Choice of Dataset:

Alzheimer's: Open Access Series of Imaging Studies (OASIS) dataset which is a set of MRI scans of Alzheimer disease patients as longitudinal studies. We would use "Alzheimer preprocessed dataset" by Sachin Kumar on Kaggle which is a cross-sectional dataset.

MS: "Multiple sclerosis" by Burak TAŞCI, cross-sectional and Brain MRI Dataset of Multiple Sclerosis with Consensus Manual Lesion Segmentation and Patient Meta Information by M Muslim, Ali (2022), Mendeley Data, V1, Doi: 10.17632/8bctsm8jz7.1

In addition to these databases, we plan find/use more about rarer diseases and continue gathering data by contacting hospitals if we need to. For example, prion related diseases such as Creutzfeldt-J or CSS.

Methodology:

Obtain medical imaging data (like MRI) with proper labels. The labels can indicate the presence or type of disease. We must ensure patient data privacy by anonymizing the dataset. Remove any personal identifiers and meta-data that might trace back to an individual.

a. Data Preprocessing:

Image Standardization: Ensure all images have the same dimensions. Normalize or standardize pixel values (e.g., scaling between 0 and 1). Adjust image contrast or brightness if necessary.
Noise Reduction: Use techniques like Gaussian blurring, median filtering, etc., to reduce noise in the images.
Augmentation: Apply transformations like rotations, flips, and shifts to increase the dataset size and improve the model's generalization.

b. ML Model:

Logistic Regression: For binary classification tasks or when a probabilistic framework is desired. Requires little computational resources, good for linearly separable data. However, limited to binary classification and sensitive to outliers. Prediction can also become slow as it computes the distance to every training to make a prediction.
K-Nearest Neighbors (KNN): Can be used for classification based on feature similarity. Easy to understand, versatile and instantaneous training. However, computationally and memory intensive.

c. Training Data:

Classification metrics like confusion matrices and precision-recall are pivotal, and offers detailed insights into diagnostic accuracy, especially crucial for imbalanced datasets. MSE would be relevant for continuous predictions, such as disease progression. The Rand Index is valuable for unsupervised tasks, like grouping patient data. While BLEU scores are niche, they're pertinent for automated report generation. The chosen metric should align with both the data nature and clinical implications, underscoring the importance of collaboration with medical experts.

Application: Our user will submit an image from an MRI. Our ML algorithm will analyze whether the brain has one (or more) of the diseases that the algorithm was trained on. It will continue to identify whether the individual is at-risk for any other diseases. We will also highlight the anomalies in the MRI.