**Data Split and Results from the Provided Notebook**

**Data Split**

* **Training-Testing Split**: The dataset is partitioned into **80% training** and **20% testing** using train\_test\_split with stratify=y to preserve class distribution[[1]](#fn1).
* **Class Mapping**:

class\_map = {'BRCA': 0, 'KIRC': 1, 'LUAD': 2, 'PRAD': 3, 'COAD': 4}

This ensures consistency across training and testing subsets[[2]](#fn2).

**Model Results**

All models were evaluated using **balanced accuracy** to account for class imbalance:

|  |  |  |
| --- | --- | --- |
| Model | Balanced Accuracy | Key Metrics (Precision/Recall) |
| **Random Forest** | 0.993 | Precision: 0.98–1.00  Recall: 0.96–1.00[[3]](#fn3) |
| **SVM** | 0.993 | Precision: 0.98–1.00  Recall: 0.96–1.00[[4]](#fn4) |
| **Decision Tree** | 0.963 | Precision: 0.96–1.00  Recall: 0.89–1.00[[5]](#fn5) |

**Key observations:**

* **Perfect performance** for KIRC, PRAD, and COAD subtypes across all models.
* **LUAD** had the lowest recall (0.96 in RF/SVM, 0.89 in Decision Tree), indicating minor misclassification[[3]](#fn3)[[4]](#fn4)[[5]](#fn5).
* **Confusion matrices** show minimal errors (e.g., 1 LUAD misclassified as BRCA in RF/SVM)[[6]](#fn6)[[7]](#fn7)[[8]](#fn8).

**Feature Importance**

* **Top 10 Genes** (Random Forest):
  + TTLL10, TBC1D22B, RPL31, AARSD1, RPL23A, RPL35A, RPL37A, RPL38, RPLP0, RPS27[[9]](#fn9).
  + These ribosomal and cancer-associated genes dominated feature importance.

**Comparison to External Studies**

* Matches the **80:20 split** used in ovarian cancer studies (e.g.,[[10]](#fn10)) but achieves higher accuracy (99% vs. 100% in[[10]](#fn10)).
* Outperforms breast cancer models (e.g., gradient boosting/neural networks in[[11]](#fn11)) in balanced accuracy.

Works Cited

1. <https://stackoverflow.com/questions/13610074/is-there-a-rule-of-thumb-for-how-to-divide-a-dataset-into-training-and-validatio>

1. <https://www.baeldung.com/cs/train-test-datasets-ratio>

1. <https://www.medrxiv.org/content/10.1101/2022.09.16.22280052.full>

1. <https://github.com/chris-lovejoy/CodingForMedicine/blob/main/exercises/Gene_classification_exercise.ipynb>

1. <https://www.kaggle.com/c/UBC-OCEAN/discussion/466455>

1. <https://www.mdpi.com/2379-139X/10/1/10>

1. <https://star-protocols.cell.com/protocols/3579>

1. <https://pubs.aip.org/aip/aml/article/1/2/026110/2887627/Label-free-identification-of-different-cancer>

1. <https://www.sciencedirect.com/science/article/pii/S2772662223001388>

1. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11272718/>

1. <https://github.com/tstran155/Multiclass-classification-of-breast-cancer-patients>

The feature importance was calculated using the **training data**.

**Explanation:**

* Feature selection steps (VarianceThreshold and SelectKBest) and model training (Random Forest) are applied to the **training set** after the data is split.
* The model is then evaluated on the **testing set**.
* Feature importances are derived from the trained Random Forest model, which was fit only on the training data and the features selected from the training data.

This approach ensures there is **no data leakage** from the test set into the feature selection or model training process, maintaining the integrity of your evaluation.