**Project Report: Multi-Omic Cancer Type Classification**

**1. Main Objective**

The primary objective of this analysis was **prediction**: to develop a machine learning model capable of accurately classifying the type of cancer based on a patient's multi-omics data (RNA sequencing, DNA methylation, and miRNA sequencing). The intended audience for this model could be researchers or clinicians seeking automated tools for cancer type identification. Accurate classification provides foundational information for understanding tumour biology and potentially guiding further diagnostic or therapeutic research directions. While high accuracy was the goal, understanding the key drivers behind the classification was also explored.

**2. Data Description**

This analysis utilised the **TCGA-PanCan dataset** available on Kaggle. This dataset comprises multi-omics profiles from The Cancer Genome Atlas (TCGA) Pan-Cancer Atlas project, covering thousands of patient samples across **18 distinct cancer types**. The specific data modalities used were:

* **RNA-seq:** Gene expression levels (tpm\_unstranded).
* **DNA Methylation:** Beta values indicating methylation levels at specific CpG sites (beta).
* **miRNA-seq:** MicroRNA expression levels (reads\_per\_million\_miRNA\_mapped).

Combining these modalities resulted in an extremely high-dimensional dataset with **548,968 features** per patient. The goal was to train a multi-class classifier to predict the CANCER\_TYPE label based on these combined features.

**3. Data Exploration, Cleaning, and Feature Engineering Summary**

* **Exploration:** Initial analysis revealed a significant **class imbalance** within the target variable (CANCER\_TYPE). The largest class (Non-Small Cell Lung Cancer) contained 790 samples, while the smallest (Ovarian Epithelial Tumour) had only 8 samples after cleaning. This imbalance necessitated the use of stratified sampling and appropriate evaluation metrics (like weighted F1-score).

Immagine che contiene testo, schermata, Diagramma, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

* **Cleaning:** The dataset, sourced from processed TCGA files, was relatively clean. Patients missing one or more of the three required data files were excluded, resulting in a final cohort of **4,106 patients** for analysis. NaN values encountered during processing were imputed with 0.
* **Feature Engineering:** The primary challenge was the extreme dimensionality (548k features), which caused memory errors and prohibitive training times. To address this, **Principal Component Analysis (PCA)** was applied. The 548k features were reduced to **300 principal components**, capturing a significant portion of the data's variance while making model training computationally feasible. Data was scaled using StandardScaler prior to PCA.

**4. Summary of Classifier Models Trained**

Three distinct classifiers were trained on the 300 PCA components, using a stratified 80/20 train/test split. Baseline models were trained first, followed by optimization attempts.

| **Model** | **Training Time** | **Overall Accuracy** | **Weighted Avg F1** | **Key Failure (Worst F1)** | **Notes** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression (Base) | 3.30 sec | 97.32% | 0.97 | Misc. Neuro. (0.25) | Strong baseline |
| Random Forest (Base) | 0.39 sec | 95.13% | 0.95 | Ovarian (0.00) | Fastest, but failed on a rare class |
| XGBoost (Base) | 3.97 sec | 95.50% | 0.95 | Misc. Neuro. (0.33) | Solid, similar to RF |
| **Logistic Regression (Opt)** | **402.82 sec (GS)** | **97.45%** | **0.97** | **Misc. Neuro. (0.33), Ovarian (0.67)** | **Best overall performance after tuning** |
| Random Forest (Balanced) | 0.49 sec | 95.38% | 0.95 | Ovarian (0.00) | class\_weight didn't fix the rare class issue |
| XGBoost (Opt) | 94.18 sec (GS) | 95.50% | 0.95 | Misc. Neuro. (0.00) | Tuning didn't improve, worsened rare class |

*(GS = GridSearchCV time)*

**Performance Comparison Visualisation:**

The application of PCA dramatically reduced training times across all models. Initial baseline results showed Logistic Regression performed surprisingly well. Optimisation using GridSearchCV slightly improved Logistic Regression and confirmed its superior performance. Attempts to optimise Random Forest (using class\_weight) and XGBoost (using GridSearchCV) did not yield significant improvements and failed to adequately address challenges with the rarest classes.

Immagine che contiene testo, schermata, diagramma, Diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

**5. Recommended Final Model**

The **Optimised Logistic Regression** model (using C=1.0, penalty='l1', solver='saga') trained on the 300 PCA components is recommended as the final model.

**Justification:**

* It achieved the **highest overall accuracy (97.45%)** and **weighted average F1-score (0.97)** among all tested models.
* While not perfect, it demonstrated the **best performance on the extremely rare classes** (e.g., Ovarian Epithelial Tumour F1=0.67, Miscellaneous Neuroepithelial Tumour F1=0.33) compared to the ensemble methods, which completely failed on one or both.
* As a linear model, Logistic Regression offers slightly **better interpretability** (in terms of feature coefficients for the principal components) compared to the more "black-box" nature of Random Forest and XGBoost.

**6. Summary Key Findings and Insights**

* **High Predictability:** Cancer type is highly predictable from multi-omic data, with the best model achieving >97% accuracy even after significant dimensionality reduction.
* **PCA Efficacy:** PCA was crucial for managing the dataset's high dimensionality. Reducing features from ~550k to 300 resulted in a >100x speedup in training time with negligible impact on overall model accuracy, demonstrating significant feature redundancy.
* **Importance of DNA Methylation:** Analysis of the feature importances (coefficients) for the top principal components in the best Logistic Regression model revealed that **DNA methylation sites (dna\_cg...) were consistently the strongest contributing original features**. This suggests that epigenetic patterns captured in the DNA methylation data hold significant discriminatory power between cancer types within this dataset.
* **Imbalance Challenge:** Extreme class imbalance remains a significant challenge. Even the best model struggled with classes having very few samples (support < 10), indicating that the available data might be insufficient to robustly model these rare types or that the PCA compression lost critical distinguishing signals for them.
* **Model Simplicity:** The relatively simple Logistic Regression model outperformed more complex ensemble methods (Random Forest, XGBoost) after PCA. This suggests that once the dimensionality was effectively reduced, a linear model was sufficient to capture the primary predictive patterns.

**7. Suggestions for Next Steps**

* **Refine PCA:** Explore alternative PCA strategies. Instead of a fixed number of components (300), re-run PCA specifying a desired variance captured (e.g., n\_components=0.95). This might retain more subtle signals relevant to rare classes, potentially improving their F1-scores, although likely increasing computational cost.
* **Advanced Imbalance Handling:** Implement more sophisticated techniques to address class imbalance *before* PCA or within the modelling step, such as:
  + **SMOTE (Synthetic Minority Over-sampling Technique):** Generate synthetic samples for rare classes.
  + **ADASYN (Adaptive Synthetic Sampling):** Similar to SMOTE but focuses on harder-to-learn minority samples.
  + **Ensemble Methods for Imbalance:** Techniques like BalancedRandomForest or EasyEnsemble are specifically designed for imbalanced datasets.
* **Feature Selection Pre-PCA:** Given the likely importance of DNA methylation, experiment with training models *only* on methylation data, or perform feature selection *before* PCA to focus on potentially more relevant biological markers.
* **Interpretability of PCA:** Further investigate the biological meaning of the most important Principal Components (e.g., PC4, PC5, PC9) by performing pathway analysis or gene set enrichment analysis on the original features with high loadings for these components.