Parkinson's Freezing of Gait Prediction

Project Checkpoint Review

vv2372, sz3211, ys3748, nk3024, al4363

Project Summary

- Freezing of Gait: Disabling symptom of Parkinson’s disease that “negatively impacts walking abilities and impinges locomotion and independence.”
- Dataset: lower-back 3D accelerometer data
  - Tdcsfog: collected in the lab with FOG provoking protocol
  - Defog: collected at home with FOG provoking protocol
  - Daily: 24/7 data from 65 subjects, 45 exhibit FOG symptoms, 20 do not
- Goal: Predict the *start* and *stop* time of each freezing episode and specify the FoG event type.

Data Format

- Features:
  - 3D Accelerometer data: on 3 axes (Vertical, Mediolateral, Anteroposterior)
  - Type (Start Hesitation, Turn, Walking)
    - Tdcsfog/Defog: One-hot encoded type
    - Daily: Binary event type
    - Subjects: Discrete Numerical Data Type(Age, YeaesSinceDx,UPDRSIII_On,UPDRSIII_Off,NFOGQ) and Binary event type(Sex)
    - Events: Continuous Numerical Data Type(Init, Completion),Binary event type (Kinetic)
- Labels:
  - Time: integer timestep (tdcsfog@128Hz, defog/daily@100Hz)

Data Exploration: Train/defog

1. 91 csv files. All files have the same column structure.
    a. 9 columns: `'Time', 'AccV', 'AccML', 'AccAP', 'StartHesitation', 'Turn', 'Walking', 'Valid', 'Task'`
2. Number of timesteps (rows/records):
    a. Max: 410k; Min: 28k; Avg: 140k
3. No missing/duplicate data across all files.
4. Features
    a. Acceleration: AccV, AccML, AccP
        i. Time vs Acceleration
        ii. Time vs Acceleration by Event Type
    b. FoG Event Types: StartHesitation, Walking, Turn
        i. Imbalanced occurrence across all files
        ii. (Fig.1) Majority of FoG event type is Turn (largest number of timesteps recorded; largest number of occurrences of all types)
    c. Indicators of whether current event type is ambiguous/unannotated: Task, Valid
        i. (Fig.2) Large amount of un-annotated records: percent between 0.5 and 0.9 across all files
            1. All files have un-annotated record labeled as ‘Normal’, except daa4d27db4.csv, which appears to be an error.
        ii. (Fig.3) Small amount of ambiguous records: percent below 0.1 across all files
            1. Majority of ambiguous records are labeled as ‘Normal’.
            2. The highest ambiguous percent is 0.08 in c50f164e00.csv, which is significantly higher than other files.

Fig. 1

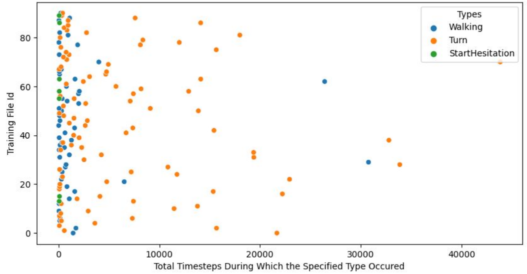


Fig. 2

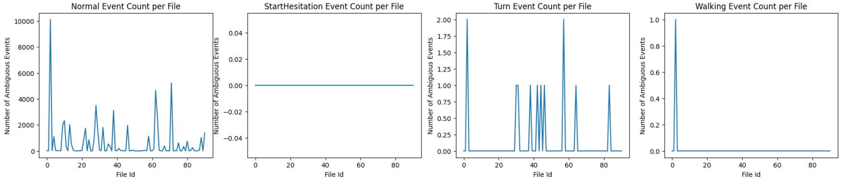
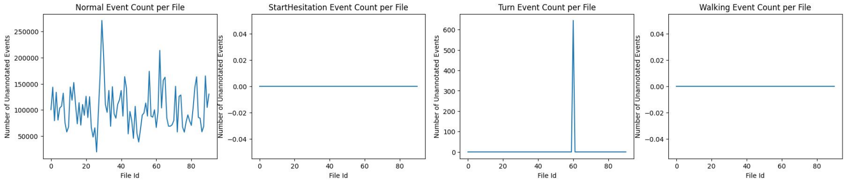


Fig. 3



Data Exploration: Defog

Fig. 1

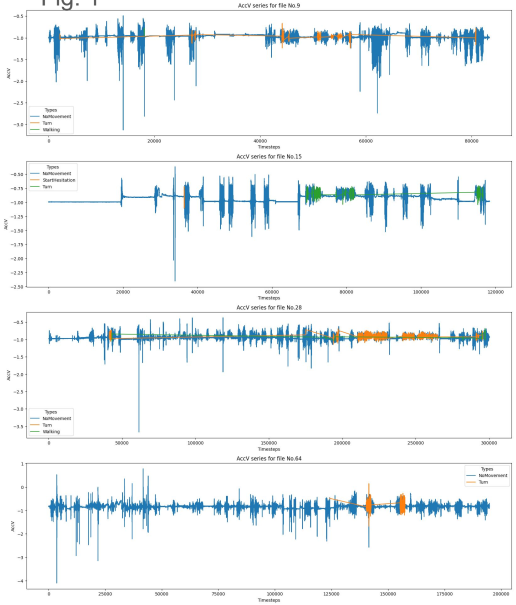


Fig. 2

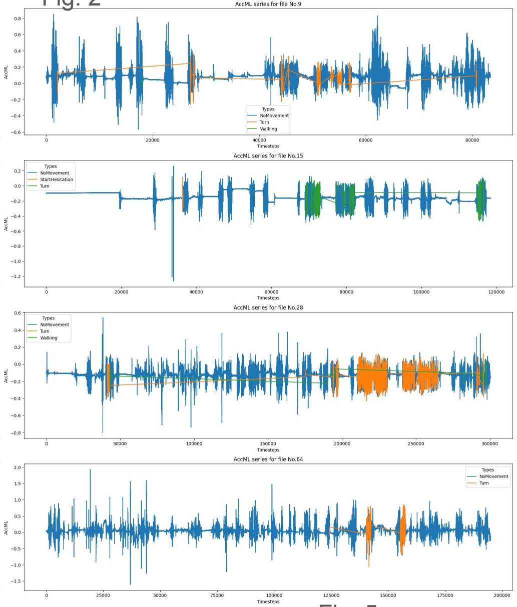


Fig. 3

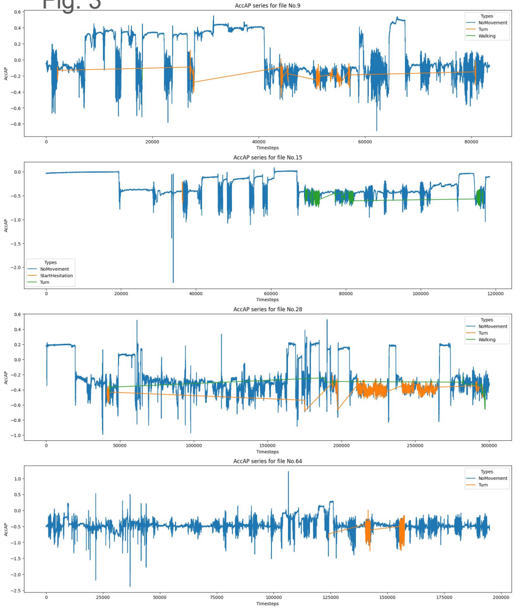


Fig. 4

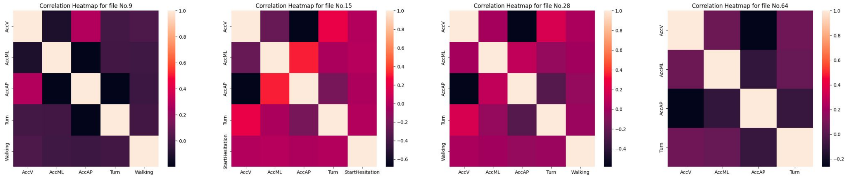
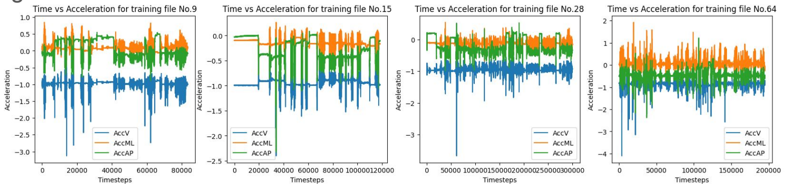
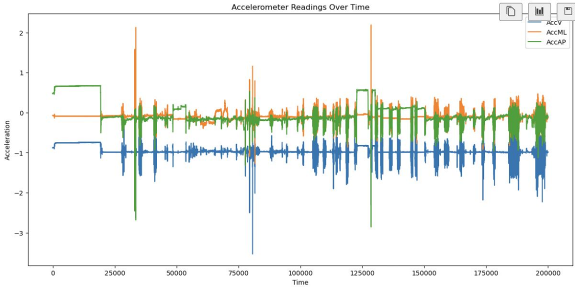


Fig. 5

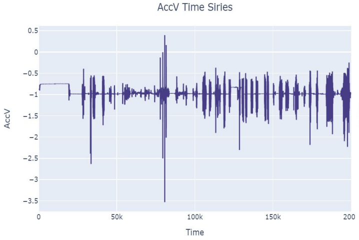


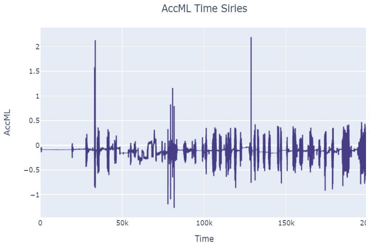
Data Exploration: Notype

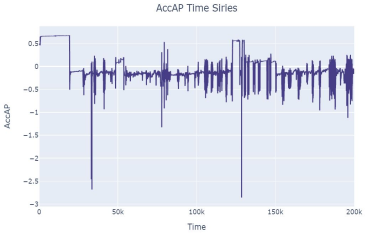
- The number of files in folder notype: 46
- Columns:Time,AccV,AccML,AccAP,Event,Valid,Task
- Number of timesteps: 281688;
- Maximum of the True probability of Valid: 0.4906 id: 2054f1d5df ;
- Maximum of the True probability of Valid Task: 0.4931 id:2054f1d5df
- There is no missing data and duplicated data
- Show one example: id-0a900ed8a2



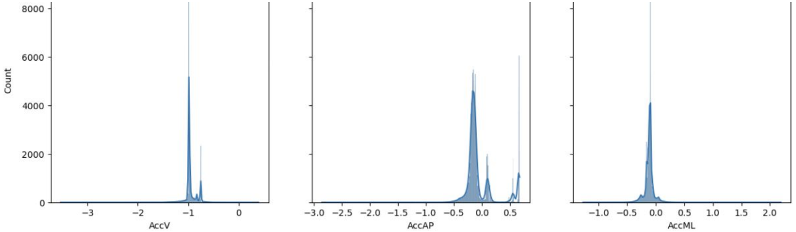
EDA Insights



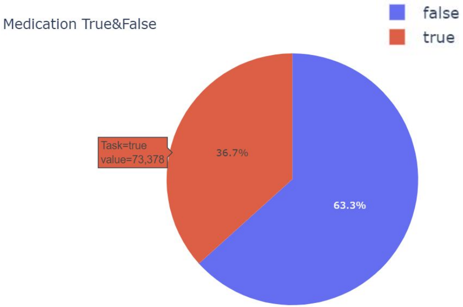




The three histogram-kde charts



Task column implied to us that almost some of the events from the recordings of a specific FoG patient aren't annotated, unlike from what we saw from the data distribution of the train_defog_df dataframe Task column.



ACCV: The right-skewed data peak shown on the AccV column hinted us that there's most events in which FoG patients walked slowly in the opposite directions when freezing while having their knees trembling.

ACCML: We found from the AccAP column hinted us that there are most events that showed decreased acceleration as well as slower gait turns, as it hinted the signs of a freezing of gait event.

ACCAP: The symmetrical distribution we saw in the AccML column implied to us that it showed the events of a specific FoG patient showing the symptoms of bradykinesia because of their slower gait velocities and step strides.

Data Exploration: Tdcsfog

Fig. 1

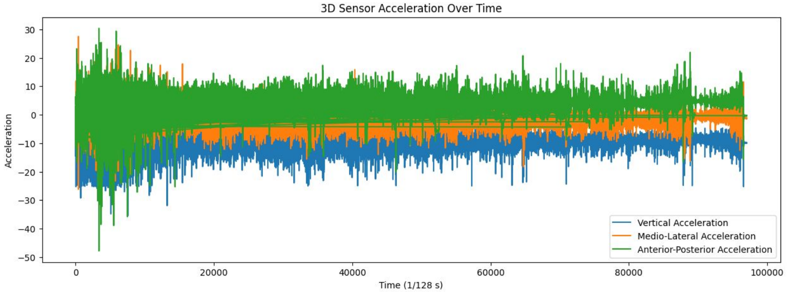
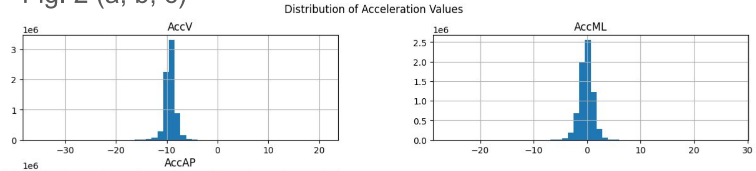
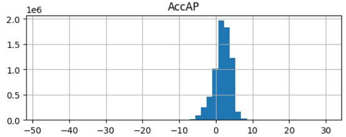


Fig. 2 (a, b, c)





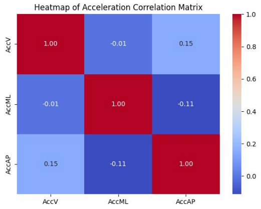


Fig. 3

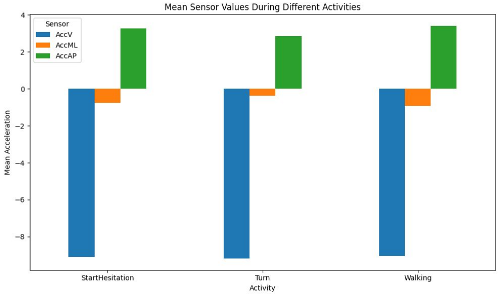
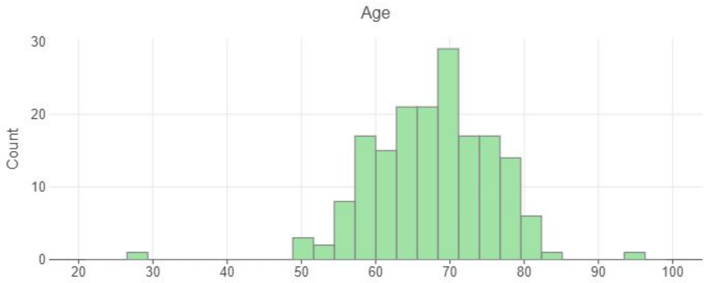


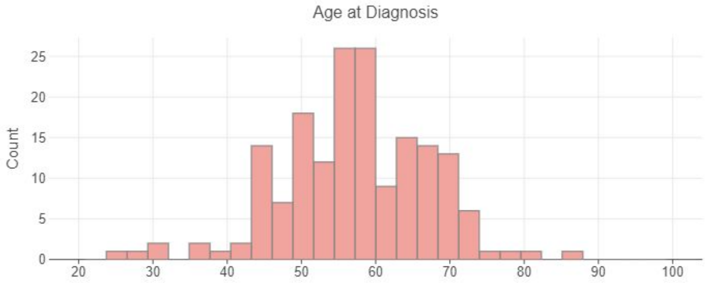
Fig. 4

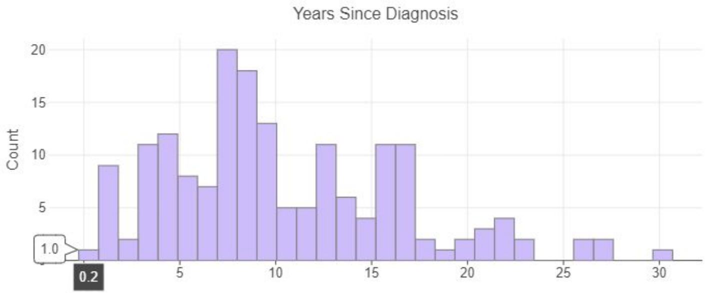
1. AccV, AccML, AccAP over time.
   a. Somewhat consistent, with regular spikes/dips, slowly decreases.
2. Distribution of AccX values
   a. Relatively focused around -10, 0, and 0 respectively.
3. Mean Sensor Values
   a. Sensor values across “events”
   b. Relatively similar, especially for AccV and AccAP
   c. AccML largest for Walking, then StartHesi, and smallest for Turn.

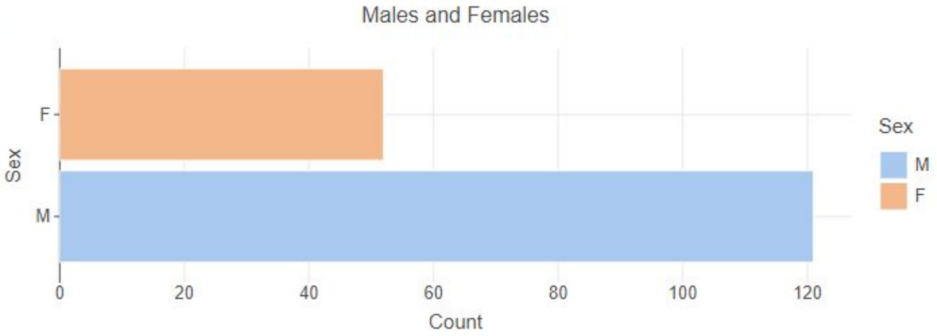
Data Exploration: Demographics

- There are far more males than females in the data
- The data contain some quite young (less than 30) and quite old (over 90) people
- Some individuals were diagnosed rather young
- The years since diagnosis range from very recent (1 year) to not (30 years)



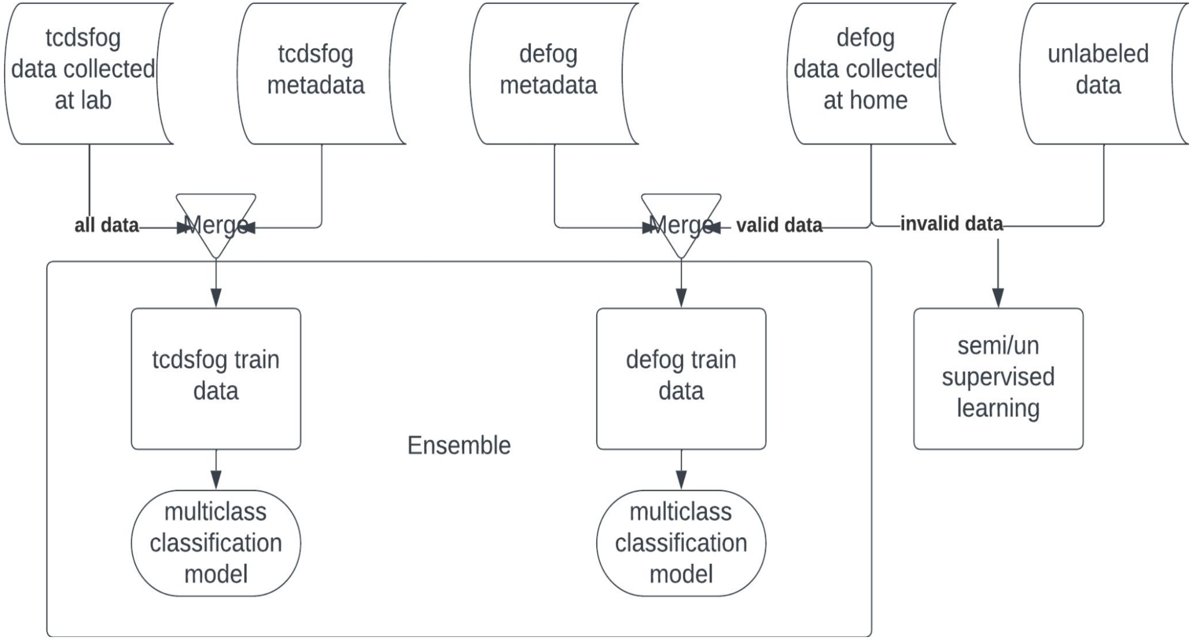




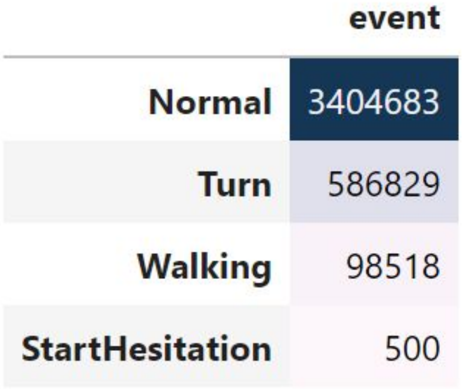


Cleaning and Sampling

The overview of data cleaning, selection and use-cases


tcdsfog data collected at lab
tcdsfog metadata
defog metadata
defog data collected at home
unlabeled data
all data
Merge
Merge
valid data
invalid data
tcdsfog train data
defog train data
semi/un supervised learning
Ensemble
multiclass classification model
multiclass classification model

As it can be seen, the data is highly imbalanced along. So we use Stratified Sampling. We refrain from oversampling due to size of dataset.



	event
Normal	3404683
Turn	586829
Walking	98518
StartHesitation	500

Proposed Techniques

Direct Data Ingestion:

1) Baseline LGBM (Implemented)

2) XGBoost

Rolling Window Ingestion:

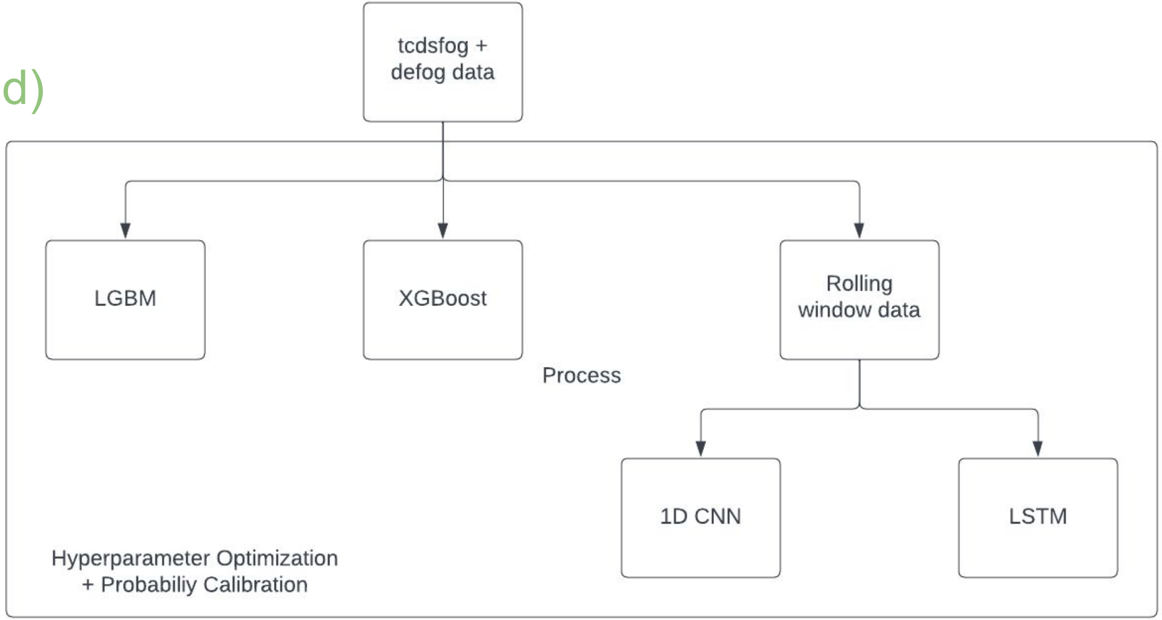
1) 1D CNN (Implemented)

2) LSTM

Hyperparameter Optimization:

1) Model optimization

2) Window size and custom weighted loss for rolling window methods


tcdsfog + defog data
LGBM
XGBoost
Rolling window data
Process
1D CNN
LSTM
Hyperparameter Optimization + Probabiliy Calibration