Syslab RES -Yilmaz
Name: Simon Thomas
Period: 1

Please have an entry for each day we meet even if you are absent (makeup as soon as you are back). Check our meeting days using the ION calendar.
If you work on your project on the days we don't meet, you can always add extra entries.
!!! Newest entry should be at the top !!!
Each entry is due at the end of each class.
More detail is better than less detail and vague statements.
Please add "Journal" subdirectory under the shared google drive folder, which you will share with me (i.e. give me edit access "syilmaz@fcpsschools.net").
4/24/25

- Tried to make datasets for CNN and NLP match
- Made weighted average function easier to use
    - Clear print functions in terminal
- Worked on poster

4/22/25 (absent)

4/21/25 (absent)

4/10/25

- Made weighted average code
    - Four inputs
    - One input of CNN accuracy
        - CNN weightage input (doctor preference)
    - One input of NLP accuracy
        - NLP weightage input (doctor pref)

4/8/25 (absent)

4/7/25 (absent)
`

4/3/25 (absent)

4/2/25

- I changed it to 100 test cases for positive template, and 100 cases for negative template (rather than 50, and 50)
    - Each test case has age, gender, doctors notes (and medicine treatment reaction, a factor I'm experimenting with)
    - Also has family history
- .012 validation loss
    - For test patient that I intended to be a positive Schizophrenia test case, the model predicted a 0.989 probability of Schizophrenia

3/27/25

- Updated Presentation
- Created Example Correlation Matrix
    - Understood how correlation matrix of regions of interest works
        - Regions of Interest within the image are selected by the scientist, according to what parts of the brain may have a relationship with schizophrenia
        - Then a correlation matrix is made, with the 20 regions on both the x and y axis
            - Correlation is measured by darkness of color overlap
                - Darkest color means 1:1 correlation
- Started creating code for weighted average
    - Most likely 50/50 weight for CNN and NLP, or I could do an input weight that the doctor wants for each algorithm

3/25/25

- Understood (more in depth) the CNN dataset (input features)
    - The fMRIs are separated into 20 Regions of Interest (of the brain,) and then a correlation matrix is created with each region overlapping each other (creating the 400 input features)
    - Layers are fully connected (each neuron to each other neuron)
- Found a study that uses brain MRIs (to put into the other solutions table)
    - https://www.nature.com/articles/s41598-023-41359-z

3/24/25

- Learned CNN architecture of code
    - 3 Layers, 400 neurons, 64 neurons, 32 neurons
        - Tanh, Relu, sigmoid activation functions

3/20/25

- Worked on updating presentation
- Looked more at BioBERT structure

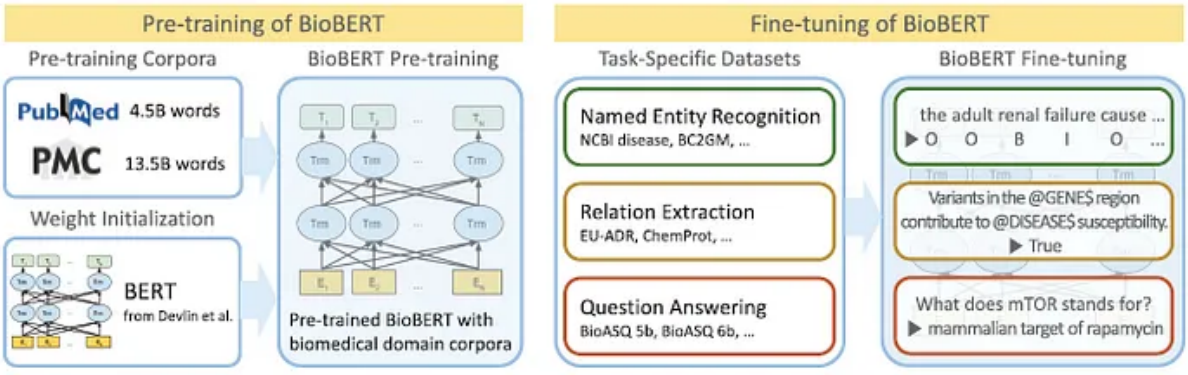
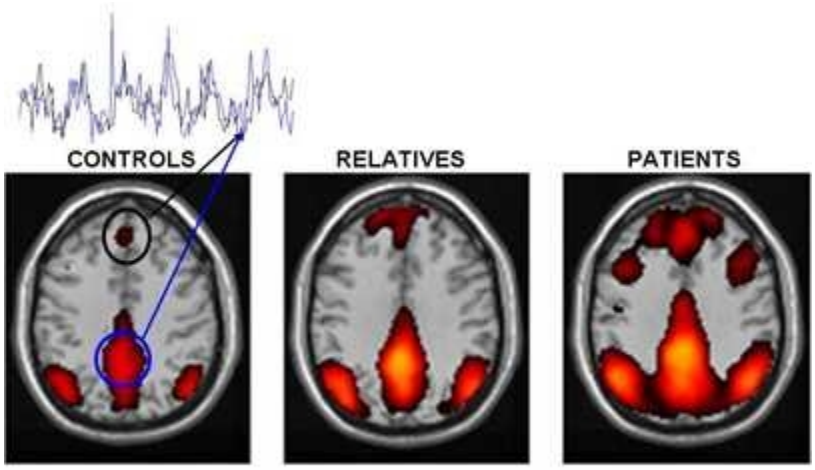


Fig. 1. Overview of the pre-training and fine-tuning of BioBERT

3/18/25

- https://www.frontiersin.org/journals/neuroimaging/articles/10.3389/fnimg.2023.1127508/full



    - https://news.mit.edu/2009/schizophrenia-0119
        - These fMRIs are what the CNN is analyzing, here you can see in the schizophrenia patients there is a large concentration in the front of the brain, while its lower in relatives and a lot lower in health controls.

3/17/25

- Worked on updating milestones progress
- Increased number of entries into NLP algorithm, tried accounting for the possibility for other mental illnesses being test cases (still trying this, using chatgpt and deepseek)

3/13/25

- Worked on fixing presentation
- Tried taking MRI images and displaying them on screen through external (online) application
    - https://mmvt.mgh.harvard.edu/gallery/

3/11/25

- I created 50 test cases for positive template, and 50 cases for negative template using deepseek
    - Each test case has age, gender, doctors notes
    - Also has family history
- .015 validation loss
    - For test patient that I intended to be a positive Schizophrenia test case, the model predicted a 0.9791 probability of Schizophrenia

3/6/25

- Did more research on the genetic factor that plays a big role in Schizophrenia being present within an individual or not (in order to potentially take family history into account for my AI assistant)
- https://www.nature.com/articles/s41380-021-01420-7
    - Schizophrenia is highly hereditable (~80%)
    - Environmental factors do still play a big role

3/4/25 (absent)

3/3/25

-

2/27/25

2/25/25

2/24/25

2/20/25 (absent)

- Combine and Add structured features

```
def combine_fields(example):
    example["text"] = f"Age: {example['age']}; Gender: {example['gender']}; Notes: {example['notes']}"
    return example

def add_structured_features(example):
    # Normalize age (divide by 100 so that typical ages fall in a smaller range)
    age_value = float(example["age"]) / 100.0
    # Encode gender: "Female" -> 1.0, "Male" -> 0.0
    gender_value = 1.0 if example["gender"].strip().lower() == "female" else 0.0
    example["structured"] = [age_value, gender_value]
    return example
```

2/18/25 (absent)

- Load in pre-trained BioBERT model
- Use BioBERT tokenizer

```
model_name = "dmis-lab/biobert-base-cased-v1.1"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = BioBERTWithStructuredData(model_name=model_name, num_labels=2, structured_dim=2)
```

```
train_dataset = Dataset.from_list(train_data)
test_dataset = Dataset.from_list(test_data)
```

2/13/25 (absent)

- Keep dataset random
- Make train and test data

```
 # Shuffle the dataset to mix positive and negative examples.
    random.shuffle(data)
    return data

# Generate synthetic data: 300 positive and 300 negative examples (total
600)
synthetic_data = generate_synthetic_data(n_positive=300, n_negative=300)

# Split the synthetic data into training and testing sets.
train_data = synthetic_data[:100]
test_data = synthetic_data[100:200]
```

2/11/25

- Need to have healthy samples for patients too
    - Negative samples

```
    # Generate negative examples.
    for _ in range(n_negative):
        age = int(round(random.gauss(45, 10)))
        age = max(30, min(age, 70))
        gender = "Male" if random.random() < 0.5 else "Female"
        notes = random.choice(negative_templates)
        data.append({"age": str(age), "gender": gender,
"notes": notes, "label": 0})
```

2/10/25

- Need to account for gender trends
    - Randomly generated patient ages and gender, with a tilt towards young males for Schizophrenia patients
    - Positive diagnosis

```
# Generate positive (schizophrenia) examples.
```

```
    for _ in range(n_positive):
        age = int(round(random.gauss(25, 5)))
        age = max(15, min(age, 40))
        # Bias toward Male.
        gender = "Male" if random.random() < 0.7 else "Female"
        notes = random.choice(positive_templates)
        data.append({"age": str(age), "gender": gender,
"notes": notes, "label": 1})
```

2/6/25 (absent)

- Schizophrenia frequency
    - Male vs Female.
    - http://pmc.ncbi.nlm.nih.gov/articles/PMC3420456/#:~:text=Using%20standard%20diagnostic%20criteria%20in,differences%20%5B5%2C%206%5D.

2/4/25 (absent)

"This paper tries to summarize the most important findings in gender differences in schizophrenia and first-psychosis episodes. Several studies indicate that the incidence of schizophrenia is higher in men. Most of the studies found the age of onset to be earlier in men than in women."

- http://pmc.ncbi.nlm.nih.gov/articles/PMC3420456/#:~:text=Using%20standard%20diagnostic%20criteria%20in,differences%20%5B5%2C%206%5D.

2/3/25

- Couldn't find desired dataset yet
    - Need a different dataset for now
- Started creating synthetic dataset (fake data)
    - "Doctor's Notes":

```
def generate_synthetic_data(n_positive, n_negative):
    positive_templates = [
        "Patient reports hearing voices and experiencing
hallucinations.",
        "Patient exhibits delusions and disorganized speech.",
        "Patient suffers from auditory hallucinations and signs of
psychosis.",
        "Patient has reported seeing things that are not there.",
        "Patient experiences frequent hallucinations and paranoid
delusions."
    ]
    negative_templates = [
        "Patient appears stable with no signs of hallucinations or
delusions.",
```

```
        "Patient reports no unusual sensory experiences or
disorganized behavior.",
        "Patient's mental state is normal with clear, coherent
thought.",
        "Patient exhibits stable mood and no symptoms of
psychosis.",
        "Patient is functioning normally with no signs of auditory
hallucinations."
```

1/27/25

- All versions of BioBERT significantly outperformed BERT and the state-of-the-art models,
    - BioBERT v1.1 (+PubMed) obtained a strict accuracy of 38.77, a lenient accuracy of 53.81 and a mean reciprocal rank score of 44.77, all of which were micro averaged.

1/23/25

- BioBERT performs much better the more that it is pre-trained, it is trained on ~4.5 billion words
- Test results of evaluation of BioBERT on biomedical named entity recognition
    - BioBERT achieves higher scores than BERT on all the datasets. BioBERT outperformed the state-of-the-art models on 6 out of 9 datasets.
    - BioBERT v1.1 (+ PubMed) outperformed the state-of-the-art models by 0.62 in terms of micro averaged F1 score.

1/21/25

- Training Steps of BioBERT
    - https://sh-tsang.medium.com/brief-review-biobert-a-pre-trained-biomedical-language-representation-model-for-biomedical-text-4b5cf07efdd7
        - First, BioBERT is intialized with weights from BERT, which was pretrained on general domain corpora (English Wikipedia and BooksCorpus).
        - Then, BioBERT is pre-trained on biomedical domain corpora (PubMed abstracts and PMC full-text articles).
        - Finally, BioBERT is fine-tuned and evaluated on three popular biomedical text mining tasks (NER, RE and QA).

1/16/25

- Continued research on methods of biomedical text mining
    - Word2Vec, ELMo, and BERT are all usually trained on datasets containing general domain texts (not texts containing biological terms)
    - BioBERT uses wordpiece tokenization
        - new words can be represented by frequent subwords (e.g. Immunoglobulin => I ##mm ##uno ##g ##lo ##bul ##in
        - Cased vocabulary results in better performances

1/14/25

BioBert methods:

- Named Entity Recognition
    - Recognizing numerous domain-specific nouns
- Relation Extraction
    - Classification of relations between named entities
- Question Answering
    - answering questions posed in natural language given related passages. strict accuracy, lenient accuracy and mean reciprocal rank (MRR) are reported
- BioBERT outperforms BERT completely
    - Recognized biomedical named entities which BERT cannot

1/13/25

- NLP for health records
    - Study on using NLP for analysis of patient health records
    - Deep Learning NLP
    - Potential for inaccuracy and misclassification of outcomes
- Bio+ClinicalBERT
    - Model is trained on biomedical literature and identified medical records
    - https://sh-tsang.medium.com/brief-review-biobert-a-pre-trained-biomedical-language-representation-model-for-biomedical-text-4b5cf07efdd7

12/19/24

- Continued research on NLP analysis of words
    - https://www.nature.com/articles/s41746-022-00589-7
    - https://www.sciencedirect.com/science/article/pii/S0920996422002742
    - Language disorders serve as a prime early indicator of schizophrenia
    - Language incoherence and disorganized speech can be measured
        - Potential language barrier for algorithm
- Listened to Presentations and Presented

12/17/24

- Did research on possible methods for NLP analysis of words
    - https://pmc.ncbi.nlm.nih.gov/articles/PMC10499191/#:~:text=Certain%20linguistic%20features%20computed%20by,and%20symptom%20severity%20in%20schizophrenia.
    - Google BERT
        - Able to make predictions about upcoming sentences
        - Would be useful for schizophrenia diagnosis as there is some level of unexpectedness with the sentences of those diagnosed with schizophrenia
    - Schizophrenia linguistic markers are highly variable, and are not necessarily a reliable method of making an accurate diagnosis of a patient
- Listened to presentations

12/16/24 (absent, did makeup work)

- Error in data processing for some reason, needed to fix
    - Quickly remade data prep code, and used source_path between cobre dataset and my code

- Code for Data Prep, iterate/read through COBRE fmri files:

```
def get_paths_and_labels(csv, source_path):

    binary_class = []
    filepaths = []
    for key, value in csv.iterrows():
        binary_class.append(value[1])
        subject_path = source_path + 'fmri_' + value[0][:-1] + '_session1_run1'
        fmri_path = subject_path + '.nii.gz'
        filepaths.append(fmri_path)
    return filepaths, binary_class

```

12/12/24

- Created source path from cobre dataset to vscode
  - COBRE mris are being processed

```
class CobreDataset(Dataset):
    def __init__(self, data, labels):
        self.x = torch.from_numpy(data)
        self.y = torch.from_numpy(labels)
        self.n_samples = labels.shape[0]
        self.targets = labels

    def __getitem__(self, index):
        return self.x[index], self.y[index]

    def __len__(self):
        return self.n_samples
```

- Increased number of EPOCHs to 30
- Overall neural network accuracy at ~90%, 15 correct healthy, 13 correct schizophrenia, 1 incorrect healthy, 2 incorrect schizophrenia
- Listened to presentations

12/10/24

- Implemented second part of training function
  - Calculating accuracies
  - Using plt to show accuracies
  - Showed losses of training, and accuracies of training

```
valid_loss = 0.0
```

```
            len_val_data = 0
            correct = 0
            for i, (inputs, labels) in enumerate(val_batches):
                outputs = model(inputs.cuda())
                loss = loss_f(outputs, labels.cuda().unsqueeze(1))
                optimizer.zero_grad()
                for o, l in zip(outputs, labels):
                    if o >= 0.5:
                        b_o = 1.0
                    else:
                        b_o = 0.0
                    if b_o == l:
                        correct += 1
                valid_loss += loss.item()
                len_val_data += batch_size
            accuracy = correct / len_data
            print(valid_loss / len_val_data)
            print("Val accuracy:", accuracy)
            val_acc.append(accuracy)
            val_losses.append(valid_loss / len_val_data)

        save_model(model)
        plt.plot(losses, label='train loss')
        plt.plot(val_losses, label='validation loss')
        plt.legend()
        plt.show()

        plt.plot(accuracies, label='train accuracy')
        plt.plot(val_acc, label='validation accuracy')
        plt.legend()
        plt.show()
        return model
```

- Listened to presentations

12/9/24

- Started implementing training function
- Need to account for accuracies, and losses

    - Losses are the errors between the model's predictions and the actual training data
- Learned from code of past convolutional neural networks

```
def training():
    losses = []
    val_losses = []
    accuracies = []
    val_acc = []
    for epoch in range(num_epochs):
        print('EPOCH:', epoch + 1)
        model.train(True)
        running_loss = 0.0
        len_data = 0
        correct = 0
        for i, (inputs, labels) in enumerate(train_batches):

            outputs = model(inputs.cuda())
            loss = loss_f(outputs, labels.cuda().unsqueeze(1))
            optimizer.zero_grad()
            loss.backward()
            optimizer.step()
            for o, l in zip(outputs, labels):
                if o >= 0.5:
                    b_o = 1.0
                else:
                    b_o = 0.0
                if b_o == l:
                    correct += 1
            running_loss += loss.item()
            len_data += batch_size
        accuracy = correct / len_data
        avg_loss = running_loss / len_data
        print(avg_loss)
        print("Accuracy:", accuracy)
        accuracies.append(accuracy)
        losses.append(avg_loss)
```

- `model.train(False)`

Listened to presentations

12/5/24

- Started learning how to create training function for neural network
    - Neural networks go through many passes
    - Forward and backward passes
        - https://stackoverflow.com/questions/4752626/epoch-vs-iteration-when-training-neural-networks#:~:text=In%20the%20neural%20network%20terminology,memory%20space%20you'll%20need.
        - EPOCHS
            - Number of times the algorithm sees the entire dataset
            - More epochs, the higher the accuracy

12/3/24

- Learning how to create convolutional neural networks
    - https://www.geeksforgeeks.org/building-a-convolutional-neural-network-using-pytorch/
    - Created my google collaborate
    - Imported important libraries & other resources
        - Pytorch tools
        - Pandas
        - Matlab
        - Confusion matrix
    - Created vscode version of code too
        - Google collab is a bit new to me

12/2/24

- Old dataset (Schizconnect) not sending me the data I requested
    - Won't have data that includes demographic information (like age and gender) and mri data
    - Found new temporary dataset with just mri data
    - https://figshare.com/articles/dataset/COBRE_preprocessed_with_NIAK_0_12_4/1160600

11/30/24

● Hippocampus dataset preprocessing investigation
○ Implemented z-score normalization for MRI images (Nyúl & Udupa, 1999, IEEE Transactions on Medical Imaging)
○ Developed normalization script standardizing image dimensions to 256x256 pixels
○ Specific preprocessing challenges identified in multimodal neuroimaging datasets
● CNN architecture research for medical image segmentation
○ Deep dive into U-Net architecture from Ronneberger et al. (2015)
○ Analyzed potential adaptations for schizophrenia-related feature detection
○ Compared segmentation performance across different network configurations

11/27/24

● Neuroimaging dataset preprocessing comparative analysis
○ Quantified image quality metrics using precise statistical methods
○ Developed comparative matrix for COBRE and MCICShare datasets
○ Specific preprocessing challenges:
- Intensity normalization variations
- Spatial resolution inconsistencies
● Data augmentation strategy development
○ Implemented rotation and noise augmentation techniques
○ Referenced medical image augmentation methods from Shorten & Khoshgoftaar (2019)
○ Created script to generate synthetic training samples while maintaining diagnostic integrity
11/25/24
● NLP approach for medical record integration
○ Comparative analysis of BERT and BioBERT model performance
○ Extracted medical terminology using specialized NLP techniques
○ Referenced Zhang et al. (2019) medical term extraction methodology
● Ethical AI framework investigation
○ Detailed analysis of "Computing Schizophrenia" ethical challenges article
○ Developed preliminary ethical compliance checklist for AI diagnostic tools
○ Identified key ethical considerations in machine learning-based diagnosis
11/24/24
● Comprehensive Kaggle hippocampus dataset analysis
○ Precise dataset composition breakdown:
- Total images: 5,921
- Subject count: 421
- Patient distribution: ~60% schizophrenic, ~40% healthy controls
○ Developed statistical analysis script for dataset characterization
○ Created visualization of demographic feature distributions
11/21/24
● Multi-modal data integration strategy development
○ Created feature extraction pipeline combining MRI and textual data
○ Implemented feature fusion technique referencing Nie et al. (2020) multimodal learning approach
○ Developed modular approach for cross-modal feature representation
○ Investigated challenges in maintaining feature semantic integrity
11/19/24
● CNN architecture comparative analysis
○ Performance evaluation of ResNet, Inception, and DenseNet architectures
○ Referenced seminal papers by He et al. (2016), Szegedy et al. (2015)
○ Developed benchmark script for architecture performance comparison
○ Analyzed architectural adaptations for neurological image classification
11/18/24
● Alternative schizophrenia dataset investigation
○ Comprehensive comparison of available neuroimaging datasets
○ Detailed analysis of Kaggle hippocampus dataset potential

○ Created comparative matrix of dataset characteristics
○ Investigated licensing and ethical considerations for data usage
11/14/24
Dataset:
https://www.kaggle.com/datasets/andrewmvd/hippocampus-segmentation-in-mri-images

11/12/24

- SchizConnect data download is taking too long
- Started looking into different data sets
    - Found a kaggle hippocampus data set, with both training and test images for schizophrenic and healthy brains
    - Link:

10/28/24
Presentation Revision (arranged an organized references system)
10/22/24
Presentation Revision (changed flow of slides, and elaborated on content)
10/21/24
School Catchup Day
10/7/24 - 10/21/24
Project Presentations

9/30/24 -10/7/24
Note: Had a lot of absences, so I don't remember the progress of individual days
Overall Progress of the Week:

- Finalized project Idea: AI Assistant for Schizophrenia diagnosis (using CNN and NLP)
- Dataset finalized (SchizConnect MRI imaging data)
- Consideration for using notes recorded by doctor in NLP, however data may be difficult to obtain so approach might be purely theoretical
- Started and finished Project Presentation 1

9/26/24 (absent 9/24/24)

- SchizConnect is back up, I requested the download of a large set of data
- Neuroimaging data, including other medical/general data such as gender, age, age of diagnosis, etc.
- Data comes from two organizations: COBRE and MCICShare
- I'm starting to think that maybe I'll focus my project on developing an "AI assistant" that will assist with the judgement and diagnosis of schizophrenic and non-schizophrenic patients
- https://www.cambridge.org/core/journals/psychological-medicine/article/computing-schizophrenia-ethical-challenges-for-machine-learning-in-psychiatry/B2E8D14AE7286E977BF9909E6DB4DF9E
    - Datasets might be too restricted to certain categories of patients, resulting in the issue where in unsupervised diagnosis of schizophrenia patients that are not the

same as the schizophrenia patients that were used in the datasets to train the AI, there may be large innaccuracies
    - Therefore it might be necessary to evaluate more factors, such as written medical records, through NLP
- The assistant might use regression, which would be evaluated by the doctors
- Use of a confusion matrix when evaluating the AI before its official use
    - This will help me to figure out if the AI should be trusted more in certain situations (such as the probability of a “true” diagnosis being correct is 99.9%, therefore if the doctor observed a false diagnosis yet the assistant asserted true, then the doctor would definitely want to reconsider
    -

9/23/24
Processing Brain Microstates Procedure:

Stage 1: Signal pre-processing for classification (Fig. 1).

1. Public dataset.
2. Signal pre-processing.
3. Transformation into microstates.

•

Stage 2: Descriptor generation (Fig. 2).

1. Random Walk generation.
    – Detrended fluctuation analysis (DFA).
2. Time series refinement
    – Phase 1: Decomposition of time series.
        * Segregation of the series into its components
        * Selection of the descriptor from the components obtained in (a).
    – Phase 2: Conversion.

*
Dickey Fuller (ADF).
*
Kwiatkowski–Phillips–Schmidt–Shin (KPSS).

•
Stage 3: Classifier (Fig. 2).
1.
Convolutional neural network (CNN).

9/19/24

- SchizConnect is down right now for some reason, I emailed a professor at USC regarding this
- COBRE dataset for schizophrenia MRI data of many patients: https://fcon_1000.projects.nitrc.org/indi/retro/cobre.html
- To access the data it requires me to create an NITRC account
- NITRC (has publicly available Neuroimaging datasets and resources): https://neuroscienceblueprint.nih.gov/resources-tools/blueprint-resources-tools-library/neuroimaging-tools-and-resources-collaboratory#:~:text=NITRC%20is%20a%20free%20one,data%20sets%2C%20or%20computing%20power.
    - I made an account, and requested to join the "1000 Functional Connectomes Project"
    - https://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html
- Article that has a project for automating Schizophrenia diagnosis using CNNs
    - https://www.sciencedirect.com/science/article/pii/S0957417422013835

9/17/24

- Used the query to search for visual data
    - Cross sectional MRI data
        - 5921 images, 421 subjects
        - COBRE: 1574 images from 198 subjects
        - MCICShare: 4347 images from 212 subjects
        - http://schizconnect.org/queries/21828
- Article about SchizConnect: https://pubmed.ncbi.nlm.nih.gov/26142271/

9/16/24

- Information Sciences Institute of USC
    - Large Schizophrenia related dataset, (containing brainscans which I will probably process in a convolutional neural network)
    - http://www.schizconnect.org/
- This dataset has a search query which allows me to filter through thousands of results

- Today I spent time trying to learn how to operate this website

9/12/24 (sick on 9/10/24)

- Microstate Analysis for processing EEG signals
    - Four microstates, all resting states
    - Resting-state auditory network
    - Resting-state visual network
    - Resting-state salience network
    - Resting state attention network
    - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11024310/#B29
- These sequences are usually evaluated in Schizophrenia diagnosis due to their rich pathological and semantic information
- An important consideration is the transferability of microstate templates between healthy individuals and SCZ patients, specifically, whether the same set of templates can effectively model EEG signals in both groups. At present, there is no publicly available research exploring this specific issue. Due to the similarity of microstates under different conditions, researchers typically model the EEG signals of healthy individuals and SCZ patients uniformly. Although this method can effectively reduce computational complexity, it overlooks the quality characteristics of microstate sequence.

9/9/24

- Starting slideshow
- Identifying the problem/State of the Art of mental illness right now
- About 1 in 5 adults in the US live with a mental illness (57.8 million in 2021) - (NIH) https://www.nimh.nih.gov/health/statistics/mental-illness
    - Schizophrenia affects approximately 24 million people worldwide - (WHO) https://www.who.int/news-room/fact-sheets/detail/schizophrenia
    - More than two out of three people with Schizophrenia do not receive specialist mental health care

9/5/24

- Specifics of Neural Networks https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414
    - Type of deep learning model
    - Modeled after the human brain, containing thousands of dense interconnected nodes
    - Learns to perform a task through training (processing large amounts of labelled data, and then being able to process unlabelled data)
    - Each node has a “weight,” as a node receives data the data is multiplied by the weight, and if the resulting number exceeds a certain threshold then the node fires (sends the number, which is usually a sum of all the weighted inputs)
    - Neural nets are Feed-Forward, data moves in one direction from the bottom input layer to the output layer
- Coding: Will probably use google colab
    - Using this dataset: https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/repod.0107441

    - Process the diagnosed schizophrenics as training data labelled as schizophrenic
    - Healthy brains will be labelled healthy
    - Then use another dataset with a mix of schizophrenic and healthy brains in order to test accuracy of predictions by the trained neural network

9/3/24

- Deep Neural Networks tend to be optimal for optimizing the accuracy of diagnosis of Schizophrenia
    - Highest specifically achieved using KNN (K nearest neighbors)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11024310/#B29

Table 4

The SCZ recognition accuracy of different features and classifier.

Studies	Feature sets	Classifier	Accuracy
Devia et al. (2019)	EEG activity	Linear discriminant analysis	71.00%
Baradits et al. (2020)	Microstate temporal parameters	Machine learning model	82.70%
Siuly et al. (2020)	EMD features	Ensemble bagged tree	89.60%
Akbari et al. (2021)	Phase space dynamic	*K*-nearest neighbor	94.80%
Kim et al. (2021)	Microstate temporal parameters	Support vector machine	75.60%
Lillo et al. (2022)	Microstate and microstate features	Convolutional neural networks	93.00%
Chen X. et al. (2023)	Linear and non-linear measures	Support vector machine	89.00%
This paper		*K*-nearest neighbor	97.20%

8/29/24

- Research on possible ML algorithms that could be used for training a model to recognize specific features of the brain that are present in Schizophrenic brains, yet differ from healthy brains - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232512/

    - Logistic Regression
    - Naive Bayes model (had the highest accuracy, however this entire study is based on the observations of the G72 protein, while I was originally focusing on EEG signals)
- Area Under the Curve (AUC) was used to compare the efficiency of the different models
- Just based off of google search, I'm seeing a lot of contradicting research regarding the implications of the G72 protein with schizophrenia diagnosis - some indicate that higher concentrations of G72 correlate with schizophrenic patients, some indicate that lower concentrations indicate this, whereas some actually indicate that there is no correlation whatsoever. This might be a point of interest - I could find out which correlation is actually true by doing my own research/simulation

8/27/24

- Consideration of a new "twist" to current research on the below project, the below project idea alone does not function as a project that I can effectively incorporate my own elements into
- One example of a fun idea (credit to Dr. Yilmaz) would be simulating the procedure that this project -NIH- follows, and evaluating whether this accuracy of the ML model that this research produced was legitimate, and any factors that they failed to take into account
    - This would involve my own creation and training of roughly the same NN that this research article depicts, in order to ensure fair comparison
- Another idea could be exploring the prospects of a ML diagnosis model which is based on a set of data other than EEG signals, and comparing the pros and cons of both models

8/22/2024

- One Project Idea: creating a model to accurately predict the presence of schizophrenia within a patient's brain based on EEG signals, using public datasets from past research
- Research specifically on EEG relation to schizophrenia: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11024310/#B29
- Dataset with EEG data of schizophrenic and healthy patients: https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/repod.0107441
- While creating a model like this would prove to be interesting, its benefits are relatively limited at the time, as there are ethical concerns revolving around the use of AI in healthcare (would a patient trust an AI with a 95% accuracy over an expert to provide a diagnosis for them?)

8/20/2024

- Background research on the intersection of Artificial Intelligence/Machine Learning and Neuroscience/healthcare
    - Systems Neuroscience: Research on how the brain implements a variety of cognitive functions - https://www.nature.com/articles/s41593-019-0520-2

        - Artificial Intelligence attempts to make computational systems based on a given task and data
        - Neural Networks (NN's) attempt to model neural computation, based on units that mimic individual neurons within the brain - largely valuable due to the recent rise in deep-learning within Machine learning
    - ML in Diagnosis/Prediction of clinical Schizophrenia - NIH
        - Models created using ML algorithms can analyze speech, behavior, and creativity for the diagnosis of mental disorders
        - Using electroencephalography (EEG) diagnosis of schizophrenia - not the typical method of diagnosis, however machine learning can significantly improve the accuracy of diagnosis using EEG signals (in this study it's stated that a hybrid neural network was built and trained to distinguish the EEG signals of a healthy person from someone with schizophrenia with 99.22% accuracy)