Data Collection and Preprocessing Phase

Date	09 July 2024
Team ID	740024
Project Title	Evolving efficient classification patterns in Lymphography
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Imaging Equipment	Inconsistent image quality (brightness, contrast, resolution)	High	*Standardize imaging protocols across machines. * Implement calibration procedures for equipment. * Perform routine maintenance and quality checks.
Annotations	Missing or inaccurate labeling of lymph nodes (normal/abnormal)	High	* Double-annotation by experienced radiologists to ensure accuracy. * Utilize consensus approach for resolving discrepancies. * Implement training programs for annotators.

Data LabelingFormat	Inconsistent labeling format (e.g., missing data points, typos)	Medium	* Develop a standardized labeling schema with clear definitions. * Implement data validation tools to catch inconsistencies during entry. * Train data entry personnel on the labeling protocol.
Class Imbalance	Unequal distribution of normal and abnormal cases	Medium	* Implement data augmentation techniques (e.g., oversampling, undersampling) to balance classes. * Explore using costsensitive learning algorithms. * Consider incorporating prior knowledge (prevalence rates) into the model.
Missing Data	Incomplete patient information or missing images	Low - Medium (depends on extent)	* Identify the cause of missing data (e.g., technical issue, patient dropout). * Impute missing values using appropriate statistical methods (e.g., mean/median imputation). * Consider excluding data points with excessive missing data.

Outliers and Anomalies	Unusual data points that deviate from expected patterns	Medium	* Implement outlier detection algorithms to identify suspicious cases. * Review outliers by medical experts to
			determine potential causes (e.g., imaging artifacts, rare conditions). * Consider excluding extreme outliers or handling them as separate cases.

Attribute Information

— NOTE: All attribute values in the database have been entered as numeric values corresponding to their index in the list of attribute values for that attribute domain as given below.

- **class**: normal find, metastases, malign lymph, fibrosis
- **lymphatics**: normal, arched, deformed, displaced
- **block of affere**: no, yes
- **bl. of lymph. c**: no, yes
- **bl. of lymph. s**: no, yes
- **by pass**: no, yes
- **extravasates**: no, yes
- **regeneration of**: no, yes
- **early uptake in**: no, yes
- **lym.nodes dimin**: 0-3
- **lym.nodes enlar**: 1-4
- **changes in lym.**: bean, oval, round

- **defect in node**: no, lacunar, lac. marginal, lac. central
- **changes in node**: no, lacunar, lac. margin, lac. central
- **changes in stru**: no, grainy, drop-like, coarse, diluted, reticular, stripped, faint
- **special forms**: no, chalices, vesicles
- **dislocation of**: no, yes
- **exclusion of no**: no, yes
- **no. of nodes in**: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, ¿=70

Missing Attribute Values

None

Class Distribution

Class	Number of Instances
normal find	2
metastases	81
malign lymph	61
fibrosis	4