MechE 395: Mechanistic Data Science HW 2

Real Life Problem: Multi-stage continuous-flow manufacturing process

Link to database:
https://www.kaggle.com/supergus/multistage-continuousflow-manufacturing-process

Problem statement
The problem presented builds on data collected from a multi stage continuous flow process. The process has 5 machines: the first 3 machines are run in parallel. Following this, their outputs are combined into a single stream. The output from the combiner is measured in 15 locations.
Following this, the output stream is transferred to a second stage where they process through machines 4 and 5 in series. The final output is measured once again in the same 15 locations.
The measurements are sampled at a rate of 1 Hz. In total, there are 14,088 measurements.

Goal: The goal is to be able to predict the values of the measurements or their errors against the target/setpoint at their respective locations based on the input parameters and operating conditions of the machines.

The dataset has 116 columns and 14088 rows (Not including data label)

Approach using the 6 steps of MDS
There are multiple types of data collected at different locations through the process. The data includes timestamps, ambient factory conditions, raw material properties, process variables and measurements at the 15 locations. It also includes the setpoint / target values for each of the 15 locations.
All the data is continuous data. It is not perfect and will require cleaning/processing. I will normalise the data where necessary. I am not sure as yet, but I am thinking of normalizing the measured values against each of the setpoints so that the relative deviations for each location can be computed more easily.
Following this, the mechanistic features that influence the data will be identified. I will draw a correlation matrix and perform PCA analyses in order to find the most relevant mechanistic features and decrease the number of random variables.
Since the dataset requires a prediction of the deviations of the measured values from the setpoints, I will employ neural networks to perform regression analyses. These will help me draw mathematical relationships between the features (input properties and process variables) and outcomes (measurements at the 15 locations). Lastly, I will tie all my findings together in order to develop a cohesive prediction model that can be used to tune the necessary input parameters and process variables to enable the most ideal production.