MCI Prediction from Audio Data

IGSA Internship Presentation
(Source Code Link)

Michael Murphy

Context

- **55 million +** people suffer from **Dementia** worldwide

- This number is expected to **increase** as the world's population grows in **size and age**

- **Mild cognitive impairment (MCI): early stage** of Dementia with symptoms such as mild memory loss

- Currently: **no cure** for Dementia, but **MCI** sometimes **reversible** if detected **early on**

The Task: Can we use machine learning for MCI detection?

- Using **AI methods** to detect the **early onset of MCI** could **save millions** of people worldwide from **irreversible cognitive decline**

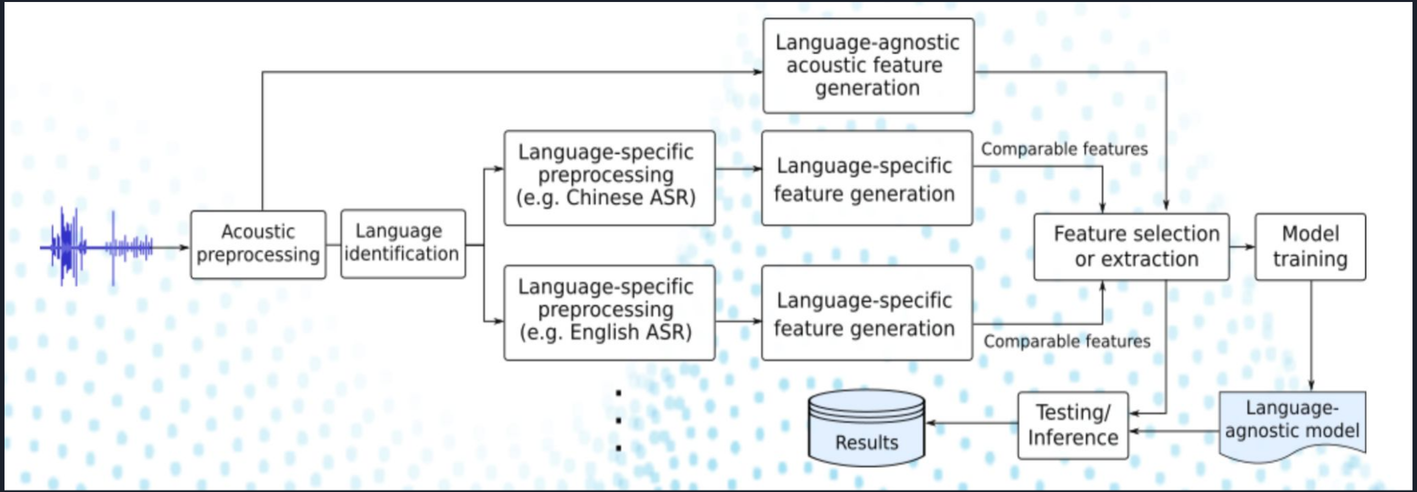
- Pros: potentially **cheaper**, more **accurate**, and more easily **scalable** than traditional diagnosis

- **ML research** has seen success in diagnosing Alzheimer's Disease and MCI using **Natural Language Processing (NLP)** and **speech data** from clinical trials

Amini S, Hao B, Yang J, et al. (2024) Prediction of Alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models. *Alzheimer's & Dementia*,. **doi:** 10.1002/alz.13886.
https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.13886

The Dataset: patient speech samples

- Data from the TAUKADIAL 2024 Interspeech Challenge
- “The training data set consists of spontaneous speech samples corresponding to audio recordings of picture descriptions produced by cognitively normal subjects and patients with MCI”
- We focused on ENGLISH ONLY



Sample Audio File (click below):

https://taukadial-luzs-69e3bf4b9878b99a6f03aea43776344580b77b9fe54725f4.gitlab.io/

Data Cleaning and Feature Extraction

- Various **speech features** were **extracted** from the audio data prior to training
- 3 categories of features: **acoustic, linguistic, and fluency** (see next slide for descriptions)
- 507 recordings: 169 patients, 3 recordings per patient

Goal: predict 'dx' (1 = MCI, 0 = Normal) using these features

	tkdname	mmse	dx	Total Duration	Mean Pitch	Jitter	Shimmer	General Silence	Mean Silence	Silence Abs Deviation	...	Word syllables 2	Repetition Frequency	Unique Word Count	Invented Word Count	Total Adjectives	Total Adverbs	Total Nouns	Total Verbs	Total Pronouns	Total Conjunction
0	002	29	0	122.440000	164.242554	0.017773	0.088550	114	0.313544	0.319138	...	14	0.0	83	52	21	9	68	59	39	15
1	002	29	0	45.660000	162.687280	0.021281	0.098618	56	0.197143	0.202367	...	8	0.0	45	23	3	8	31	29	7	7
2	002	29	0	62.690000	179.818570	0.021813	0.098348	35	0.733257	0.874998	...	4	0.0	43	37	4	4	30	35	14	5
3	003	23	1	21.691521	111.579973	0.017167	0.098349	27	0.312099	0.305310	...	0	0.0	5	9	5	0	2	13	0	0
4	003	23	1	29.917625	114.257428	0.023902	0.140785	108	0.152198	0.108852	...	0	0.0	3	7	1	1	4	8	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
502	166	28	1	57.910000	174.392947	0.026509	0.144390	30	0.806400	0.562347	...	0	0.0	28	15	4	7	21	30	8	7
503	166	28	1	40.190000	182.735181	0.022918	0.120985	26	0.793846	0.764876	...	0	0.0	23	14	1	2	15	10	3	2
504	168	29	0	113.081167	107.070285	0.024537	0.098584	166	0.428466	0.433658	...	2	0.0	19	35	12	6	32	27	6	13
505	168	29	0	139.926437	107.509813	0.024251	0.103465	180	0.517215	0.584536	...	12	0.0	27	33	8	11	51	41	20	13
506	168	29	0	61.045437	106.611775	0.024256	0.099281	69	0.592386	0.727462	...	0	0.0	12	11	2	4	13	16	5	1

507 rows × 24 columns

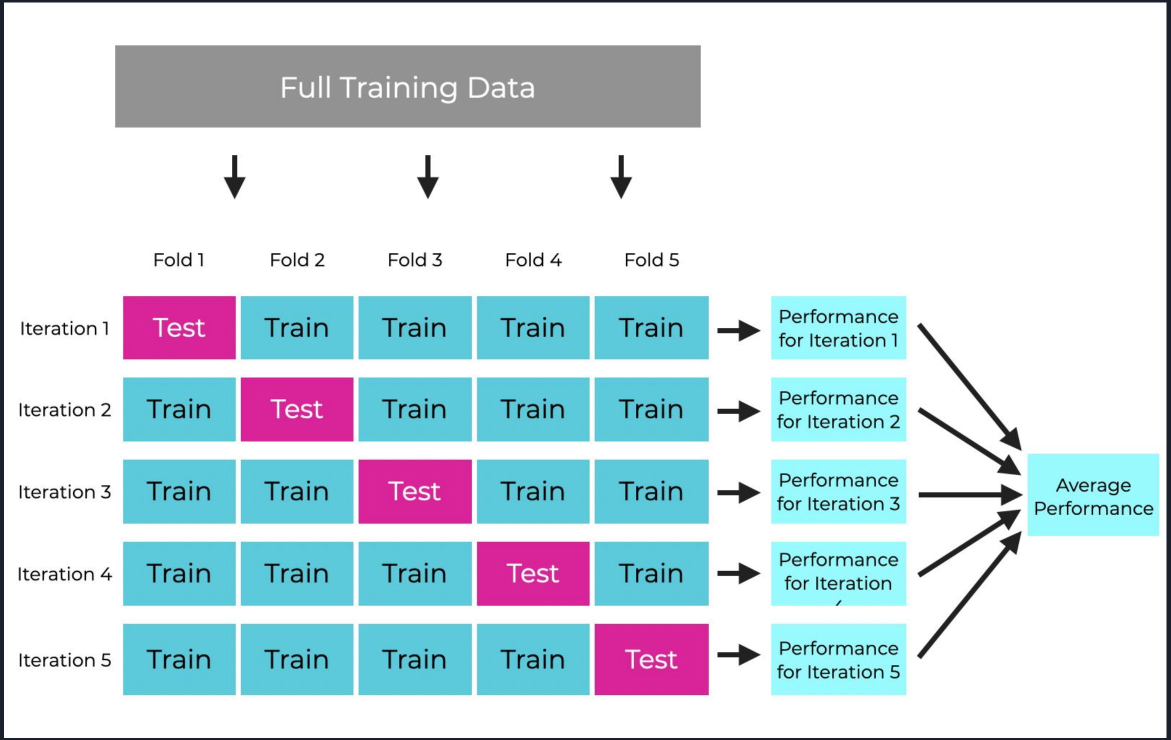
Feature Descriptions:

Category	Features	Description	Methods
Acoustic features	Total duration	Duration of audio	Librosa
	Mean pitch	Mean of the pitch of the audio	Parselmouth
	Jitter	Variations of pitch	Parselmouth
	Shimmer	Variations of amplitude	Parselmouth
Linguistic content features	Unique word count	Total count of unique words (ignore words of length 3 or smaller)	nltk, numpy
	Invented word count	Total count of invented words	nltk, numpy
	Total adjectives	Total count of adjectives	nltk, numpy
	Total adverbs	Total count of adverbs	nltk, numpy
	Total nouns	Total count of nouns	nltk, numpy
	Total verbs	Total count of verbs	nltk, numpy
	Total pronouns	Total count of pronouns	nltk, numpy
	Total conjunction	Total count of conjunction	nltk, numpy
	Number of subject	Total count of subject	nltk, numpy
	Number of object	Total count of direct objects	nltk, numpy
	Depth of syntax tree	Depth of syntax tree of the text	nltk, numpy
Fluency features	Filler rate	Number of fillers (uh, um) per second	numpy, textgrids
	General silence	Number of silences where silent duration between two words is greater than 0.145 seconds	numpy, textgrids
	Mean silence	Mean duration of silence in seconds	numpy, textgrids
	Silence abs deviation	Mean absolute difference of silent durations	numpy, textgrids
	Silence rate 1	Number of silences divided by total number of words	numpy, textgrids
	Silence rate 2	Number of silences divided by total duration in seconds	numpy, textgrids
	Speaking rate	Number of words per second in total duration	numpy, textgrids
	Articulate rate	Number of words per second in total articulation time ((i.e. the resulting length of subtracting the time of silences and filled pauses from the total response duration)	numpy, textgrids
	Avg. syllables in words	Get average count of syllables in words after removing all stop words and pause words.	numpy, textgrids
	Word syllables 2	Number of words with syllables greater than two	numpy, textgrids
	Repetition frequency	Frequency of repetition by calculating number of repetition divided by total number of words.	numpy, textgrids

Train-test-split

- **5 fold cross-validation**

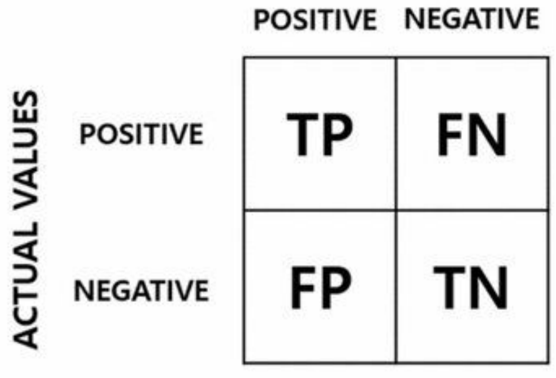
- Model performance evaluated on **unseen data**



https://www.sharpsightlabs.com/blog/cross-validation-explained/

Performance Metrics

- In the context of disease classification, some prediction errors are worse than others
- To ensure our model was providing false negatives as infrequently as possible but also providing mostly true positives, we tried to **maximize performance** on the following 4 metrics:



$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

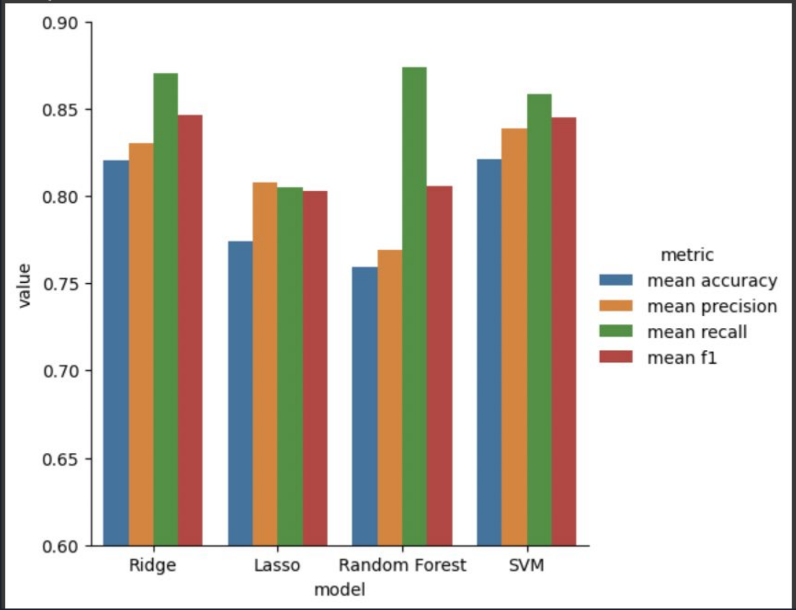
Cross-Validated Model Performances

- Compared performance of **L2 -regularized logistic regression (Ridge)**, **L1-regularized logistic regression** (LASSO), **random forest**, and **SVM**

- **Best performance** (no fine-tuning): **SVM**

- **All** models have **> 80% recall**: they **detect MCI** most of the time

```
LogisticRegression()
{'mean accuracy': 0.8209, 'mean precision': 0.8301, 'mean recall': 0.8705, 'mean f1': 0.8464}
LogisticRegression(penalty = 'l1', solver = 'liblinear')
{'mean accuracy': 0.7745, 'mean precision': 0.8082, 'mean recall': 0.8051, 'mean f1': 0.8032}
RandomForestClassifier()
{'mean accuracy': 0.7597, 'mean precision': 0.7695, 'mean recall': 0.8741, 'mean f1': 0.8056}
svm.SVC(kernel = 'linear')
{'mean accuracy': 0.8212, 'mean precision': 0.8385, 'mean recall': 0.8587, 'mean f1': 0.8449}
(0.6, 0.9)
```



Training Approaches (aggregation methods)

- There are **multiple rows** in the dataframe **per patient** (corresponding to multiple recordings), but we only want to make **1 prediction per patient**

Option 1:

- Train a model that predicts the MCI status of every row in the dataset
- Each patient is classified as the most frequent prediction in their corresponding rows

Option 2:

- Aggregate (using mean) the information from all the rows corresponding to a given patient
- Train a model that predicts the MCI status of every patient using this new aggregated data

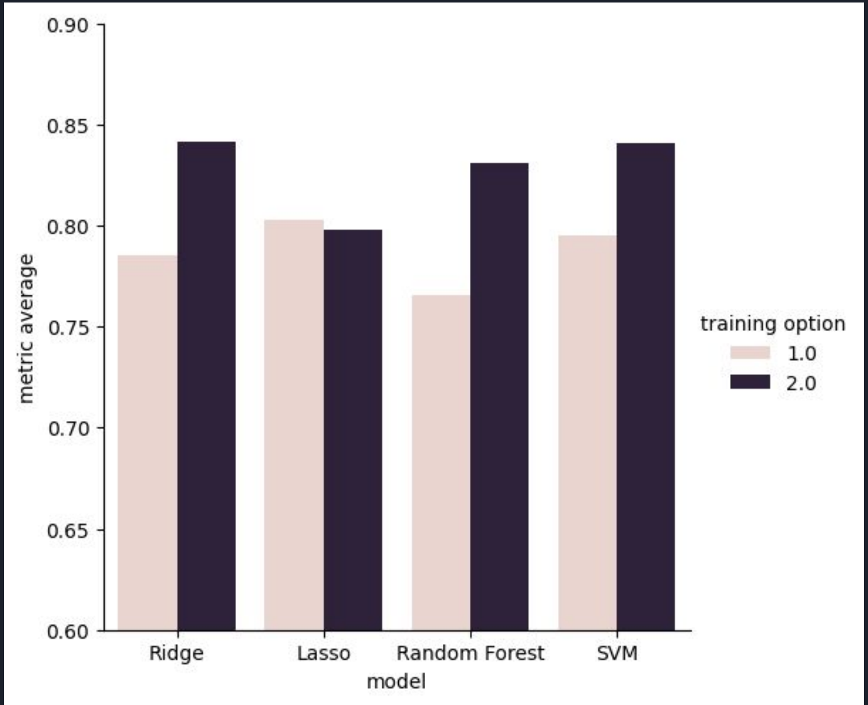
	tkdname	mmse	dx	Total Duration	Mean Pitch	Jitter	Shimmer	General Silence	Mean Silence	Silence Abs Deviation	...
0	002	29	0	122.440000	164.242554	0.017773	0.088550	114	0.313544	0.319138	...
1	002	29	0	45.660000	162.687280	0.021281	0.098618	56	0.197143	0.202367	...
2	002	29	0	62.690000	179.818570	0.021813	0.098348	35	0.733257	0.874998	...
3	003	23	1	21.691521	111.579973	0.017167	0.098349	27	0.312099	0.305310	...
4	003	23	1	29.917625	114.257428	0.023902	0.140785	108	0.152198	0.108852	...
...	...	...	...	...	...	...	...	...	...	...	...
502	166	28	1	57.910000	174.392947	0.026509	0.144390	30	0.806400	0.562347	...
503	166	28	1	40.190000	182.735181	0.022918	0.120985	26	0.793846	0.764876	...
504	168	29	0	113.081167	107.070285	0.024537	0.098584	166	0.428466	0.433658	...
505	168	29	0	139.926437	107.509813	0.024251	0.103465	180	0.517215	0.584536	...
506	168	29	0	61.045437	106.611775	0.024256	0.099281	69	0.592386	0.727462	...

507 rows × 24 columns

Training Approaches (cont.)

- **Option 2** (aggregating each patient's data and making a single prediction per patient) **performs better** for every model (except LASSO, for unknown reasons)

- Possible explanation:
    - Every row for a patient corresponds to the audio response for a different question
    - Difficult: using a single model to try to classify 3 different sets of speech feature values per patient
    - Easier: using a single model to classify one aggregated set of speech features per patient



Feature importances

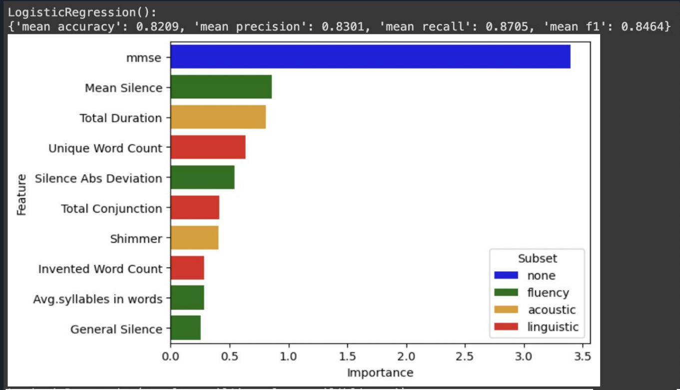
- Which **features** were the **best predictors** of MCI status?

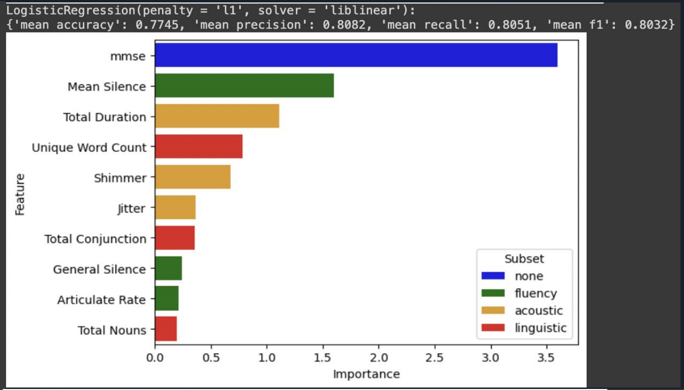
- Which **subset** of features (fluency, acoustic, linguistic) were the **best?**

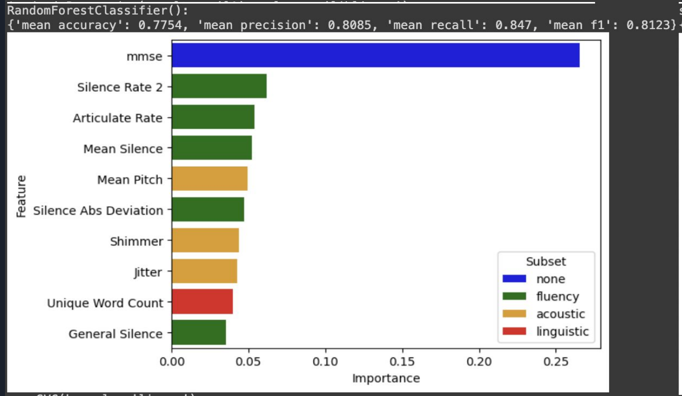
	tkdname	mmse	dx	Total Duration	Mean Pitch	Jitter	Shimmer	General Silence	Mean Silence	Silence Abs Deviation	...
0	002	29	0	122.440000	164.242554	0.017773	0.088550	114	0.313544	0.319138	...
1	002	29	0	45.660000	162.687280	0.021281	0.098618	56	0.197143	0.202367	...
2	002	29	0	62.690000	179.818570	0.021813	0.098348	35	0.733257	0.874998	...
3	003	23	1	21.691521	111.579973	0.017167	0.098349	27	0.312099	0.305310	...
4	003	23	1	29.917625	114.257428	0.023902	0.140785	108	0.152198	0.108852	...
...	...	...	...	...	...	...	...	...	...	...	...
502	166	28	1	57.910000	174.392947	0.026509	0.144390	30	0.806400	0.562347	...
503	166	28	1	40.190000	182.735181	0.022918	0.120985	26	0.793846	0.764876	...
504	168	29	0	113.081167	107.070285	0.024537	0.098584	166	0.428466	0.433658	...
505	168	29	0	139.926437	107.509813	0.024251	0.103465	180	0.517215	0.584536	...
506	168	29	0	61.045437	106.611775	0.024256	0.099281	69	0.592386	0.727462	...

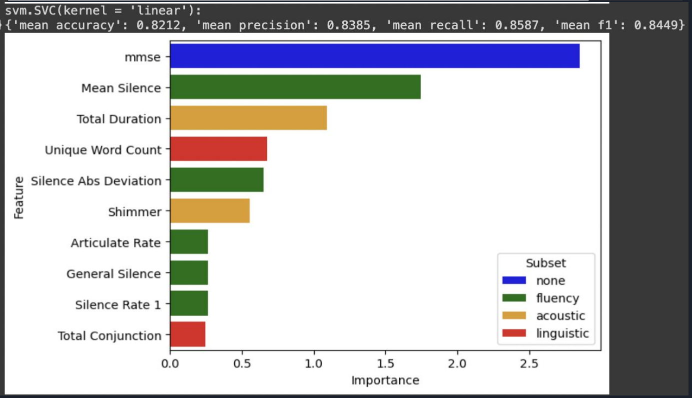
507 rows × 24 columns

Approach 1: Analyze coefficients for each feature in different models





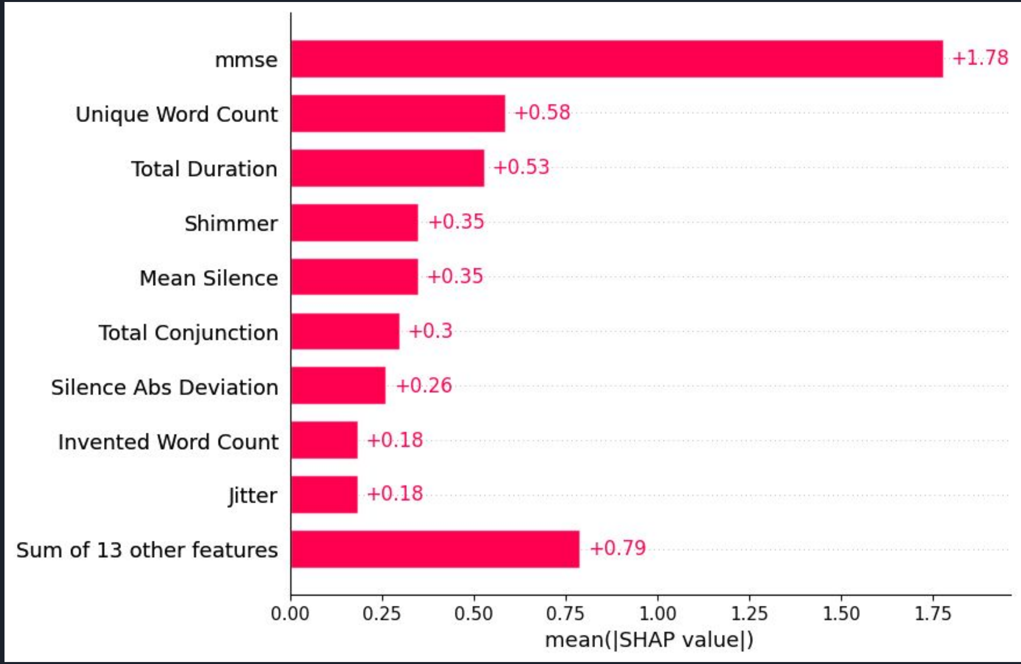




Approach 2: SHAP (SHapley Additive exPlanations)

- SHAP is a library that uses a game-theoretic approach to evaluate feature importances

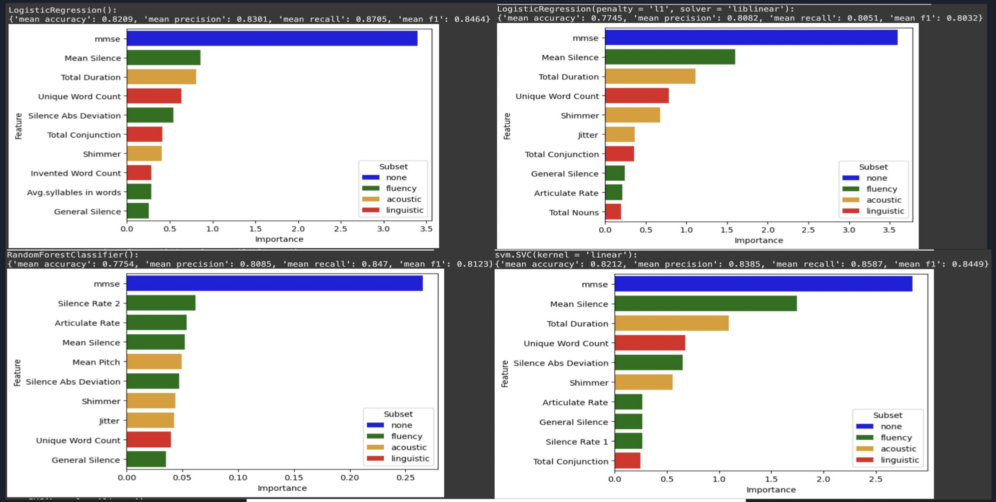
SHAP values for the ridge model:

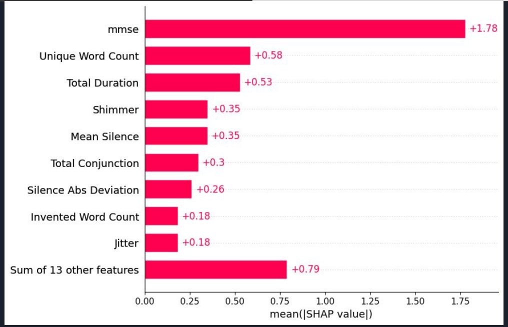


Which subset of features was the best?

- **Fluency** features were the most important, followed by **acoustic** and then **linguistic**

- According to these findings, **how** the patient speaks may be more important than what they actually **say**





Features that best predicted MCI status:

1. **mmse**
   - mmse denotes the patient’s score on the Mini-Mental State Examination, which tests memory, language and other skills

2. Features related to the amount of **silence** in the recording (Mean Silence, General Silene, Silence Rate 1 / 2)

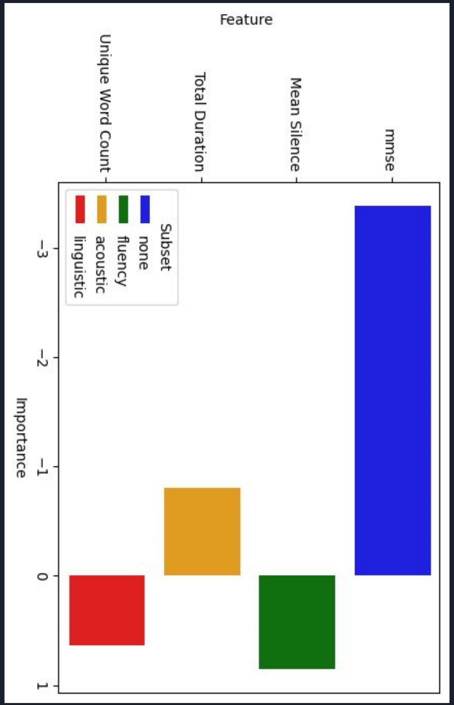
3. **Unique Word Count**

4. **Total Duration** (length of the recording)

Feature importances: align with prior research?

1. **mmse - Yes**
   - mmse has been shown to be a “modest” predictor of MCI, with higher scores corresponding to higher mental acuity
   - This translates to **lower mmse scores** for **MCI patients,** hence the **negative coefficients** found by the models

Mitchell, A. J. (2015). Can the MMSE help clinicians predict progression from mild cognitive impairment to dementia?: Commentary on… Cochrane Corner. *BJPsych Advances*, *21*(6), 363–366. doi:10.1192/apt.21.6.363



2. **Silence - Yes**
   - Patients with mild and modest Alzheimer’s disease tend to have issues with **word retrieval**, causing them to **pause more frequently** while speaking
   - This translates to **higher values** for **silence features** for **MCI patients,** hence the **positive coefficients** found by the models

Lofgren, M., & Hinzen, W. (2022). Breaking the flow of thought: increase of empty pauses in the connected speech of people with mild and moderate Alzheimer's disease. *Journal of Communication Disorders*, *97*, 106214.

Feature importances: align with prior research?

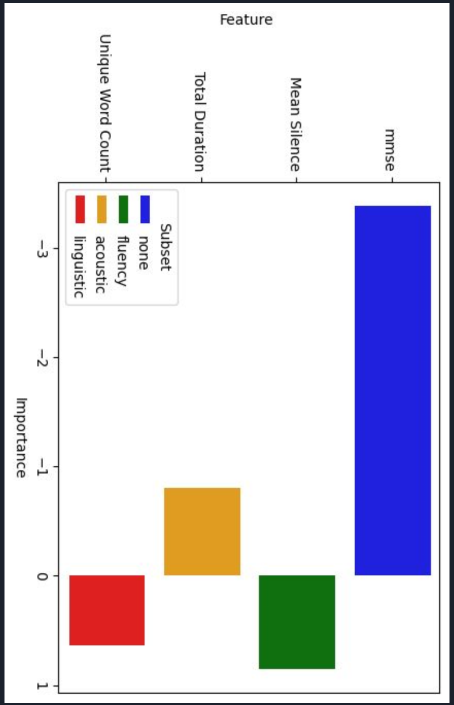
3. **Unique Word Count - No?**

    - A recent study found that MCI patients "spoke less, **produced fewer** and more abstract **nouns**"
    - Expected: **Negative coefficient** for **unique word count**
    - Found: **Positive coefficient** for **unique word count**
    - Requires **further investigation**

Cao, L., Han, K., Lin, L., Hing, J., Ooi, V., Huang, N., ... & Bao, Z. (2024). Reversal of the concreteness effect can be detected in the natural speech of older adults with amnestic, but not non-amnestic, mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *16*(2), e12588.

4. **Total Duration - No**

    - Potentially just noise
    - More research necessary



Summary

- With the prevalence of Dementia expected to increase, using AI methods to detect the disease early in the MCI stage (via audio recordings and other data) may be a **reliable** and **scalable** response

- Models that train on **aggregated statistics** for each patient instead of **multiple recordings** per patient are **more effective**

- **Ridge Logistic Regression** and **SVM** can **detect MCI** from speech and mmse data with **over 82% accuracy** and **over 85% recall**

- A patient's **MMSE score** and **silence patterns** are among the **most effective predictors**

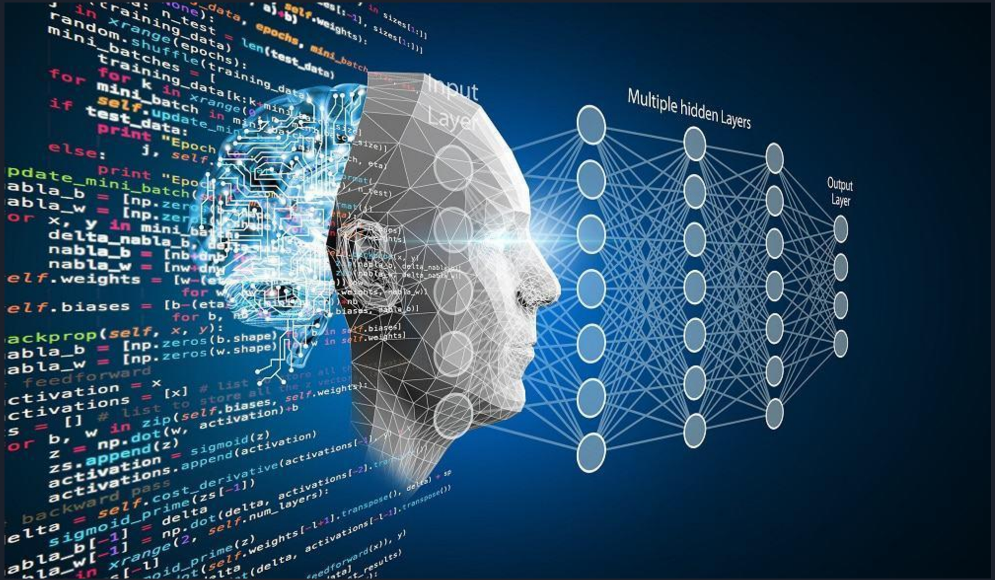
Future Goals

- Further analysis on impact of ‘unique word count’

- Hyperparameter fine tuning

- Examine / fix feature collinearity

- Deep learning with more data



Thanks for listening!