**Data description:**

Data.zip contains two files: data\_fiber.csv and data\_baseline.csv.

**data\_fiber.csv:**

Contains 1092 rows and 301 columns. The rows correspond to different subjects with different conditions and given treatments. The rows in the **'subject'** column that have the same ID correspond to one subject. The **'condition'** column corresponds to the condition each subject have. There are five conditions: Crohn’s disease (**'CD'**), Parkinson’s disease (**'PD'**), ulcerative colitis (**'UC'**), human immunodeficiency virus (**'HIV'**), and healthy control (**'HC'**). The **'Treatment'** column corresponds to the type of fiber treatment given to patients. There are 14 treatment types **'FosPsProCgWb'**, **'Feces'**, **'Blank12'**, **'Blank24'**, **'PsWb'**, **'PsCg'**, **'ProWb'**, **'PsProWb'**, **'FosPro'**, **'FOS'**, **'bGlucan'**, **'Pectin'**, **'SAX'**, **'Mixture'**. The **'Feces'** treatments correspond to the baseline data, where treatment is given. The **'time'** column corresponds to the time at which measurements are done after the treatments are given. Columns **'Acetate'** to **'Gas'** correspond to the concentration of the by-products in the gut. Columns **'Abiotrophia'** to **'unassigned unassigned'** correspond to different bacterial classes at the genus level with OTU-based values. The **'fiber'** columns demonstrate the type of fiber treatments based on the concentration of fibers (high\_fiber, medium\_fiber, no\_fiber), and the columns **'Treatment\_codes'**, **'condition\_codes'**, and the **'fiber\_codes'** are the numeric representation of the categorical **'Treatment'**, **'condition'**and **'fiber'** columns.

|  |
| --- |
| Table  Description automatically generated |
| Table  Description automatically generated |

Figure 1 The fiber data

**data\_baseline.csv:**

It is similar to the data\_fiber.csv except that only conditions with Feces value are present.

**data\_baseline\_augmented.csv:**

It is similar to the data\_fiber.csv except that has more HIV cases.

**Prediction\_data.csv:**

Contains the predictions and actual values for each fold in the 5-fold cross-validation for different machine learning algorithms and datasets. For example, the columns for predictions\_vs\_actual\_ANN\_fiber\_HC\_NH.csv are shown in Figure 2.

Table

Description automatically generated with medium confidence

Figure 2 predictions vs actual data

The columns **'y\_test\_fold\_i'**, '**y\_pred\_fold\_i'** and **'y\_score\_class\_j\_fold\_i**' represent the predicted label, actual labels and prediction score for fold i and class j.

**Size\_variation\_data.csv:**

Contains accuracy data as a function of size for the fiber and baseline data. For example, the columns for accuracies\_CNN\_baseline.csv are shown in Figure3.

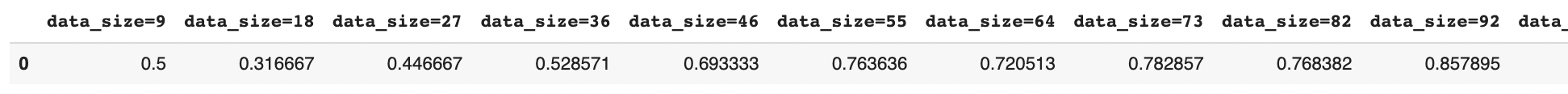


Figure 3 predictions vs actual data