publicly available on GitHub (Nowak (2025)). The experiments were conducted in a Python 3.12 environment on a Virtual Machine equipped with dual Intel Xeon Silver 4114 CPUs and 58.6 GB of RAM DDR4-3200.

4.1 Methodological Assumptions and Scope

The analytical framework presented here is predicated on several foundational assumptions that circumscribe its application and define its scope:

- **Availability of a Labelled Target:** A defined clinical endpoint or outcome variable (e.g., tumour stage, disease subtype) is required to train and evaluate the models.
- **Multi-View Data Structure:** The methodology is designed for multi-view data, assuming inputs are structured as distinct yet complementary modalities (e.g., genomics) that can be integrated via intermediate fusion.
- **Quantitative Feature Space:** The computational algorithms in this Feature Space require that all input modalities must be transformed into a numerical feature matrix to serve as vectorised input.
- **Existence of a Predictive Signal:** The framework presupposes the existence of a non-random, predictive signal within the data. Its objective is to optimise the extraction and modelling of this signal, rather than to test for its statistical significance.

4.2 Data Preprocessing and Quality Control

Effective data preprocessing is critical for mitigating the challenges of high dimensionality and biological variability inherent in multi-omics data (Rappoport and Shamir (2018)). By following the state-of-the-art, the preprocessing workflow, illustrated in Figure 4.1, begins with aligning patient samples across all modalities using the TCGA patient ID convention to create a unified master patient list. Each modality then underwent a tailored quality control and preprocessing sequence. (Figure 4.1a&b)

Fuzzy Matching Recovery was employed to correct minor formatting differences in patient IDs and features across various data files, a necessary step to prevent sample loss when combining them. Next, Class Distribution Optimisation addressed severe class imbalances by consolidating rare categories, a crucial step for ensuring robust cross-validation. (Figure 4.1c) Before more advanced feature engineering, a Median Absolute Deviation (MAD) threshold was applied to remove low-variability features, followed by a `SelectKBest` algorithm using mutual information for gene/miRNA expression and an F-test for DNA methylation (compares explained and unexplained variance). Then, MAD-based Feature Filtering is applied to keep only the best features with the highest MAD score to reduce high dimensionality. Afterwards, Sparsity Filter removes all features with a high percentage of zero values and NaNs. (Figure 4.1d) For gene and miRNA expression data, a $\ln(1+x)$ transformation was applied to reduce skewness, followed by normalisation using a `RobustScaler`. DNA methylation data, already bounded between 0 and 1, required no scaling. Outliers were clipped adaptively for each modality based on their standard deviation. (Figure 4.1e) Additionally, automated checks at the end of preprocessing verified dimensional consistency and data quality. (Figure 4.1f) Table 4.1 provides a detailed overview of the sample and feature counts for each dataset, together with the missing data percentage for every modality, after the completion of this preprocessing stage.

Table 4.1: Overview of sample, feature and missingness retention for every omics modality after preprocessing.

Dataset	Target Samples	Sample Retention (%)	Modality	Samples per Modality	Retention per Modality (%)	Features per Modality	Feature Retention (%)	Missing (%) After	Combined Missing (%) After
AML	170	85.00	Gene Expression	170	98.84	1418	6.91	0.18	0.19
			miRNA Expression	170	90.91	142	20.14	0.19	
			DNA Methylation	170	88.08	1900	38.00	0.18	
Sarcoma	226	83.39	Gene Expression	226	85.61	1423	6.93	0.00	0.00
			miRNA Expression	226	86.26	142	13.58	0.00	
			DNA Methylation	226	84.33	1900	38.00	0.00	
Breast	704	56.46	Gene Expression	704	58.13	1425	6.94	0.14	0.14
			miRNA Expression	704	83.02	142	13.58	0.14	
			DNA Methylation	704	79.64	1900	38.00	0.14	
Colon	201	35.93	Gene Expression	201	61.47	1419	6.91	0.50	0.52
			miRNA Expression	201	91.36	142	20.14	0.50	
			DNA Methylation	201	60.18	1900	38.00	0.54	
Kidney	245	24.87	Gene Expression	245	40.50	1419	6.91	0.40	0.41
			miRNA Expression	245	75.38	142	13.58	0.40	
			DNA Methylation	245	51.15	1900	38.00	0.42	
Liver	409	93.38	Gene Expression	409	96.92	1422	6.93	0.24	0.24
			miRNA Expression	409	96.69	142	13.58	0.25	
			DNA Methylation	409	95.56	1900	38.00	0.24	
Lung	332	53.04	Gene Expression	332	60.25	1422	6.93	0.30	0.30
			miRNA Expression	332	86.01	142	13.58	0.30	
			DNA Methylation	332	80.78	1900	38.00	0.30	
Melanoma	421	87.53	Gene Expression	421	89.19	1420	6.92	0.24	0.24
			miRNA Expression	421	93.35	142	13.58	0.24	
			DNA Methylation	421	88.82	1900	38.00	0.24	
Ovarian	290	46.03	Gene Expression	290	94.77	1424	6.94	0.00	0.00
			miRNA Expression	290	63.04	142	20.14	0.00	
			DNA Methylation	290	47.46	1900	38.00	0.00	

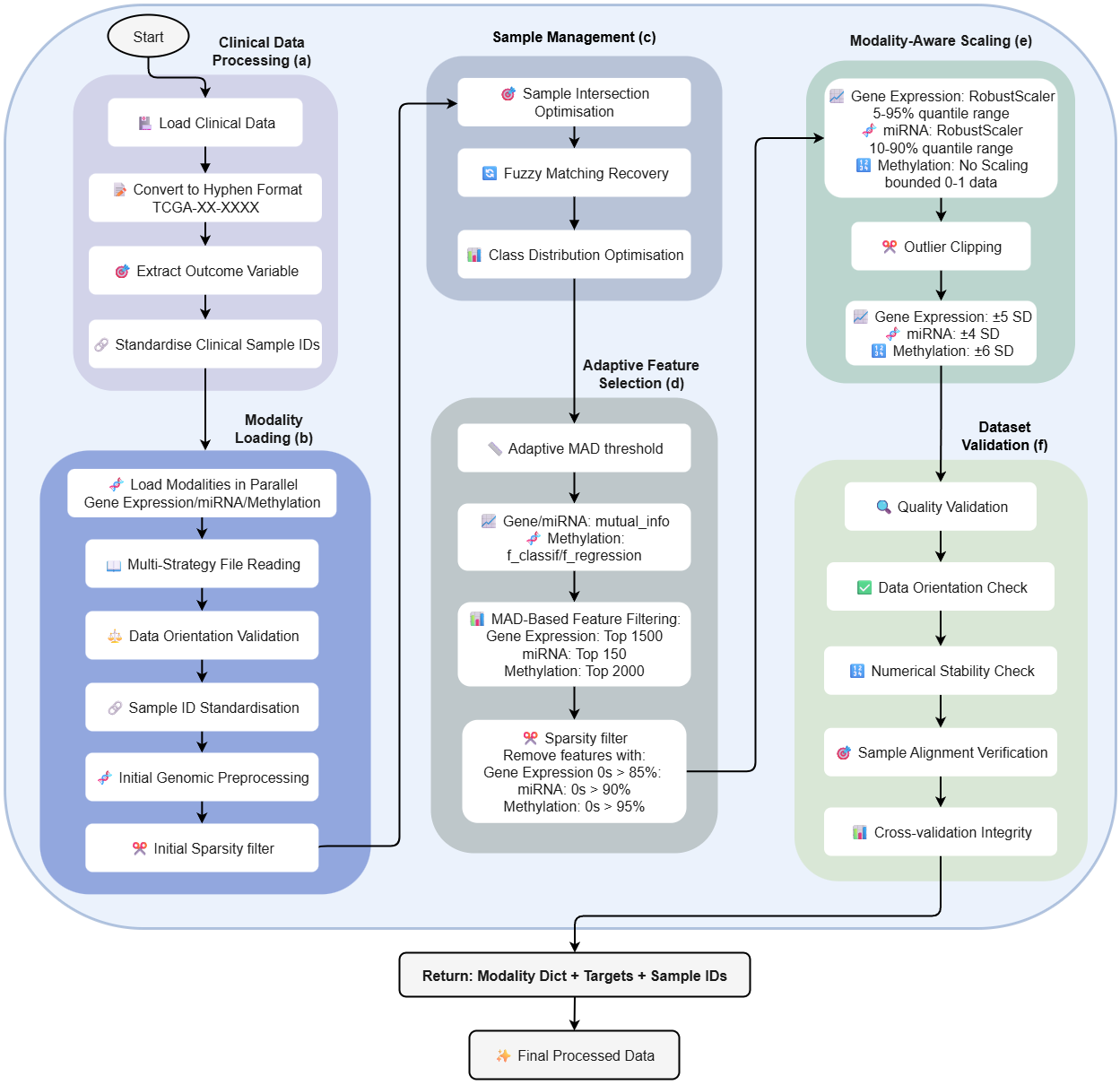


Figure 4.1: End-to-end preprocessing pipeline for multi-omics data.

4.3 Experimental Pipeline Design and Missing Data Simulation

The experimental workflow, presented in Figure 4.2, was designed to systematically test all combinations of the selected algorithms. A central component of this design is the simulation of missing data to model real-world clinical scenarios where patient records are often incomplete. Entire omics data blocks are randomly removed from a proportion of samples before model training across three completeness scenarios: 0% missing (complete data), 20% missing, and 50% missing. (Figure 4.2a) To ensure reproducibility, fixed random seeds were used across all cross-validation folds.

The pipeline employed a conditional approach to data imputation and fusion based on the level of missingness, which allows for balancing complexity and performance with computing time for different scenarios. For missing values within a modality, an adaptive rule was applied: mean imputation for $\leq$ 10% gaps, k-nearest neighbours ($k = 5$) for 10-50% gaps, and an iterative ExtraTrees regressor for $>$ 50% gaps (Geurts et al. (2006)). When an entire modality block was absent, its features were imputed using a cross-modality k-nearest neighbours approach. The choice of fusion strategy was also dependent on data completeness, with certain methods like attention-based fusion being reserved for complete-data scenarios only.

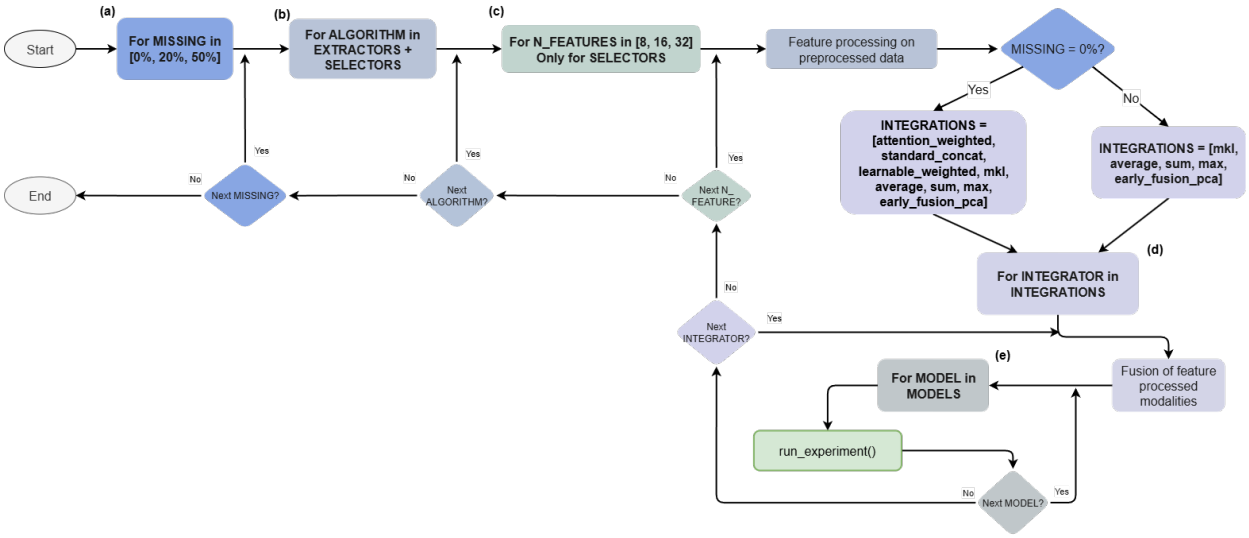


Figure 4.2: The Pipeline Structure Flowchart

4.4 Feature Engineering, FusioModellingodeling Strategies

The pipeline evaluated a diverse toolkit of algorithms at each stage, chosen to capture a wide range of linear and non-linear relationships and to assess both supervised and unsupervised approaches.

4.4.1 Feature Extraction and Selection

Two complementary strategies were applied independently to each modality: **feature extraction**, which creates new latent features, and **feature selection**, which retains a subset of original features. (Figure 4.2b) For the extractors, the number of latent components was treated as a hyperparameter. After tuning based on the MCC and R^2 scores, the optimal values were most frequently found to be 8, 16, or 32. (4.5 Hyperparameter Tuning) For the selectors, all models were trained using predefined subsets of the top 8, 16, and 32 features to assess performance across different levels of dimensionality. (Figure 4.2c) The pipeline tests fourteen algorithms in total, which are:

- **Feature Extractors:**
  - **PCA, KPCA, FA:** Unsupervised methods for dimensionality reduction.
  - **PLS, KPLS, SparsePLS:** Supervised methods for regression that find latent variables correlated with the outcome. SparsePLS was also optimised and tested for classification tasks.
  - **LDA, PLS-DA:** Supervised methods for classification that find latent variables maximising class separability.
- **Feature Selectors:**
  - **ElasticNetFS, LASSO, LogisticL1:** Embedded methods that perform selection via regularisation, where the LogisticL1 is used only in classification tasks.
  - **RFImportance:** Ensemble-based method using feature importance from a Random Forest.
  - **Variance F-Test, FRegressionFS:** Univariate filters based on statistical F-tests, where FRegressionFS is used only in regression tasks.

4.4.2 Intermediate Fusion Strategies

After feature engineering, one of eight fusion strategies was used to merge the modality-specific matrices. (Figure 4.2d)

- **Attention-weighted & Learnable-weighted fusion:** Adaptive methods that learn modality-specific weights. Used only if the missing data simulation was 0%, as these are concatenation techniques and can't be used with missing data.
- **Multiple Kernel Learning (MKL):** A non-linear, kernel-based method that learns an optimal combination of modality-specific kernels and is robust to missing data Gönen and Alpaydın (2011).
- **Simple Averaging, Summation, & Maximum:** Arithmetic baselines for rapid signal aggregation.
- **Standard Concatenation & Early-fusion PCA:** Concatenation-based approaches. Standard Concatenation was used only if the missing data simulation was 0%. Early-fusion PCA was used in every scenario, because of being specifically optimised.

4.4.3 Predictive Modeling

The final stage used a portfolio of models selected for their proven effectiveness in biomedical research (Breiman (2001); Tibshirani (1996); Cortes and Vapnik (1995); Zou and Hastie (2005)). (Figure 4.2e)

- **Classification Models:**
  - **Logistic Regression:** A linear, probabilistic baseline model.
  - **Random Forest Classifier:** A non-linear ensemble model robust to noise.
  - **Support Vector Classifier (SVC):** A kernel-based model effective in high-dimensional spaces.
- **Regression Models:**
  - **Linear Regression:** A fundamental linear baseline.
  - **ElasticNet:** A regularised linear model that manages collinearity.
  - **Random Forest Regressor:** A non-linear ensemble model for regression tasks.

4.5 Evaluation Framework and Statistical Analysis

A rigorous evaluation framework was established to ensure the reliability and reproducibility of the results. This framework was built on three core components: hyperparameter tuning, standardised evaluation metrics, and adaptive cross-validation.

4.5.1 Hyperparameter Tuning

Prior to the main experimental runs, hyperparameters for all extraction algorithms and predictive models were optimised for each dataset using Halving Grid Search and Bayesian optimisation. Based on the MCC and R^2 scores, the best-performing parameter sets were stored and reused for all subsequent experiments, ensuring that each algorithm operated at its optimal level while significantly reducing the overall computational burden.

4.5.2 Evaluation Metrics

Predictive performance was assessed using standard, robust metrics appropriate for the given task.

- **Classification Metric:** The primary metric was the **Matthews Correlation Coefficient (MCC)**, chosen for its reliability on imbalanced datasets. MCC measures the quality of classifications by considering all confusion matrix elements.

- **Regression Metric:** The **Coefficient of Determination (R^2)** was used as the primary metric. R^2 indicates the proportion of variance in the target explained by the model.

In addition to predictive accuracy, computational efficiency was evaluated by recording the model fitting and scoring times, serving as a key secondary criterion for assessing clinical feasibility.

4.5.3 Adaptive Cross-Validation Strategy

An adaptive cross-validation strategy was employed to ensure statistical robustness by automatically selecting the most appropriate CV method based on dataset characteristics. Split counts was implemented based on sample size, where small datasets (< 100 samples) have 2 splits, medium datasets (100-200 samples) have 3 splits and large datasets (> 200 samples) have 5 splits The system implements a decision tree that chooses exactly one CV strategy per whole dataset:

- **KFold:** Used for regression tasks without patient replicates.
- **StratifiedKFold:** Used for classification tasks when there are sufficient samples per class and no patient replicates are present. It maintains class proportions across folds.
- **GroupKFold:** Used for regression tasks with patient replicates to prevent data leakage by ensuring all measurements from a single patient remain in the same fold.
- **StratifiedGroupKFold:** Used for classification tasks with both viable stratification and patient replicates, combining both constraints to maintain class proportions while preventing patient-level data leakage.

For each selected CV strategy, the model is evaluated across all folds using comprehensive metrics. All reported metrics are represented by the mean and standard deviation of scores obtained across the folds of the single selected CV strategy, ensuring robust performance estimation while maintaining statistical validity.

5. Results

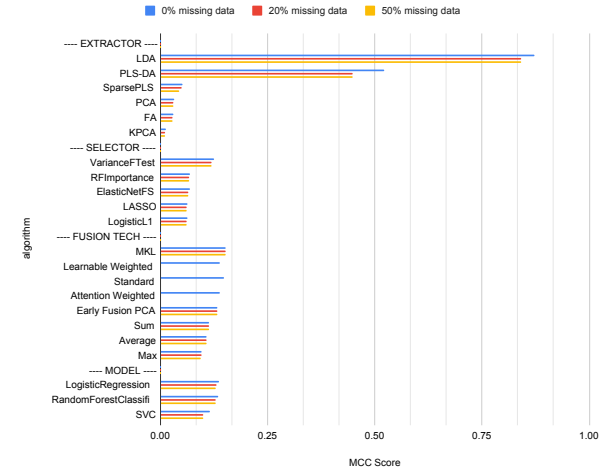
The comprehensive results are created from the systematic benchmarking of 10,206 unique multi-omics integration combinations across nine cancer datasets. The findings are analysed for both classification and regression tasks, focusing on predictive performance, robustness to missing data, and computational efficiency. They are first presented as aggregated results across all datasets to establish general principles, followed by a detailed dataset-specific analysis to highlight the nuances of each cancer type.

5.1 Overall Performance Trends

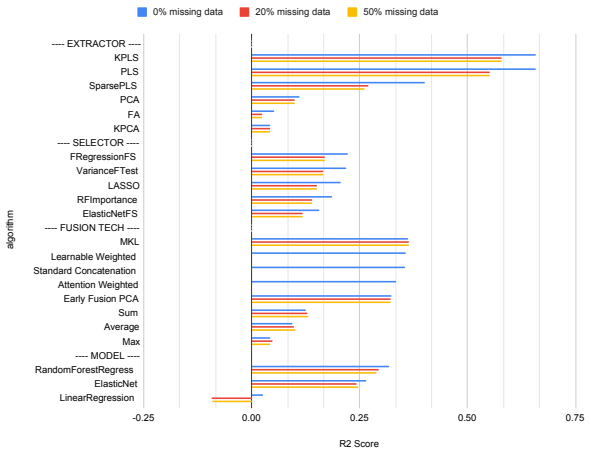
When averaged across all datasets, clear and consistent performance trends emerge for each component of the pipeline, as summarised in Figure 5.1.

5.1.1 Comparative Performance of Fusion Strategies

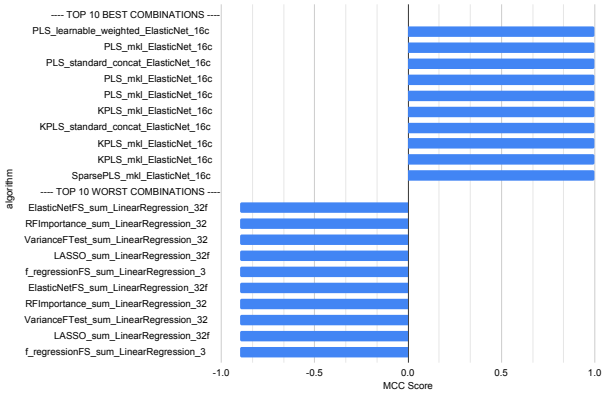
The choice of intermediate fusion strategy was found to be **a significant determinant of model performance**, based on MCC and R^2 scores. Across both classification and regression tasks, **adaptive fusion methods that learn to weight modalities consistently outperform simpler** arithmetic or concatenation-based approaches. As shown in Figures 5.1a and 5.1b, Multiple Kernel Learning (MKL) and Learnable Weighted fusion achieved the highest average MCC and R^2 scores. This indicates their superior ability to discern and prioritise the most informative data sources. In contrast, naive arithmetic methods such as Sum, Average, and Max fusion yielded significantly lower performance, highlighting the inadequacy of simple signal aggregation. The top-performing combinations, shown in Figures 5.1c and 5.1d, almost exclusively featured MKL, confirming that **sophisticated, learned integration is a key component** of a high-performing combination.



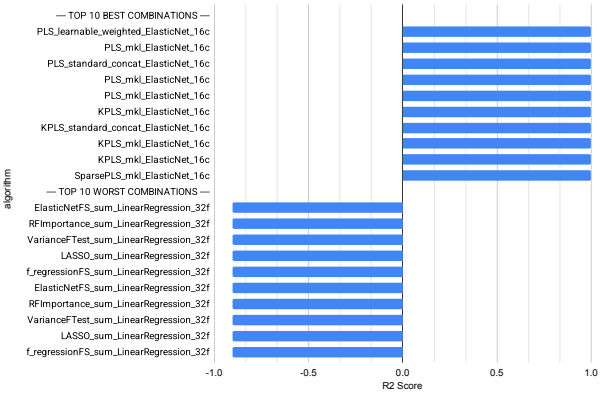
(a) Classification: MCC vs. missing data



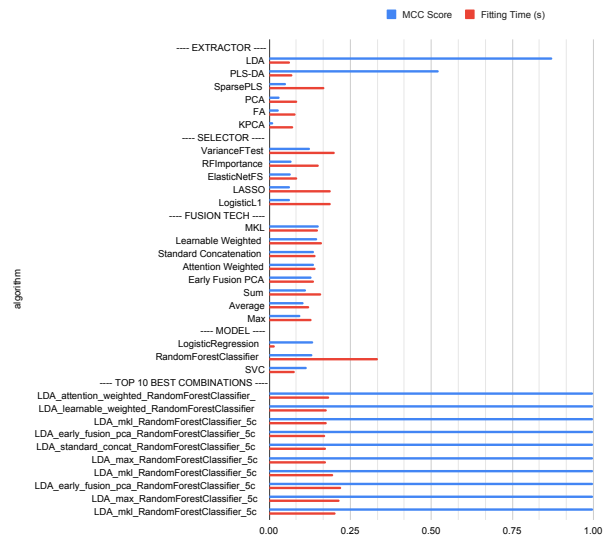
(b) Regression: R^2 vs. missing data



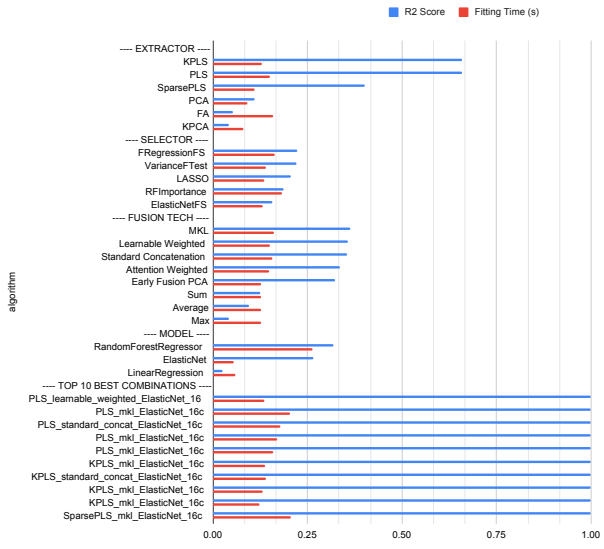
(c) Classification: best & worst combinations



(d) Regression: best & worst combinations



(e) Classification: MCC vs. fit time



(f) Regression: R^2 vs. fit time

Figure 5.1: Summary of classification and regression results. Each plot shows the average performance of individual components (Extractors, Selectors, Fusion Tech, Models) and the specific performance of the top-10 best and worst pipeline combinations. Combination's name legend: `Extractor/Selector_Fusion_technique_Model_Number_of_components/features`

5.1.2 The Impact of Feature Engineering on Predictive Accuracy

The results indicate the substantial **superiority of feature extraction methods over feature selection techniques**. The feature extraction algorithms achieved mean scores of 0.85 for the Matthews Correlation Coefficient (MCC) and 0.70 for the coefficient of determination (R^2). In contrast, feature selection algorithms only achieved average scores of 0.15 for MCC and 0.23 for R^2. (Figures 5.1a and 5.1b)

A critical determinant of predictive accuracy was the feature engineering strategy, particularly the distinction between supervised and unsupervised feature extraction. The findings reveal that **supervised methods**, which incorporate clinical outcome labels in the dimensionality reduction process, are crucial for **achieving high-performance results**.

For classification tasks, **Linear Discriminant Analysis (LDA) and Partial Least Squares Discriminant Analysis (PLS-DA)** emerged as the most effective feature extraction methods by a significant margin. (Figure 5.1a) Similarly, in the context of regression, **Partial Least Squares (PLS) and Kernel PLS (KPLS)** were the leading performers. (Figure 5.1b) Conversely, unsupervised feature extractors such as Principal Component Analysis (PCA), Kernel PCA (KPCA), and Factor Analysis (FA) consistently yielded suboptimal outcomes, with average MCC and R^2 scores frequently approaching or falling below zero.

5.1.3 Robustness to Missing Data

The systematic introduction of missing data revealed significant differences in the robustness of the various pipeline components. The **superiority of supervised feature extraction was maintained even under high levels of data incompleteness**. As illustrated in Figures 5.1a and 5.1b, extractors like LDA and PLS retained high predictive scores, with only a marginal decrease in performance as missingness increased from 0% to 50%. This demonstrates their ability to extract stable, clinically relevant latent features that are resilient to the loss of entire data modalities.

Among the fusion methods, **MKL exhibited exceptional robustness**, maintaining its high performance across all missing data scenarios. This is a critical finding, as MKL's kernel-based framework appears inherently tolerant of missing data blocks, making it **particularly well-suited for real-world clinical applications** where complete multi-omics panels are rare. In contrast, while simpler fusion methods also demonstrated consistent performance, their overall effectiveness was significantly lower (Figures 5.1a and 5.1b).

5.1.4 Computational Efficiency Analysis

A key finding of this study is that **high predictive performance does not require high computational cost**. The trade-off between accuracy and efficiency, visualised in Figures 5.1e and 5.1f, reveals that the **most accurate combinations were also highly efficient** in terms of fitting time. For classification, the top-performing combinations, typically comprising LDA, MKL, and a Random Forest Classifier, achieved near-perfect MCC scores with average model fitting times consistently under 0.2 seconds. In contrast, other combinations required longer training times, generally up to 3 seconds, while the slowest combination (SparsePLS with Sum Fusion and a Random Forest Classifier) took approximately 43 seconds to train. (Figure 5.1e) A similar trend was observed for regression, where the top-performing combinations (KPLS or PLS with MKL and an ElasticNet model) achieved near-perfect R^2 scores in 0.1–0.15 seconds, while less effective combinations took up to 0.9 seconds and yielded significantly lower predictive accuracy. (Figure 5.1f) This demonstrates that robust, high-performance multi-omics integration is **computationally feasible**, a finding that has **significant positive implications for the potential deployment** of these models in clinical decision support systems.

5.2 Dataset-specific Performance Trends

5.2.1 Regression Tasks: AML and Sarcoma Cancer Types

For both the **AML** and **Sarcoma** regression tasks, the results were unambiguous. The top-performing combinations, which achieved near-perfect R^2 scores approaching 1.0, were consistently composed of a supervised extractor (**KPLS** or **PLS**), an adaptive fusion method (**MKL** or **Learnable Weighted**), and a regularised or ensemble model (**ElasticNet** or **RandomForestRegressor**). This confirms that for these cancer types, tumour characteristics are highly predictable from multi-omics data when an optimised pipeline is employed. The detailed performance metrics for AML and Sarcoma are provided in Appendix B and C, respectively.

5.2.2 Classification Tasks: Breast, Colon, Kidney, Liver, Lung, Melanoma, and Ovarian Cancer Types

The seven classification tasks demonstrated a greater degree of variability, yet the core principles remained consistent.

- For **Breast**, **Kidney**, **Lung**, and **Ovarian** cancers, the results were clear. The optimal combinations, achieving the highest MCC scores, invariably combined the **LDA** extractor with **MKL** or **Learnable Weighted** fusion and a **Random Forest** or **SVC** model. For these datasets, the failure of unsupervised methods was absolute, with MCC scores of near zero.
- The **Colon** cancer dataset highlighted the outsized importance of the fusion strategy. While the LDA and PLS-DA extractors were still superior, **MKL** was the dominant fusion method by a remarkable margin, suggesting a particularly complex and heterogeneous interplay between the omics modalities in this cancer type.
- The **Liver** and **Melanoma** datasets represented the most challenging classification problems. Overall performance was lower, but this low-signal environment made the distinction between effective and ineffective methods even more stark. For these datasets, the combination of **LDA** and **MKL** was not merely the best strategy. It was the only strategy capable of producing a predictive signal greater than random chance, but still better than the other combinations.

This dataset-specific analysis confirms that while the ideal formula is a robust general guideline, the relative importance of its components can shift, and for challenging predictive tasks, the choice of an advanced fusion method like MKL becomes paramount. The detailed performance metrics for each cancer type are provided in Appendices D through J.

6. Discussion

The findings are interpreted in the context of the initial research question and the existing scientific literature, with a focus on both the general principles and the dataset-specific nuances that emerged from the analysis. It is concluded by addressing the methodological limitations of the study and its broader implications for clinical practice.

6.1 Interpretation of Key Findings

The results of this comprehensive benchmarking study deliver a message: **the combination of supervised feature extraction and adaptive intermediate fusion is the most effective strategy** for multi-omics cancer prediction. The performance gap between supervised extractors (e.g., LDA, PLS) and their unsupervised counterparts (e.g., PCA) underscores the **critical importance of leveraging clinical outcome labels** during the dimensionality reduction phase. Supervised methods are inherently guided to find and preserve latent features that are maximally correlated with the biological question

at hand, whereas unsupervised methods, which only consider the variance within the feature space, consistently fail to retain these vital, outcome-relevant signals. This was not a minor effect; for most cancer types studied, including Kidney, Liver, and Melanoma, unsupervised methods produced models with no predictive performance, where MCC and $R^2 \approx 0$.

Furthermore, the superiority of adaptive fusion methods like MKL highlights the **necessity of intelligently integrating data modalities**. These techniques can dynamically learn to up-weight informative modalities and down-weight noisy or less relevant ones. This capability is particularly crucial in heterogeneous diseases, as evidenced by the **outsized importance of MKL in the Colon cancer dataset**, where it was a near-mandatory component of any high-performing combination. This suggests that the interplay between omics modalities in colon cancer is especially complex, requiring a sophisticated non-linear fusion approach.

The dataset-specific results also reveal a spectrum of "predictability". For the AML and Sarcoma regression tasks, the optimal combinations achieved near-perfect prediction with $R^2 \approx 1.0$, indicating a strong, clear signal in the data. In contrast, for more challenging classification tasks like Liver and Melanoma, the best formula was not just about achieving the highest score, but about being the **only strategy capable of extracting any meaningful predictive signal** from the noise. This reinforces the robustness of the identified principles across a range of problem difficulties.

6.2 Comparison with Existing Literature

The findings of this thesis **both corroborate and extend the existing body of work** in multi-omics integration. The demonstrated strength of MKL aligns with the work of Gönen and Alpaydın (2011), who first detailed its power for integrating heterogeneous data sources. Similarly, the observation that adaptive fusion outperforms naive arithmetic methods confirms patterns reported in other benchmarking studies, as Cai et al. (2022) and Leng et al. (2022). The critical role of supervised feature extraction methods identified in this study is consistent with the conclusions of Cantini et al. (2021), who emphasised the need for label-informed dimensionality reduction in multi-omics cancer analysis.

This work also adds important context to the current field. While many recent studies focus on novel deep learning architectures for fusion (ex. Leng et al. (2022)), this thesis confirms that well-established machine learning methods like MKL and PLS, when combined in an optimised pipeline, can achieve **exceptional performance and robustness, often with lower computational demands and higher interpretability**. This provides a **crucial, high-quality baseline** against which the true added value of more complex models can be rigorously assessed. The finding that different cancer types benefit from slight variations in the optimal pipeline, such as the heightened importance of MKL for Colon cancer, also supports the broader conclusion from studies like Rappoport and Shamir (2018) that there is no single method in multi-omics analysis that fits all.

6.3 Limitations and Methodological Considerations

Despite the rigorous experimental design, several limitations must be acknowledged. First, the study relies on datasets from public repositories like TCGA, which, while invaluable, often have **modest sample sizes** relative to the high dimensionality of omics data. This increases the risk of overfitting, though this was mitigated through robust cross-validation and the use of regularised models.

Second, the missing data patterns were **simulated synthetically** by removing parts of modality blocks at random. While this approach allows for controlled assessment of robustness, it may not fully capture the complexity of real-world missingness, which can be non-random (e.g., related to patient subgroups or disease severity) or occur at the individual feature level.

Finally, the scope of this research was intentionally focused on establishing a strong benchmark for traditional and kernel-based machine learning methods. As such, the **exclusion of state-of-the-art deep learning-based fusion techniques** represents a key boundary of this work.

6.4 Broader Implications and Generalizability

The findings have significant implications that extend beyond the immediate clinical application in oncology. The demonstrated robustness and efficiency of the identified combinations suggest a path toward practical implementation, while the underlying principles of the framework are generalizable to other fields.

6.4.1 Generalizability within Biomedicine

While this study focused on nine cancer types, the principles and the framework itself are highly generalizable to other complex diseases where multi-omics data is generated. Conditions such as Alzheimer's disease, cardiovascular disease, and diabetes are increasingly studied using multi-modal data. The challenge of integrating genomics, proteomics, and clinical data to predict disease onset or progression in these areas is directly analogous to the problem solved in this thesis. The best formula identified here serves as a powerful starting hypothesis for researchers in those fields.

6.4.2 Applicability to Other Domains

Beyond medicine, the core challenge of multi-view data fusion is universal. The benchmarking framework presented in this thesis can be adapted to any domain where insights are derived from integrating heterogeneous data sources:

- **Finance:** Fusing economic indicators, market sentiment from text data, and company financial statements to predict asset performance.
- **E-commerce:** Combining user browsing history, demographic data, and purchase records to build robust recommendation engines or churn prediction models.
- **Climate Science:** Integrating satellite imagery, sensor data, and climate model outputs to improve weather forecasting.

In these contexts, the framework provides a rigorous methodology for determining the best way to combine different data types, while the core finding, that supervised feature extraction and adaptive fusion are critical, offers a valuable guiding principle for any multi-view learning problem. By providing a transparent, reproducible, and high-performing benchmark, this work lays a **strong foundation for the future development and validation of robust predictive tools** across a wide range of scientific and industrial domains.

7. Conclusion and Future Work

This thesis makes two principal contributions to the field of multi-omics integration. These contributions, along with directions for future research, are discussed in detail below.

7.1 Conclusion

First, it delivers a **large-scale, systematic, and reproducible framework with a foundational, high-quality performance baseline** for benchmarking intermediate fusion combinations. By evaluating 10,206 model configurations across nine cancer datasets and three missing data scenarios, this work addresses critical gaps in the existing literature, which has often overlooked the comparative performance of intermediate fusion strategies and their robustness to incomplete data (Rappoport and Shamir (2018); Duan et al. (2021)). This framework provides the empirical evidence needed to guide the development of more effective predictive models and enables a **rigorous, fair comparison for future innovations** in the field.

Second, by deploying the proposed framework, this study identifies an optimal pipeline architecture among the combinations tested. The evidence overwhelmingly confirms that this architecture consists of **supervised feature extraction combined with adaptive intermediate fusion**. This optimal formula was validated across both regression and classification tasks and on datasets with varying degrees of predictive difficulty. For highly predictable tasks like AML and Sarcoma, this approach yielded near-perfect results ($R^2 \approx 1.0$), while for more challenging tasks such as Liver and Melanoma, this combination was the only one capable of extracting a meaningful predictive signal from the data. Furthermore, the results highlight that while the general formula is consistent, the relative importance of its components can vary. For instance, the non-linear fusion capabilities of **MKL were particularly crucial for the Colon cancer dataset**.

Critically, this work demonstrates that the most accurate and robust combinations are also **computationally efficient** compared to other Machine Learning techniques, with sub-second fitting times. This combination of **high accuracy, robustness to incomplete data, and efficiency** makes these combinations prime candidates for clinical translation and establishes a new, high-quality performance baseline for the field.

7.2 Future Research Directions

Building on the foundation established by this work, several promising avenues for future research emerge.

- **Benchmarking Against Deep Learning Models:** A logical next step is to extend the developed benchmarking framework to include state-of-the-art deep learning and graph-based fusion strategies (Leng et al. (2022); Yang et al. (2024)). This would allow for a direct and fair comparison of their performance, robustness, and computational cost against the strong traditional baselines established here.
- **Advanced Feature Selection:** Future work should incorporate more advanced feature selection methods. An important candidate is the **Minimum Redundancy Maximum Relevance (MRMR)** selector, which was not tested here due to its higher computational cost, it is well-regarded in bioinformatics for its ability to select a compact set of features that are highly correlated with the clinical outcome while being minimally correlated with each other. This could further enhance model performance and interpretability.
- **External and Clinical Validation:** The top-performing combinations identified in this study should be rigorously validated on independent, external clinical cohorts. This is a critical step to assess their generalizability and confirm their practical applicability beyond the TCGA datasets used in this research.
- **Advanced Missing Data Scenarios:** Future investigations should explore more complex and realistic missing data patterns. Moving beyond the synthetic, random removal used in this study to model feature-level missingness and structured, non-random patterns of data absence would better reflect the challenges of real-world clinical practice.
- **Expansion to Additional Modalities:** The modular architecture of the pipeline is designed for extension. Future work could incorporate additional, increasingly prevalent data types in cancer research, such as proteomics, metabolomics, and digital pathology images (histomics), to build even more comprehensive and powerful predictive models.

References

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Cai, Z., Poulos, R., Liu, J. and Zhong, Q. (2022), 'Machine learning for multi-omics data integration in cancer', *iScience* **25**(2), 103798.

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E. and Baudot, A. (2021), 'Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer', *Nature Communications* **12**(1), 104.

Cortes, C. and Vapnik, V. (1995), 'Support-vector networks', *Machine Learning* **20**(3), 273–297.

Duan, R., Gao, L., Gao, Y., Hu, Y., Xu, H., Huang, M., Song, K., Wang, H., Dong, Y., Jiang, C., Zhang, C. and Jia, S. (2021), 'Evaluation and comparison of multi-omics data integration methods for cancer subtyping', *PLoS Computational Biology* **17**(8), e1009224.

Geurts, P., Ernst, D. and Wehenkel, L. (2006), 'Extremely randomized trees', *Machine Learning* **63**(1), 3–42.

Gönen, M. and Alpaydın, E. (2011), 'Multiple kernel learning algorithms', *Journal of Machine Learning Research* **12**, 2211–2268.
URL: *http://jmlr.org/papers/v12/gonen11a.html*

Hernández-Lemus, E. and Ochoa, S. (2024), 'Methods for multi-omic data integration in cancer research', *Frontiers in Genetics* **15**, 1425456.

Leng, D., Zheng, L., Wen, Y., Zhang, Y., Wu, L., Wang, J., Wang, M., Zhang, Z., He, S. and Bo, X. (2022), 'A benchmark study of deep learning-based multi-omics data fusion methods for cancer', *Genome biology* **23**(1), 1–23.

Nowak, K. P. (2025), 'Multi-omics data integration for cancer prediction: A comparative analysis of intermediate fusion techniques', https://github.com/kpnowak/OUH-Internship-Krzysztof-Nowak. Bachelor's thesis, VU Amsterdam.

Ous-research.no (2024), 'Ouh - computational systems medicine', https://ous-research.no/aittokallio/. [Accessed 3 Jul. 2025].

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O. and Droit, A. (2021), 'Integration strategies of multi-omics data for machine learning analysis', *Computational and Structural Biotechnology Journal* **19**, 3735–3746.

Rappoport, N. and Shamir, R. (2018), 'Multi-omic and multi-view clustering algorithms: review and cancer benchmark', *Nucleic Acids Research* **46**(20), 10546–10562.

Sammut, S.-J., Crispin-Ortuzar, M., Chin, S.-F., Provenzano, E., Bardwell, H. A., Ma, W., Cope, W., Dariush, A., Dawson, S.-J., Abraham, J. E., Dunn, J., Hiller, L., Thomas, J., Cameron, D. A., Bartlett, J. M. S., Hayward, L., Pharoah, P. D., Markowetz, F., Rueda, O. M., Earl, H. M. and Caldas, C. (2022), 'Multi-omic machine learning predictor of breast cancer therapy response', *Nature* **601**(7894), 623–629.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B (Methodological)* **58**(1), 267–288.

Yang, H., Yang, M., Chen, J., Yao, G., Zou, Q. and Jia, L. (2024), 'Multimodal deep learning approaches for precision oncology: a comprehensive review', *Briefings in Bioinformatics* **26**(1), bbae699.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**(2), 301–320.

A. Appendix - Data Modalities

Dataset	Modality	Samples	Features	Missing (%)	Combined missing (%)
AML	Gene Expression	172	20531	20.19	16.97
	miRNA Expression	187	705	54.83	
	DNA Methylation	193	5000	0.00	
Sarcoma	Gene Expression	264	20531	15.78	14.39
	miRNA Expression	262	1046	57.30	
	DNA Methylation	268	5000	0.00	
Breast	Gene Expression	1211	20531	14.28	13.41
	miRNA Expression	848	1046	55.86	
	DNA Methylation	884	5000	0.00	
Colon	Gene Expression	327	20531	15.30	12.80
	miRNA Expression	220	705	42.44	
	DNA Methylation	334	5000	0.00	
Kidney	Gene Expression	605	20531	14.04	12.84
	miRNA Expression	325	1046	59.43	
	DNA Methylation	479	5000	0.00	
Liver	Gene Expression	423	20531	17.88	15.93
	miRNA Expression	423	1046	54.65	
	DNA Methylation	428	5000	0.00	
Lung	Gene Expression	551	20531	13.07	12.26
	miRNA Expression	386	1046	52.03	
	DNA Methylation	411	5000	0.00	
Melanoma	Gene Expression	472	20531	16.21	14.45
	miRNA Expression	451	1046	50.98	
	DNA Methylation	474	5000	0.00	
Ovarian	Gene Expression	306	20531	12.98	9.97
	miRNA Expression	460	705	45.67	
	DNA Methylation	611	5000	0.00	

Table A.1: Overview of sample, feature and missingness statistics for every omics modality for every dataset.

B. Appendix - AML Results

Type	Ranking	Algorithm	Missing Data (%)	R2 Score	Standard Deviation	Fitting Time (s)
			0%	0.608	0.345	0.131
			20%	0.471	0.360	0.129
	1	KPLS	50%	0.471	0.360	0.129
			0%	0.614	0.332	0.151
			20%	0.461	0.327	0.172
	2	PLS	50%	0.461	0.327	0.174
			0%	0.426	0.353	0.116
			20%	0.350	0.324	0.089
	3	SparsePLS	50%	0.263	0.356	0.082
			0%	0.190	0.157	0.159
			20%	0.189	0.206	0.153
	4	FA	50%	0.189	0.206	0.166
			0%	0.198	0.170	0.088
			20%	0.176	0.153	0.057
	5	PCA	50%	0.176	0.153	0.057
			0%	0.115	0.192	0.081
			20%	0.110	0.213	0.058
Extractor	6	KPCA	50%	0.110	0.213	0.070
			0%	0.304	0.338	0.171
			20%	0.222	0.381	0.167
	1	FRegressionFS	50%	0.222	0.381	0.164
			0%	0.300	0.337	0.147
			20%	0.218	0.379	0.126
	2	VarianceFTest	50%	0.218	0.379	0.127
			0%	0.278	0.331	0.139
			20%	0.195	0.369	0.143
	3	LASSO	50%	0.195	0.369	0.143
			0%	0.253	0.340	0.187
			20%	0.182	0.376	0.160
	4	RFImportance	50%	0.182	0.376	0.150
			0%	0.208	0.333	0.137
			20%	0.148	0.361	0.119
Selector	5	ElasticNetFS	50%	0.148	0.361	0.116
			0%	0.495	0.222	0.173
			20%	0.497	0.223	0.164
	1	MKL	50%	0.497	0.224	0.164
	2	Standard Concatenation	0%	0.485	0.217	0.167
	3	Learnable Weighted	0%	0.457	0.249	0.157
	4	Attention Weighted	0%	0.444	0.252	0.155
			0%	0.396	0.259	0.128
			20%	0.396	0.270	0.124
	5	Early Fusion PCA	50%	0.396	0.270	0.123
			0%	0.101	0.381	0.128
			20%	0.113	0.386	0.125
	6	Sum	50%	0.108	0.384	0.123
			0%	0.094	0.231	0.133
			20%	0.107	0.234	0.129
	7	Max	50%	0.094	0.234	0.130
			0%	0.062	0.367	0.129
			20%	0.075	0.374	0.125
Fusion Technique	8	Average	50%	0.070	0.371	0.124
			0%	0.440	0.140	0.270
		Random Forest	20%	0.400	0.120	0.260
	1	Regressor	50%	-0.030	0.490	0.060
			0%	0.340	0.260	0.060
			20%	0.320	0.230	0.050
	2	ElasticNet	50%	-0.150	0.440	0.060
			0%	-0.030	0.490	0.060
			20%	-0.150	0.440	0.060
Model	3	Linear Regression	50%	-0.150	0.440	0.060
	1	PLS_learnable_weighted _ElasticNet_16c	0%	1.000	0.000	0.136
	2	PLS_mkl_ElasticNet_16c	0%	0.998	0.000	0.208
	3	PLS_standard_concat_El asticNet_16c	0%	0.998	0.000	0.204
	4	PLS_mkl_ElasticNet_16c	20%	0.998	0.000	0.175
	5	PLS_mkl_ElasticNet_16c	50%	0.998	0.000	0.184
	6	SparsePLS_mkl_ElasticN et_16c	50%	0.996	0.000	0.184
	7	SparsePLS_mkl_ElasticN et_16c	20%	0.996	0.000	0.210
	8	PLS_early_fusion_pca_El asticNet_16c	0%	0.994	0.000	0.039
	9	PLS_early_fusion_pca_El asticNet_16c	20%	0.994	0.000	0.036
Best Combinations	10	PLS_early_fusion_pca_El asticNet_16c	50%	0.994	0.000	0.040
	1	ElasticNetFS_sum_Linea rRegression_32f	0%	-0.901	0.480	0.110
	2	RFImportance_sum_Line arRegression_32f	0%	-0.901	0.480	0.130
	3	VarianceFTest_sum_Line arRegression_32f	0%	-0.901	0.480	0.100
	4	LASSO_sum_LinearRegr ession_32f	0%	-0.901	0.480	0.110
	5	f_regressionFS_sum_Lin earRegression_32f	0%	-0.901	0.480	0.100
	6	ElasticNetFS_sum_Linea rRegression_32f	20%	-0.901	0.480	0.100
	7	RFImportance_sum_Line arRegression_32f	20%	-0.901	0.480	0.100
	8	VarianceFTest_sum_Line arRegression_32f	20%	-0.901	0.480	0.110
	9	LASSO_sum_LinearRegr ession_32f	20%	-0.901	0.480	0.100
Worst Combinations	10	f_regressionFS_sum_Lin earRegression_32f	20%	-0.901	0.480	0.100

Figure B.1: Overview of the results from the AML dataset. The table shows the average performance of individual components (Extractors, Selectors, Fusion Tech, Models) and the specific performance of the 10 best and worst pipeline combinations.

C. Appendix - Sarcoma Results

Type	Ranking	Algorithm	Missing Data (%)	R2 Score	Standard Deviation	Fitting Time (s)
			0%	0.710	0.190	0.130
			20%	0.690	0.190	0.120
	1	KPLS	50%	0.690	0.190	0.120
			0%	0.700	0.200	0.150
			20%	0.640	0.210	0.160
	2	PLS	50%	0.640	0.210	0.160
			0%	0.380	0.350	0.100
			20%	0.190	0.310	0.080
	3	SparsePLS	50%	0.260	0.280	0.080
			0%	0.020	0.050	0.090
			20%	0.020	0.040	0.050
	4	PCA	50%	0.020	0.040	0.050
			0%	-0.030	0.030	0.080
			20%	-0.020	0.040	0.060
	5	KPCA	50%	-0.020	0.040	0.060
			0%	-0.090	0.250	0.160
			20%	-0.150	0.310	0.150
Extractor	6	FA	50%	-0.150	0.310	0.140
			0%	0.130	0.160	0.150
			20%	0.110	0.190	0.150
	1	FRegressionFS	50%	0.110	0.190	0.150
			0%	0.130	0.160	0.130
			20%	0.110	0.190	0.120
	2	VarianceFTest	50%	0.110	0.190	0.120
			0%	0.130	0.160	0.130
			20%	0.100	0.190	0.140
	3	LASSO	50%	0.100	0.190	0.140
			0%	0.110	0.170	0.180
			20%	0.090	0.200	0.140
	4	RFImportance	50%	0.090	0.200	0.140
			0%	0.100	0.170	0.130
			20%	0.090	0.200	0.110
Selector	5	ElasticNetFS	50%	0.090	0.200	0.110
	1	Learnable Weighted	0%	0.260	0.260	0.140
			0%	0.240	0.230	0.130
			20%	0.240	0.230	0.120
	2	Early Fusion PCA	50%	0.240	0.230	0.120
			0%	0.230	0.290	0.150
			20%	0.230	0.280	0.140
	3	MKL	50%	0.230	0.290	0.140
	4	Attention Weighted	0%	0.230	0.260	0.140
	5	Standard Concatenation	0%	0.220	0.280	0.150
			0%	0.160	0.250	0.130
			20%	0.150	0.250	0.120
	6	Sum	50%	0.160	0.250	0.120
			0%	0.130	0.260	0.120
			20%	0.130	0.260	0.120
	7	Average	50%	0.140	0.260	0.120
			0%	-0.010	0.300	0.120
			20%	-0.010	0.300	0.120
Fusion Technique	8	Max	50%	-0.010	0.300	0.120
			0%	0.200	0.140	0.250
			20%	0.190	0.140	0.240
	1	Random Forest Regressor	50%	0.190	0.140	0.240
			0%	0.190	0.280	0.050
			20%	0.170	0.250	0.040
	2	ElasticNet	50%	0.180	0.250	0.040
			0%	0.100	0.470	0.050
			20%	-0.010	0.460	0.040
Model	3	Linear Regression	50%	-0.010	0.460	0.040
	1	PLS_mkl_ElasticNet_16c	0%	1.000	0.000	0.200
	2	PLS_standard_concat_ElasticNet_16c	0%	1.000	0.000	0.180
	3	PLS_mkl_ElasticNet_16c	20%	1.000	0.000	0.170
	4	PLS_mkl_ElasticNet_16c	50%	1.000	0.000	0.160
	5	KPLS_mkl_ElasticNet_16c	0%	1.000	0.000	0.140
	6	KPLS_standard_concat_ElasticNet_16c	0%	1.000	0.000	0.140
	7	KPLS_mkl_ElasticNet_16c	20%	1.000	0.000	0.130
	8	KPLS_mkl_ElasticNet_16c	50%	1.000	0.000	0.120
	9	SparsePLS_mkl_ElasticNet_16c	50%	1.000	0.000	0.210
Best Combinations	10	SparsePLS_mkl_ElasticNet_16c	20%	1.000	0.000	0.200
	1	ElasticNetFS_max_LinearRegression_8f	0%	-0.790	0.730	0.080
	2	RFImportance_max_LinearRegression_8f	0%	-0.790	0.730	0.080
	3	VarianceFTest_max_LinearRegression_8f	0%	-0.790	0.730	0.080
	4	LASSO_max_LinearRegression_8f	0%	-0.790	0.730	0.080
	5	f_regressionFS_max_LinearRegression_8f	0%	-0.790	0.730	0.070
	6	ElasticNetFS_max_LinearRegression_8f	20%	-0.790	0.730	0.070
	7	RFImportance_max_LinearRegression_8f	20%	-0.790	0.730	0.080
	8	VarianceFTest_max_LinearRegression_8f	20%	-0.790	0.730	0.080
	9	LASSO_max_LinearRegression_8f	20%	-0.790	0.730	0.080
Worst Combinations	10	f_regressionFS_max_LinearRegression_8f	20%	-0.790	0.730	0.080

Figure C.1: Overview of the results from the Sarcoma dataset. The table shows the average performance of individual components (Extractors, Selectors, Fusion Tech, Models) and the specific performance of the 10 best and worst pipeline combinations.

D. Appendix - Breast Results

Type	Ranking	Algorithm	Missing Data (%)	MCC Score	Standard Deviation	Fitting Time (s)
			0%	0.877	0.141	0.070
			20%	0.877	0.141	0.065
	1	LDA	50%	0.877	0.141	0.066
			0%	0.479	0.165	0.073
			20%	0.479	0.165	0.069
	2	PLS-DA	50%	0.479	0.165	0.069
			0%	0.032	0.028	0.081
			20%	0.032	0.028	0.078
	3	FA	50%	0.032	0.028	0.079
			0%	0.021	0.050	0.108
			20%	0.021	0.050	0.105
	4	SparsePLS	50%	0.021	0.050	0.106
			0%	0.017	0.028	0.094
			20%	0.017	0.028	0.091
	5	PCA	50%	0.017	0.028	0.092
			0%	0.007	0.024	0.084
			20%	0.007	0.024	0.080
Extractor	6	KPCA	50%	0.007	0.024	0.081
			0%	0.127	0.029	0.155
			20%	0.127	0.029	0.141
	1	VarianceFTest	50%	0.127	0.029	0.142
			0%	0.070	0.027	0.105
			20%	0.070	0.027	0.103
	2	RFImportance	50%	0.070	0.027	0.104
			0%	0.062	0.029	0.112
			20%	0.062	0.029	0.110
	3	LogisticL1	50%	0.062	0.029	0.111
			0%	0.062	0.029	0.127
			20%	0.062	0.029	0.122
	4	LASSO	50%	0.062	0.029	0.123
			0%	0.056	0.027	0.054
			20%	0.056	0.027	0.056
Selector	5	ElasticNetFS	50%	0.056	0.027	0.057
			0%	0.153	0.227	0.102
			20%	0.153	0.227	0.099
	1	MKL	50%	0.153	0.227	0.100
	2	Standard Concatenation	0%	0.152	0.226	0.107
	3	Learnable Weighted	0%	0.152	0.228	0.095
	4	Attention Weighted	0%	0.147	0.228	0.098
			0%	0.145	0.201	0.095
			20%	0.145	0.201	0.092
	5	Early Fusion PCA	50%	0.145	0.201	0.093
			0%	0.110	0.191	0.105
			20%	0.110	0.191	0.102
	6	Sum	50%	0.110	0.191	0.103
			0%	0.107	0.192	0.083
			20%	0.107	0.192	0.082
	7	Average	50%	0.107	0.192	0.082
			0%	0.100	0.174	0.090
			20%	0.100	0.174	0.088
Fusion Technique	8	Max	50%	0.100	0.174	0.088
			0%	0.152	0.226	0.212
			20%	0.152	0.226	0.208
	1	RandomForestClassifier	50%	0.152	0.226	0.210
			0%	0.140	0.215	0.009
			20%	0.140	0.215	0.008
	2	LogisticRegression	50%	0.140	0.215	0.009
			0%	0.107	0.172	0.066
			20%	0.107	0.172	0.065
Model	3	SVC	50%	0.107	0.172	0.066
	1	LDA_early_fusion_pca_RandomForestClassifier	0%	0.996	0.005	0.090
	2	LDA_early_fusion_pca_RandomForestClassifier	20%	0.996	0.005	0.088
	3	LDA_early_fusion_pca_RandomForestClassifier	50%	0.996	0.005	0.089
	4	LDA_learnable_weighted_RandomForestClassifier	0%	0.996	0.005	0.096
	5	LDA_mkl_RandomForestClassifier	0%	0.996	0.005	0.097
	6	LDA_standard_concat_RandomForestClassifier	0%	0.996	0.005	0.124
	7	LDA_mkl_RandomForestClassifier	20%	0.996	0.005	0.097
	8	LDA_mkl_RandomForestClassifier	50%	0.996	0.005	0.098
	9	LDA_attention_weighted_RandomForestClassifier	0%	0.993	0.007	0.097
Best Combinations	10	LDA_early_fusion_pca_LogisticRegression	0%	0.912	0.020	0.004
	1	KPCA_average_SVC	0%	-0.019	0.029	0.024
	2	FA_sum_SVC	0%	-0.018	0.020	0.042
	3	FA_sum_SVC	20%	-0.018	0.020	0.042
	4	FA_sum_SVC	50%	-0.018	0.020	0.042
	5	ElasticNetFS_sum_SVC	0%	-0.014	0.018	0.027
	6	RFImportance_sum_SVC	0%	-0.014	0.018	0.028
	7	LASSO_sum_SVC	0%	-0.014	0.018	0.040
	8	LogisticL1_sum_SVC	0%	-0.014	0.018	0.030
	9	ElasticNetFS_sum_SVC	20%	-0.014	0.018	0.028
Worst Combinations	10	RFImportance_sum_SVC	20%	-0.014	0.018	0.029

Figure D.1: Overview of the results from the Breast dataset. The table shows the average performance of individual components (Extractors, Selectors, Fusion Tech, Models) and the specific performance of the 10 best and worst pipeline combinations.

E. Appendix - Colon Results

Type	Ranking	Algorithm	Missing Data (%)	MCC Score	Standard Deviation	Fitting Time (s)
			0%	0.788	0.184	0.074
			20%	0.788	0.184	0.071
	1	LDA	50%	0.788	0.184	0.071
			0%	0.427	0.165	0.078
			20%	0.427	0.165	0.075
	2	PLS-DA	50%	0.427	0.165	0.076
			0%	0.027	0.025	0.083
			20%	0.027	0.025	0.081
	3	FA	50%	0.027	0.025	0.082
			0%	0.022	0.048	0.113
			20%	0.022	0.048	0.111
	4	SparsePLS	50%	0.022	0.048	0.111
			0%	0.021	0.025	0.096
			20%	0.021	0.025	0.093
	5	PCA	50%	0.021	0.025	0.094
			0%	0.008	0.024	0.085
			20%	0.008	0.024	0.083
Extractor	6	KPCA	50%	0.008	0.024	0.083
			0%	0.116	0.024	0.163
			20%	0.116	0.024	0.155
	1	VarianceFTest	50%	0.116	0.024	0.156
			0%	0.066	0.024	0.111
			20%	0.066	0.024	0.109
	2	RFImportance	50%	0.066	0.024	0.109
			0%	0.057	0.028	0.119
			20%	0.057	0.028	0.117
	3	LogisticL1	50%	0.057	0.028	0.117
			0%	0.057	0.028	0.135
			20%	0.057	0.028	0.132
	4	LASSO	50%	0.057	0.028	0.133
			0%	0.052	0.025	0.056
			20%	0.052	0.025	0.058
Selector	5	ElasticNetFS	50%	0.052	0.025	0.059
			0%	0.143	0.223	0.107
			20%	0.143	0.223	0.105
	1	MKL	50%	0.143	0.223	0.105
	2	Standard Concatenation	0%	0.139	0.218	0.113
	3	Learnable Weighted	0%	0.139	0.225	0.100
	4	Attention Weighted	0%	0.135	0.225	0.104
			0%	0.132	0.189	0.100
			20%	0.132	0.189	0.098
	5	Early Fusion PCA	50%	0.132	0.189	0.098
			0%	0.103	0.185	0.110
			20%	0.103	0.185	0.108
	6	Sum	50%	0.103	0.185	0.109
			0%	0.100	0.186	0.088
			20%	0.100	0.186	0.086
	7	Average	50%	0.100	0.186	0.087
			0%	0.091	0.169	0.095
			20%	0.091	0.169	0.093
Fusion Technique	8	Max	50%	0.091	0.169	0.093
			0%	0.138	0.213	0.225
			20%	0.138	0.213	0.220
	1	RandomForestClassifier	50%	0.138	0.213	0.222
			0%	0.131	0.209	0.009
			20%	0.131	0.209	0.009
	2	LogisticRegression	50%	0.131	0.209	0.009
			0%	0.099	0.167	0.069
			20%	0.099	0.167	0.068
Model	3	SVC	50%	0.099	0.167	0.068
	1	LDA_early_fusion_pca_RandomForestClassifier	0%	0.995	0.007	0.090
	2	LDA_early_fusion_pca_RandomForestClassifier	20%	0.995	0.007	0.089
	3	LDA_early_fusion_pca_RandomForestClassifier	50%	0.995	0.007	0.089
	4	LDA_learnable_weighted_RandomForestClassifier	0%	0.995	0.007	0.097
	5	LDA_mkl_RandomForestClassifier	0%	0.995	0.007	0.098
	6	LDA_standard_concat_RandomForestClassifier	0%	0.995	0.007	0.125
	7	LDA_mkl_RandomForestClassifier	20%	0.995	0.007	0.098
	8	LDA_mkl_RandomForestClassifier	50%	0.995	0.007	0.099
	9	LDA_attention_weighted_RandomForestClassifier	0%	0.992	0.007	0.098
Best Combinations	10	LDA_early_fusion_pca_LogisticRegression	0%	0.906	0.020	0.004
	1	KPCA_average_SVC	0%	-0.018	0.028	0.025
	2	FA_sum_SVC	0%	-0.017	0.020	0.042
	3	FA_sum_SVC	20%	-0.017	0.020	0.043
	4	FA_sum_SVC	50%	-0.017	0.020	0.043
	5	ElasticNetFS_sum_SVC	0%	-0.013	0.017	0.028
	6	RFImportance_sum_SVC	0%	-0.013	0.017	0.029
	7	LASSO_sum_SVC	0%	-0.013	0.017	0.040
	8	LogisticL1_sum_SVC	0%	-0.013	0.017	0.030
	9	ElasticNetFS_sum_SVC	20%	-0.013	0.017	0.028
Worst Combinations	10	RFImportance_sum_SVC	20%	-0.013	0.017	0.029

Figure E.1: Overview of the results from the Colon dataset. The table shows the average performance of individual components (Extractors, Selectors, Fusion Tech, Models) and the specific performance of the 10 best and worst pipeline combinations.

F. Appendix - Kidney Results

Type	Ranking	Algorithm	Missing Data (%)	MCC Score	Standard Deviation	Fitting Time (s)
			0%	0.859	0.154	0.074
			20%	0.859	0.154	0.070
	1	LDA	50%	0.859	0.154	0.071
			0%	0.488	0.161	0.077
			20%	0.488	0.161	0.074
	2	PLS-DA	50%	0.488	0.161	0.075
			0%	0.032	0.027	0.082
			20%	0.032	0.027	0.080
	3	FA	50%	0.032	0.027	0.080
			0%	0.025	0.049	0.112
			20%	0.025	0.049	0.109
	4	SparsePLS	50%	0.025	0.049	0.110
			0%	0.024	0.027	0.095
			20%	0.024	0.027	0.092
	5	PCA	50%	0.024	0.027	0.092
			0%	0.008	0.023	0.084
			20%	0.008	0.023	0.082
Extractor	6	KPCA	50%	0.008	0.023	0.082
			0%	0.126	0.029	0.163
			20%	0.126	0.029	0.156
	1	VarianceFTest	50%	0.126	0.029	0.157
			0%	0.068	0.027	0.112
			20%	0.068	0.027	0.109
	2	RFImportance	50%	0.068	0.027	0.110
			0%	0.061	0.028	0.120
			20%	0.061	0.028	0.117
	3	LogisticL1	50%	0.061	0.028	0.118
			0%	0.061	0.028	0.136
			20%	0.061	0.028	0.132
	4	LASSO	50%	0.061	0.028	0.134
			0%	0.055	0.027	0.057
			20%	0.055	0.027	0.059
Selector	5	ElasticNetFS	50%	0.055	0.027	0.059
			0%	0.148	0.225	0.109
			20%	0.148	0.225	0.106
	1	MKL	50%	0.148	0.225	0.107
	2	Standard Concatenation	0%	0.144	0.219	0.114
	3	Learnable Weighted	0%	0.142	0.226	0.102
	4	Attention Weighted	0%	0.138	0.226	0.105
			0%	0.136	0.191	0.102
			20%	0.136	0.191	0.099
	5	Early Fusion PCA	50%	0.136	0.191	0.100
			0%	0.108	0.187	0.112
			20%	0.108	0.187	0.109
	6	Sum	50%	0.108	0.187	0.110
			0%	0.105	0.187	0.089
			20%	0.105	0.187	0.087
	7	Average	50%	0.105	0.187	0.088
			0%	0.097	0.170	0.096
			20%	0.097	0.170	0.094
Fusion Technique	8	Max	50%	0.097	0.170	0.094
			0%	0.145	0.215	0.229
			20%	0.145	0.215	0.224
	1	RandomForestClassifier	50%	0.145	0.215	0.225
			0%	0.138	0.211	0.009
			20%	0.138	0.211	0.009
	2	LogisticRegression	50%	0.138	0.211	0.009
			0%	0.102	0.169	0.069
			20%	0.102	0.169	0.069
Model	3	SVC	50%	0.102	0.169	0.069
	1	LDA_early_fusion_pca_RandomForestClassifier	0%	0.996	0.005	0.091
	2	LDA_early_fusion_pca_RandomForestClassifier	20%	0.996	0.005	0.089
	3	LDA_early_fusion_pca_RandomForestClassifier	50%	0.996	0.005	0.090
	4	LDA_learnable_weighted_RandomForestClassifier	0%	0.996	0.005	0.098
	5	LDA_mkl_RandomForestClassifier	0%	0.996	0.005	0.098
	6	LDA_standard_concat_RandomForestClassifier	0%	0.996	0.005	0.126
	7	LDA_mkl_RandomForestClassifier	20%	0.996	0.005	0.099
	8	LDA_mkl_RandomForestClassifier	50%	0.996	0.005	0.100
	9	LDA_attention_weighted_RandomForestClassifier	0%	0.993	0.007	0.099
Best Combinations	10	LDA_early_fusion_pca_LogisticRegression	0%	0.909	0.020	0.004
	1	KPCA_average_SVC	0%	-0.019	0.028	0.025
	2	FA_sum_SVC	0%	-0.017	0.020	0.043
	3	FA_sum_SVC	20%	-0.017	0.020	0.043
	4	FA_sum_SVC	50%	-0.017	0.020	0.043
	5	ElasticNetFS_sum_SVC	0%	-0.014	0.017	0.028
	6	RFImportance_sum_SVC	0%	-0.014	0.017	0.029
	7	LASSO_sum_SVC	0%	-0.014	0.017	0.041
	8	LogisticL1_sum_SVC	0%	-0.014	0.017	0.030
	9	ElasticNetFS_sum_SVC	20%	-0.014	0.017	0.029
Worst Combinations	10	RFImportance_sum_SVC	20%	-0.014	0.017	0.030

Figure F.1: Overview of the results from the Kidney dataset. The table shows the average performance of individual components (Extractors, Selectors, Fusion Tech, Models) and the specific performance of the 10 best and worst pipeline combinations.

G. Appendix - Liver Results

Type	Ranking	Algorithm	Missing Data (%)	MCC Score	Standard Deviation	Fitting Time (s)
			0%	0.845	0.163	0.072
			20%	0.845	0.163	0.069
	1	LDA	50%	0.845	0.163	0.069
			0%	0.466	0.166	0.075
			20%	0.466	0.166	0.073
	2	PLS-DA	50%	0.466	0.166	0.073
			0%	0.030	0.027	0.080
			20%	0.030	0.027	0.078
	3	FA	50%	0.030	0.027	0.079
			0%	0.023	0.048	0.110
			20%	0.023	0.048	0.108
	4	SparsePLS	50%	0.023	0.048	0.108
			0%	0.021	0.026	0.093
			20%	0.021	0.026	0.090
	5	PCA	50%	0.021	0.026	0.090
			0%	0.008	0.023	0.082
			20%	0.008	0.023	0.080
Extractor	6	KPCA	50%	0.008	0.023	0.080
			0%	0.120	0.028	0.160
			20%	0.120	0.028	0.153
	1	VarianceFTest	50%	0.120	0.028	0.154
			0%	0.065	0.026	0.109
			20%	0.065	0.026	0.107
	2	RFImportance	50%	0.065	0.026	0.108
			0%	0.058	0.028	0.117
			20%	0.058	0.028	0.115
	3	LogisticL1	50%	0.058	0.028	0.116
			0%	0.058	0.028	0.133
			20%	0.058	0.028	0.130
	4	LASSO	50%	0.058	0.028	0.131
			0%	0.053	0.026	0.056
			20%	0.053	0.026	0.058
Selector	5	ElasticNetFS	50%	0.053	0.026	0.058
			0%	0.142	0.222	0.106
			20%	0.142	0.222	0.104
	1	MKL	50%	0.142	0.222	0.105
	2	Standard Concatenation	0%	0.138	0.217	0.112
	3	Learnable Weighted	0%	0.137	0.224	0.100
	4	Attention Weighted	0%	0.133	0.223	0.103
			0%	0.131	0.188	0.100
			20%	0.131	0.188	0.097
	5	Early Fusion PCA	50%	0.131	0.188	0.098
			0%	0.104	0.185	0.109
			20%	0.104	0.185	0.107
	6	Sum	50%	0.104	0.185	0.108
			0%	0.102	0.185	0.087
			20%	0.102	0.185	0.085
	7	Average	50%	0.102	0.185	0.086
			0%	0.094	0.168	0.094
			20%	0.094	0.168	0.092
Fusion Technique	8	Max	50%	0.094	0.168	0.092
			0%	0.138	0.212	0.224
			20%	0.138	0.212	0.220
	1	RandomForestClassifier	50%	0.138	0.212	0.221
			0%	0.134	0.208	0.009
			20%	0.134	0.208	0.009
	2	LogisticRegression	50%	0.134	0.208	0.009
			0%	0.100	0.167	0.068
			20%	0.100	0.167	0.067
Model	3	SVC	50%	0.100	0.167	0.068
	1	LDA_early_fusion_pca_RandomForestClassifier	0%	0.995	0.006	0.089
	2	LDA_early_fusion_pca_RandomForestClassifier	20%	0.995	0.006	0.088
	3	LDA_early_fusion_pca_RandomForestClassifier	50%	0.995	0.006	0.088
	4	LDA_learnable_weighted_RandomForestClassifier	0%	0.995	0.006	0.096
	5	LDA_mkl_RandomForestClassifier	0%	0.995	0.006	0.097
	6	LDA_standard_concat_RandomForestClassifier	0%	0.995	0.006	0.122
	7	LDA_mkl_RandomForestClassifier	20%	0.995	0.006	0.097
	8	LDA_mkl_RandomForestClassifier	50%	0.995	0.006	0.098
	9	LDA_attention_weighted_RandomForestClassifier	0%	0.992	0.007	0.097
Best Combinations	10	LDA_early_fusion_pca_LogisticRegression	0%	0.910	0.020	0.004
	1	KPCA_average_SVC	0%	-0.018	0.028	0.025
	2	FA_sum_SVC	0%	-0.017	0.020	0.042
	3	FA_sum_SVC	20%	-0.017	0.020	0.042
	4	FA_sum_SVC	50%	-0.017	0.020	0.042
	5	ElasticNetFS_sum_SVC	0%	-0.013	0.017	0.027
	6	RFImportance_sum_SVC	0%	-0.013	0.017	0.028
	7	LASSO_sum_SVC	0%	-0.013	0.017	0.040
	8	LogisticL1_sum_SVC	0%	-0.013	0.017	0.030
	9	ElasticNetFS_sum_SVC	20%	-0.013	0.017	0.028
Worst Combinations	10	RFImportance_sum_SVC	20%	-0.013	0.017	0.029

Figure G.1: Overview of the results from the Liver dataset. The table shows the average performance of individual components (Extractors, Selectors, Fusion Tech, Models) and the specific performance of the 10 best and worst pipeline combinations.

H. Appendix - Lung Results

Type	Ranking	Algorithm	Missing Data (%)	MCC Score	Standard Deviation	Fitting Time (s)
			0%	0.830	0.158	0.075
			20%	0.776	0.176	0.071
	1	LDA	50%	0.776	0.176	0.074
			0%	0.468	0.162	0.078
			20%	0.379	0.141	0.077
	2	PLS-DA	50%	0.379	0.141	0.075
			0%	0.045	0.069	0.116
			20%	0.017	0.049	0.086
	3	SparsePLS	50%	0.022	0.037	0.087
			0%	0.030	0.027	0.084
			20%	0.032	0.028	0.070
	4	FA	50%	0.032	0.028	0.077
			0%	0.028	0.026	0.094
			20%	0.025	0.028	0.077
	5	PCA	50%	0.025	0.028	0.095
			0%	0.007	0.023	0.083
			20%	0.006	0.027	0.070
Extractor	6	KPCA	50%	0.006	0.027	0.082
			0%	0.112	0.024	0.155
			20%	0.112	0.024	0.149
	1	VarianceFTest	50%	0.112	0.024	0.150
			0%	0.061	0.026	0.108
			20%	0.061	0.026	0.106
	2	RFImportance	50%	0.061	0.026	0.106
			0%	0.052	0.028	0.116
			20%	0.052	0.028	0.114
	3	LogisticL1	50%	0.052	0.028	0.114
			0%	0.052	0.028	0.131
			20%	0.052	0.028	0.128
	4	LASSO	50%	0.052	0.028	0.129
			0%	0.046	0.025	0.055
			20%	0.046	0.025	0.056
Selector	5	ElasticNetFS	50%	0.046	0.025	0.057
			0%	0.140	0.219	0.105
			20%	0.140	0.219	0.102
	1	MKL	50%	0.140	0.219	0.103
	2	Standard Concatenation	0%	0.137	0.214	0.111
	3	Learnable Weighted	0%	0.135	0.221	0.098
	4	Attention Weighted	0%	0.132	0.220	0.101
			0%	0.129	0.186	0.098
			20%	0.129	0.186	0.096
	5	Early Fusion PCA	50%	0.129	0.186	0.096
			0%	0.102	0.183	0.108
			20%	0.102	0.183	0.105
	6	Sum	50%	0.102	0.183	0.106
			0%	0.098	0.183	0.086
			20%	0.098	0.183	0.084
	7	Average	50%	0.098	0.183	0.085
			0%	0.090	0.166	0.093
			20%	0.090	0.166	0.091
Fusion Technique	8	Max	50%	0.090	0.166	0.091
			0%	0.135	0.208	0.220
			20%	0.135	0.208	0.215
	1	RandomForestClassifier	50%	0.135	0.208	0.217
			0%	0.128	0.204	0.009
			20%	0.128	0.204	0.009
	2	LogisticRegression	50%	0.128	0.204	0.009
			0%	0.097	0.164	0.067
			20%	0.097	0.164	0.066
Model	3	SVC	50%	0.097	0.164	0.066
	1	LDA_early_fusion_pca_RandomForestClassifier	0%	0.994	0.008	0.088
	2	LDA_early_fusion_pca_RandomForestClassifier	20%	0.994	0.008	0.086
	3	LDA_early_fusion_pca_RandomForestClassifier	50%	0.994	0.008	0.087
	4	LDA_learnable_weighted_RandomForestClassifier	0%	0.994	0.008	0.095
	5	LDA_mkl_RandomForestClassifier	0%	0.994	0.008	0.095
	6	LDA_standard_concat_RandomForestClassifier	0%	0.994	0.008	0.122
	7	LDA_mkl_RandomForestClassifier	20%	0.994	0.008	0.096
	8	LDA_mkl_RandomForestClassifier	50%	0.994	0.008	0.097
	9	LDA_attention_weighted_RandomForestClassifier	0%	0.991	0.008	0.096
Best Combinations	10	LDA_early_fusion_pca_LogisticRegression	0%	0.902	0.020	0.004
	1	KPCA_average_SVC	0%	-0.018	0.027	0.024
	2	FA_sum_SVC	0%	-0.016	0.019	0.041
	3	FA_sum_SVC	20%	-0.016	0.019	0.041
	4	FA_sum_SVC	50%	-0.016	0.019	0.041
	5	ElasticNetFS_sum_SVC	0%	-0.012	0.017	0.027
	6	RFImportance_sum_SVC	0%	-0.012	0.017	0.028
	7	LASSO_sum_SVC	0%	-0.012	0.017	0.039
	8	LogisticL1_sum_SVC	0%	-0.012	0.017	0.029
	9	ElasticNetFS_sum_SVC	20%	-0.012	0.017	0.027
Worst Combinations	10	RFImportance_sum_SVC	20%	-0.012	0.017	0.028

Figure H.1: Overview of the results from the Lung dataset. The table shows the average performance of individual components (Extractors, Selectors, Fusion Tech, Models) and the specific performance of the 10 best and worst pipeline combinations.

I. Appendix - Melanoma Results

Type	Ranking	Algorithm	Missing Data (%)	MCC Score	Standard Deviation	Fitting Time (s)
			0%	0.830	0.158	0.075
			20%	0.776	0.176	0.071
	1	LDA	50%	0.776	0.176	0.074
			0%	0.468	0.162	0.078
			20%	0.379	0.141	0.077
	2	PLS-DA	50%	0.379	0.141	0.075
			0%	0.045	0.069	0.116
			20%	0.017	0.049	0.086
	3	SparsePLS	50%	0.022	0.037	0.087
			0%	0.030	0.027	0.084
			20%	0.032	0.028	0.070
	4	FA	50%	0.032	0.028	0.077
			0%	0.028	0.026	0.094
			20%	0.025	0.028	0.077
	5	PCA	50%	0.025	0.028	0.095
			0%	0.007	0.023	0.083
			20%	0.006	0.027	0.070
Extractor	6	KPCA	50%	0.006	0.027	0.082
			0%	0.114	0.023	0.163
			20%	0.112	0.024	0.147
	1	VarianceFTest	50%	0.112	0.024	0.143
			0%	0.066	0.025	0.111
			20%	0.059	0.027	0.109
	2	RFImportance	50%	0.059	0.027	0.109
			0%	0.058	0.029	0.120
			20%	0.050	0.028	0.123
	3	LogisticL1	50%	0.050	0.028	0.111
			0%	0.058	0.029	0.135
			20%	0.050	0.028	0.127
	4	LASSO	50%	0.050	0.028	0.112
			0%	0.053	0.025	0.056
			20%	0.046	0.025	0.064
Selector	5	ElasticNetFS	50%	0.046	0.025	0.081
			0%	0.144	0.214	0.106
			20%	0.144	0.213	0.108
	1	MKL	50%	0.145	0.213	0.106
	2	Standard Concatenation	0%	0.143	0.214	0.109
	3	Learnable Weighted	0%	0.139	0.216	0.096
	4	Attention Weighted	0%	0.135	0.215	0.099
			0%	0.127	0.187	0.097
			20%	0.127	0.187	0.097
	5	Early Fusion PCA	50%	0.127	0.187	0.097
			0%	0.101	0.180	0.107
			20%	0.100	0.180	0.107
	6	Sum	50%	0.099	0.180	0.780
			0%	0.098	0.183	0.082
			20%	0.098	0.183	0.083
	7	Average	50%	0.097	0.183	0.080
			0%	0.089	0.166	0.089
			20%	0.088	0.166	0.091
Fusion Technique	8	Max	50%	0.087	0.166	0.090
			0%	0.140	0.212	0.220
			20%	0.126	0.202	0.210
	1	RandomForestClassifier	50%	0.126	0.202	0.606
			0%	0.128	0.209	0.009
			20%	0.121	0.195	0.008
	2	LogisticRegression	50%	0.119	0.195	0.008
			0%	0.098	0.167	0.066
			20%	0.087	0.159	0.074
Model	3	SVC	50%	0.088	0.159	0.078
	1	LDA_early_fusion_pca_RandomForestClassifier	0%	0.995	0.006	0.091
	2	LDA_early_fusion_pca_RandomForestClassifier	20%	0.995	0.006	0.107
	3	LDA_early_fusion_pca_RandomForestClassifier	50%	0.995	0.006	0.104
	4	LDA_learnable_weighted_RandomForestClassifier	0%	0.995	0.007	0.098
	5	LDA_mkl_RandomForestClassifier	0%	0.995	0.007	0.099
	6	LDA_standard_concat_RandomForestClassifier	0%	0.995	0.007	0.122
	7	LDA_mkl_RandomForestClassifier	20%	0.995	0.007	0.124
	8	LDA_mkl_RandomForestClassifier	50%	0.995	0.007	0.113
	9	LDA_attention_weighted_RandomForestClassifier	0%	0.992	0.007	0.100
Best Combinations	10	LDA_early_fusion_pca_LogisticRegression	0%	0.906	0.020	0.004
	1	KPCA_average_SVC	0%	-0.014	0.028	0.024
	2	FA_average_SVC	0%	-0.014	0.020	0.042
	3	FA_average_SVC	20%	-0.014	0.020	0.042
	4	FA_average_SVC	50%	-0.014	0.020	0.042
	5	ElasticNetFS_average_SVC	0%	-0.011	0.015	0.027
	6	RFImportance_average_SVC	0%	-0.011	0.015	0.028
	7	LASSO_average_SVC	0%	-0.011	0.015	0.040
	8	LogisticL1_average_SVC	0%	-0.011	0.015	0.030
	9	ElasticNetFS_average_SVC	20%	-0.011	0.015	0.028
Worst Combinations	10	RFImportance_average_SVC	20%	-0.011	0.015	0.029

Figure I.1: Overview of the results from the Melanoma dataset. The table shows the average performance of individual components (Extractors, Selectors, Fusion Tech, Models) and the specific performance of the 10 best and worst pipeline combinations.

J. Appendix - Ovarian Results

Type	Ranking	Algorithm	Missing Data (%)	MCC Score	Standard Deviation	Fitting Time (s)
			0%	0.830	0.158	0.075
			20%	0.776	0.176	0.071
	1	LDA	50%	0.776	0.176	0.074
			0%	0.468	0.162	0.078
			20%	0.379	0.141	0.077
	2	PLS-DA	50%	0.379	0.141	0.075
			0%	0.045	0.069	0.116
			20%	0.017	0.049	0.086
	3	SparsePLS	50%	0.022	0.037	0.087
			0%	0.030	0.027	0.084
			20%	0.032	0.028	0.070
	4	FA	50%	0.032	0.028	0.077
			0%	0.028	0.026	0.094
			20%	0.025	0.028	0.077
	5	PCA	50%	0.025	0.028	0.095
			0%	0.007	0.023	0.083
			20%	0.006	0.027	0.070
Extractor	6	KPCA	50%	0.006	0.027	0.082
			0%	0.208	0.130	0.091
			20%	0.184	0.128	0.100
	1	VarianceFTest	50%	0.184	0.128	0.089
			0%	0.081	0.064	0.065
			20%	0.088	0.062	0.068
	2	RFImportance	50%	0.088	0.062	0.065
			0%	0.074	0.076	0.035
			20%	0.072	0.079	0.035
	3	ElasticNetFS	50%	0.072	0.079	0.037
			0%	0.052	0.067	0.085
			20%	0.053	0.067	0.090
	4	LASSO	50%	0.053	0.067	0.083
			0%	0.052	0.067	0.083
			20%	0.053	0.067	0.097
Selector	5	LogisticL1	50%	0.053	0.067	0.086
			0%	0.176	0.254	0.062
			20%	0.175	0.251	0.068
	1	MKL	50%	0.175	0.251	0.064
	2	Standard Concatenation	0%	0.171	0.247	0.068
	3	Learnable Weighted	0%	0.144	0.249	0.065
	4	Attention Weighted	0%	0.139	0.249	0.066
			0%	0.138	0.242	0.061
			20%	0.140	0.241	0.064
	5	Early Fusion PCA	50%	0.140	0.241	0.060
			0%	0.123	0.184	0.065
			20%	0.125	0.183	0.070
	6	Sum	50%	0.126	0.182	0.065
			0%	0.108	0.184	0.058
			20%	0.110	0.182	0.062
	7	Average	50%	0.110	0.182	0.059
			0%	0.096	0.177	0.060
			20%	0.096	0.177	0.064
Fusion Technique	8	Max	50%	0.096	0.177	0.061
			0%	0.143	0.234	0.007
			20%	0.136	0.216	0.003
	1	RandomForestClassifier	50%	0.137	0.216	0.003
			0%	0.134	0.231	0.169
			20%	0.141	0.219	0.178
	2	LogisticRegression	50%	0.141	0.219	0.168
			0%	0.134	0.212	0.013
			20%	0.111	0.194	0.015
Model	3	SVC	50%	0.110	0.194	0.015
	1	LDA_attention_weighted_RandomForestClassifier	0%	1.000	0.000	0.089
	2	LDA_learnable_weighted_RandomForestClassifier	0%	1.000	0.000	0.089
	3	LDA_mkl_RandomForestClassifier	0%	1.000	0.000	0.087
	4	LDA_early_fusion_pca_RandomForestClassifier	0%	1.000	0.000	0.085
	5	LDA_standard_concat_RandomForestClassifier	0%	1.000	0.000	0.113
	6	LDA_mkl_RandomForestClassifier	20%	1.000	0.000	0.090
	7	LDA_early_fusion_pca_RandomForestClassifier	20%	1.000	0.000	0.085
	8	LDA_mkl_RandomForestClassifier	50%	1.000	0.000	0.094
	9	LDA_early_fusion_pca_RandomForestClassifier	50%	1.000	0.000	0.091
Best Combinations	10	LDA_attention_weighted_LogisticRegression	0%	0.934	0.016	0.001
	1	KPCA_sum_SVC	0%	-0.084	0.020	0.011
	2	PCA_attention_weighted_SVC	0%	-0.083	0.054	0.018
	3	SparsePLS_early_fusion_pca_LogisticRegression	0%	-0.083	0.064	0.001
	4	PCA_early_fusion_pca_SVC	0%	-0.080	0.077	0.014
	5	PCA_early_fusion_pca_SVC	20%	-0.080	0.077	0.015
	6	PCA_early_fusion_pca_SVC	50%	-0.080	0.077	0.016
	7	PCA_learnable_weighted_SVC	0%	-0.080	0.054	0.016
	8	PCA_attention_weighted_LogisticRegression	0%	-0.074	0.071	0.005
	9	KPCA_sum_LogisticRegression	0%	-0.073	0.063	0.003
Worst Combinations	10	KPCA_sum_LogisticRegression	20%	-0.073	0.063	0.003

Figure J.1: Overview of the results from the Ovarian dataset. The table shows the average performance of individual components (Extractors, Selectors, Fusion Tech, Models) and the specific performance of the 10 best and worst pipeline combinations.