Proper Dataset Documentation

“Alzheimer’s Disease Prediction Using Handwriting and AI Models for Cognitive Assessment”

Dataset Overview

Dataset Link: [Kaggle Dataset] (https://www.kaggle.com/datasets/taeefnajib/handwriting-data-to-detect-alzheimers-disease)

Dataset Collection

- Source: provided by the corresponding supervisor.
- Collection Method: Not explicitly mentioned, but typically involves collection of handwriting data from participants, possibly in a clinical or research setting.

Corresponding Authors (Supervisors)

1. Md Sipon Miah
    - IVR Low-Carbon Research Institute, Chang’an University, Shaanxi, 710018, China.
2. Mingbo Niu
    - Dept. of Information and Communication Technology, Islamic University, Kushtia, 7003, Bangladesh.
    - Dept. of Signal Theory and Communications, University Carlos III of Madrid (UC3M), Leganes, 28911, Madrid, Spain.

Variables

- Total Variables: 452 (including class label)
- Key Variables: ID, air_time1, disp_index1, gmrt_in_air1, gmrt_on_paper1, max_x_extension1, max_y_extension1, mean_acc_in_air1, mean_acc_on_paper1, mean_gmrt1, etc.
- Target Variable: 'class' (not explicitly described but likely indicating the presence or absence of Alzheimer’s Disease)

Statistical Test for Validation

- Method Used: Correlation matrix to understand the relationship between different handwriting features.
- Observations:
- A large number of NaN values, indicating a lack of correlation or absence of data between many pairs of variables.
- Certain variables show a significant correlation, suggesting potential relevance in predicting Alzheimer’s disease.

Correlation Analysis for a Specific Variable (air_time1)

- shows the highest correlation with 'total_time1' and 'paper_time1'.
- Some variables have negative correlations, indicating an inverse relationship.

Dataset Usage

This dataset can be used to develop and test AI models for cognitive assessment, particularly in predicting Alzheimer's Disease through handwriting analysis. The dataset's extensive variables offer a comprehensive set of features for exploring potential handwriting markers associated with cognitive decline.

Limitations and Considerations

- The dataset appears to have many NaN values in the correlation matrix, suggesting the need for careful preprocessing and feature selection.
- The specific methodology of data collection and the demographic information of the participants are not detailed, which are crucial for understanding the generalizability of the findings.

Variables:

```
Index(['ID', 'air_time1', 'disp_index1', 'gmrt_in_air1',
'gmrt_on_paper1', 'max_x_extension1', 'max_y_extension1',
'mean_acc_in_air1', 'mean_acc_on_paper1', 'mean_gmrt1', ...
'mean_jerk_in_air25', 'mean_jerk_on_paper25', 'mean_speed_in_air25',
'mean_speed_on_paper25', 'num_of_pendown25', 'paper_time25',
'pressure_mean25', 'pressure_var25', 'total_time25', 'class'],
dtype='object', length=452)
```

Statistical Test for validation the dataset:

Correlation matrix:

```
                   ID  air_time1  disp_index1  gmrt_in_air1
gmrt_on_paper1  \
ID                NaN        NaN          NaN           NaN
NaN
air_time1         NaN   1.000000     0.361259     -0.232732       -
0.207819
disp_index1       NaN   0.361259     1.000000     -0.246372       -
0.385838
gmrt_in_air1      NaN  -0.232732    -0.246372      1.000000
0.587802
gmrt_on_paper1    NaN  -0.207819    -0.385838      0.587802
1.000000
...                ..        ...          ...           ...
...
num_of_pendown25  NaN   0.023792     0.039369     -0.128569       -
0.052636
paper_time25      NaN   0.110482     0.153675     -0.170229       -
0.076708
pressure_mean25   NaN  -0.041226    -0.111902      0.088930
0.113275
pressure_var25    NaN   0.091019     0.066591     -0.119840       -
0.045748
total_time25      NaN   0.018695     0.097970     -0.084335       -
0.093643

                   max_x_extension1  max_y_extension1  mean_acc_in_air1
\
ID                              NaN               NaN               NaN
air_time1                  0.225954          0.014668          0.119806
disp_index1                0.127803          0.605181         -0.136236
gmrt_in_air1               0.133476          0.206822          0.342540
gmrt_on_paper1             0.245255          0.203036          0.294892
...                             ...               ...               ...
num_of_pendown25           0.066281         -0.012220         -0.076861
paper_time25              -0.026867          0.015182         -0.096416
pressure_mean25           -0.043523          0.060988          0.027171
pressure_var25             0.074755         -0.099829          0.001428
total_time25               0.054187         -0.002383         -0.031624
```

ID	mean_acc_on_paper1	mean_gmrt1	...	mean_gmrt25	\
	NaN	NaN	...	NaN	
air_time1	-0.128735	-0.248715	...	-0.114213	
disp_index1	-0.283894	-0.333046	...	-0.158105	
gmrt_in_air1	0.455927	0.940325	...	0.192780	
gmrt_on_paper1	0.875478	0.828012	...	0.169428	
...	...	...	...	...	
num_of_pendown25	-0.067840	-0.111249	...	-0.511888	
paper_time25	-0.045188	-0.150248	...	-0.645693	
pressure_mean25	0.069150	0.109281	...	0.473979	
pressure_var25	-0.002497	-0.102302	...	-0.470591	
total_time25	-0.060157	-0.097839	...	-0.200133	

ID	mean_jerk_in_air25	mean_jerk_on_paper25	\
	NaN	NaN	
air_time1	-0.068343	-0.117830	
disp_index1	-0.135669	0.038488	
gmrt_in_air1	0.187074	0.023682	
gmrt_on_paper1	0.149880	-0.121669	
...	...	...	
num_of_pendown25	-0.484924	-0.085739	
paper_time25	-0.509041	-0.332458	
pressure_mean25	0.436551	0.231458	
pressure_var25	-0.473727	0.013114	
total_time25	-0.131651	-0.173553	

ID	mean_speed_in_air25	mean_speed_on_paper25	\
	NaN	NaN	
air_time1	-0.109298	-0.088565	
disp_index1	-0.110590	-0.152034	
gmrt_in_air1	0.206675	0.127726	
gmrt_on_paper1	0.160714	0.111510	
...	...	...	
num_of_pendown25	-0.478635	-0.390462	
paper_time25	-0.551218	-0.600186	
pressure_mean25	0.430762	0.447979	
pressure_var25	-0.428663	-0.359515	
total_time25	-0.159280	-0.193496	

ID	num_of_pendown25	paper_time25	pressure_mean25	\
	NaN	NaN	NaN	
air_time1	0.023792	0.110482	-0.041226	
disp_index1	0.039369	0.153675	-0.111902	
gmrt_in_air1	-0.128569	-0.170229	0.088930	
gmrt_on_paper1	-0.052636	-0.076708	0.113275	
...	...	...	...	
num_of_pendown25	1.000000	0.538560	-0.237773	
paper_time25	0.538560	1.000000	-0.363841	
pressure_mean25	-0.237773	-0.363841	1.000000	
pressure_var25	0.565881	0.418078	-0.142470	
total_time25	0.126791	0.245180	-0.093414	

ID	pressure_var25	total_time25
	NaN	NaN
air_time1	0.091019	0.018695
disp_index1	0.066591	0.097970
gmrt_in_air1	-0.119840	-0.084335
gmrt_on_paper1	-0.045748	-0.093643

```
...                          ...          ...
num_of_pendown25        0.565881     0.126791
paper_time25            0.418078     0.245180
pressure_mean25        -0.142470    -0.093414
pressure_var25          1.000000     0.124291
total_time25            0.124291     1.000000
```

Sort Correlations for a Specific Variable:

```
air_time1               1.000000
total_time1             0.972902
paper_time1             0.585619
num_of_pendown1         0.509712
paper_time23            0.490596
                          ...
mean_jerk_on_paper23   -0.210078
mean_speed_on_paper1   -0.219192
gmrt_in_air1           -0.232732
pressure_mean14        -0.244316
mean_gmrt1             -0.248715
```