Five-Fold Cross-Validation

Five-fold cross-validation (5-fold CV) is a statistical method used to **evaluate the performance of a machine learning model**. It helps ensure the model generalizes well to unseen data.

What Is It?

* The dataset is randomly split into **5 equal parts (folds)**. * The model is trained on **4 folds** and tested on the **remaining 1 fold**. * This process repeats **5 times**, each time using a different fold as the test set. * Final performance is the **average** of the 5 results.

Why Use It?

* Reduces **overfitting risk** * Provides a **more reliable estimate** of model performance than a single train/test split * Especially useful when data is **limited**

Example Workflow

1. Total samples = 100 patients with MRI radiomics
2. Split into 5 folds of 20 patients each:
    - Iteration 1: Train on folds 1–4, test on fold 5
    - Iteration 2: Train on folds 1,2,3,5, test on fold 4
    - …
    - Iteration 5: Train on folds 2–5, test on fold 1
3. Compute metrics (e.g., accuracy, AUC) at each step
4. Final result: **mean of all 5 metrics**

Advantages

- Uses **all data** for both training and testing
- Provides **variance estimate** of model performance
- Good balance between **bias and variance**

⚠ Notes

- Use **stratified** 5-fold CV when class labels are imbalanced
- Avoid data leakage: normalize or extract features **within each fold**

Python Snippet

```python
from sklearn.model_selection import cross_val_score
from xgboost import XGBClassifier

model = XGBClassifier()
scores = cross_val_score(model, X, y, cv=5, scoring='accuracy')

print("5-fold CV Accuracy:", scores.mean())
```