**Predicting Alzheimer's Disease Status Using Hippocampal Data**

**Background:**

Alzheimer's Disease (AD) is a neurodegenerative disorder characterized by progressive cognitive decline. Atrophy in the hippocampus, a brain region critical for memory formation, is one of the early biomarkers of AD. In this assignment, you are provided with left and right hippocampus matrices in compressed CSV format for 369 subjects from ADNI downloadable from Google Drive (<https://drive.google.com/drive/folders/1X1HPde-5snvjZhUyYKu5BHtqoJyCLrYb?usp=share_link> ). Each matrix represents imaging data preprocessed and extracted from MRI scans. Your task is to develop a Convolutional Neural Network (CNN) model to predict the AD status of subjects based on their hippocampal data.

**Data Description:**

* **Hippocampal Matrices:** Compressed CSV files containing numerical data representing the hippocampal regions for each subject.
  + **Left Hippocampus Matrix:** Dimensions may vary per subject.
  + **Right Hippocampus Matrix:** Dimensions may vary per subject.
* **AD Status Labels:** A separate CSV file indicating the AD status (AD or Control) for each subject.

**Objectives:**

**1. Data Exploration and Preprocessing (15 points)**

a. **Data Loading**:

* Decompress and load the CSV files containing the hippocampal matrices.
* Describe the dimensions and structure of the data.

b. **Preprocessing**:

* Check for missing or corrupted data and handle appropriately.
* Normalize or standardize the data if necessary.
* Justify each preprocessing step taken.

**2. Model Development (30 points)**

a. **CNN Architecture Design**:

* Propose a CNN architecture suitable for this classification task.
* Explain your choices regarding:
  + Number and type of layers (convolutional, pooling, fully connected).
  + Kernel sizes, strides, and padding.
  + Activation functions.
  + Regularization techniques (e.g., dropout, batch normalization).

b. **Implementation**:

* Implement the CNN using a deep learning framework of your choice (e.g. PyTorch, TensorFlow).
* Include code snippets highlighting key components of your model.

**3. Training and Validation (20 points)**

a. **Data Splitting**:

* Split the training data into training and validation sets (e.g., 80% training, 20% validation).
* Explain your rationale behind the chosen split.

b. **Model Training**:

* Train your CNN on the training set.
* Monitor training and validation loss and accuracy over epochs.
* Plot these metrics to visualize the training process.

c. **Overfitting Prevention**:

* Describe any techniques used to prevent overfitting.
* Justify the choice of techniques and their implementation.

**4. Model Evaluation (15 points)**

a. **Performance Metrics**:

* Evaluate your model on the validation set using:
  + Accuracy
  + Precision
  + Recall
  + F1-Score
  + ROC-AUC Curve

b. **Results Interpretation**:

* Discuss the performance of your model.
* Identify strengths and weaknesses based on the evaluation metrics.

**5. Test Set Prediction (10 points)**

a. **Inference**:

* Use your trained model to predict AD status for the test set subjects.

b. **Submission File**:

* Create a CSV file with two columns: SubjectID and PredictedLabel (0 for healthy, 1 for AD).
* Ensure the SubjectID matches the identifiers provided in the test set.

**6. Discussion and Conclusion (10 points)**

a. **Biomedical Implications**:

* Interpret your findings in the context of AD diagnosis.
* Discuss how hippocampal structures contribute to AD prediction.

b. **Challenges and Limitations**:

* Reflect on any difficulties encountered during the assignment.
* Suggest potential improvements or alternative approaches.

**Submission Guidelines:**

* **Report:**
  + A concise report (maximum 5 pages, excluding references and appendices) documenting your methodology, results, and interpretations.
  + Include figures, tables, and code snippets as necessary.
  + Use IEEE or APA citation style for references.
* **Code:**
  + Submit all code files used in the analysis.
  + Ensure code is well-documented with comments.
* **Data Confidentiality:**
  + Do not share the data with unauthorized individuals.
  + Ensure compliance with all data protection regulations.

**Evaluation Criteria:**

* Understanding and Exploration: Clarity in data exploration and justification of preprocessing steps.
* Model Design and Implementation: Appropriateness of the CNN architecture and correctness of the implementation.
* Analysis and Interpretation: Depth of evaluation and insightful interpretation of results.
* Communication: Quality of writing, organization of the report, and adherence to formatting guidelines.

**Deadline:** [Insert Deadline Here]

**Academic Integrity:**

Adhere to the highest standards of academic integrity. Plagiarism or any form of academic dishonesty will not be tolerated.

**Additional Instructions:**

* You are encouraged to explore advanced techniques and discuss any novel approaches.
* Consider computational efficiency in your implementations.
* If you encounter any issues with the data, document them and explain how you addressed them.

**Resources:**

* Recent research articles on machine learning applications in neuroimaging.
* Documentation for machine learning libraries (e.g., scikit-learn, TensorFlow, PyTorch).
* Tutorials on handling and analyzing high-dimensional biomedical data.

**Good luck!**