Data Splitting & Stratification

Patient ID	Class	Images
001	A	5
002	B	3
003	A	4
004	C	6
005	B	3
006	C	2
007	A	7
008	C	3
009	B	4
010	A	5

Class ratio preserved

Class ratio preserved

Class ratio preserved

Training Set (70%)

Pat ID	Class
001	A
003	A
007	A
002	B
005	B
006	C
008	C

Grouped by Patient ID

Validation Set (15%)

Pat ID	Class
010	A
009	B

Test Set (15%)

Pat ID	Class
-	A
-	B
004	C

Key Principles:

- Stratify by class to maintain class distribution
- Group by patient to prevent data leakage
- Lock test set - never use for hyperparameter tuning
- Document split methodology and random seeds

Image ID	Class
001_1	A
001_2	A
002_1	B
003_1	A

Random image-level split