# Cancer Type Classification Using Clinical and Gene-Expression Data

MD Shahadat Hossain Shamim 22101174
Golam Mostofa Mahin 21301564
Ishrak Hamim Mahi 24241312

18-09-2025

**Abstract**

Cancer type classification is a crucial problem in bioinformatics, since accurate detection of cancer subtypes can guide treatment decisions and improve survival outcomes. In this project, we combined clinical and gene-expression data from seven cancer types, applied dimensionality reduction (PCA) to gene features, and merged them with clinical features to build a master dataset. We trained Logistic Regression and Random Forest models, achieving a test accuracy of 93.07% and 87.59% respectively. Results indicate that Logistic Regression with PCA features generalizes best across cancer types, while Random Forest provides robust but slightly lower performance.

## 1 Introduction

Cancer remains one of the leading causes of death worldwide. Different cancer types show unique gene-expression signatures and clinical patterns, but overlapping features make accurate classification difficult. Early and precise identification of cancer type is vital for personalized treatment strategies.

Recent pan-cancer classification studies consistently tackle the high dimensionality of RNA-seq by combining feature reduction with robust learners or deep nets. Divate et al. proposed a deep learning pan-cancer model on gene-expression achieving 97% accuracy across types [1]. Reviews surveying ML for cancer classification on gene expression highlight the effectiveness of regularized linear models, tree ensembles, and neural architectures while warning about overfitting and sample imbalance [2]. XGBoost-based tumor-type classifiers from genomic profiles demonstrate strong performance and interpretability via feature importance [3, 4]. RNA-seq-focused surveys catalog pan-cancer approaches and data handling choices (PCA/autoencoders, stratified evaluation) that directly align with our pipeline [5]. Interpretable deep generative models (e.g., VAE variants) further address dimensionality while retaining biological

signal [6]. Deep stacking/ensemble designs on RNA-seq improve multi-class accuracy and stability [7]. More recently, occlusion-enhanced deep frameworks have been used for pan-cancer classification and marker discovery [8]. In parallel, Mao et al. introduced Cox-Sage, an interpretable graph neural network that enhances the Cox proportional hazards model for cancer prognosis, demonstrating the importance of combining omics data with clinical information in interpretable deep learning frameworks [9]. Together these works motivate our use of PCA $\rightarrow$ (clinical + PCs) $\rightarrow$ stratified ML, and our comparison of linear vs. tree-based baselines on seven TCGA-like cancer types.

In this project, we build upon these works by combining gene-expression PCA features with clinical features for classification across seven cancer types using Logistic Regression and Random Forest models. Compared to prior studies, we ensure stratified train/test splits, report cross-validation results, and provide confusion matrix analyses.

## 1.1 Our Contribution

1. Data loading (clinical + gene-expression CSVs for seven cancers).
2. Gene alignment and PCA dimensionality reduction.
3. Clinical preprocessing (imputation + categorical encoding).
4. Merging PCA and clinical features into a master dataset.
5. Stratified train-test split.
6. Model training and evaluation.

## 2 Material and Methods

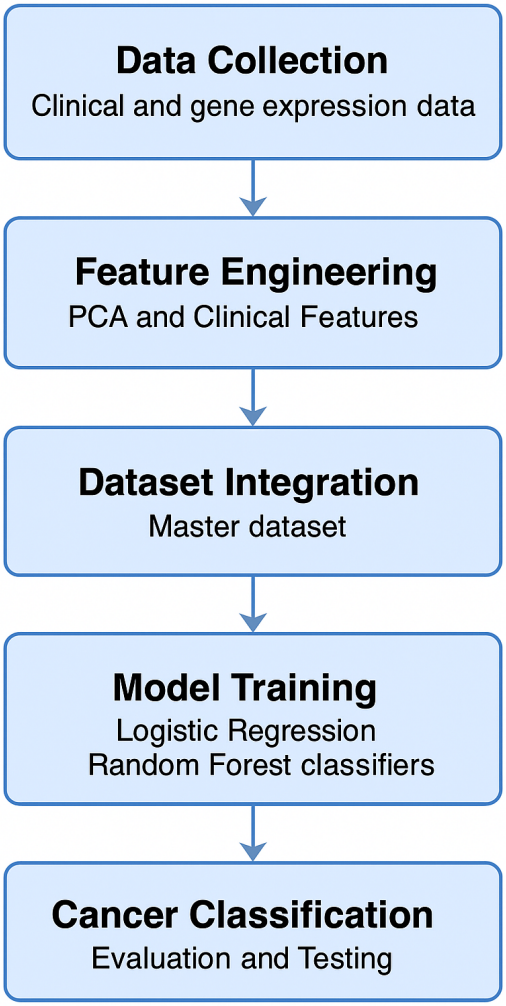


Figure 1: Block diagram of the cancer classification pipeline.

**Detailed Description:**

1. **Data Collection:** Clinical and gene expression data were collected for seven cancer types from Kaggle.
2. **Gene Expression Alignment & PCA:** Common genes were aligned across cancers, followed by PCA to reduce 19,938 genes to 1,313 principal components (95% variance).

3. **Clinical Feature Integration:** Clinical numeric (age, survival, vital status) and categorical (gender, race, stage, cancer status) features were retained.

4. **Master Dataset Construction:** A unified dataset with 1,369 samples and 1,322 features (1,313 PCs + 9 clinical) was built.

5. **Model Training:** Logistic Regression and Random Forest classifiers were trained using an 80/20 stratified split.

6. **Evaluation:** Models were evaluated with cross-validation, classification reports, and confusion matrices.

## 2.1 Dataset Description

We used the Kaggle dataset "Processed Gene and Clinical Data" [10], which contains both gene expression profiles and associated clinical features for seven cancer types: COAD, ESCA, HNSC, LIHC, LUAD, LUSC, and STAD. Each cancer has two data files: one for gene expression and one for clinical attributes.

**Gene Expression Data:** Each cancer dataset contains 19,939 gene expression features (columns) measured across different patient samples (rows). Table 1 summarizes the sample counts per cancer type.

| Cancer Type | Samples (rows) | Gene Features (columns) |
|---|---|---|
| COAD | 453 | 19,939 |
| ESCA | 184 | 19,939 |
| HNSC | 520 | 19,939 |
| LIHC | 368 | 19,939 |
| LUAD | 485 | 19,939 |
| LUSC | 489 | 19,939 |
| STAD | 406 | 19,939 |

Table 1: Gene expression dataset dimensions per cancer type.

**Clinical Data:** The clinical datasets contain patient-level demographic and health-related attributes such as age, gender, vital status, race, tumor stage, and cancer status. The number of features ranges from 10–12 columns depending on cancer type. Table 2 provides details.

**Merged Master Dataset:** After aligning the gene-expression matrices to retain only common genes and merging with clinical features, the master dataset contained **1,369 samples** with **1,322 features** (1,313 PCA gene components + 9 clinical features). This dataset was used for all subsequent classification experiments.

## 2.2 Tools / Models / Algorithms Used

In this project we did not employ Neural Networks; instead, we focused on two strong classical machine learning baselines that are widely used in cancer

| Cancer Type | Samples (rows) | Clinical Features (columns) |
|---|---|---|
| COAD | 453 | 10 |
| ESCA | 184 | 10 |
| HNSC | 520 | 12 |
| LIHC | 368 | 12 |
| LUAD | 485 | 10 |
| LUSC | 489 | 10 |
| STAD | 406 | 10 |

Table 2: Clinical dataset dimensions per cancer type.

classification tasks: Logistic Regression and Random Forest. Below we describe the models, their architectures, parameters, and motivations.

### 2.2.1 Logistic Regression (Multinomial)

Logistic Regression is a linear model used for multi-class classification. It models the probability that a sample belongs to a class using the softmax function. This model was chosen as a baseline due to its interpretability, efficiency, and robustness when combined with dimensionality reduction such as PCA.

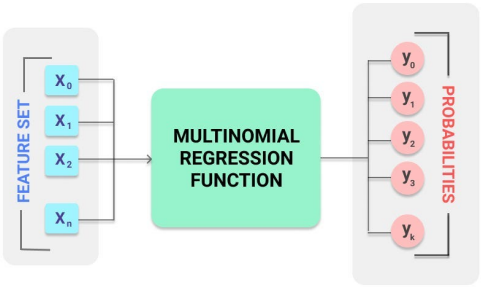


Figure 2: Conceptual diagram of Logistic Regression for multi-class classification.

**Parameters and Hyperparameters:**

- Solver: saga (supports multinomial loss and large feature sets)
- Multi-class setting: multinomial
- Regularization: default $L_2$ penalty with $C = 1.0$
- Maximum iterations: 5000

We used Logistic Regression to evaluate how well a simple linear boundary can separate cancer types in the PCA-transformed gene expression space. It achieved the highest accuracy among all tested models (93.07% test accuracy).

### 2.2.2 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees, each trained on a random subset of the data and features, to produce a robust classifier. It was chosen for its ability to capture nonlinear interactions between clinical and gene features, and for its robustness against noise and overfitting.

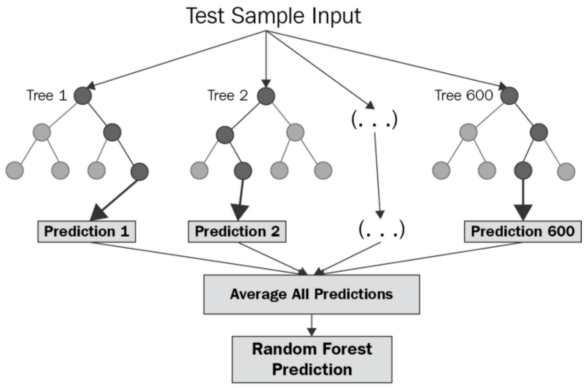


Figure 3: Conceptual diagram of a Random Forest classifier.

**Parameters and Hyperparameters:**

- Number of trees: 400
- Criterion: Gini impurity
- Class weight: balanced_subsample (to handle class imbalance)
- Random state: 42
- Parallel jobs: -1 (all cores)

Random Forest performed well (87.59% test accuracy) and provided feature importance scores, which can be used in future work to identify key PCA components or clinical attributes contributing to classification.

**Why These Models?** Both models were chosen because they represent two complementary approaches: Logistic Regression offers interpretability and efficiency in high-dimensional reduced spaces, while Random Forest captures nonlinear relationships and interactions. Together, they provide a strong baseline against which more complex models (e.g., Neural Networks, XGBoost) can be compared in future work.

## 2.3 Performance Evaluation

We evaluated the models using accuracy, precision, recall, and F1-score as performance metrics. Confusion matrices were also generated to analyze class-level

performance. The dataset was divided into training (80%) and testing (20%) subsets using a stratified split to preserve the distribution of cancer types. Cross-validation with 5 folds on the training set was applied to ensure robust evaluation and reduce variance in performance estimation.

# 3 Experimental Analysis

Table 3 summarizes the performance of the two baseline models, Logistic Regression and Random Forest, on the stratified train-test split. Logistic Regression achieved the highest accuracy and F1-score overall, while Random Forest performed slightly lower but remained competitive.

| **Model** | **CV Accuracy** | **Test Accuracy** | **Macro F1** |
|---|---|---|---|
| Logistic Regression | $0.9388 \pm 0.0074$ | 0.9307 | 0.9261 |
| Random Forest | $0.8749 \pm 0.0131$ | 0.8759 | 0.8630 |

Table 3: Baseline model comparison results.

Figures 4 and 5 show that Logistic Regression demonstrates strong separation across all seven cancer types, while Random Forest shows higher variance and more misclassification, particularly between ESCA and STAD, as well as confusion involving LUSC.

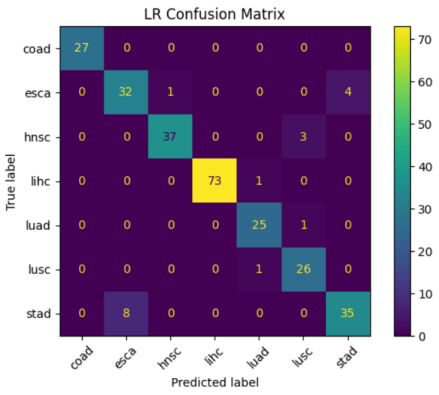


Figure 4: Logistic Regression Confusion Matrix

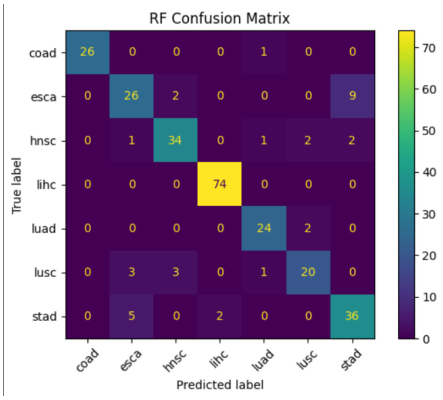


Figure 5: Random Forest Confusion Matrix

## Ablation Study: PCA-only vs PCA+Clinical

To investigate the contribution of clinical features, we performed an ablation study comparing models trained with PCA-only features and PCA+clinical features. Results are shown in Table 4.

**Observations:**

| Model | Feature Set | Test Accuracy | Macro F1 |
|---|---|---|---|
| Logistic Regression | PCA-only | 0.9307 | 0.9261 |
| Logistic Regression | PCA + Clinical | 0.9307 | 0.9261 |
| Random Forest | PCA-only | 0.8394 | 0.8248 |
| Random Forest | PCA + Clinical | 0.8759 | 0.8613 |

Table 4: Ablation study results: comparison of PCA-only vs PCA+clinical features.

- Logistic Regression achieved identical performance with PCA-only and PCA+clinical, indicating that clinical features added little additional predictive power in this setup.
- Random Forest showed noticeable improvement when clinical features were included (Accuracy: 83.94% $\rightarrow$ 87.59%, Macro F1: 0.8248 $\rightarrow$ 0.8613).
- This suggests that Random Forest benefited from heterogeneous feature sets, while Logistic Regression relied primarily on PCA-reduced gene expression features.

The complete implementation of data preprocessing, dimensionality reduction, model training, and evaluation is available as a Google Colab notebook at:
`Colab Project Notebook`.

# 4 Conclusion

We developed a cancer classification pipeline combining gene-expression PCA features with clinical data for seven cancer types. Logistic Regression achieved the best performance (93% accuracy, macro F1 0.926). Random Forest performed lower (87.6% accuracy, macro F1 0.863). Limitations include small dataset size. Future work includes testing deep learning models and feature importance analyses to identify key genes contributing to classification.

# References


[1] M. Divate, A. Agarwal, S. John, and S. Sengupta, "Deep learning-based pan-cancer classification model using gene expression," Available at PubMed Central, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8909043/

[2] F. Alharbi, A. Alahmadi, M. Almalki, and R. Meo, "Machine learning methods for cancer classification using gene expression data: A review," Available at PubMed Central, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9952758/

[3] V. Zelli, F. Grizzi, S. D. Giacomo, A. Coppola, G. Feo, and M. M. Barbareschi, "Classification of tumor types using xgboost machine learning," Available at PubMed Central, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10664515/

[4] Y. Zhang, J. Li, H. Wang, and X. Wang, "A novel xgboost method to identify cancer tissue-of-origin," Available at PubMed Central, 2020. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7716814/

[5] P. Štancl, M. Švéda, P. Gnip, V. Janouš, and P. Šimoník, "Machine learning for pan-cancer classification based on rna-seq data: A survey," Available at PubMed Central, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10667476/

[6] E. Withnell, A. H. Marshall, J. J. G. Leek, and A. A. Margolin, "Xomivae: An interpretable deep learning model for cancer classification using high-dimensional omics," Available at PubMed Central, 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8575033/

[7] M. Mohammed, D. M. Khan, and R. Garcia, "A stacking ensemble deep learning approach to cancer classification based on rna-seq data," *Scientific Reports*, vol. 11, no. 1, p. 21339, 2021. [Online]. Available: https://www.nature.com/articles/s41598-021-95128-x

[8] X. Zhao, Y. Liu, Z. Wang, and X. Li, "Occlusion-enhanced pan-cancer classification via deep learning (geneso)," BMC Bioinformatics (Open Access), 2024. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-024-05870-y

[9] R. Mao, L. Wan, M. Zhou, and D. Li, "Cox-sage: enhancing cox proportional hazards model with interpretable graph neural networks for cancer prognosis," *Briefings in Bioinformatics*, vol. 26, no. 2, p. bbaf108, 03 2025. [Online]. Available: https://doi.org/10.1093/bib/bbaf108

[10] RidgieMo, "Processed gene and clinical data," https://www.kaggle.com/datasets/ridgiemo/processed-gene-and-clinical-data, 2023, accessed: 2025-09-18.