Data Collection and Preprocessing Phase

Date	22 October 2024
Team ID	SWTID1727274979
Project Title	Deep Learning Techniques for Breast Cancer Risk Prediction
Maximum Marks	2 Marks

Data Quality Report Template

This Data Quality report identifies key data quality issues in the Breast Histopathology Images Dataset obtained from Kaggle. High-severity issues include Incomplete metadata, Class imbalance, and Labeling inaccuracies, which must be prioritized to ensure the dataset's reliability . Medium-severity concerns, such as image quality inconsistencies and duplicate entries, also require systematic resolution which is mention as follows :

Data Source	Data Quality Issue	Severity	Resolution Plan
Dataset : Breast Histopathology Images **Source :** Kaggle	**1) Incomplete metadata :** The dataset lacks critical metadata fields such as patient ID, diagnosis details, or imaging conditions. (The dataset only contains breast histopathology images , other fields are not mentioned)	High	Collaborate with dataset contributors to add missing metadata ; create assumptions or placeholders where metadata cannot be retrieved .
	2) Duplicate entries : Duplicate images or near-identical images exist, potentially leading to overfitting in model training.	Medium	Use image hashing or similarity-checking algorithms to identify and remove duplicates.

	3) Label inaccuracies : Misclassifications or errors in labeling affect the dataset's reliability for training a supervised learning model.	High	Conduct a thorough review of labels using domain experts or cross-validation methods; re-label incorrect entries.
	4) Class Imbalance : Diagnostic categories such as malignant (indicated as 1) and benign (indicated as 0) cases are unequally distributed, leading to model bias and inaccuracy .	High	Employing data augmentation techniques for minority classes; consider oversampling or synthetic data generation methods .
	5) Artifacts and noise : Some images contain irrelevant stains, marks, or scanning artifacts that may confuse the model.	Medium	Implement artifact-removal techniques or exclude noisy images based on domain expert feedback.
	6) Image quality variability : Images show inconsistent resolution, brightness, and contrast, which can degrade model performance.	Medium	Apply image preprocessing techniques such as importing ImageDataGenerator , histogram equalization and normalization to standardize quality.