Project Summary: Head CT Hemorrhage Detection

Author: Gnanasudharsan Ashokumar

Project Summary:

Head CT Hemorrhage Detection

This project uses a dataset of head CT scans to detect hemorrhages. Various machine learning tools were employed to classify images and identify the most effective method in terms of accuracy and computational efficiency. The tools include K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Convolutional Neural Networks (CNN), which proved to be the most suitable models for the task.

Dataset:

The dataset was sourced from Kaggle and contains 200 head CT images-100 with hemorrhage and 100 without. Additionally, the dataset includes a CSV file where each image is labeled as either 0 (no hemorrhage) or 1 (hemorrhage).

Feature Extraction:

The images were first processed to extract feature vectors using various techniques:

- Simple Method (SIMPLE): The image was resized and flattened into a single vector of pixel intensities.
- Histogram Method (HISTOGRAM): The image was converted into a color histogram.
- Shape Descriptor Method (HUMOMENTS): Images were described using seven Hu moments, which capture the shapes in the image.
- Principal Component Analysis (PCA): PCA was applied for dimensionality reduction.
- Canny Edge Detection (EDGES): The Canny algorithm was used to detect edges in the image.

Machine Learning Models:

1. KNN: Implemented with two distance functions: Euclidean distance and Earth Mover's Distance (EMD).
2. SVM: Evaluated with four kernels: linear, polynomial, radial basis function (RBF), and sigmoid.
3. Decision Tree: A decision tree algorithm (CART) with the Gini index was used.
4. AdaBoost: Implemented using decision trees as weak classifiers.
5. Random Forest: A model with 100 decision trees.

Deep Learning:

A CNN was trained separately from the other models as it is the most suitable for image classification. The CNN consisted of convolutional and pooling layers to automatically extract the most important features from the images, followed by fully connected layers for classification.

Challenges:

One of the key challenges was working with the DICOM format used in medical imaging. Due to installation issues with the Pydicom library, the team decided to use regular image formats instead. Another challenge was choosing the optimal size for feature vectors. Increasing the number of features sometimes reduced the performance of certain models due to the "curse of dimensionality."

Results:

- The SIMPLE method yielded the highest accuracy, with the Random Forest and SVM models achieving 90% accuracy.
- The PCA method also performed well, with the SVM model achieving 80-87.5% accuracy.
- The HISTOGRAM method produced moderate accuracy but was computationally efficient.
- The EDGES method resulted in lower accuracy across all models, while the HUMOMENTS method yielded the worst performance.

Conclusion:

The CNN outperformed the other models, as expected, achieving nearly 100% accuracy in classifying the images. However, the simpler models, such as SVM and Random Forest, also achieved high accuracy with much less computational cost, making them suitable alternatives in cases where computational efficiency is a priority.