Machine Learning Prediction in mHealth Data

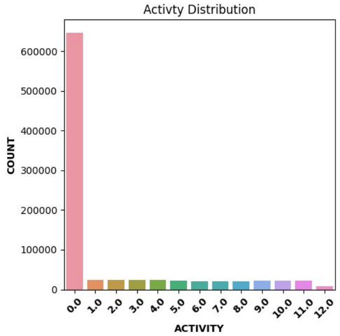
Team 15: Xiaoxuan Han, Hunter Ponzzebon, Gwen Eagle

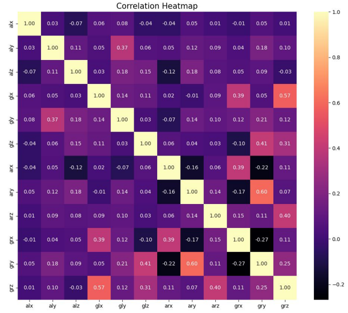
Introduction:

- ❖ **Dataset:**
  - ➢ mHealth data (999,999 records, 14 variables);
  - ➢ Time-series data on linear and angular motion on the x-,y-,z-axis during *13 different activities*
- ❖ **Goal:** To build a machine learning model that can accurately ascertain activity type based on multidimensional time series data
  - ➢ Recurrent Neural Network (RNN)
  - ➢ Long Short-Term Memory Neural Network (LSTMNN)

Data Engineering & EDA :

- ❖ **Data Engineering:**
  - ➢ Data split into 80% for training and 20% for test;
  - ➢ Data normalization;
  - ➢ Addressed class imbalance: undersampled *class 0* and oversampled *class 12;*
  - ➢ Sequence generation
- ❖ **EDA:**
  - ➢ No missing data;
  - ➢ Correlation heatmap on *12 sensors*





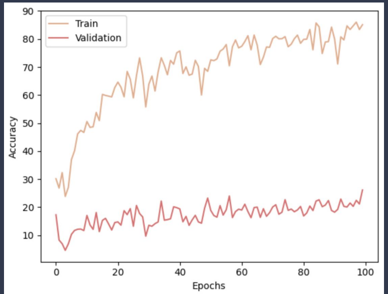
	alx	aly	alz	glx	gly	glz	arx	ary	arz	grx	gry	grz
alx	1.00	0.03	-0.07	0.06	0.08	-0.04	-0.04	0.05	0.01	-0.01	0.05	0.01
aly	0.03	1.00	0.11	0.05	0.37	0.06	0.05	0.12	0.09	0.04	0.18	0.10
alz	-0.07	0.11	1.00	0.03	0.18	0.15	-0.12	0.18	0.08	0.05	0.09	-0.03
glx	0.06	0.05	0.03	1.00	0.14	0.11	0.02	-0.01	0.09	0.39	0.05	0.57
gly	0.08	0.37	0.18	0.14	1.00	0.03	-0.07	0.14	0.10	0.12	0.21	0.12
glz	-0.04	0.06	0.15	0.11	0.03	1.00	0.06	0.04	0.03	-0.10	0.41	0.31
arx	-0.04	0.05	-0.12	0.02	-0.07	0.06	1.00	-0.16	0.06	0.39	-0.22	0.11
ary	0.05	0.12	0.18	-0.01	0.14	0.04	-0.16	1.00	0.14	-0.17	0.60	0.07
arz	0.01	0.09	0.08	0.09	0.10	0.03	0.06	0.14	1.00	0.15	0.11	0.40
grx	-0.01	0.04	0.05	0.39	0.12	-0.10	0.39	-0.17	0.15	1.00	-0.27	0.11
gry	0.05	0.18	0.09	0.05	0.21	0.41	-0.22	0.60	0.11	-0.27	1.00	0.25
grz	0.01	0.10	-0.03	0.57	0.12	0.31	0.11	0.07	0.40	0.11	0.25	1.00

Model Comparisons

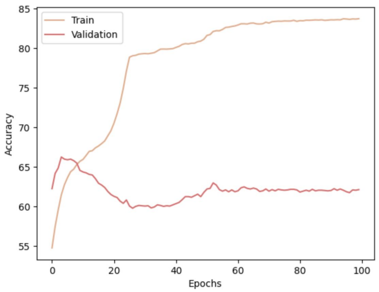
Features: 12 motions (linear and angular motion on the x-,y-,z-axis)
Target: Activity (multi-classes vs. binary classes)

RNN:

- 256 hidden units, 100 epochs, learning rate = 0.001/0.00001, batch size = 512, regularization term = 0.0001/0.00001



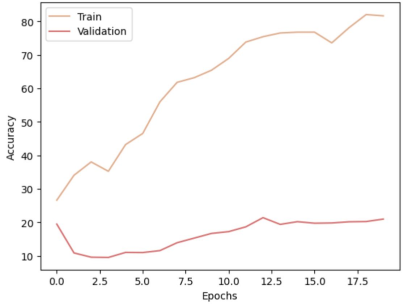
Multi-class (13 classes)



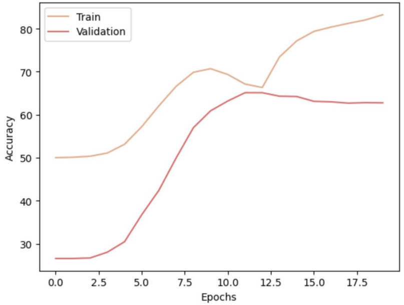
Binary class

LSTMNN:

- 256 hidden units, 20 epochs, learning rate = 0.001/0.00001, a batch size = 512, regularization term = 0.0001/0.00001



Multi-class (13 classes)



Binary class

Findings: The RNN and LSTMNN model for the activity type performed similarly (validation accuracy of ~20%); the RNN and LSTMNN model for the presence of activity performed similarly (validation accuracy of ~62%).
Limitations: A high degree of overfitting might have occurred in both models, potentially due to the models trying to predict a high number of classes based on a low number of subjects, as well as a high degree of class imbalance.