Professional Certificate in AI for Digital Pathology

Data Preprocessing for AI

Data Preprocessing for AI in Digital Pathology: A comprehensive glossary of terms

Anonymization: The process of removing or encrypting personally identifiable information (PII) from data to protect the privacy and confidentiality of individuals. In digital pathology, anonymization is crucial when sharing data for AI model development and collaboration.

Augmentation: A technique used to increase the size of a dataset by generating modified versions of existing images. Augmentation can include rotations, translations, zooming, flipping, and other transformations to create a more diverse and robust dataset for AI model training.

Balanced dataset: A dataset containing roughly equal numbers of samples from each class or category to avoid bias in AI model training. In digital pathology, creating a balanced dataset might be necessary when dealing with imbalanced class distributions, such as rare diseases.

Batch normalization: A technique used to normalize the activations of the neural network layers during training, reducing internal covariate shift and improving model convergence. It can help AI models in digital pathology generalize better and learn more efficiently.

Convolutional Neural Network (CNN): A type of deep learning architecture designed to process grid-like data, such as images. CNNs are widely used in digital pathology for tasks such as tumor detection, segmentation, and classification.

Cross-validation: A model evaluation technique used to estimate model performance by partitioning the dataset into k folds, where k-1 folds are used for training, and the remaining fold is used for testing. This process is repeated k times, with each fold serving as the test set once. The average performance across all iterations is then calculated.

Data augmentation: (see Augmentation)

Data balancing: (see Balanced dataset)

Data exploration: The process of examining and understanding the characteristics of a dataset through statistical analysis and visualization. Data exploration in digital pathology can help identify trends, outliers, and data quality issues.

Data leakage: An issue that occurs when information from the test or validation set is unintentionally used during the training phase, leading to overly optimistic performance estimates. Data leakage can result from

poor data splitting, inappropriate feature engineering, or other preprocessing steps.

Data normalization: A technique used to transform features to a similar scale, typically between 0 and 1, to avoid features with larger scales dominating the learning process. Common normalization methods include min-max scaling and z-score normalization.

Data preprocessing: The process of cleaning, transforming, and preparing raw data for AI model training. Data preprocessing in digital pathology includes tasks such as image acquisition, enhancement, segmentation, feature extraction, and data splitting.

Data splitting: The process of dividing a dataset into training, validation, and test sets to evaluate model performance and prevent overfitting. Proper data splitting ensures that each set is independent and representative of the entire dataset.

Deep learning: A subset of machine learning using artificial neural networks with multiple hidden layers to learn and represent complex patterns in data. Deep learning models can automatically learn features and representations from raw data, making them particularly suitable for digital pathology applications.

Digital pathology: The practice of converting glass slides into digital images, allowing pathologists to view, analyze, and share cases electronically. Digital pathology enables the application of AI techniques to improve diagnostic accuracy, efficiency, and consistency.

Feature engineering: The process of creating, selecting, and transforming features from raw data to improve AI model performance. In digital pathology, feature engineering can involve tasks such as image segmentation, texture analysis, or morphological feature extraction.

Feature extraction: The process of deriving meaningful features or attributes from raw data to represent the underlying patterns and relationships. In digital pathology, feature extraction can involve tasks such as color analysis, shape detection, or texture quantification.

Feature selection: The process of identifying and retaining the most relevant and informative features for AI model training. Feature selection can help improve model performance, reduce overfitting, and decrease training time.

Generalization: The ability of an AI model to perform well on unseen data, making accurate predictions and avoiding overfitting. In digital pathology, generalization is crucial for deploying AI models in clinical settings, where they will encounter new, unseen cases.

Ground truth: The true or accepted label or value for a given sample or feature. In digital pathology, ground truth is often established through manual annotation by expert pathologists.

Histopathology: The study of diseased tissues at the microscopic level, typically using stained glass slides.

Histopathology is a critical component of diagnostic pathology, and digital histopathology images are a primary data source for AI model development in digital pathology.

Hyperparameter tuning: The process of adjusting the configuration parameters of an AI model to optimize performance. Common hyperparameters include learning rate, batch size, number of layers, and regularization coefficients.

Image acquisition: The process of capturing digital images from glass slides using whole-slide imaging (WSI) scanners. Image acquisition in digital pathology involves considerations such as resolution, color representation, and file format.

Image annotation: The process of manually labeling regions of interest (ROIs) in digital pathology images, such as tumor regions, mitotic figures, or other relevant structures. Annotation is typically performed by expert pathologists and serves as the ground truth for AI model training and evaluation.

Image enhancement: The process of improving the visual quality of digital pathology images through techniques such as contrast adjustment, noise reduction, or sharpening. Image enhancement can help improve AI model performance by making features more discernible.

Image segmentation: The process of partitioning a digital pathology image into distinct regions or objects, such as nuclei, cells, or tumor regions. Segmentation is an essential preprocessing step for many AI applications in digital pathology, including tumor detection, grading, and survival prediction.

Instance normalization: A normalization technique that normalizes each image instance independently, instead of normalizing across the entire dataset. Instance normalization can help improve generalization and reduce the effects of domain shift in digital pathology applications.

Label noise: The presence of incorrect or inconsistent labels in a dataset, which can negatively impact AI model performance. In digital pathology, label noise can result from variability in manual annotation by different pathologists.

Machine learning: A subset of artificial intelligence that focuses on developing algorithms that automatically learn patterns and relationships from data, without explicit programming. Machine learning models can be categorized as supervised, unsupervised, or reinforcement learning.

Overfitting: A phenomenon where an AI model learns the training data too closely, capturing noise and idiosyncrasies, and performs poorly on unseen data. Overfitting can be mitigated through techniques such as regularization, early stopping, or increasing the dataset size.

Pathologist variability: The natural variation in diagnostic accuracy, interpretation, and annotation among pathologists. Pathologist variability can impact AI model performance and generalization, particularly in tasks requiring subjective judgment or expertise.

Pixel-level annotation: A type of image annotation where individual pixels are labeled according to their class or category, creating a pixel-wise mask for the region of interest in digital pathology images.

Preprocessing: (see Data preprocessing)

Region of interest (ROI): A specific area or structure within a digital pathology image that is relevant for a given task, such as tumor regions, mitotic figures, or necrotic areas. Identifying and extracting ROIs is a critical step in digital pathology AI applications.

Slide scanning: The process of converting glass slides into digital images using whole-slide imaging (WSI) scanners. Slide scanning is an essential step in digital pathology, enabling the application of AI techniques for image analysis and diagnosis.

Stain normalization: A technique used to adjust the color and intensity of digital pathology images to account for variations in staining protocols and scanner settings. Stain normalization can help improve AI model performance by reducing color and stain variability.

Supervised learning: A machine learning approach where a model is trained using labeled