**Project Report**

**Team:** vv2372, sz3211, ys3748,nk3024, al4363

**Problem Statement:**

We will develop a machine learning model trained on data collected from a wearable 3D lower back sensor to detect freezing of gait (FOG), a debilitating symptom that afflicts many people with Parkinson’s disease.

**Background:**

Parkinson's disease affects 7 to 10 million people globally, with many experiencing freezing of gait (FOG), a condition where their feet seem stuck during movement attempts. FOG significantly impairs quality of life, leading to depression, increased fall risk, wheelchair dependence, and reduced independence. Despite multiple theories on FOG, its causes remain unclear. Accurate quantification is crucial for understanding and treating FOG. Data science skills can aid in collecting and analyzing FOG events, potentially leading to treatments. Machine learning, particularly with lower back accelerometers, shows promise in accurate FOG detection.

**Dataset:**

We will be using the dataset from kaggle competition : <https://www.kaggle.com/competitions/tlvmc-parkinsons-freezing-gait-prediction/data>

The dataset involves 3D accelerometer data from individuals experiencing freezing of gait (FOG) in Parkinson's disease. The goal is to identify the beginning and end of each freezing episode and categorize them into three types: Start Hesitation, Turn, and Walking.

Brief details of dataset:

* 1044 files (csv and parquet) of 70GB
* 65 subjects
* 3 axis motion data AccV, AccML, and AccAP as primary input
* 3 output classes StartHesitation, Turn, Walking Indicator variables
* Age, sex, years since diagnosis, parkinson’s score as additional information

**Data Preprocess**

The data is divided in tabular, time series and metadata which requires merging to be useful for models.

1. Meta info
   1. Merged
      1. Merged metadata.csv with subjects.csv
   2. Feature engineering
      * 1. Removed features: 'UPDRSIII\_On', 'UPDRSIII\_Off’; missing data in Tdcsfog in these features
        2. Removed features: 'UPDRSIII\_On', 'UPDRSIII\_Off’, ‘Test’; ‘Test’ feature does not exist in Defog
        3. Categorical values encoded: ‘Sex’, ‘Medication’
2. Preprocess time series: only considered defog and tdcsfog in train data.
   1. Defog
      1. Extracted only sequences with Task == True
   2. Both defog and tdcsfog
      1. If ’StartHesitation’, ‘Turn’, ‘Walking’ labeling have multiple ones, remove the sequence
   3. For 1D CNN and LSTM, segment the time series in 256 and 32 length windows
3. Merged meta info with each time series
4. Separated samples into features (combined meta info and time series) and labels (’StartHesitation’, ‘Turn’, ‘Walking’)

**Models**

We are working with multiclass classification problem with inputs being in categorical as well as sequence values. We explore the following 4 popular models used for such data modalities:

* XGBoost
* LGBM
* 1D CNN
* LSTM

**XGBoost**

XGBoost is well-suited for the Parkinson's Freezing of Gait Prediction challenge, especially when dealing with tabular data that includes features derived from accelerometer readings, metadata, and subject information. Its ability to handle structured data makes it effective in capturing intricate relationships within the dataset. XGBoost can excel in scenarios where various features contribute to predicting freezing of gait events. We use 100 trees and keep rest of the hyperparameters to default. Due to size of training data, we keep number of trees low to reduce training time.

**LGBM**

In the context of the challenge's dataset, LightGBM shines when there's a need for efficient processing of large datasets. It's particularly useful for handling categorical features, making it applicable to cases where subject-specific or categorical metadata plays a significant role. LightGBM's speed and scalability make it a pragmatic choice for quick and effective modeling. Compared to XGBoost, since LGBM is slower to converge, we use an ensemble approach of 5 estimators trained on different folds to data to predict final test results.

**1D CNN**

The accelerometer data in the Parkinson's Freezing of Gait Prediction challenge is sequential and time-dependent. Utilizing a 1D CNN becomes advantageous in this scenario as it can learn hierarchical representations of temporal patterns within the accelerometer signals. It excels at capturing local dependencies and is well-suited for tasks where recognizing nuanced temporal patterns is crucial. For 1D CNN, we segment the sequence data from sensors into vectors of length 256 and batch of 64. We train it for 10 epochs using NLL loss and adam optimizer.

**LSTM**

Given the nature of sequential data in the challenge, LSTMs prove beneficial in capturing long-term dependencies in accelerometer readings over extended periods. These neural networks are adept at understanding the temporal dynamics within the data. LSTMs are particularly valuable when the challenge involves predicting freezing of gait events where the context over time is pivotal for accurate predictions. For LSTM, we shorten the window segment to 32 and increase the batch to 1024. We train it with BCELogit loss and adam optimizer for 10 epochs.

**Analysis**

**Note: Since we are working with a kaggle challenge dataset, we do not have access to testing labels. So, the result analysis we have only covers the quantitative metric provided by the challenge platform.**

As can be seen from the table, ensemble LGBM performs the best with AP 0.266 (1st place 0.51 AP). XGBoost performs particularly worse, especially when compared to 1D CNN which is trained for only 10 epochs. For LSTM with 10 epochs, it performed better than 1D CNN. However, the higher AP is not proportional to increase in training time. This can be attributed to significantly higher segments to process for LSTM with small window of 32, compared to 256 for 1D CNN.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **XGBoost** | **1D CNN** | **LSTM** | **LGBM** |
| **Average Precision** | 0.107 | 0.167 | 0.195 | 0.266 |
| **Training Time** | ~25mins | ~6mins | ~30mins | ~40mins |

**Conclusion and Future Work**

In this project, we faced challenges in complexity of modality of dataset and adapting it to various models. After analysis of results and performance of the models, we can conclude that the models require significant higher training times, which was prohibitively expensive for us considering significant wall time for posting results (exceeded 1 hour for 2 models). In future, we would like to explore signal processing on sensor data to refine the features reduce the dimensionality and try hyperparameter optimization to increase the precision of the models. Considering that our best model achieved 0.266 AP with limited training, we believe it has potential to break into top leaderboard (> AP 0.4).