Foundations of Machine Learning
CS – 725
(Autumn 2023)

Sleep Onset and Wake Detection

A KAGGLE COMPETITION

Prepared by:

Frederic J Maliakkal (23M0745)
Varn Gupta (23M0749)
Shivang Sharma (23M0752)
Harshvivek Kashid (23M0762)
Anuj Asati (23M0763)
Tanisha Chawada (23M1071)

Abstract

Human health depends critically on sleep, which has a major impact on emotional stability and cognitive performance, particularly in children and young adults. However, due to the shortcomings of traditional annotation techniques, effective sleep state detection remains a difficulty. The goal of the project is to create reliable models for accurate sleep state diagnosis by utilising data science and wrist-worn accelerometer data. By enhancing the analysis of sleep data, we aim to facilitate extensive sleep studies and provide insightful information on the relationship between sleep and health.

We will perform thorough exploratory data analysis (EDA), to reveal patterns and hidden relationships within the dataset. Unbalances in the dataset will be addressed by data preprocessing, which includes operations like feature engineering. The dataset contains 12,79,46,340 rows (127 million) and 5 columns.

To guarantee accurate sleep state predictions, a variety of machine learning models, including Decision Trees, RNN, XGBoost, Neural Nets, etc will be skillfully used and optimised.

In summary, this project merges data science with exploratory data analysis to reveal actionable insights for enhanced sleep state detection, with the potential to advance sleep research and its applications in healthcare while focusing on the crucial relationship between sleep and health.

Problem Statement:

The competition's goal is to develop a model that properly detects the start of sleep and wake periods from wrist-worn accelerometer data is the competition's main objective. This work could result in more dependable large-scale sleep studies and focused interventions, which could completely transform the area of sleep research by providing more dependable and expandable techniques for gathering data. Crucially, it can have a significant effect on kids' mental health by making the relationship between kids' sleep habits and emotional health clearer.

Solution Approach:

The data at hand exhibits a temporal nature, characterised by a time-series structure where the current output is intricately linked not solely to the present input but also to the historical input patterns. Consequently, we can leverage sophisticated machine learning methodologies such as Neural Networks, Decision Trees, and Random Forests to construct a predictive model that can capture the intricate dependencies and patterns within the data, enabling us to make informed predictions and decisions over time.

Before applying algorithms, we have done data analysis, data preprocessing and data cleaning. Data analysis involves visualising and summarising the data to gain a better understanding of its characteristics, such as trends, seasonality, outliers, and any underlying patterns. It conducts statistical tests and analyses to identify key statistical properties and relationships within the data.

In data preprocessing, we have done data handling, standardisation and feature engineering, which is covered in this report below.

Code Survey:

During the initial stages of this project, we conducted a thorough survey of existing codebases, repositories, and libraries relevant to time series analysis, accelerometer data processing, and machine learning applied to sleep research.

The following resources have been particularly valuable for shaping our approach:

- **Time Series Analysis:**

  https://www.kaggle.com/code/prashant111/complete-guide-on-time-series-analysis-in-python

  This open-source codebase provided a collection of time series analysis tools, including various preprocessing techniques, feature extraction methods, and visualisation functions. We referred to this codebase to gain insights into time series data handling.

- **Research Paper on Sleep Classification (Reference):**

  We read a scientific paper titled ‘Sleep classification from wrist-worn accelerometer data using random forests’.
  This paper helped us understand how random forests can be used to classify sleep patterns from data collected by wrist-worn accelerometers. It provided us with valuable ideas about how to build our model and select important features for our project. While we didn't use the paper's code directly, it influenced our approach to solving the problem.

- **Baseline solution on Kaggle:**

  https://www.kaggle.com/code/zulqarnainali/explained-baseline-solution

  This notebook is designed to help participants in the competition and to Detect sleep onset and wake from wrist-worn accelerometer data. It has explained the baseline solution which is provided for all.

These resources have contributed to the formulation of our project's methodology and have helped us gain insights into best practices for accelerometer data analysis.

Please note that these codebases have been referenced for guidance and inspiration but have not been directly incorporated into our solution.

Datasets:

The data used for this competition was provided by the Healthy Brain Network, a landmark mental health study based in New York City that will help children around the world.
The dataset comprises about 500 multi-day recordings of wrist-worn accelerometer data annotated with two event types: *onset*, the beginning of sleep, and *wakeup*, the end of sleep. Each data series represents this continuous (multi-day/event) recording for a unique experimental subject.

train_series.parquet -Series to be used as training data. Each series is a continuous recording of accelerometer data for a single subject spanning many days. This file contains 12,79,46,340 rows (127 million) and 5 columns.

1. series_id - Unique identifier for each accelerometer series.
2. step - An integer timestep for each observation within a series.
3. timestamp - A corresponding datetime with ISO 8601 format %Y-%m-%dT%H:%M:%S%z.
4. anglez - As calculated and described by the GGIR package, z-angle is a metric derived from individual accelerometer components that is commonly used in sleep detection, and refers to the angle of the arm relative to the vertical axis of the body
5. enmo - As calculated and described by the GGIR package, ENMO is the Euclidean Norm Minus One of all accelerometer signals, with negative values rounded to zero. While no standard measure of acceleration exists in this space, this is one of the several commonly computed features.

Column Name	Datatype	Number of NaN
series_id	object	0
step	uint32	0
timestamp	object	0
anglez	float32	0
enmo	float32	0

Table 1: Columns and Data Types of train_series.parquet

train_events.csv - Sleep logs for series in the training set recording onset and wake events.This is the output file for the trained data.

1. series_id - Unique identifier for each series of accelerometer data in train_series.parquet.
2. night - An enumeration of potential onset / wakeup event pairs. At most one pair of events can occur for each night.
3. event - The type of event, whether onset or wakeup.
4. step and timestamp - The recorded time of occurrence of the event in the accelerometer series

Column Name	**Datatype**
series_id	object
night	uint32
event	object
Step and timestamp	object

Table 2: Columns and Data Types of train_events.csv

Lightweight training data :-
https://www.kaggle.com/datasets/carlmcbrideellis/zzzs-lightweight-training-dataset-target

There are 277 unique series in the train events (labels) file but out of these only 35 series have complete data without any missing labels. Thus we will be using these series ids for further EDA, feature engineering and model training. This dataset will save time in model training and processing. It has a total of 1,31,65,560 rows and 5 columns.

Implementation details:

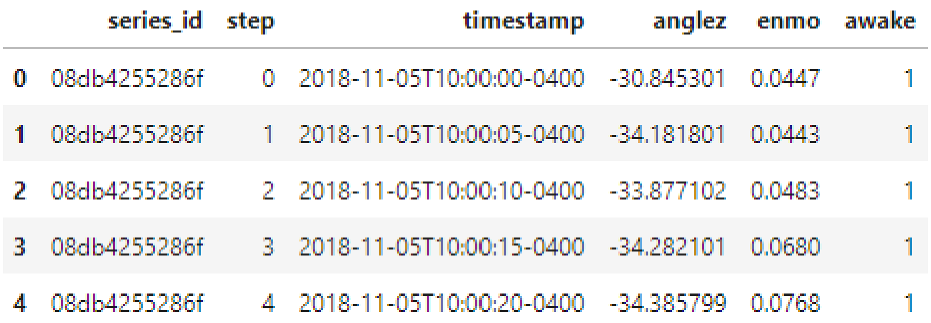
Our code implementation is in Python, utilising popular libraries such as NumPy, Pandas, TensorFlow, and scikit-learn. We have set up the project structure and are currently working on data loading, preprocessing, feature engineering and initial modelling.

Data Pre-processing:

1. Converting the 'series_id' column to a categorical data type to optimise memory usage
2. Standardising the 'timestamp' column by removing time zone information
3. Extracting the hour component from the timestamp to create a new 'hour' feature.
4. Scaling the values in the 'enmo' column by a factor of 1000 and changes the data type of both the 'enmo' and 'anglez' columns to 16-bit integers, which can lead to memory savings if the original data types were larger

Data Analysis of training data on lightweight data

- **Dataset view of first 5 rows**

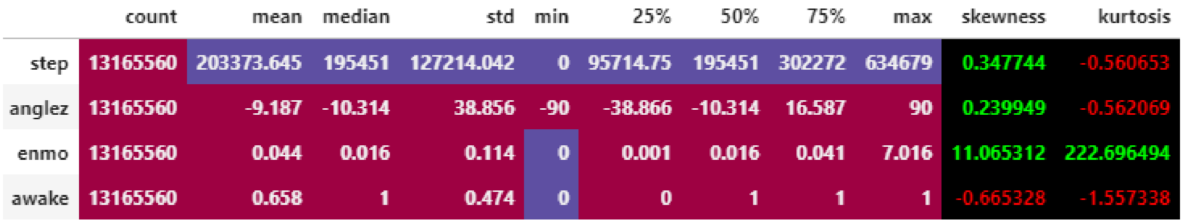


	series_id	step	timestamp	anglez	enmo	awake
0	08db4255286f	0	2018-11-05T10:00:00-0400	-30.845301	0.0447	1
1	08db4255286f	1	2018-11-05T10:00:05-0400	-34.181801	0.0443	1
2	08db4255286f	2	2018-11-05T10:00:10-0400	-33.877102	0.0483	1
3	08db4255286f	3	2018-11-05T10:00:15-0400	-34.282101	0.0680	1
4	08db4255286f	4	2018-11-05T10:00:20-0400	-34.385799	0.0768	1

Fig 1: First 5 row of the Dataset

1. Dataset shape = **(13165559, 4)**
2. This is the lightweight dataset used to increase the number of rows and perform EDA better. Here, 'awake' column is introduced to tell when the child is awake (=1) and when asleep (=0).
3. This data has 0 missing or null values

- **Description of numerical features**



	count	mean	median	std	min	25%	50%	75%	max	skewness	kurtosis
step	13165560	203373.645	195451	127214.042	0	95714.75	195451	302272	634679	0.347744	-0.560653
anglez	13165560	-9.187	-10.314	38.856	-90	-38.866	-10.314	16.587	90	0.239949	-0.562069
enmo	13165560	0.044	0.016	0.114	0	0.001	0.016	0.041	7.016	11.065312	222.696494
awake	13165560	0.658	1	0.474	0	0	1	1	1	-0.665328	-1.557338

Fig 2: Description of numerical features

1. It can be seen that the 'awake' column is not equally partitioned. It only contains 25% 0 values and the rest is 1.
2. The angle varies from -90 to 90.
3. Value of enmo varies from 0 to 7.016.

- **Correlation between features**

1. There is no correlation visible between the features.
2. But some asymmetric association can be seen between some columns.
3. 'enmo' column provides some information to the 'awake' column as both are associated asymmetrically.

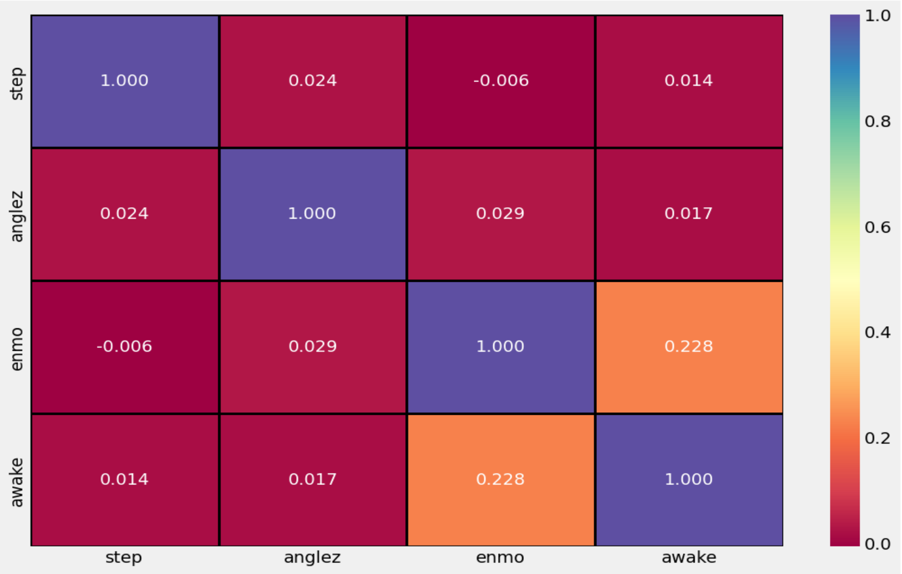


Fig 3: Correlation

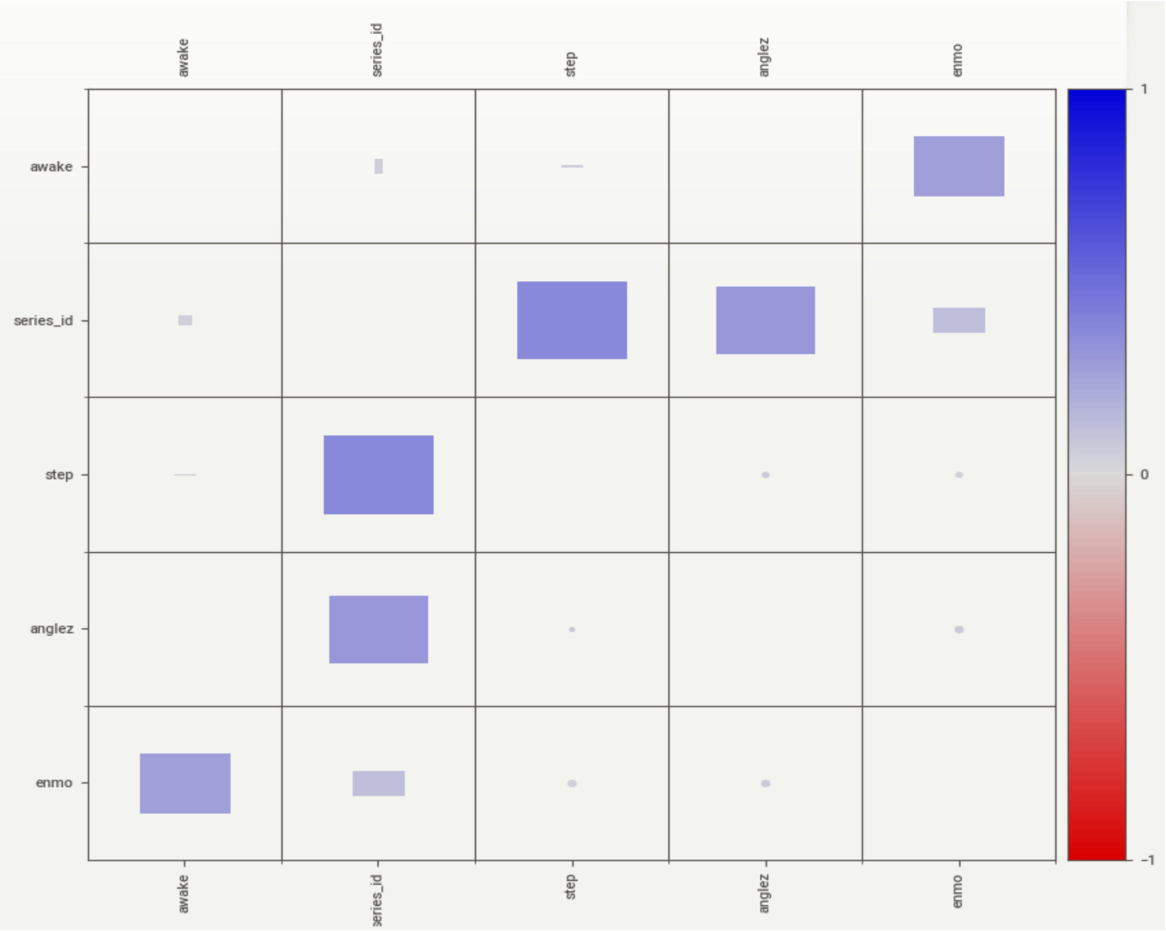
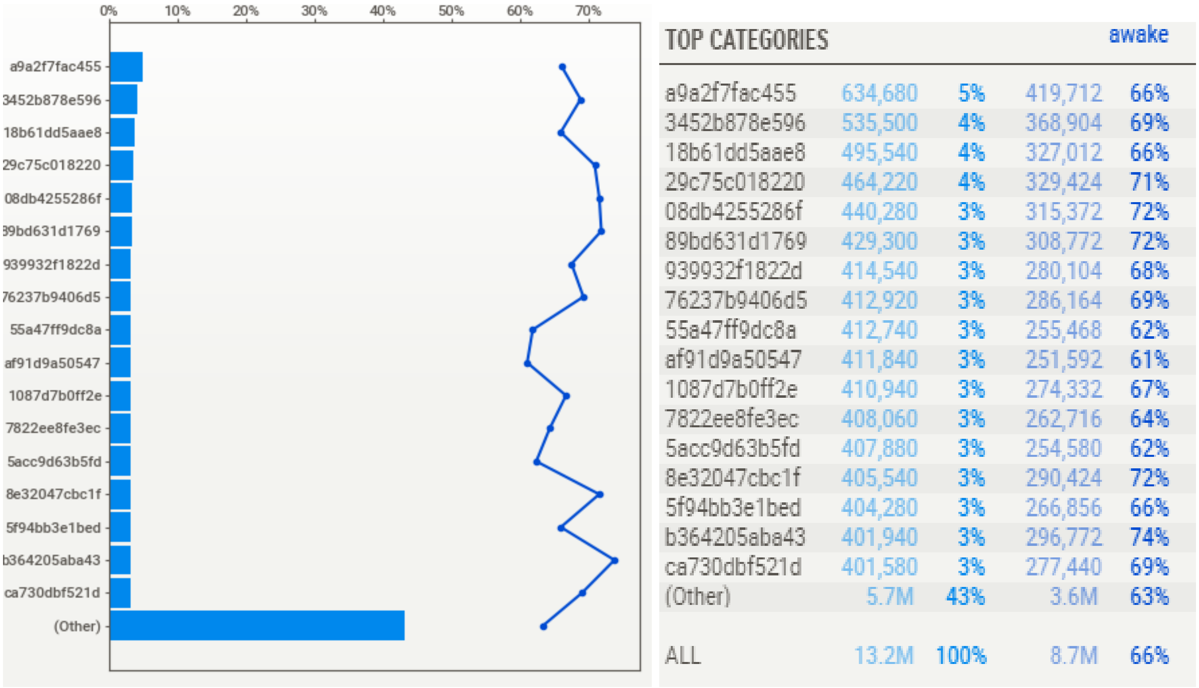


Fig 4: Association

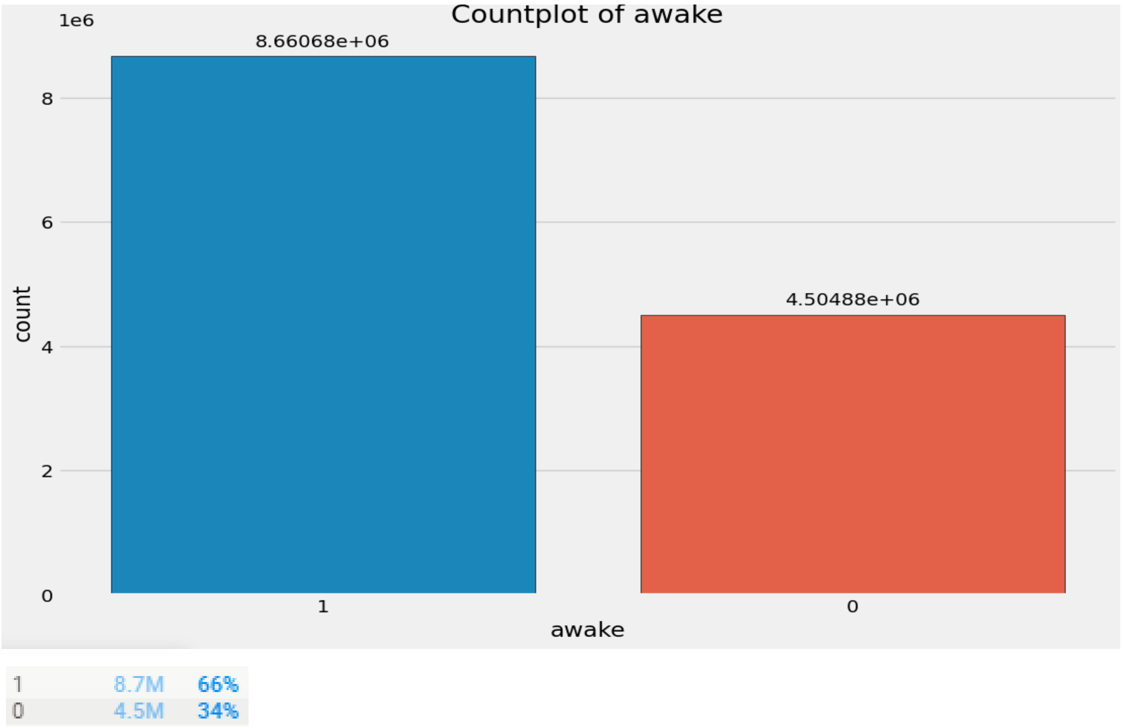
- **Series_id**



TOP CATEGORIES				awake
a9a2f7fac455	634,680	5%	419,712	66%
3452b878e596	535,500	4%	368,904	69%
18b61dd5aae8	495,540	4%	327,012	66%
29c75c018220	464,220	4%	329,424	71%
08db4255286f	440,280	3%	315,372	72%
89bd631d1769	429,300	3%	308,772	72%
939932f1822d	414,540	3%	280,104	68%
76237b9406d5	412,920	3%	286,164	69%
55a47ff9dc8a	412,740	3%	255,468	62%
af91d9a50547	411,840	3%	251,592	61%
1087d7b0ff2e	410,940	3%	274,332	67%
7822ee8fe3ec	408,060	3%	262,716	64%
5acc9d63b5fd	407,880	3%	254,580	62%
8e32047cbc1f	405,540	3%	290,424	72%
5f94bb3e1bed	404,280	3%	266,856	66%
b364205aba43	401,940	3%	296,772	74%
ca730dbf521d	401,580	3%	277,440	69%
(Other)	5.7M	43%	3.6M	63%
ALL	13.2M	100%	8.7M	66%

Fig 5: There are some series with awake time atmost upto 72% and atleast 61%.

- **Awake**



1	8.7M	66%
0	4.5M	34%

Fig 6: Countplot of awake

- **Anglez**

As calculated and described by the GGIR package, z-angle is a metric derived from individual accelerometer components that is commonly used in sleep detection, and refers to the angle of the arm relative to the vertical axis of the body.

GGIR package provides an example of what ANGLEZ-records might look like:

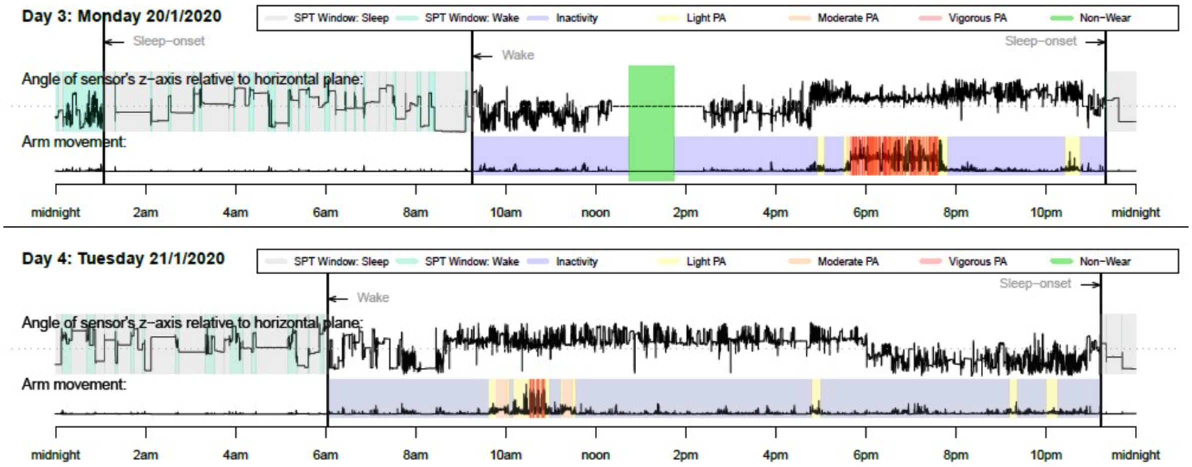


Fig 7: Expected changes in the arm angle data

The image above shows the naturally expected changes in the arm angle data during periods of sleep and activity. It is natural to observe more frequent changes of the 'anglez' values during non-sleeping period when someone's arm position changes more often than during sleep period.

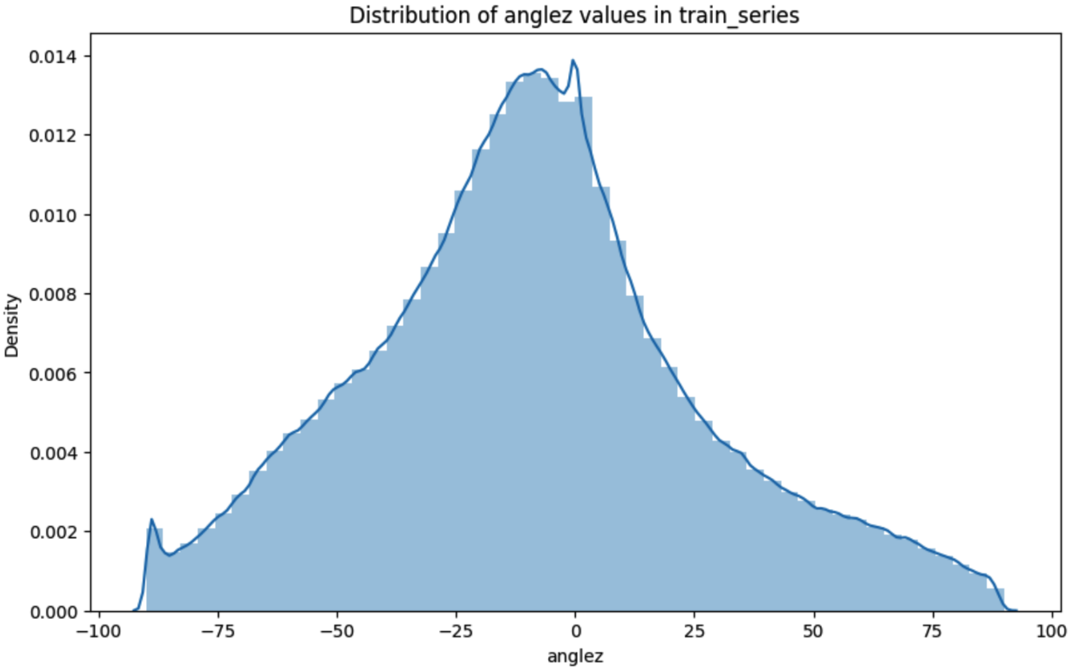


Fig 8: Distribution of angelz values in train_series

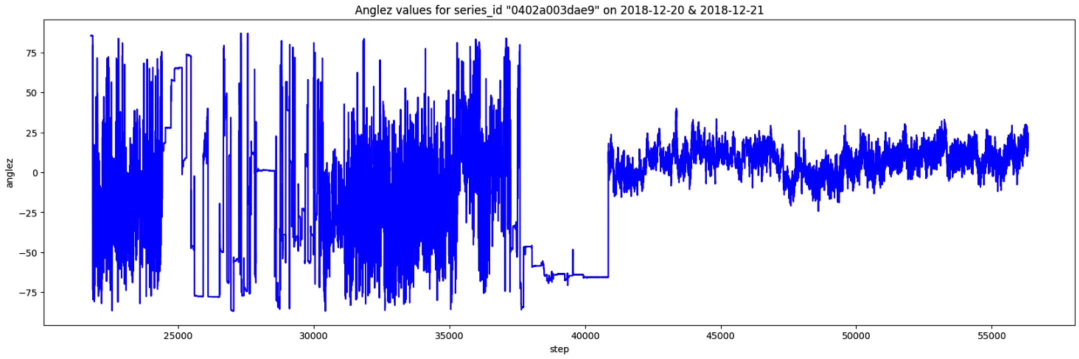


Fig 9: angelz values change with the pattern

We see that this visual corresponds to the expected picture from the GGIR package: anglez values change with the following pattern - switching between periods with small & large range of changes in anglez values. Less changes should belong to sleep periods. We can check it after adding the data about sleep and active periods from the train_events data. For now we can *add ENMO values to the graph with anglez values.*

Few Visualisations of the ENMO & ANGLEZ for different series_id so that we can see how the dataset is with respect to different step sizes.

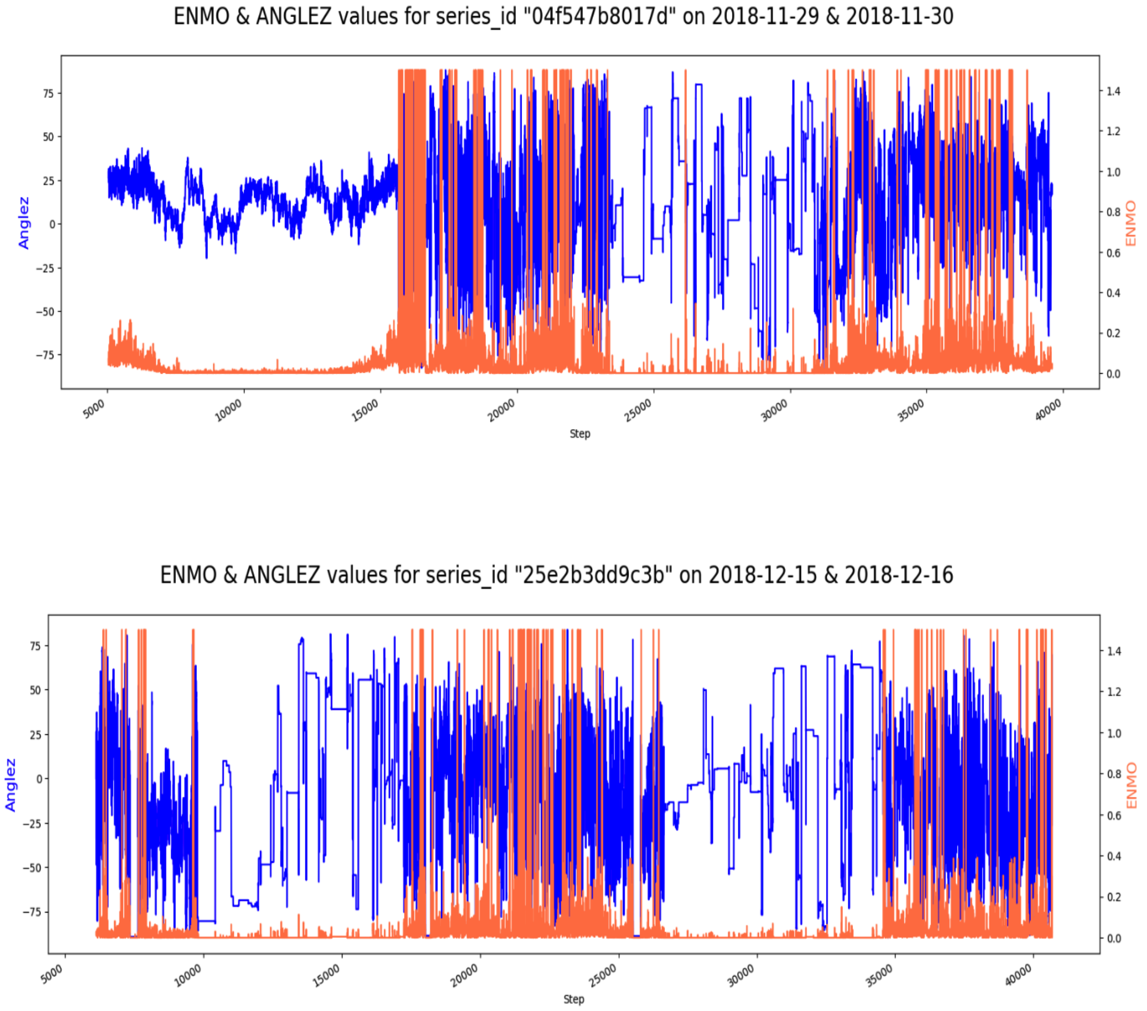


Fig 10: Visualisations of the ENMO & ANGLEZ for different series_id

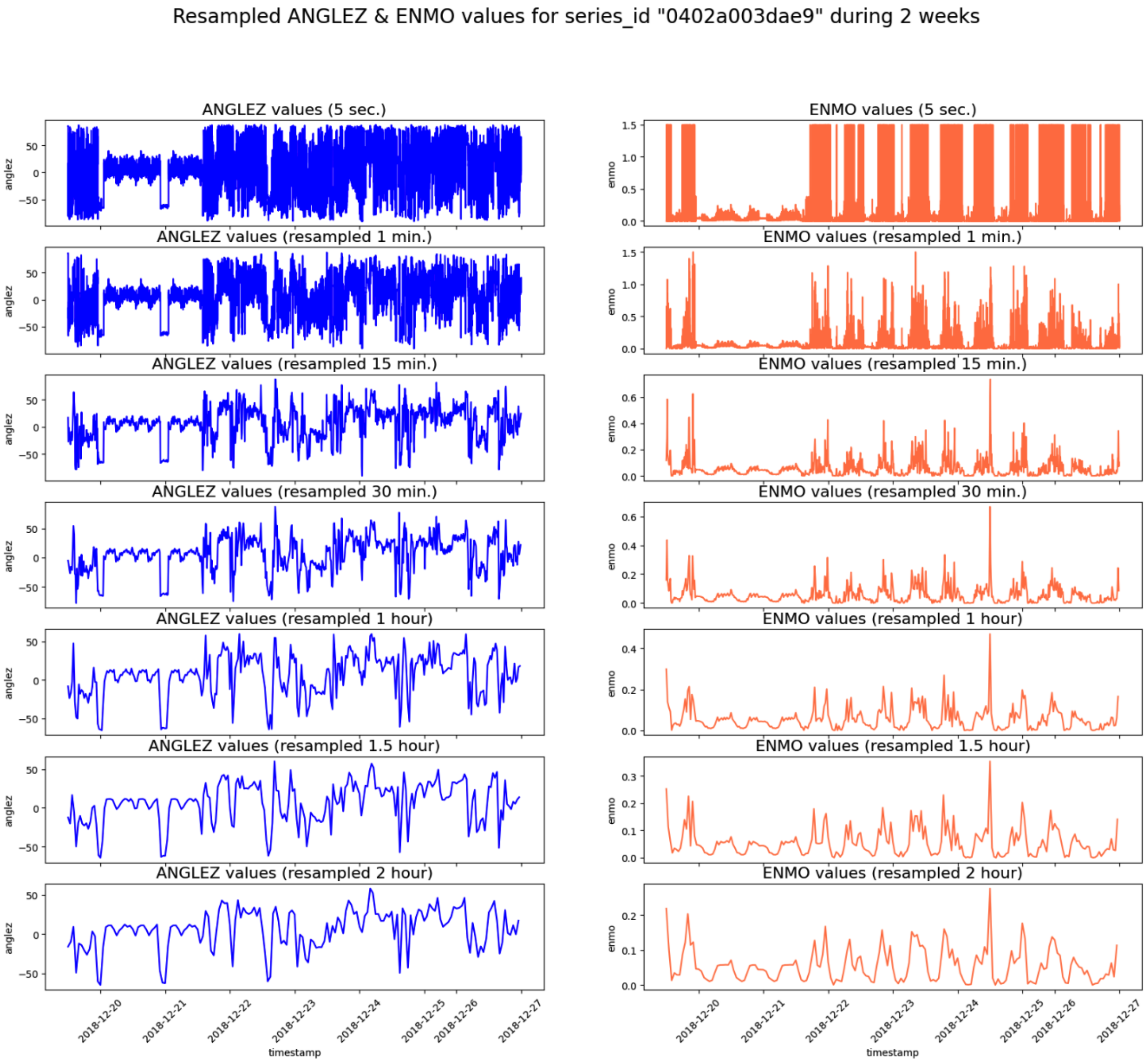


Fig 11: Resampled ANGELZ & ENMO values for series_id

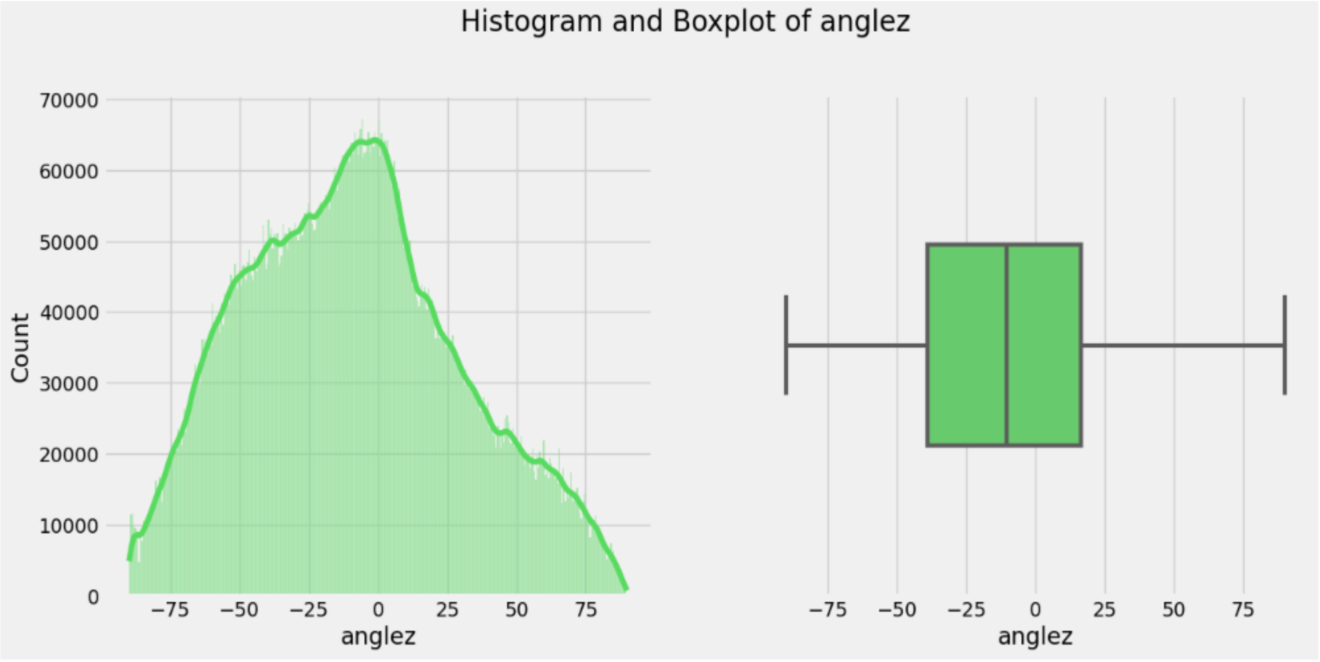
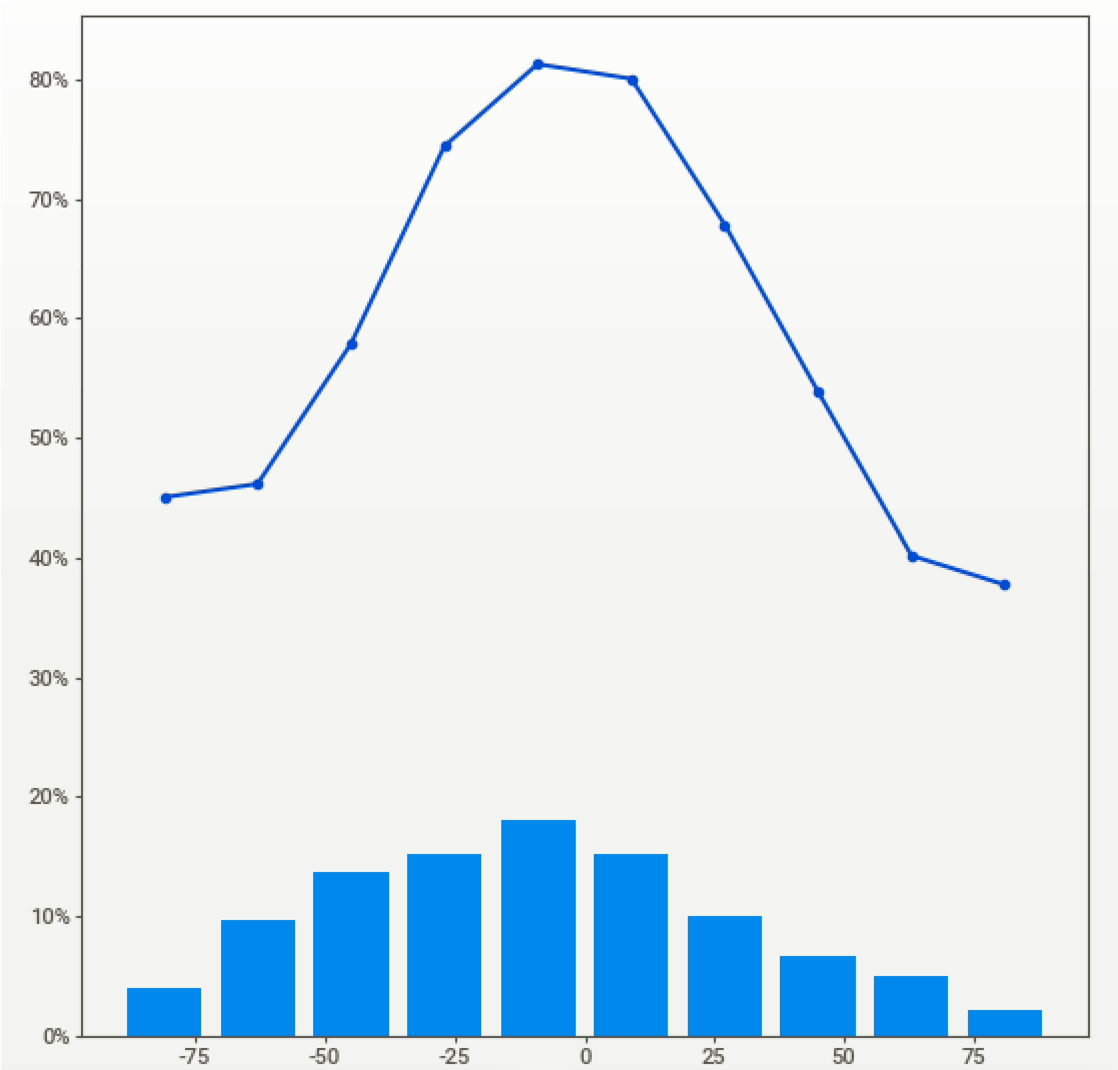
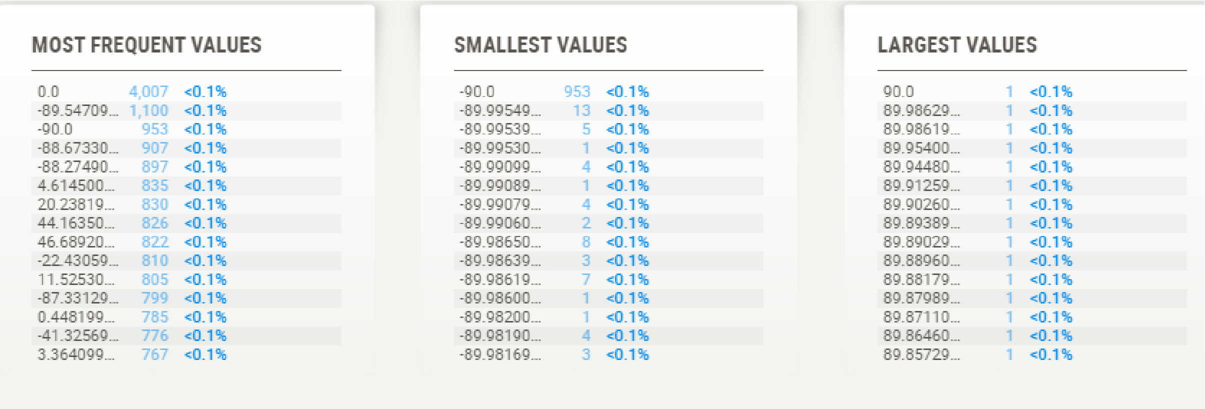


Fig 12: Histogram and Boxplot of angelz





MOST FREQUENT VALUES

0.0	4,007	<0.1%
-89.54709...	1,100	<0.1%
-90.0	953	<0.1%
-88.67330...	907	<0.1%
-88.27490...	897	<0.1%
4.614500...	835	<0.1%
20.23819...	830	<0.1%
44.16350...	826	<0.1%
46.68920...	822	<0.1%
-22.43059...	810	<0.1%
11.52530...	805	<0.1%
-87.33129...	799	<0.1%
0.448199...	785	<0.1%
-41.32569...	776	<0.1%
3.364099...	767	<0.1%

SMALLEST VALUES

-90.0	953	<0.1%
-89.99549...	13	<0.1%
-89.99539...	5	<0.1%
-89.99530...	1	<0.1%
-89.99099...	4	<0.1%
-89.99089...	1	<0.1%
-89.99079...	4	<0.1%
-89.99060...	2	<0.1%
-89.98650...	8	<0.1%
-89.98639...	3	<0.1%
-89.98619...	7	<0.1%
-89.98600...	1	<0.1%
-89.98200...	1	<0.1%
-89.98190...	4	<0.1%
-89.98169...	3	<0.1%

LARGEST VALUES

90.0	1	<0.1%
89.98629...	1	<0.1%
89.98619...	1	<0.1%
89.95400...	1	<0.1%
89.94480...	1	<0.1%
89.91259...	1	<0.1%
89.90260...	1	<0.1%
89.89389...	1	<0.1%
89.89029...	1	<0.1%
89.88960...	1	<0.1%
89.88179...	1	<0.1%
89.87989...	1	<0.1%
89.87110...	1	<0.1%
89.86460...	1	<0.1%
89.85729...	1	<0.1%

Fig 13: Bar Graph of angelz roughly following normal distribution

1. Anglez column roughly follows normal distribution with mean at -9.18 and standard deviation 38.856.
2. Here 0.0 is the most frequently occurring angle.
3. And 3.364 is the least occurring angle.

- **Enmo**

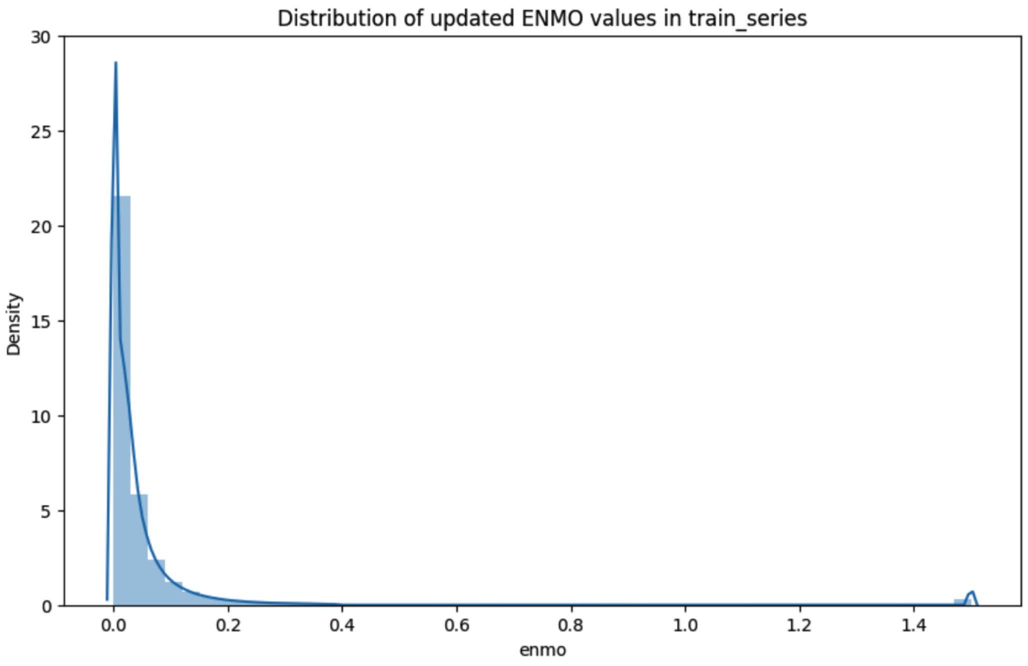


Fig 14: Distribution of updated ENMO values in train_series

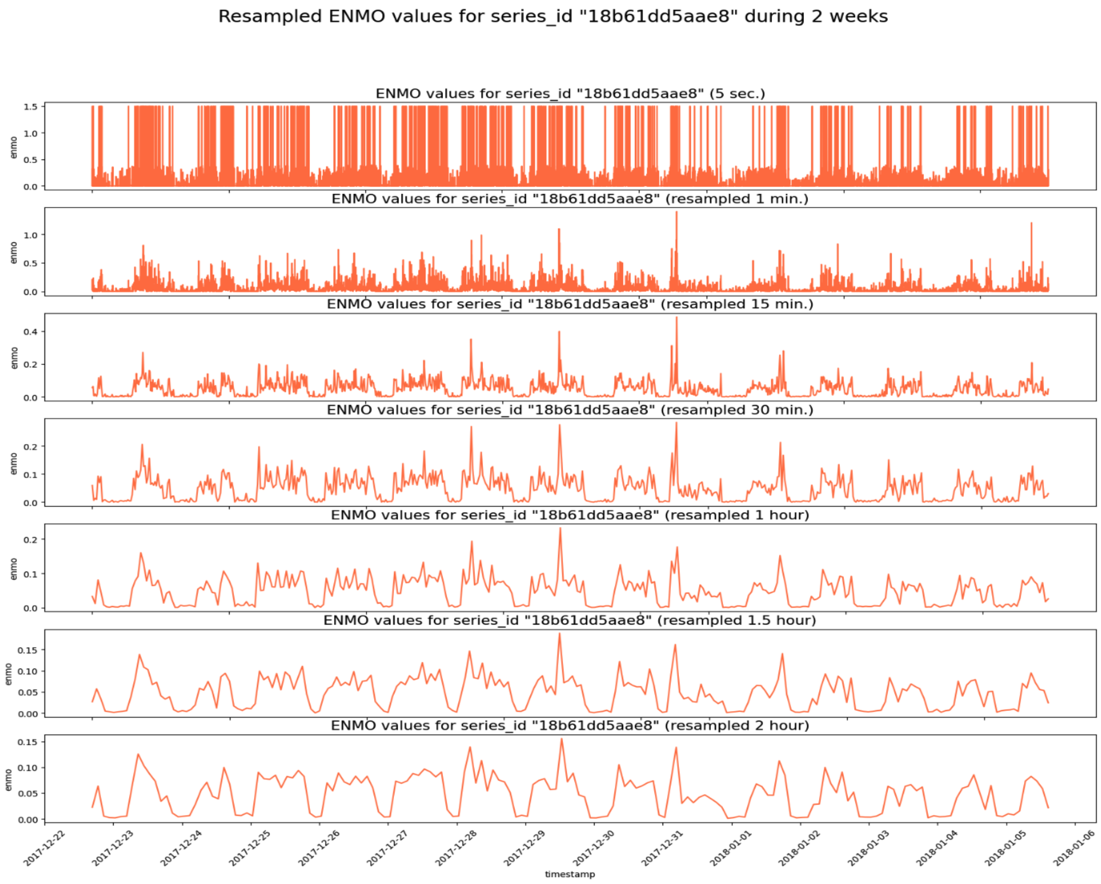
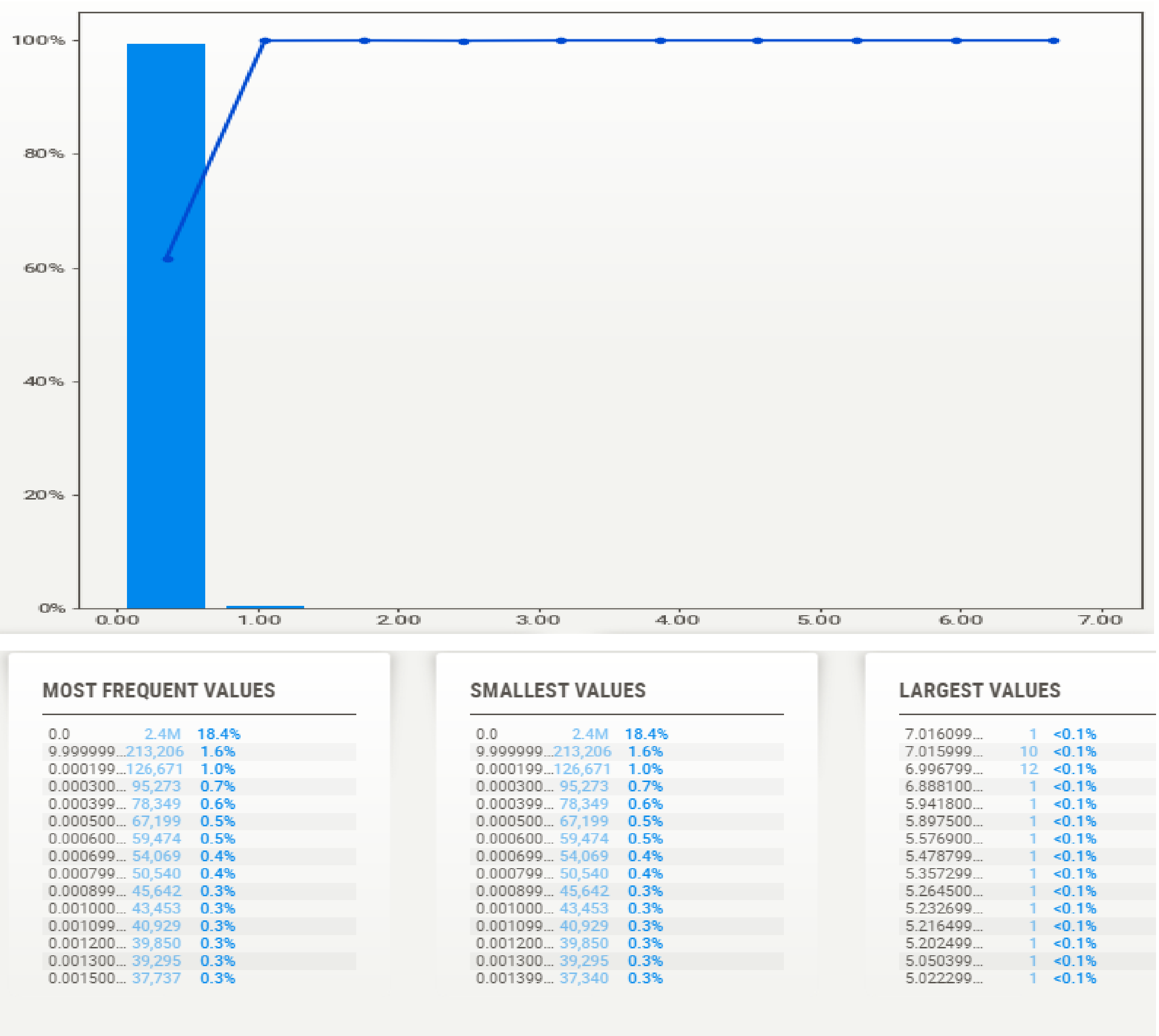


Fig 15: Resampled ENMO values for series_id



MOST FREQUENT VALUES

0.0	2.4M	18.4%
9.999999...	213,206	1.6%
0.000199...	126,671	1.0%
0.000300...	95,273	0.7%
0.000399...	78,349	0.6%
0.000500...	67,199	0.5%
0.000600...	59,474	0.5%
0.000699...	54,069	0.4%
0.000799...	50,540	0.4%
0.000899...	45,642	0.3%
0.001000...	43,453	0.3%
0.001099...	40,929	0.3%
0.001200...	39,850	0.3%
0.001300...	39,295	0.3%
0.001500...	37,737	0.3%

SMALLEST VALUES

0.0	2.4M	18.4%
9.999999...	213,206	1.6%
0.000199...	126,671	1.0%
0.000300...	95,273	0.7%
0.000399...	78,349	0.6%
0.000500...	67,199	0.5%
0.000600...	59,474	0.5%
0.000699...	54,069	0.4%
0.000799...	50,540	0.4%
0.000899...	45,642	0.3%
0.001000...	43,453	0.3%
0.001099...	40,929	0.3%
0.001200...	39,850	0.3%
0.001300...	39,295	0.3%
0.001399...	37,340	0.3%

LARGEST VALUES

7.016099...	1	<0.1%
7.015999...	10	<0.1%
6.996799...	12	<0.1%
6.888100...	1	<0.1%
5.941800...	1	<0.1%
5.897500...	1	<0.1%
5.576900...	1	<0.1%
5.478799...	1	<0.1%
5.357299...	1	<0.1%
5.264500...	1	<0.1%
5.232699...	1	<0.1%
5.216499...	1	<0.1%
5.202499...	1	<0.1%
5.050399...	1	<0.1%
5.022299...	1	<0.1%

Fig 16: ENMO frequency analysis

1. Most of the values of enmo lie near 0 and 1.
2. 0.0 is the most frequent value with 2.4 M instances

- **Pair plot between features**

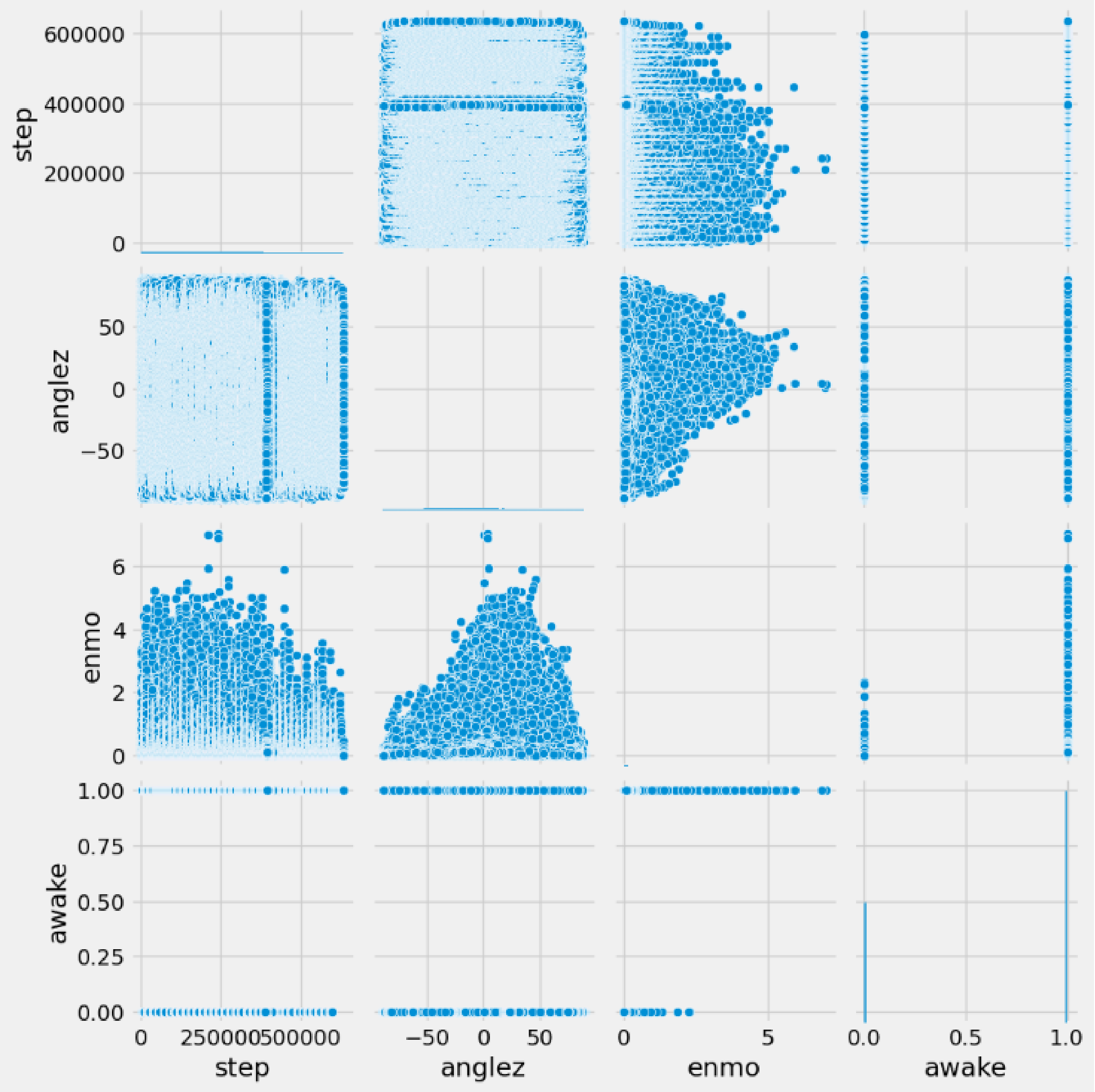


Fig 17: Pair plot between features

1. We can see that in most of the step the value of angle would be either less than -50 or greater than 50 but at particular step near 400000 the value of angle takes different values in each series.
2. Initial step showed less awake time which means in most of the series in initial steps the child is asleep.
3. We don't see any sleeping time with higher enmo values, which indicates sleeping time occurs when the accelerometer enmo values lie near 0.

- **Scatterplot between anglez and enmo with awake as hue**

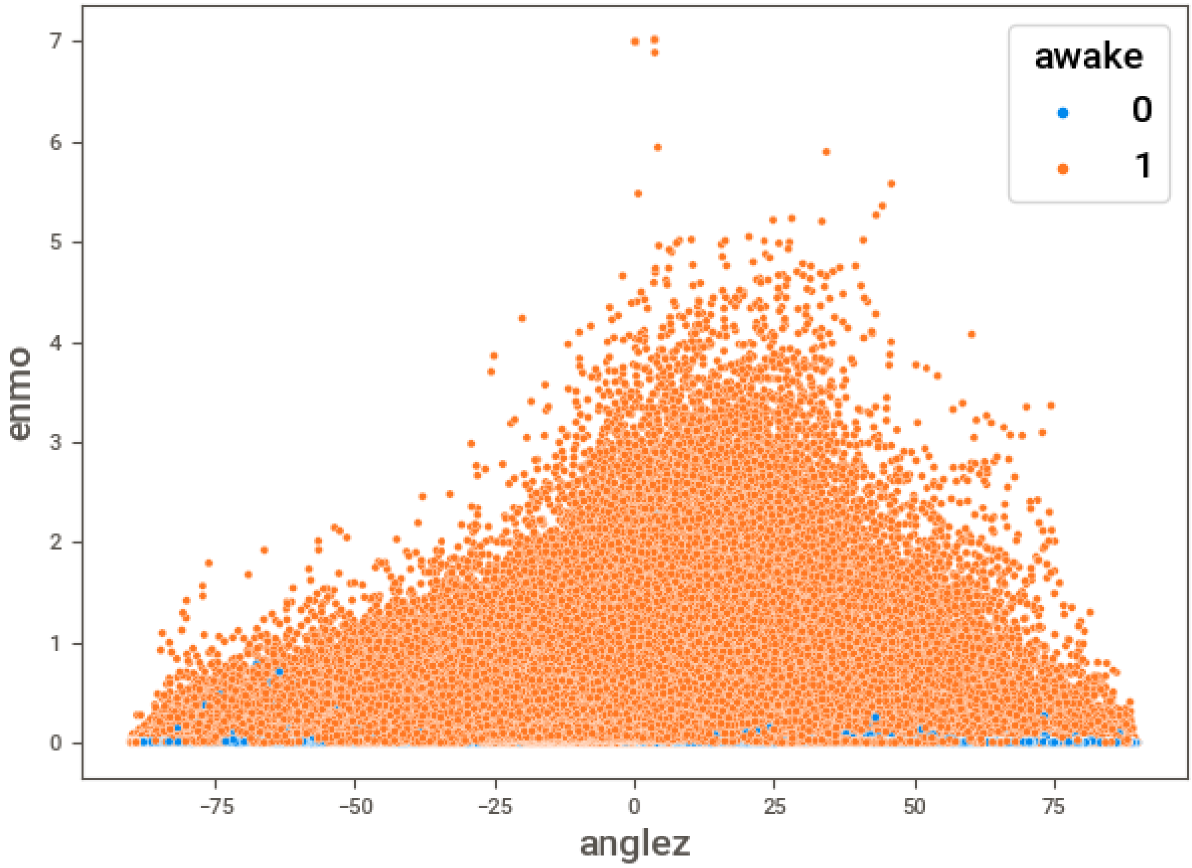


Fig 18: Scatterplot between anglez and enmo with awake as hue

1. It is visible that the time of sleeping is very less compared to that of being awake but the most important thing is that the most of the sleep happened at the value of enmo near 0 and value angle of accelerometer at -75 and 75.

Feature engineering

We added a new column based on ‘anglez’ named ‘anglezdiffabs’. It represents the absolute differences between consecutive values of the ‘anglez’ column and stores the results as 32-bit floating-point numbers.

Further, we create a set of new columns in the DataFrame df based on the columns 'enmo', 'anglez' and 'anglezdiffabs'. These new columns capture various statistics and transformations over different rolling time windows.

1. **For Each Column** ('enmo', 'anglez', 'anglezdiffabs'):
   We perform the same set of operations for each of the three columns.

2. **For Each Time Period** (60, 360, 720, 3600 seconds):
   For each time period specified (60 seconds, 360 seconds, 720 seconds, and 3600 seconds), the code calculates various rolling statistics and creates new columns to store the results.

3. **New Columns for Rolling Statistics:**
   For each combination of the original column, time period, and rolling window parameters (window size, minimum periods, centering), the following new columns are created:
   - {col}_{n}_{agg}: This column stores the rolling statistic (e.g., median, mean, max, min, var) for the original column (col) over the specified time period (n) and with the given aggregation method (agg).
4. **Additional Features:**
   New columns are created in the DataFrame with names that include information about the column name, time period, and aggregation method (e.g., 'enmo_60_mean', 'anglez_360_max').
   The statistics are stored as values with reduced data types (e.g., float16 or float32) to optimise memory usage.

5. **Amplitude and Difference Features:**
   Additional features are calculated based on the rolling statistics:

   - {col}_{n}_amplit: This column captures the amplitude, which is the difference between the rolling maximum and minimum values for the specified time period.

   - {col}_diff_{n}_max: This column calculates the maximum absolute difference between consecutive rolling maximum values for the specified time period.

   - {col}_diff_{n}_mean: This column calculates the mean absolute difference between consecutive rolling minimum values for the specified time period.

In summary, we performed feature engineering by creating a set of new features that capture various statistics and transformations of the original columns ('enmo', 'anglez', and 'anglezdiffabs') over different rolling time windows.

These new features can be used for data analysis, modelling, or other data-driven tasks to potentially capture patterns and relationships in the data over different time scales and to provide information about amplitude and variations within each time window.

Initial Modelling:

The objective can be considered as a binary classification task or multitask classification task or sleep pattern recognition as we are required to predict the longest sleep cycle everyday given the new dataset (say test data).

But the data given for training have far too many limitations like out of the 200+ series ids of data given only 37 of them are complete with the longest sleep cycle on each day defined. Furthermore there is an issue of daylight saving time(for some of the training data) since the data is taken from US people. Which makes it hard to deal with the objective as a classification problem.

One major issue is that if both onset and wakeup is given we can confidently say that the participant is sleeping in the window but it's hard to comment on the other parts as the participant may be sleeping for a short duration and in other cases he might not be even wearing the wrist watch. So even if the participant is not in the particular sleep hours given, it's hard to conclude he is awake. This insight gives rise to considering it as a pattern recognition objective and effective use of RNN and related models.

As an initial approach RNN will give a higher score but if we do correct data manipulation and filter out the data we can actually use without misleading the model and feature engineer it to a particular extent other classification based models will also start to give higher scores and in fact it might be the ones which gives higher score than RNN.

Obtained Result:

Feature Engineering:

After doing feature engineering, we have obtained the following columns. Thus our new dataset contains 120 columns.

Some correlation heatmaps associated with these features :

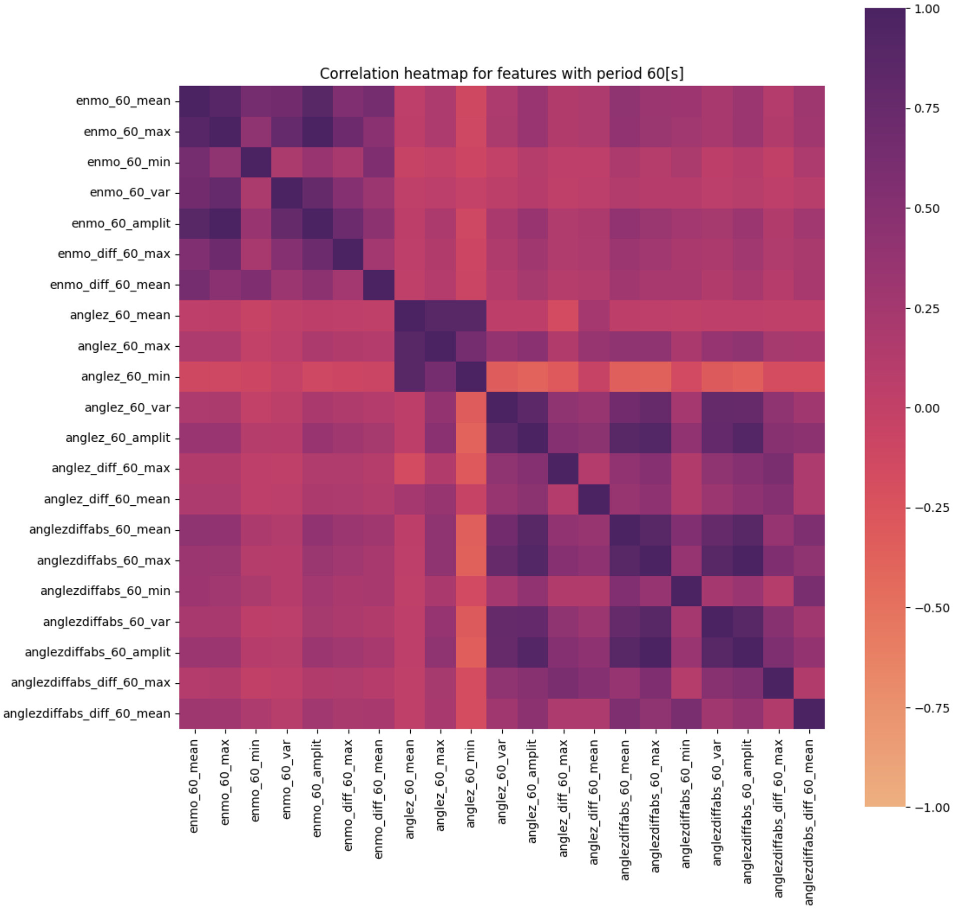


Fig 19: Correlation heatmap for features with period 60(s)

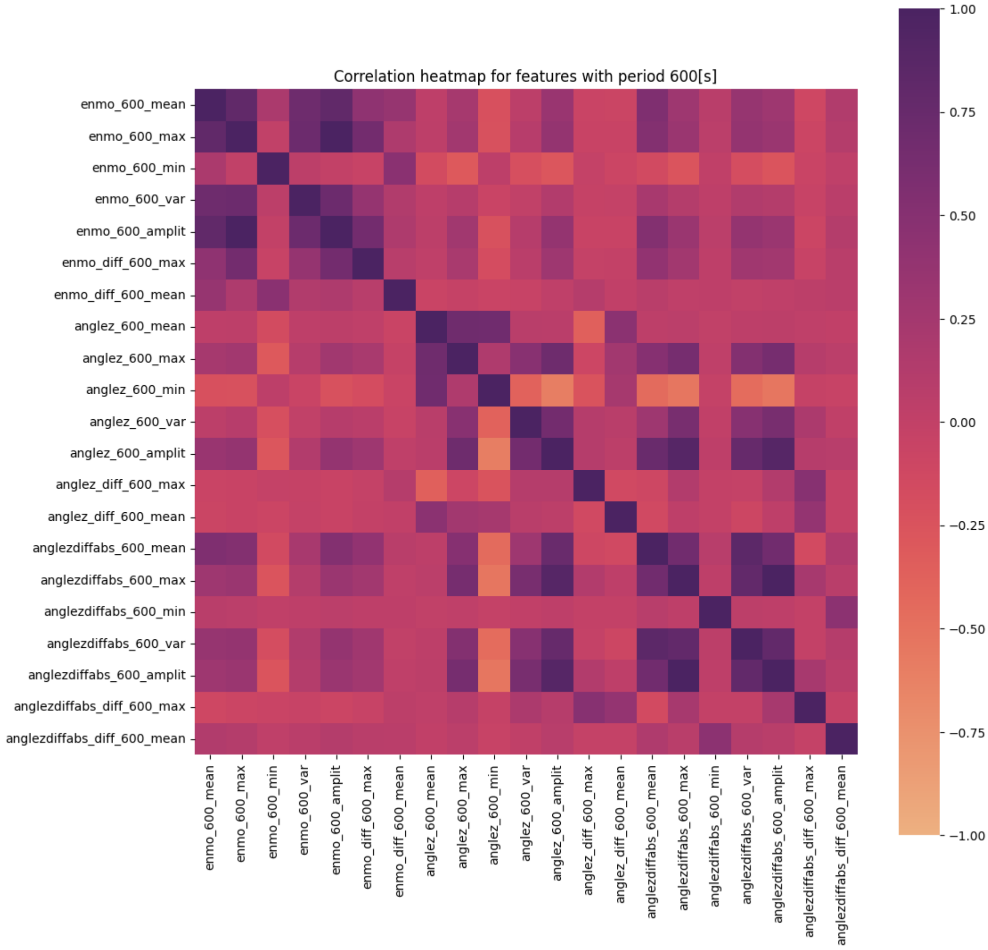


Fig 20: Correlation heatmap for features with period 600(s)

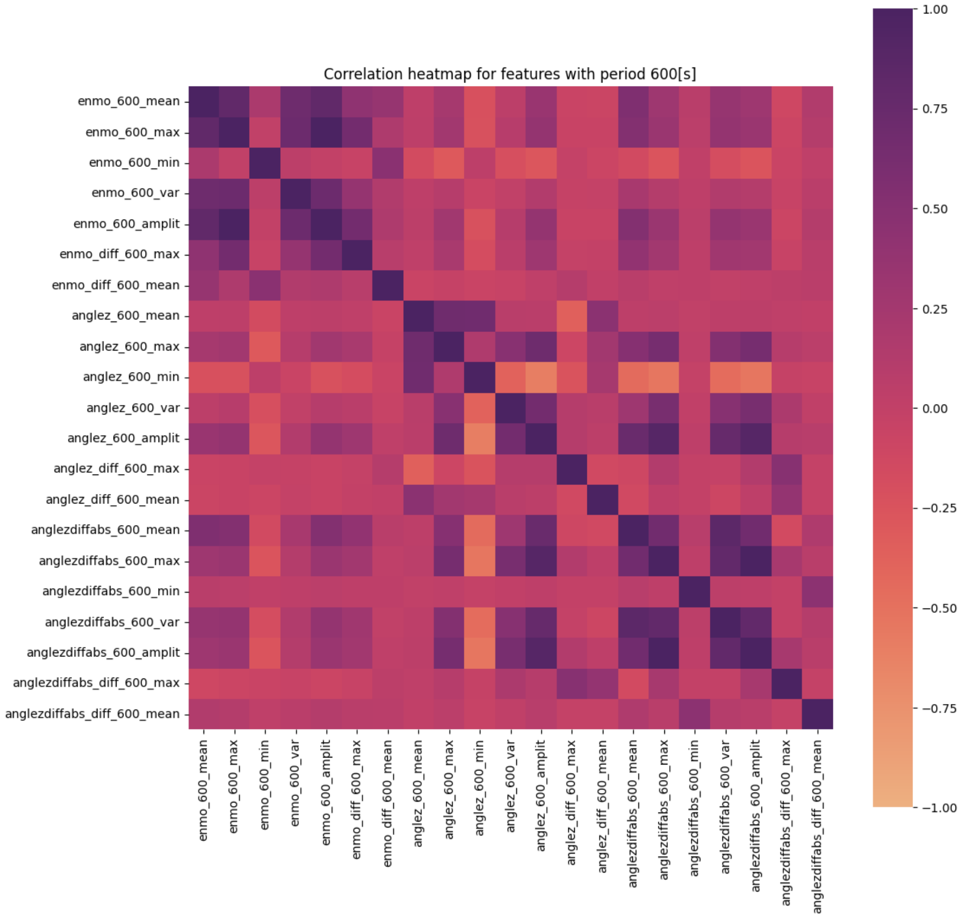


Fig 21: Correlation heatmap for features with period 60(s)

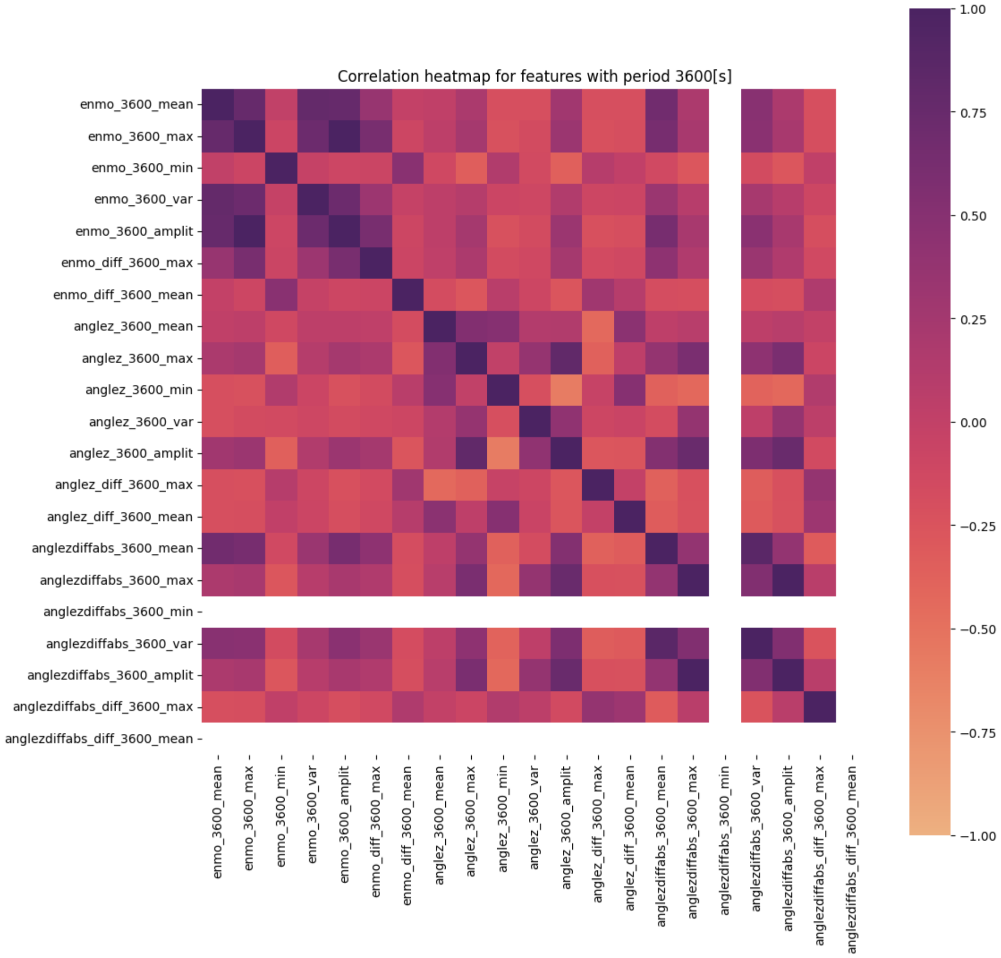


Fig 22: Correlation heatmap for features with period 3600(s)

From these correlation heatmaps we can say that:

1. enmo_{p}_max and enmo_{p}_amplit are highly correlated for each aggregation window, which is logical since enmo is normally fluctuating around zero, with a limited number of sharp spikes. It's worth dropping one of them.
2. anglezdiffabs_{p}_max and anglezdiffabs_{p}_amplit are highly correlated for each aggregation window.
3. anglezdiffabs_3600_min and anglezdiffabs_600_min are just columns of zeros (see below), useless feature aggregation for large periods.
4. anglezdiffabs_diff_3600_mean and anglezdiffabs_diff_600_mean - same issue.
5. enmo_3600_min and enmo_600_min values are also all close to zero and make no sense

Top 50 features by correlation with the target:

anglez_3600_min	0.300191
anglez_3600_max	0.301631
anglez_diff_3600_mean	0.302583
enmo_60_max	0.305192
anglezdiffabs_diff_60_max	0.314172
enmo_diff_300_max	0.317025
anglezdiffabs_diff_3600_max	0.320073

```
anglezdiffabs_60_var           0.322677
enmo_60_mean                   0.325824
anglez_diff_60_max             0.337915
anglez_diff_60_mean            0.339530
anglezdiffabs_diff_60_mean     0.344148
anglezdiffabs                  0.347098
enmo_diff_600_max              0.348327
anglez_300_min                 0.360647
anglez_diff_3600_max           0.363111
enmo_300_max                   0.368225
enmo_300_amplit                0.369672
anglez_600_min                 0.380604
anglez_300_max                 0.385129
enmo_300_mean                  0.387068
enmo_600_max                   0.404006
anglez_600_max                 0.405358
enmo_600_amplit                0.405517
enmo_diff_3600_max             0.408280
enmo_600_mean                  0.416992
anglez_3600_amplit             0.421008
anglezdiffabs_300_var          0.479343
anglezdiffabs_600_amplit       0.509603
anglezdiffabs_600_max          0.509626
hour                           0.518312
enmo_3600_max                  0.519454
enmo_3600_amplit               0.519514
enmo_3600_mean                 0.526689
anglezdiffabs_60_amplit        0.536042
anglezdiffabs_600_var          0.539413
anglezdiffabs_60_max           0.540071
anglezdiffabs_300_amplit       0.553711
anglezdiffabs_300_max          0.553931
anglez_60_amplit               0.570143
anglezdiffabs_60_mean          0.585693
anglez_600_amplit              0.604697
anglez_300_amplit              0.626600
anglezdiffabs_3600_var         0.637513
anglezdiffabs_300_mean         0.692686
anglezdiffabs_600_mean         0.729521
anglezdiffabs_3600_mean        0.805335
lids                           0.811964
```

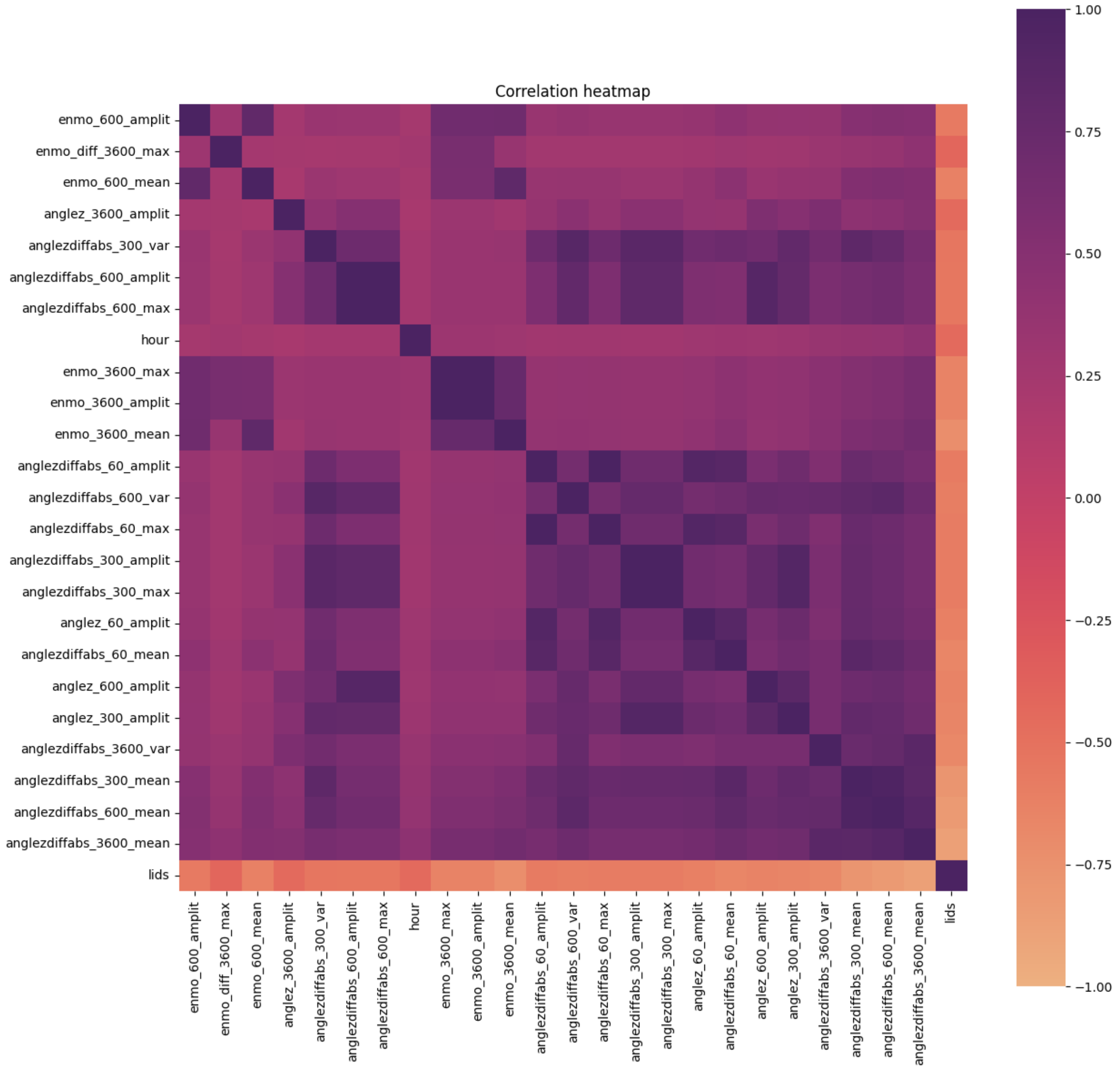


Fig 23: Correlation heatmap

We will be using the top features and exclude the least important features for training our model.

Strategy for further work:

Looking ahead, the roadmap for the remaining stages of the project includes the following steps:

Currently we are using only 35 of the available 277 parquet data because the other participants from whom the data is recorded is not complete and we need to do further data processing and enhancement so that we can successfully use that to train our model**.**

- **DATA PROCESSING:**

We will still spend a significant amount of time in engineering and processing the data so that we can get usable data from the ambiguous data that is provided to us. The major issue we facing with the data is that there is so many NAN values in the gold labels so that we can't effectively take the data in those nights and other major issue we are facing is that the awake information is not persistent (ie; The gold labels only mention the longest sleep data of the participants so if the participant slept for 2hrs in the middle and had a 6 hr sleep later the 2 hr sleep info is not there in gold label so we can’t effectively conclude that a participant is awake even if he is not sleeping).

- **RNN:**

The RNN will continue to work well in this context of our problem as we can continue to recognize the sleeping parts in the data as we have the gold labels for the same from the csv file provided to train the data. As we do more feature engineering and have confidence in the other part of the data we can try to classify the parts in the data which are awake and try to use that to further enhance the accuracy of RNN.

- **CLASSIFICATION PROBLEM:**

We can even consider the problem as a classification problem as we need to predict if the candidate is sleeping or not from the input variables. We can consider this as either multi class classification problem with three classes say { sleep, awake, other} or we can consider it as a binary classification problem with only two classes say { Sleep, Not sleep}.

Currently with the processing we did till now on the data, the classification problems are not giving good scores. But if we can successfully find a way to use more of the training data given with future approaches in data processing we can carry on to use those data to improve the score using classification problems.

Then we can use various other enhancement techniques too if we consider the problem as classification problem like Random forest multiclass problem, Other ensemble methods with enhancements using XGBoost.