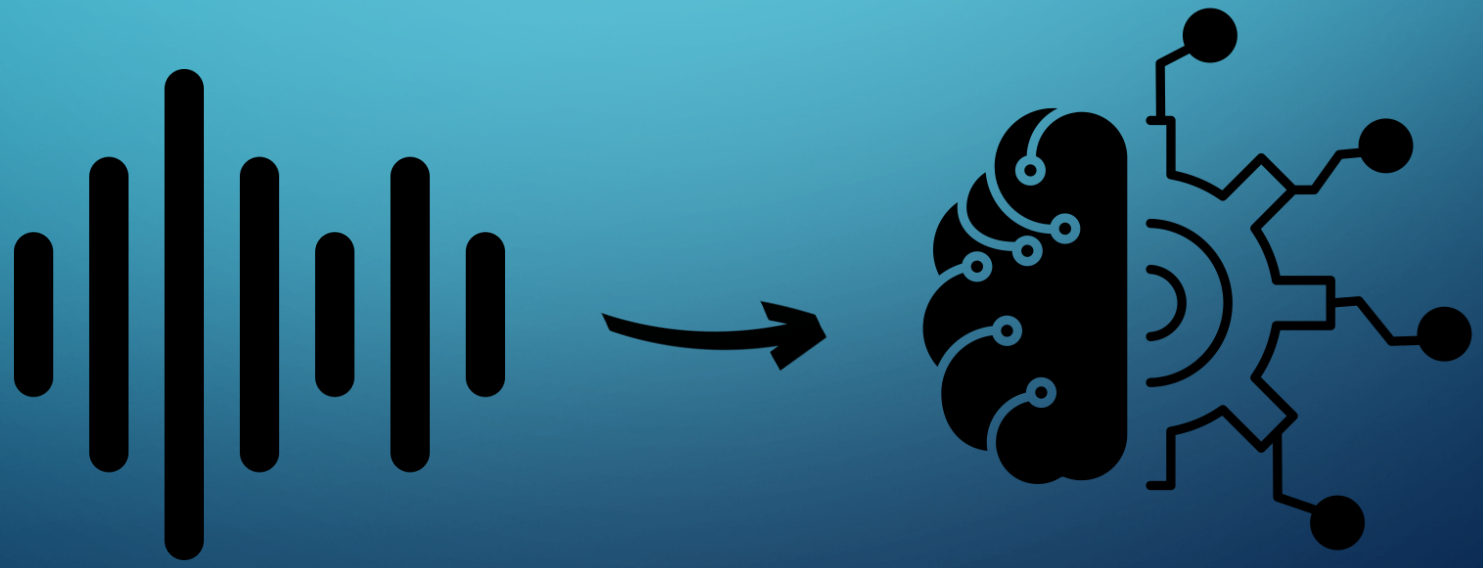
DIAGNOSING PARKINSONS DISEASE USING VOICE

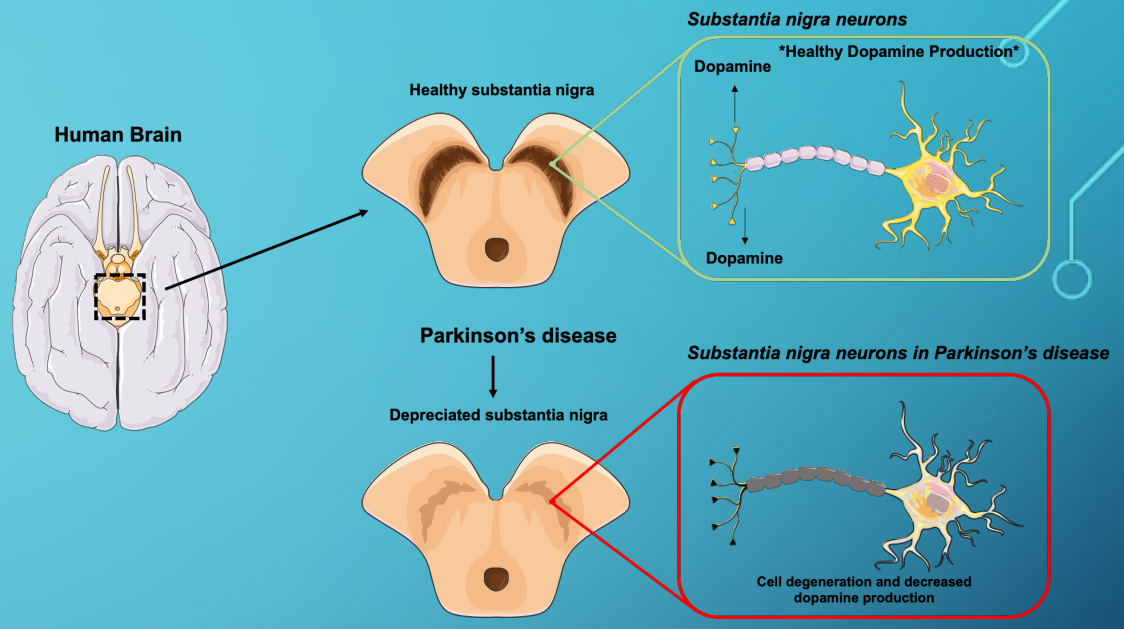
BY: HUY PHI

CSE 163 PROJECT



BACKGROUND:

- Parkinson's Disease (PD) is the 2nd most common neurodegenerative disease
- While mostly known for motor symptoms, PD can also affect various acoustic parameters of voice
- Voice features extracted from voice samples have demonstrated potential in diagnosing PD

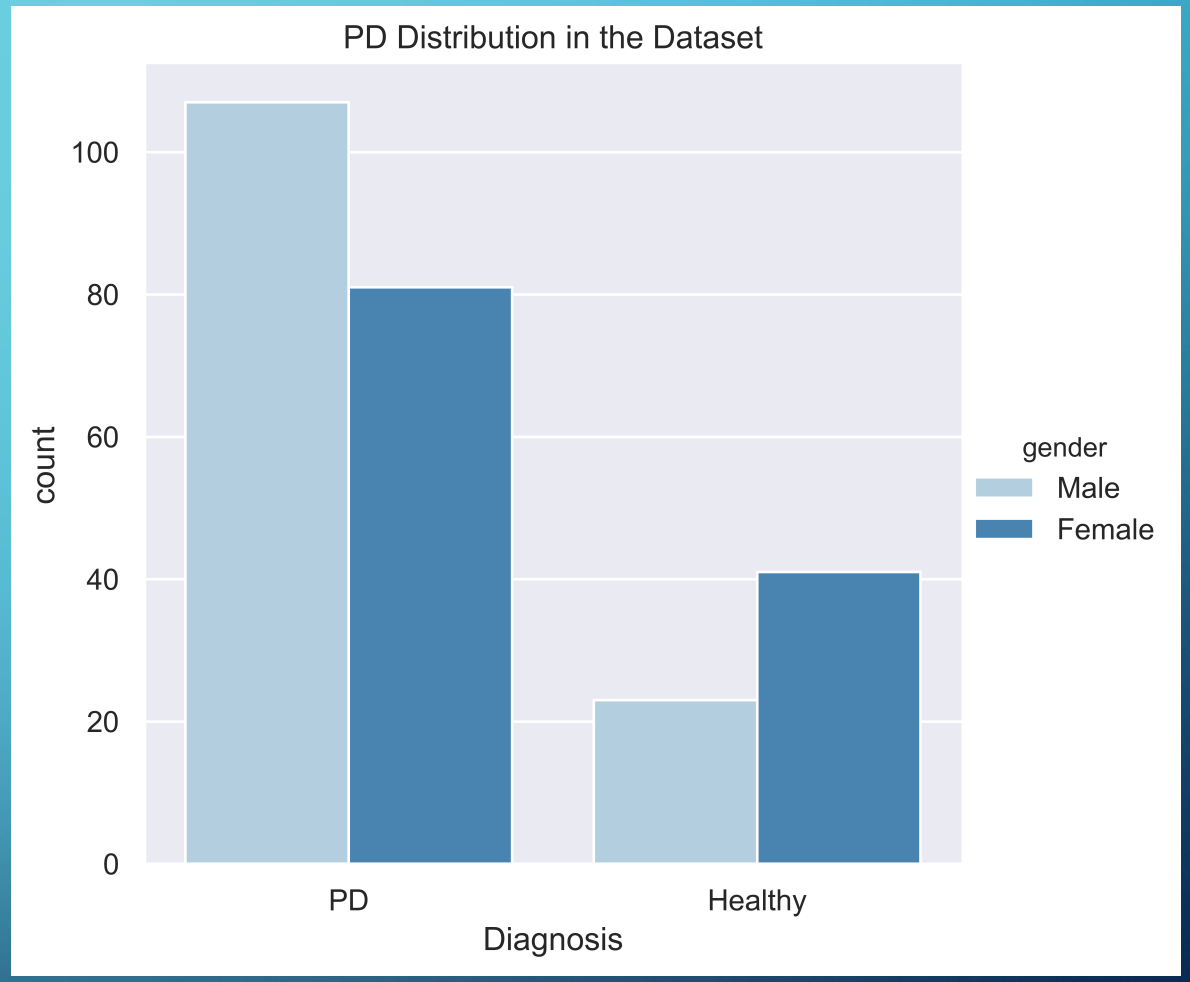


RESEARCH QUESTIONS:

- What voice features are most distinguishable between PD and Healthy voice samples?
- Which machine learning method will be most accurate in diagnosing PD from voice?
- What other methods can we use to assess our machine learning models?

DATASET:

- The dataset was obtained from the UCI machine learning repository
- It included voice samples from 188 participants with PD, and 64 age-controlled normals
- Each participant contributed 3 voice samples, and were converted into various feautre sets using speech signal processing



METHODOLOGY:

STATISTICAL TESTING FOR FEATURE SELECTION

- Read in the Data using Pandas library
- Perform independent T-tests using the SciPy library, comparing feature means between PD and Healthy
- Drop all features that were above a threshold of 0.05 to put into models

BUILDING/EVALUATING MACHINE LEARNING MODELS

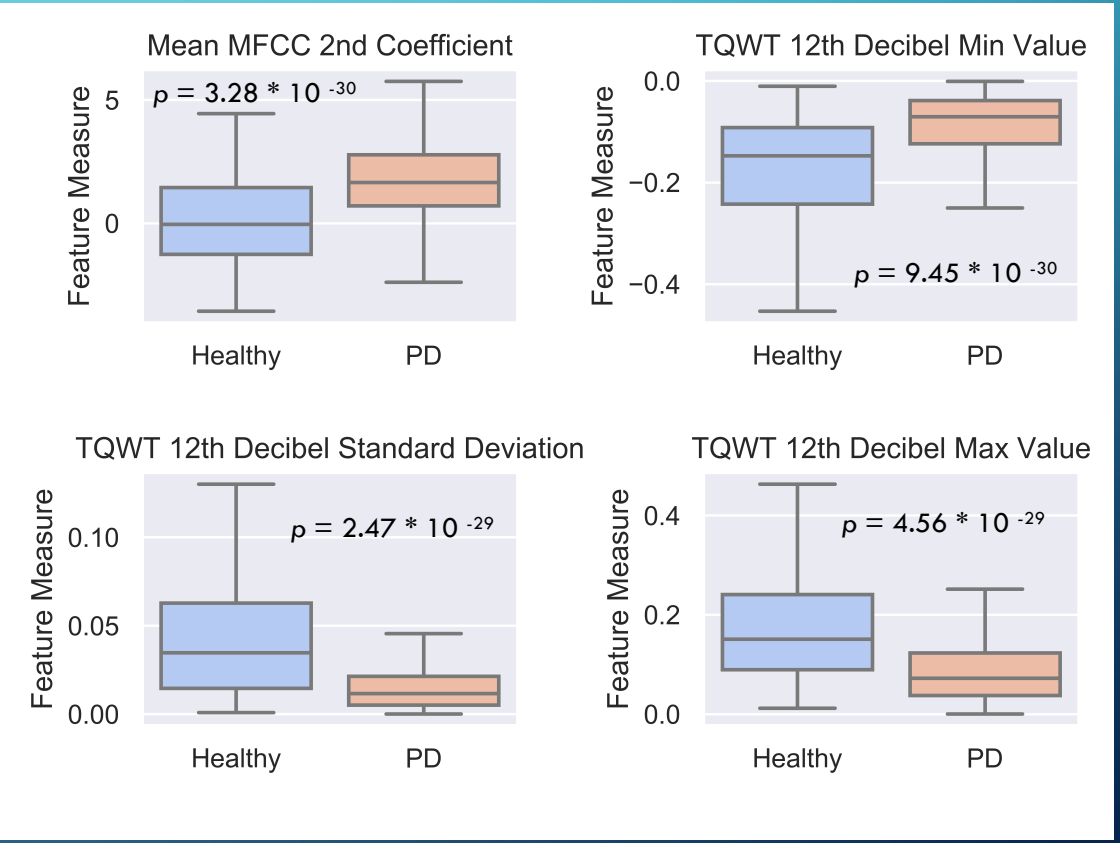
- Use Sklearn library to build machine learning models, including Decision Tree, Random Forest, Logistic Regression and K-neighbors
- Assess accuracy using 5-fold cross-validation and confusion matrix all from Sklearn

FURTHER EVALUATION VIA ROC CURVES

- Generate ROC curves using Sklearn
- Calculate area under the curve (AUC) for each model and compare

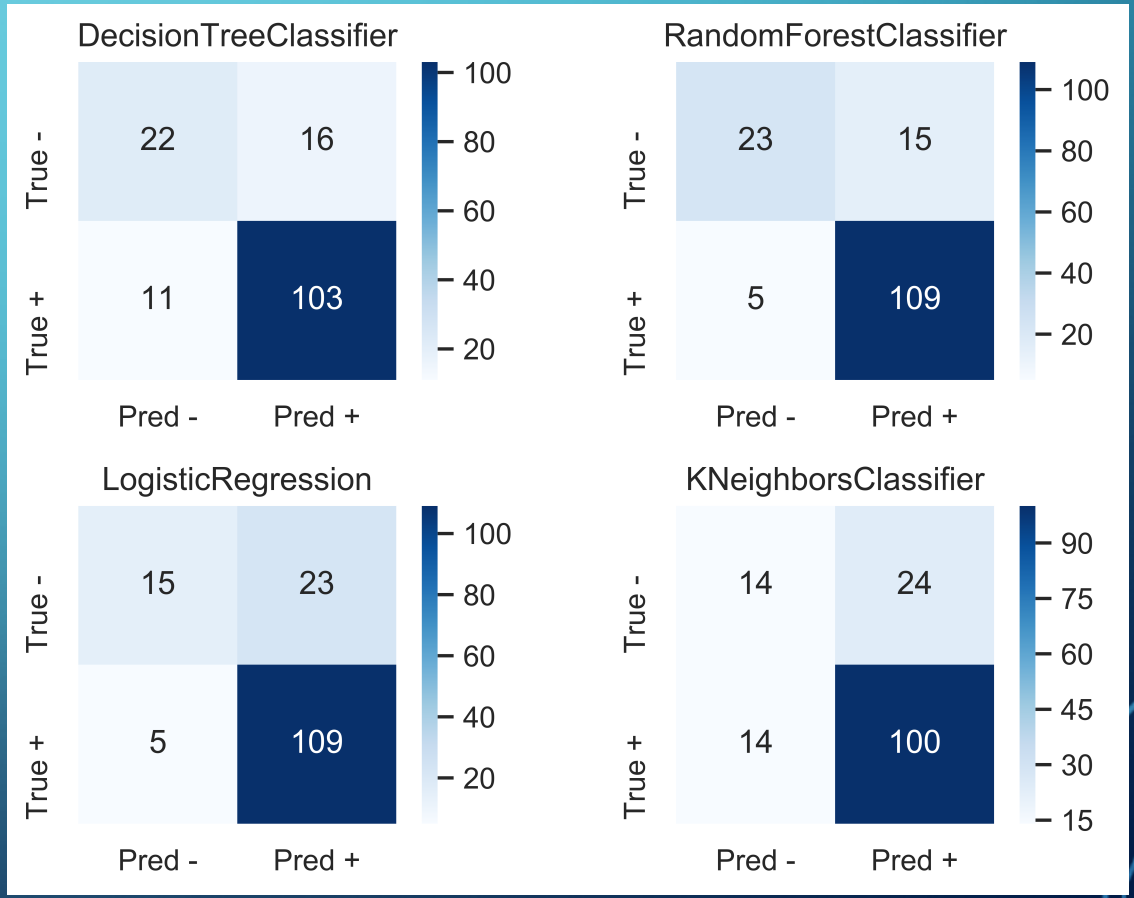
RESULTS: STATISTICAL TESTING

- Independent T-tests found *455/755* voice features to have statistically significant differences bewteen PD and Healthy Groups



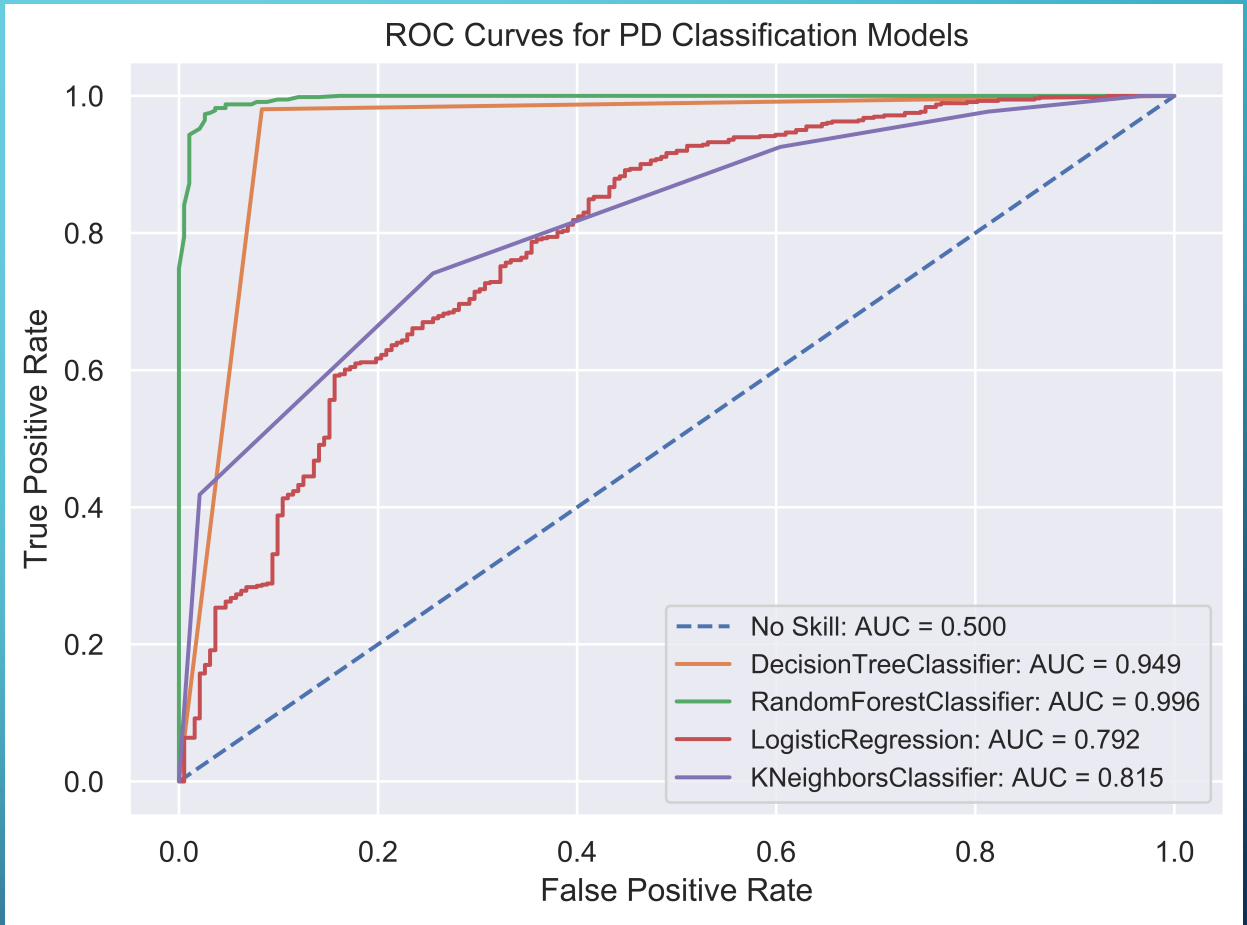
RESULTS:
RANDOM FOREST MODEL PERFORMED BEST

Model	Recall (Mean ± SD)	Precision (Mean ± SD)	F1-Score (Mean ± SD)	Accuracy (Mean ± SD)
Decision Tree	0.85 ± 0.06	0.84 ± 0.03	0.84 ± 0.03	0.76 ± 0.05
Random Forest	**0.95 ± 0.02**	**0.86 ± 0.02**	**0.89 ± 0.02**	**0.84 ± 0.02**
Logistic Regression	0.96 ± 0.02	0.78 ± 0.01	0.86 ± 0.01	0.77 ± 0.01
K-Neighbors	0.87 ± 0.05	0.77 ± 0.01	0.82 ± 0.03	0.71 ± 0.04



RESULTS: ROC CURVE

- ROC curves show that the Random Forest model performs better than all models with an AUC OF 0.996
- Logistic regression is now the worst performing model



NEXT STEPS:

- Compare between various feature sets (i.e MFCC vs. TQWT etc.)
- Aggregate other datasets together to work on an overall larger dataset
- Evaluate other machine learning/deep learning models