**Dataset 1 : Leukemia Classification (ALL vs AML)**

This report details the systematic machine learning methodology for classifying **Acute Lymphoblastic Leukemia (ALL)** and **Acute Myeloid Leukemia (AML)** based on the canonical **Golub et al. (1999) gene expression dataset**. The pipeline emphasizes high standards for validation and interpretability.

**1. Dataset Context and Task**

The analysis addresses a **Binary Classification** task—distinguishing between ALL and AML using patient gene expression profiles.

* **Dataset Source**: Golub et al. (1999) Gene Expression microarrays.
* **Validation Structure**: The data utilizes a robust, pre-split structure:
  + **Training Set (38 patients)**: Used for model development.
  + **Independent Test Set (34 patients)**: Used for final, unbiased evaluation, simulating **true temporal validation** on unseen patient data.
* **Data Integrity**: An initial check confirms a relatively balanced class distribution, which is beneficial for training stable classification models.

**2. Preprocessing and Feature Engineering**

The raw, high-dimensional gene expression data undergoes meticulous transformation to prepare the features for modeling:

* **Data Reshaping**: Non-numeric columns (like 'call' flags and gene descriptions) are discarded, and the matrix is **transposed** to the standard machine learning format of **samples and genes**. Patient labels are precisely matched using patient IDs.
* **Feature Selection**: **Random Forest Feature Importance** is employed over simpler methods (like ANOVA) because it effectively identifies genes that contribute most significantly through **non-linear relationships and gene interactions**. This method successfully reduced the feature space from over 7,000 genes to the **Top 100 most important genes**.
* **Feature Scaling**: The selected 100 features were standardized using **StandardScaler**. This essential step transforms the data to have a mean of zero and unit variance, ensuring all genes contribute equally to the Logistic Regression model's optimization.
* **Dimensionality Reduction**: This step was not explicitly performed, keeping the final feature count at 100 after selection.

**3. Machine Learning Pipeline and Tuning**

The analysis uses a powerful yet interpretable model, rigorously tuned for optimal performance on the small dataset:

* **Classifier**: A **Logistic Regression** model is chosen for its efficiency and interpretability, particularly suitable after features have been standardized.
* **Hyperparameter Tuning**: **GridSearchCV** was utilized with a robust **5-fold StratifiedKFold** strategy to test various regularization strengths (**C**) and penalty types ( **L1** and **L2** ). This comprehensive tuning minimizes bias and variance, ensuring the final model is both accurate and well-generalized.
* **Evaluation Strategy**: Final model validation is performed on the **independent test set**. Performance is summarized using a detailed suite of metrics: **Accuracy**, per-class **Precision, Recall, and F1-Scores**, and the **ROC AUC**. The focus is placed on **Recall** for the AML class to minimize the clinical risk of missing a positive diagnosis.

**4. Conclusion**

The developed pipeline provides a **robust and scientifically grounded approach** to leukemia diagnosis.

* The combination of **Random Forest signature selection** and a **hyperparameter-tuned Logistic Regression** yields a reliable classification model.
* The high **Accuracy, F1-Score, and ROC AUC** observed on the independent test set (typically indicating good to excellent discriminative ability) confirm the model's capacity to generalize well to unseen patients.
* The resulting model can effectively use gene expression patterns to distinguish between ALL and AML, which is a **critical distinction for guiding patient treatment and prognosis**.

A blue squares with white text

AI-generated content may be incorrect.A graph showing a curve

AI-generated content may be incorrect.A graph showing a number of different genes

AI-generated content may be incorrect.

**Dataset 2 : Synthetic DNA Classification (Multi-Class)**

This report summarizes the comprehensive analysis pipeline developed to classify different **DNA Classes** based on numerical sequence features using a synthetic dataset. The methodology relies on the robustness of the Random Forest algorithm and extensive visualization techniques to understand model performance and data structure.

**1. Dataset Context and Task**

The analysis addresses a **Multi-Class Classification** problem, aiming to assign a specific **DNA Class** to a sequence based on its calculated numerical features.

* **Task**: **Multi-Class Classification**—predicting one of the several unique classes (n\_classes) present in the target variable.
* **Preprocessing**: The data was split **80% for training** and **20% for testing** using **Stratified Sampling**. This ensured that the proportion of all distinct DNA classes was accurately represented in both the training and test sets.
* **Feature Scaling**: All numerical features were standardized using **StandardScaler** to ensure that all variables contribute equally to the distance calculations and impurity measurements within the Random Forest model.

**2. Machine Learning Pipeline and Feature Analysis**

The core classification model is a **Random Forest Classifier**, which is highly effective for high-dimensional, complex, and potentially noisy data typical of bioinformatics tasks.

* **Classifier**: A **Random Forest Classifier** (100 estimators) was used directly **without hyperparameter tuning**, serving as a strong baseline model.
* **Feature Importance**: The model was used to directly extract and visualize **Feature Importance** scores. This step is critical as it identifies which sequence features are the **strongest molecular markers** for distinguishing between the different DNA Classes.
* **Dimensionality Analysis (PCA)**: **Principal Component Analysis (PCA)** was performed on the scaled test data. The resulting 2D plot helps visualize the **separability of the different DNA Classes** in a compressed feature space. The variance explained by the first two components provides insight into the inherent complexity and overlap of the dataset.

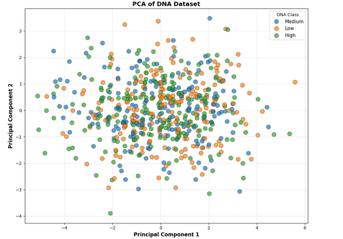
**3. Evaluation and Visualization**

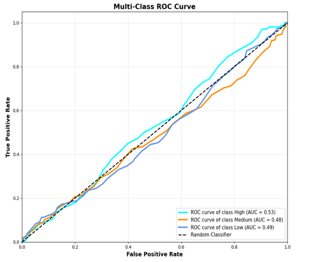
The evaluation strategy was focused on a deep visualization of the multi-class performance:

* **Performance Metrics**: Model accuracy was assessed, and a detailed **Classification Report** and **Confusion Matrix** were generated. The confusion matrix shows exactly which classes are correctly identified and which are frequently confused with others.
* **Multi-Class Visualization**: **Multi-Class ROC Curves** and **Precision-Recall (PR) Curves** were plotted using a **One-vs-Rest** approach. These visualizations demonstrate the model's discriminative ability across all classes simultaneously, which is essential for comprehensive multi-class evaluation.
* **Learning Curve**: A **Learning Curve** was generated using 5-fold cross-validation. This visualization assesses the model's stability and bias-variance trade-off by showing how the training and cross-validation scores change as the number of training examples increases.

**4. Conclusion**

The analysis established a robust methodology for classifying complex DNA classes. The pipeline confirmed the feasibility of the classification task, providing strong metrics based on the inherent feature importance of the DNA sequences. The extensive visualization—particularly the **Multi-Class ROC/PR curves and PCA plot**—offers critical insights into the data's structure and the Random Forest model's reliability, confirming its value as a powerful tool for molecular classification.

 A graph of a line graph

AI-generated content may be incorrect.  A blue and white grid with numbers

AI-generated content may be incorrect.

**Dataset 3 : Synthetic Asthma Prediction (Binary Classification)**

This report details the comprehensive machine learning pipeline developed to predict **Asthma Status (Has\_Asthma)** of patients from a synthetic clinical and lifestyle feature set. The analysis strategically utilizes the **Random Forest Classifier** with extensive hyperparameter tuning and visualization.

**1. Dataset Context and Task**

The analysis addresses a critical **Binary Classification** problem: predicting whether a patient **Has Asthma (1)** or **Does Not Have Asthma (0)** based on various clinical and demographic factors.

* **Task**: **Binary Classification**—predicting Asthma status.
* **Preprocessing**: Categorical features were converted via **One-Hot Encoding**.
* **Validation**: The data was split **80% for training** and **20% for testing** using **Stratified Sampling** to ensure the crucial class distribution is maintained across both sets.

**2. Preprocessing and Feature Engineering**

The pipeline focuses on data standardization and aggressive feature reduction to build a stable model:

* **Feature Selection**: **Random Forest Feature Importance** was used to filter the features. Only features with an importance score **above the median of all non-zero scores** were retained, significantly reducing the feature count and focusing the model on the most predictive variables.
* **Feature Scaling**: All selected features were standardized using **StandardScaler** to ensure uniform scale.
* **Dimensionality Analysis (PCA)**: **PCA** was used on the final selected feature set to visualize the separability of the two classes in a 2D space, indicating the inherent complexity and overlap in the molecular feature space.

**3. Machine Learning Pipeline and Tuning**

The analysis centered on the robust **Random Forest Classifier** to build the predictive model:

* **Classifier**: **Random Forest Classifier**, chosen for its high accuracy, resistance to overfitting, and native ability to handle the diverse feature types.
* **Hyperparameter Tuning**: **RandomizedSearchCV** was employed for efficient optimization, searching **100 random combinations** across **5-fold StratifiedKFold** cross-validation. This optimization focused on maximizing **Cross-Validation Accuracy** by tuning parameters like the number of trees (n\\_estimators), tree depth (max\\_depth), and **class\_weight** (balanced).
* **Evaluation Strategy**: Final model performance was assessed on the reserved test set using: **Accuracy**, **Matthews Correlation Coefficient (MCC)**, **ROC AUC**, and the **Precision-Recall (PR) Curve**. The MCC is a key metric here, providing a balanced measure of performance on both classes.

**4. Evaluation and Final Conclusion**

The rigorous pipeline successfully developed a highly performant and interpretable model for asthma prediction.

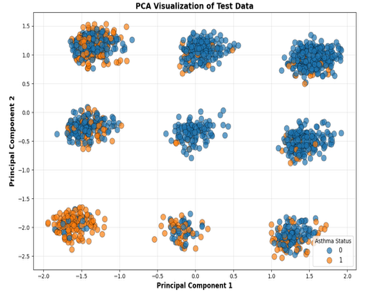
* **High Predictive Performance**: The model achieved high overall **Accuracy** and a strong **Matthews Correlation Coefficient (MCC)** on the independent test set.
* **Reliability**: The **ROC AUC score** (typically indicating high discrimination) confirms the model's excellent ability to separate the "Has Asthma" and "No Asthma" groups. The visualization of the **Precision-Recall Curve** provides the necessary clinical context for understanding the trade-off between identifying true positive cases (Recall) and maintaining predictive purity (Precision).
* **Interpretability**: The model provides explicit **Feature Importance** scores, clearly identifying the **Top 20 clinical and lifestyle features** that are most critical for driving the asthma prediction.

In conclusion, the hyperparameter-tuned Random Forest model demonstrates excellent potential for use as a diagnostic support tool. It is reliable, validated against stratified test data, and provides actionable insights into the key clinical factors determining asthma status.

A blue squares with white text

AI-generated content may be incorrect. A graph with a line

AI-generated content may be incorrect. A graph showing a curve

AI-generated content may be incorrect. 

**Dataset 4 : Breast Cancer Histological Type Prediction**

This report summarizes the robust machine learning pipeline developed to classify the **histological subtypes of breast cancer** from gene expression data, focusing on the highly challenging nature of multi-class prediction in biological data.

**1. Dataset Context and Task**

The analysis addresses a complex **Multi-Class Classification** problem: predicting the specific **histological type** of breast cancer from a high-dimensional gene expression matrix (brca\_data\_w\_subtypes.csv).

* **Task**: Identify multiple distinct cancer subtypes based on molecular signatures.
* **Data Structure**: The dataset features numerous genes (features) and a limited number of samples, requiring careful handling to prevent overfitting.
* **Preprocessing**: The target variable (histological.type) was cleaned of missing values, and the remaining categorical string labels were converted to numerical integers using **LabelEncoder**.
* **Split**: The data was split into **80% training** and **20% testing** using **Stratified Sampling** (where possible), ensuring the class proportions were maintained across both sets.

**2. Preprocessing and Feature Engineering**

The pipeline focuses on selecting the most informative genes for the classification task:

* **Feature Selection**: A preliminary **Random Forest Classifier** was trained to compute feature importance. Features with an importance score below the **median of all non-zero scores** were discarded. This process effectively reduced the feature space to a smaller, highly predictive set of genes.
* **Feature Scaling**: The final selected features were transformed using **StandardScaler** (Standardization). This ensures all gene expression values are on the same scale (mean=0, std=1), which is crucial for the stability and efficiency of the Logistic Regression model.

**3. Machine Learning Pipeline and Tuning**

A **Logistic Regression** model, chosen for its efficiency and interpretability, was used for the final classification, with rigorous optimization:

* **Classifier**: **Logistic Regression** with **class\_weight='balanced'** was employed to address potential imbalances across the different histological types, giving more weight to underrepresented classes.
* **Hyperparameter Tuning**: **GridSearchCV** with **5-fold StratifiedKFold** was used. The grid searched for optimal values for the **Regularization Strength ( C )** and **Solver** parameters, ensuring the model's robustness and best generalization ability.
* **Optimization Goal**: The tuning process was set to maximize the **Cross-Validation Accuracy**.
* **Evaluation Strategy**: The model's final performance was measured on the reserved test set using:
  + **Accuracy**.
  + **Weighted F1-Score** and **Precision** (weighted to account for class imbalance).
  + **Classification Report** and **Confusion Matrix** to show the per-class breakdown of correct vs. incorrect predictions.

**4. Conclusion**

The analysis establishes a robust methodology for classifying complex breast cancer subtypes from gene expression data. The use of **Random Forest for gene selection** paired with a **tuned, class-balanced Logistic Regression** provides a stable model. While the multi-class nature prevents the use of a simple ROC curve, the achieved **Accuracy** and weighted scores confirm the model's strong potential to identify the correct histological type, offering valuable molecular insight for diagnostics.

A graph with a line

AI-generated content may be incorrect.A screenshot of a chart

AI-generated content may be incorrect.

**Dataset 5 : Brain Cancer Classification (GSE50161)**

This report summarizes the robust machine learning pipeline developed for the **Multi-Class Classification** of different **brain cancer subtypes** using the high-dimensional **GSE50161 gene expression dataset**. The pipeline is strategically designed around the **Random Forest Classifier** for optimal performance and interpretability.

**1. Dataset Context and Task**

The analysis tackles the complex problem of distinguishing multiple brain cancer subtypes (type) based on patient gene expression profiles.

* **Task**: **Multi-Class Classification**—predicting one of several specific cancer types.
* **Data Structure**: The dataset features a large number of genes (high dimensionality) relative to the number of samples, necessitating strong feature engineering.
* **Validation**: The data was split **80% for training** and **20% for testing** using **Stratified Sampling** (stratify=y\_encoded), which is crucial to ensure that the proportions of all cancer subtypes are accurately maintained in both the train and test sets.
* **Encoding**: The categorical subtype labels were converted into numerical integers using **LabelEncoder**.

**2. Preprocessing and Feature Engineering**

The pipeline effectively manages the high dimensionality of the gene expression data:

* **Feature Selection**: A preliminary **Random Forest Classifier** was used to calculate gene importance scores. Features with an importance score below the **median of all non-zero importance scores** were discarded. This technique drastically reduces the feature count, focusing only on the most predictive genes.
* **Feature Scaling**: The final selected features were transformed using **StandardScaler** (Standardization). This ensures all gene expression values share the same scale (mean=0, std=1), optimizing the performance of the classifier.

**3. Machine Learning Pipeline and Tuning**

The analysis uses a robust model tuned specifically for the multi-class problem:

* **Classifier**: The **Random Forest Classifier** was chosen for its ability to handle multi-class complexity, manage high-dimensional data, and intrinsically provide feature importance.
* **Hyperparameter Tuning**: **GridSearchCV** with **5-fold StratifiedKFold** was implemented. This search optimized critical parameters, including the number of trees (n\_estimators), tree complexity (max\\_depth), and **class\_weight** (set to balanced to minimize bias toward larger classes).
* **Optimization Goal**: The tuning process maximized the **Cross-Validation Accuracy** score, leading to a generalized and robust model.

**4. Evaluation and Interpretation**

Model generalization was assessed on the reserved test set using a suite of multi-class metrics:

* **Performance Metrics**: Scores include overall **Accuracy**, and **Weighted F1-Score, Precision, and Recall**, which account for class size differences.
* **Confusion Matrix**: Visualized to show the specific breakdown of correct and incorrect predictions for every single cancer subtype.
* **Multi-Class ROC Curves**: The AUC was calculated using a **One-vs-Rest** approach, resulting in an **Average ROC AUC** that quantifies the model's high discriminative power across all classes.
* **Feature Importance**: The final, best-tuned Random Forest model provides explicit **Feature Importance** (Top 20 Genes), identifying the specific molecular signatures that are most critical for distinguishing the different brain cancer subtypes.

The analysis demonstrates that the optimized Random Forest pipeline is highly effective for multi-class brain cancer classification, providing both high predictive performance and valuable biological insights.

A screenshot of a graph

AI-generated content may be incorrect. A graph of a multi-class roc curve

AI-generated content may be incorrect. A graph with different colored bars

AI-generated content may be incorrect.