Model To Predict Alzheimer's Disease Utilizing a Neural Network

University of Massachusetts Boston

Team 8:

Steven Carr (Student ID: 01993037), Jordan Tong (Student ID: 01867192), Karthikmohan Jala (Student ID: 01973054), Avanith Kanamarlapudi (Student ID: 02181663), Davis Jordan (Student ID: 01991124), HariKishanRao Madhavaram (Student ID: 02147792)

Presentation Date: December 11th, 2024

Submission Date: December 21st, 2024

Abstract

Alzheimer's Disease is a dreadful form of Dementia that affects an individual's cognitive ability- that is, affects a person's ability to perform tasks or process information. We seek to build a machine learning model which can predict Alzheimer's Disease in a patient, given information similar to that of our dataset. Our methodology involves using a neural network architecture in the Python programming language to train on our dataset and subsequently perform predictions on our test dataset. After successfully building our model, we were able to predict whether a patient in our test dataset had Alzheimer's Disease or not to an accuracy of 85%. Although this is a high percentage of accuracy, we concluded that because there is barely any sort of correlation between the features in our dataset and whether or not an individual has Alzheimer's Disease, it is unsure whether our model would currently assist that much in a real-world application outside of our dataset.

Table Of Contents

1.0. Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that affects the brain, resulting in changes in behavior, memory loss and cognitive decline (*What You Need to Know About Alzheimer's and Brain Health,* 2023). AD is currently the most common form of Dementia, which is a term which contains a variety of cognitive impairments. The terminology "progressive" is used with respect to AD to show it is a disease which can become progressively worse over time.

1.1. Background and Motivation

Due to the progressive nature of the disease, accompanied by the fact that there is currently no cure, it is essential that AD is able to be detected as early as possible, so that early treatment can mitigate how bad the disease could potentially become. Theoretically, the best way to detect a disease in its earliest stages would be to use an Artificial Intelligence (AI) model to take in a series of information about a patient and having the ability to accurately predict whether this patient has the disease. The predictive nature can be useful to assist in the diagnosis of patients regardless of the reason they are being seen to ensure they are afforded the best possible care at that moment.

1.2. Problem Statement

Utilizing a neural network, we aim to develop a machine learning model capable of being trained on a medical dataset with information regarding a given individual's medical history and their corresponding Alzheimer's Disease diagnosis (that is, whether they have been diagnosed).

1.3. Objectives

- Accurately predict whether a patient has Alzheimer's Disease based on their medical history
- Provide insight into how AI can assist in diagnosing during the early stages of Alzheimer's Disease

---

2.0. Literature Review

2.1. Related Work

A study was conducted titled "Machine Learning Predictive Models Can Improve Efficacy of Clinical Trials for Alzheimer's Disease", which was overseen by Ali Ezzati, Richard B Lipton and the Alzheimer's Disease Neuroimaging Initiative (Ezzati et al, 2021). Their work sought to efficiently predict progression in Alzheimer's Disease (worsening cognitive decline) in patients who suffer from mild to moderate symptoms of the disease. Data was collected from 202 patients, where they used machine learning classifiers to differentiate between declining cognitive function and those with stable cognitive function. The results of their research showed their classifier had a positive predictive value of 80.8%, concluding that machine learning models can indeed be an effective tool used in clinical trials.

While the research from Ezzati, Lipton and the Alzheimer's Disease Neuroimaging Initiative is similar to our objective, we seek to assist in the diagnosis process itself, as opposed to their research in detecting which patients who already have Alzheimer's Disease had declining cognitive ability compared to others.

3.0. Methodology

3.1. Data Description

Our data contains information on 2,149 different patients, all anonymous due to regulations around an individual's medical history. The types of information in this dataset are as follows:

- PatientID, a unique identifier for each patient due to the anonymity of our data
- Age, ranging from 60 to 90 years old
- Gender, with a 0 representing male and 1 representing female
- Ethnicity, with 0 representing Caucasian, 1 representing African American, 2 representing Asian and 3 representing "Other" (meaning, any other ethnicity besides the three previously named)
- EducationLevel, with 0 being None, 1 being High School, 2 being Bachelors and 3 being higher
- BMI (Body Mass Index), ranging from 15 to 40
- Smoking, with 0 representing a "No" (the patient does NOT smoke) and a 1 representing "Yes"
- AlcoholConsumption, which is based on a weekly period, ranging from 0 to 20
- PhysicalActivity, which is based on a weekly period (in hours), ranging from 0 to 10
- DietQuality, ranging from 0 to 10
- SleepQuality, ranging from 4 to 10
- FamilyHistoryAlzheimers, where 0 indicates "No" (there is no family history) and 1 indicates a "Yes"
- CardiovascularDisease, where 0 indicates "No" and 1 indicates "Yes"
- Diabetes, where 0 indicates "No", and 1 indicates "Yes"
- Depression, where 0 indicates "No" and 1 indicates "Yes"
- HeadInjury, which represents a history of a head injury, with 0 representing "No" and 1 representing "Yes"
- Hypertension, where 0 indicates "No" and 1 indicates "Yes"
- SystolicBP, which represents Systolic Blood Pressure, ranging from 90 mmHg (millimeters of mercury) to 180 mmHg
- DiastolicBP, representing Diastolic Blood Pressure, ranging from 60 mmHg to 120 mmHg
- CholesterolTotal, which is the patient's total cholesterol levels, ranging from 150 mg/dL (milligrams per deciliter) to 300 mg/dL
- CholesterolLDL, the low-density lipoprotein cholesterol levels, ranging from 50 mg/dL to 200 mg/dL
- CholesterolHDL, the high-density lipoprotein cholesterol levels, ranging from 20 mg/dL to 100 mg/dL

- CholesterolTriglycerides, triglycerides levels, ranging from 50 mg/dL to 400 mg/dL
- MMSE, Mini-Mental State Examination score, ranging from 0 to 30, with a lower score representing a higher cognitive impairment
- FunctionalAssessment, ranging from 0 to 10, with a lower score indicating a greater functional impairment
- MemoryComplaints (the presence of), with 0 indicating "No" and 1 indicating "Yes"
- BehavioralProblems (the presence of), with 0 indicating "No" and 1 indicating "Yes"
- ADL, which is Activities of Daily Living score, ranging from 0 to 10, with a lower score indicating a greater impairment
- Confusion (the presence of), with 0 indicating "No" and 1 indicating "Yes"
- Disorientation (the presence of), with 0 indicating "No", and 1 indicating "Yes"
- PersonalityChanges (the presence of), with 0 indicating "No", and 1 indicating "Yes"
- DifficultyCompletingTasks, with 0 indicating "No", and 1 indicating "Yes"
- Forgetfulness (the presence of), with 0 indicating "No", and 1 indicating "Yes"
- Diagnosis (of Alzheimer's Disease), with 0 indicating "No" (the patient ***does not*** have the disease) and 1 indicating "Yes" (the patient ***does*** have Alzheimer's Disease).
- DoctorInCharge, this feature has the same value for every patient, which is "XXXConfid" due to anonymity

3.1.1. Data Source

Our dataset to train and test our model was obtained via a website named *Kaggle* (Kharoua, 2024).

3.1.2. Data Preprocessing

Within our dataset, the PatientID and DoctorInCharge features were determined to be unnecessary, because these values don't affect the resulting diagnosis of Alzheimer's Disease. The data was normalized using "StandardScaler()", which originates from the Scikit-learn library for preprocessing data. Utilizing StandardScaler ensures all our input features have a mean of 0 and a standard deviation of 1, which allows our model to efficiently converge during training. The process of normalizing our data also prevents features we may have with larger ranges from dominating other features which have smaller ranges. We then took our dataset and divided it, with 80% being used for training our model and the remaining 20% used for testing our model's efficiency in predicting.

3.2. Model Selection

3.2.1. Algorithms and Techniques

As previously stated, we decided to build a neural network model to feed our training and test data into. Neural networks are suited well for our goal due to their capabilities modeling non-linear relationships to show patterns between features that other models might be unable to otherwise. The customizability of a neural network allowed us to fine-tune parameters to achieve different results in accuracy. We also utilized different dropout rates during training, along with using ReLu and Sigmoid activation functions.

3.2.2. Justification

Utilizing ReLu and Sigmoid activation functions ensured that our model was able to effectively make a binary prediction based on our data (such as whether a given patient has Alzheimer's Disease). The changing of our dropout rates while altering different parameters also allowed us to account for any overfitting problems that might have arisen.

3.3. Implementation Details

3.3.1. Tools and Libraries

- Pandas library – used for data manipulation and analysis
- NumPy library – numerical operations; mathematical computations
- PyTorch library – the deep learning library used to build and train our neural network model
- PyTorch Data Utilities – to create datasets and data loaders for batch processing
- Matplotlib – visualizing our data, such as plotting learning curves, confusion matrices and ROC-AUC curves
- Seaborn – an additional visualization library for heatmaps and feature correlations
- Scikit-learn – For splitting our dataset, normalizing our data, and assisting in performance evaluation through metrics

3.3.2. Hyperparameters, Training, Fine-Tuning

- Hyperparameters:
  - Learning Rate (.001 – step size for weight updates for training)
  - Epochs (initially 30 – the number of times the model goes over the entire dataset)
  - Batch size (the number of samples processed before the model updates its parameters, set to 32)
  - Dropout rates (initially set to .4 and .3)

- Training:
  - Utilized Binary Cross-Entropy loss function (good for binary classification)
  - Adam Optimizer chosen for adaptability and effective gradient updates

- Fine-Tuning:
  - Initially, our model had 128 neurons in its first layer, and 64 in its second layer. These were changed to 256, 128 and a third layer was added with 64.
  - Epochs were adjusted from 30 to 90 to attempt a higher accuracy
  - Increasing the epochs to 90 led to an overfitting problem with our results, so the dropout rates were adjusted in each layer to .6, .5 and .4 respectively to account for this problem

4.0. Experiments and Results

4.1. Experimental Setup

- Hardware:
  - Virtual Environment via Google Colab – the free version has a default CPU that is an Intel Xeon CPU with two virtual CPUs and 13 GB of RAM. Also included for free is a default GPU which is a NVIDIA Tesla K80 with 12 GB of VRAM.

- Software:
  - Python 3 runtime environment (Google Colab)
  - Pre-installed libraries (stated previously)
  - Interactive python script interface

4.2. Performance Metrics

- Accuracy – how correct our model's predictions were

- Precision – proportion of our model's true positive in comparison to all predicted positives (avoid false positives)

- Recall – proportion of positives that were correctly identified (avoid false negatives)

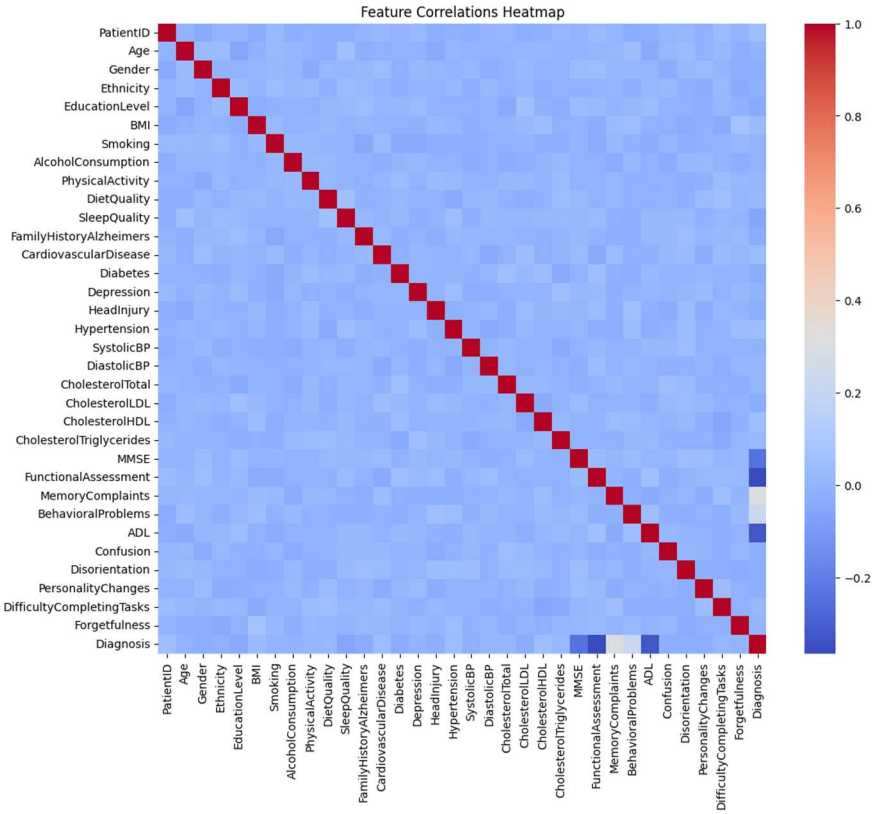
- F1-Score – Harmonic mean of precision and recall

- Confusion Matrix – Table showing our model's count of true positives, true negatives, false positives and false negatives

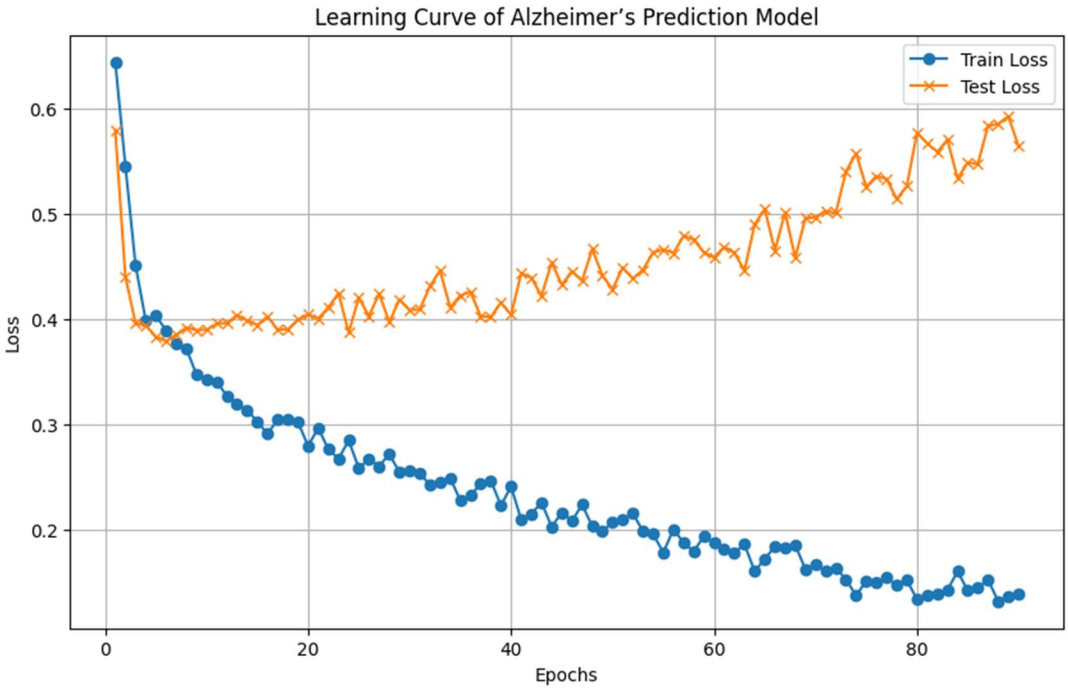
- ROC-AUC Curve (Receiver Operating Characteristic/Area Under the Curve) – plots the true positive rate against the false positive rate at different thresholds to summarize the model's overall performance

4.3. Results

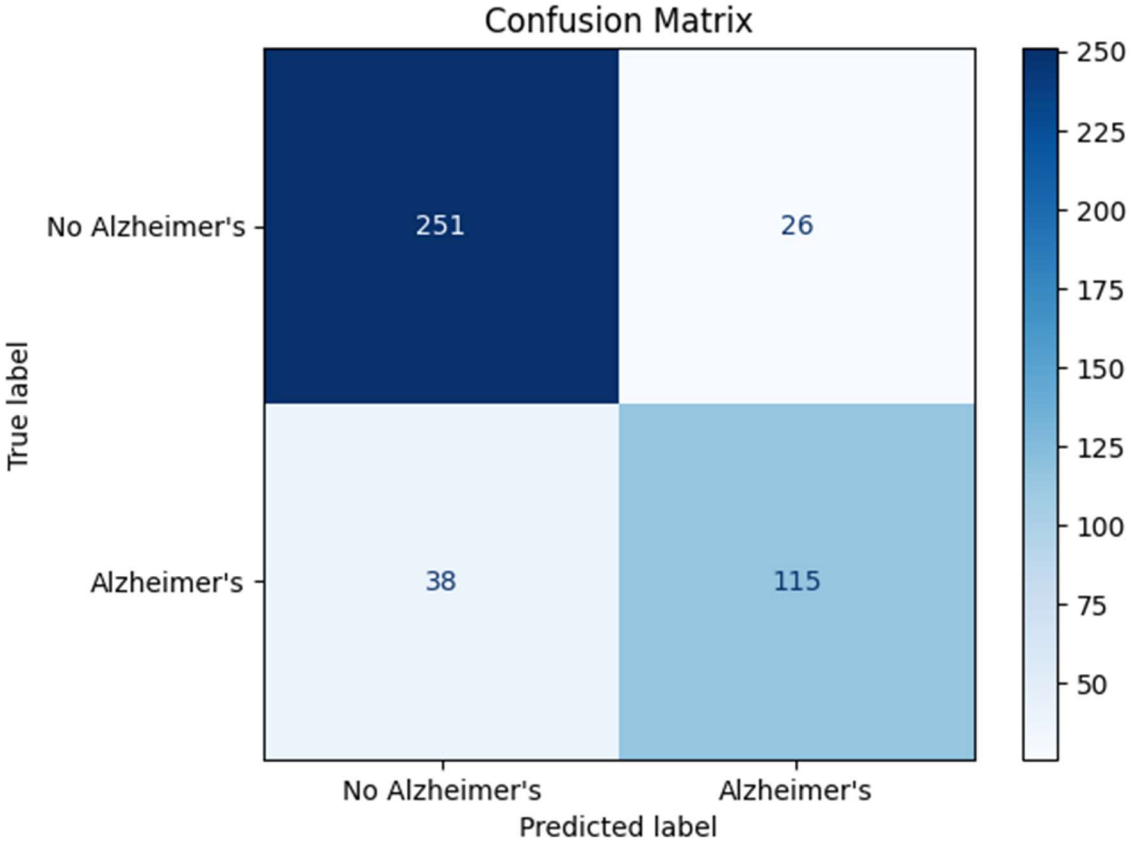
Feature Correlation Heatmap



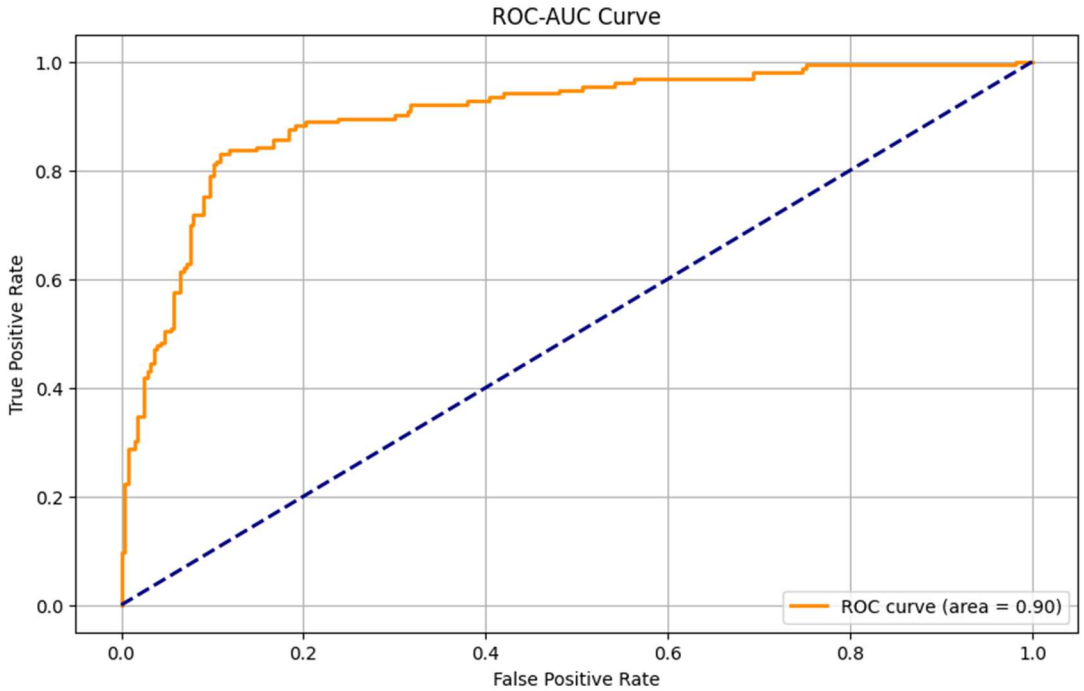
Learning Curve



Confusion Matrix



ROC-AUC Graph



Classification Report

```
=== Classification Report ===
                precision    recall  f1-score   support

No Alzheimer's       0.87      0.91      0.89       277
   Alzheimer's       0.82      0.75      0.78       153

      accuracy                           0.85       430
     macro avg       0.84      0.83      0.83       430
  weighted avg       0.85      0.85      0.85       430
```

4.4. Comparison with Baselines

- The only "baseline" experiment we could compare our result to was that of Ezzati, Lipton and the Alzheimer's Disease Neuroimaging Initiative, where they achieved an accuracy of about 80% while our model achieved an accuracy of 85%. However, in our objective there are harder obstacles we must overcome, and I will explain later on why our model might not be quite that effective in predicting Alzheimer's in a patient even though our accuracy is 85%.

- While we achieved an accuracy of about 85%, it is possible we could have achieved a higher accuracy using a different machine learning model or technique, such as:
  - Logistic Regression (effective for binary classification)
  - Random Forest (multiple decision trees, great against overfitting)
  - Support Vector Machine (SVM – effective for high dimensional spaces)

4.5. Analysis

- Heatmap:
  - The red diagonal line represents a one-to-one correlation with features and themselves
  - The only other correlation we can potentially observe is an Alzheimer's Disease diagnosis and MMSE, Functional Assessment, Memory Complaints, Behavioral Problems and ADL – all things that can be linked to symptoms or showings of Alzheimer's Disease in a patient

- Learning Curve:
  - Our model encounters a little bit of overfitting, but the overfitting used to be much worse before adjusting the dropout rates – previously, the overfitting had the test loss towards the end being greater than 1, but after adjusting the dropout the test loss stayed under .6

- Confusion Matrix:
  - 251 patients predicted to not have Alzheimer's Disease

    - 26 patients incorrectly predicted to have Alzheimer's when they did not
    - 38 patients incorrectly predicted to not have Alzheimer's when they did
    - 115 patients correctly predicted to have Alzheimer's Disease

- ROC-AUC Graph:
    - Shows our model's true positive rate in comparison to false positive – the fact that our true positive rate stays high above false positive is a good indication of our model's effectiveness in its predictability

- Classification Report:
    - Our model has an accuracy of approximately 85%. However, because we can see that there is virtually no correlation between our provided features and an Alzheimer's Disease diagnosis, it is hard to say that our model would be effective in real-world applications, or if it is able to simply study our dataset efficiently well.
    - There seemingly being no correlation between features also lines up with our current understanding of Alzheimer's Disease – that is, there is still much that needs to be discovered about it to fully understand the causes

---

5.0. Discussion

5.1. Insights

Strictly speaking from a data analysis standpoint, our model was able to accurately predict whether a patient did or did not have Alzheimer's Disease at 85%. However, due to the nature of the disease and what the underlying causes are or potential factors in being diagnosed, it is unclear whether our model would be able to accurately predict the disease in patients for real-world applications

5.2. Limitations

The limitations in our current knowledge of Alzheimer's Disease in comparison to our results from our model leave us unsure of the effectiveness of our model in professional use.

5.3. Future Work

As stated previously, improvements in our model could potentially be made by using a different machine learning approach (with respect to the techniques listed above) as well as a better understanding of the underlying causes of Alzheimer's Disease.

6.0. Conclusion

6.1. Summary of Findings

- Neural Network model used to learn patterns in predicting Alzheimer's Disease in patients

- Accuracy of 85% achieved (among predicting Alzheimer's in patients and Alzheimer's not being in patients)
- Lack of correlation between features shows we're unsure of the practicality of our model in real-world applications

6.2. Contributions

- Steven Carr (Team Lead, Dataset Researcher, Programming Contributor, Presentation Contributor, Main Report Developer)
- Jordan Tong (Programming Contributor, Architecture Fine-Tuning, Presentation Contributor, Report Contributor)
- Karthikmohan Jala (Programming Contributor, Architecture Fine-Tuning, Presentation Contributor, Report Contributor)
- Avanith Kanamarlapudi (Main Hands-On Programmer, Architecture Fine-Tuning, Presentation Contributor, Report Contributor)
- Davis Jordan (Programming Contributor, Architecture Fine-Tuning, Presentation Contributor, Report Contributor)
- HariKishanRao Madhavaram (Programming Contributor, Architecture Fine-Tuning, Presentation Contributor, Report Contributor)

7.0. References

Cloud, S. (2023, June 12). *What's the Hardware Spec for Google Colaboratory.* SaturnCloud. https://saturncloud.io/blog/whats-the-hardware-spec-for-google-colaboratory/#:~:text=CPU%20and%20RAM,-The%20CPU%20(Central&text=The%20default%20CPU%20for%20Colab,vCPUs%20and%20624GB%20of%20RAM

Ezzati, A., Lipton, R. B., & Alzheimer's Disease Neuroimaging Initiative. (2021, January 1). *Machine Learning Predictive Models Can Improve Efficacy of Clinical Trials for Alzheimer's Disease*. National Library of Medicine. https://pmc.ncbi.nlm.nih.gov/articles/PMC7201366/#:~:text=In%20this%20study%20we%20showed,trial%20of%20AD%20(LFAN)

Kharoua, R. E. (2024). *Alzheimer's disease dataset*. Kaggle. https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset/data

What is Overfitting? (2021, October 15). IBM. https://www.ibm.com/think/topics/overfitting#:~:text=In%20machine%20learning%2C%20overfitting%20occurs,of%20the%20machine%20learning%20model

What You Need to Know About Alzheimer's and Brain Health. (2023, June 14). Rappahannock Westminster-Canterbury. https://rw-c.org/alzheimers-brain-health/?gad_source=1&gclid=CjwKCAiAjeW6BhBAEiwAdKltMg7W-EsygLXskgm49QpnQ2wgeic_XNV6s5mUxVTMnUPqNsR4eOmVNBoCzVEQAvD_BwE

7.1. Other Sources

ChatGPT – assisted in fundamental understanding of our neural network architecture and data analysis

Google Colaboratory – free Python environment provided by Google