Project_2_1

July 28, 2020

1 Project 2 Exercise 1. Audiology databases preprocessing

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from os import walk
import re
```

1.1 1. Read files

[2]:
```
data_path = './data'
```

[3]:
```
for root, dirs, files in walk(data_path):
    all_files = [data_path+'/'+file for file in files if file.split('.')[-1] in␣
 ↪['data', 'names']]

for file_num in range(len(all_files)):
    print(f'#{file_num} ->', all_files[file_num])
```

```
#0 -> ./data/audiology.names
#1 -> ./data/audiology.data
#2 -> ./data/audiology.standardized.data
#3 -> ./data/audiology.standardized.names
```

[4]:
```
# in here we choose the index from above result. for example here we want data␣
 ↪file with index 2
# and data names file with index 3

# open names file
names_file = [line.strip() for line in open(all_files[3], 'r').readlines()]

print_flag = False
attrs = []
for line in names_file:
    if 'Missing attributes' in line:
        break
    if print_flag:
        attr = re.findall("(\w*)\(*\)*:", line)

        if len(attr) > 0:
            attrs.append(attr[0])

    if 'Attribute' in line:
        print_flag = True

attrs.insert(-1,'indentifier')

# print all attributes
print('attributes count: ', len(attrs), '\n'*2)
print(attrs)
```

```
attributes count:  71


['age_gt_60', 'air', 'airBoneGap', 'ar_c', 'ar_u', 'bone', 'boneAbnormal',
'bser', 'history_buzzing', 'history_dizziness', 'history_fluctuating',
'history_fullness', 'history_heredity', 'history_nausea', 'history_noise',
'history_recruitment', 'history_ringing', 'history_roaring', 'history_vomiting',
'late_wave_poor', 'm_at_2k', 'm_cond_lt_1k', 'm_gt_1k', 'm_m_gt_2k', 'm_m_sn',
'm_m_sn_gt_1k', 'm_m_sn_gt_2k', 'm_m_sn_gt_500', 'm_p_sn_gt_2k', 'm_s_gt_500',
'm_s_sn', 'm_s_sn_gt_1k', 'm_s_sn_gt_2k', 'm_s_sn_gt_3k', 'm_s_sn_gt_4k',
'm_sn_2_3k', 'm_sn_gt_1k', 'm_sn_gt_2k', 'm_sn_gt_3k', 'm_sn_gt_4k',
'm_sn_gt_500', 'm_sn_gt_6k', 'm_sn_lt_1k', 'm_sn_lt_2k', 'm_sn_lt_3k',
'middle_wave_poor', 'mod_gt_4k', 'mod_mixed', 'mod_s_mixed', 'mod_s_sn_gt_500',
'mod_sn', 'mod_sn_gt_1k', 'mod_sn_gt_2k', 'mod_sn_gt_3k', 'mod_sn_gt_4k',
'mod_sn_gt_500', 'notch_4k', 'notch_at_4k', 'o_ar_c', 'o_ar_u', 's_sn_gt_1k',
's_sn_gt_2k', 's_sn_gt_4k', 'speech', 'static_normal', 'tymp',
'viith_nerve_signs', 'wave_V_delayed', 'waveform_ItoV_prolonged', 'indentifier',
'class']
```

[5]:

```
data_df = pd.read_csv(all_files[2])
data_df.columns = attrs
data_df = data_df.replace('?', np.NaN)
data_df
```

[5]:

```
    age_gt_60       air airBoneGap    ar_c    ar_u        bone boneAbnormal  \
0           f  moderate          f  normal  normal         NaN            t
1           t      mild          t     NaN  absent        mild            t
2           t      mild          t     NaN  absent        mild            f
3           t      mild          f  normal  normal        mild            t
4           t      mild          f  normal  normal        mild            t
..        …       …        …     …     …        …          …
```

```
194         t      mild          f  absent  normal        mild           t
195         t      mild          f  normal  absent        mild           f
196         f    normal          f  normal  normal  unmeasured           f
197         t      mild          f  normal  normal  unmeasured           f
198         t    normal          f  normal  normal  unmeasured           f

         bser history_buzzing history_dizziness  … s_sn_gt_2k s_sn_gt_4k  \
0         NaN               f                 f  …          f          f
1         NaN               f                 f  …          f          f
2         NaN               f                 f  …          f          f
3         NaN               f                 f  …          f          f
4         NaN               f                 f  …          f          f
..        …             …               …  …        …        …
194       NaN               f                 f  …          f          f
195       NaN               f                 f  …          f          f
196  degraded               f                 f  …          f          f
197       NaN               f                 f  …          f          f
198       NaN               f                 f  …          f          f

        speech static_normal tymp viith_nerve_signs wave_V_delayed  \
0       normal             t    a                 f              f
1       normal             t   as                 f              f
2       normal             t    b                 f              f
3         good             t    a                 f              f
4    very_good             t    a                 f              f
..         …            … …               …            …
194  very_good             t    a                 f              f
195  very_good             t    c                 f              f
196     normal             f    a                 f              f
197  very_good             t    a                 f              f
198     normal             t    a                 f              f

    waveform_ItoV_prolonged indentifier                            class
0                          f          p2                 cochlear_unknown
1                          f          p3      mixed_cochlear_age_fixation
2                          f          p4  mixed_cochlear_age_otitis_media
3                          f          p5                     cochlear_age
4                          f          p6                     cochlear_age
..                      …         …                              …
194                        f        p196                     cochlear_age
195                        f        p197  mixed_cochlear_age_otitis_media
196                        f        p198      possible_brainstem_disorder
197                        f        p199                     cochlear_age
198                        f        p200                     cochlear_age

[199 rows x 71 columns]
```

1.2 2. Columns and Rows counts

```
[6]: print('Columns counts: ',len(data_df.columns))
     print('Row counts: ', len(data_df))
```

```
Columns counts:  71
Row counts:  199
```

1.3 3 & 4 . Miss values count

```
[7]: pd.DataFrame(data_df.isna().any())
```

```
[7]:                              0
     age_gt_60                False
     air                      False
     airBoneGap               False
     ar_c                      True
     ar_u                      True
     …                        …
     viith_nerve_signs        False
     wave_V_delayed           False
     waveform_ItoV_prolonged  False
     indentifier              False
     class                    False

     [71 rows x 1 columns]
```

```
[8]: pd.DataFrame(data_df.isna().sum())
```

```
[8]:                          0
     age_gt_60                0
     air                      0
     airBoneGap               0
     ar_c                     4
     ar_u                     3
     …                     ..
     viith_nerve_signs        0
     wave_V_delayed           0
     waveform_ItoV_prolonged  0
     indentifier              0
     class                    0

     [71 rows x 1 columns]
```

1.4 5. Fill missing values by imputing them

[9]:

```
cols_with_missing = [col for col in data_df.columns if data_df[col].isnull().
 ↪any()]
print('Columns with missing values: ', cols_with_missing)

# for missing values we can use simple imputer to fill missing values...
# from sklearn.impute import SimpleImputer
# my_imputer = SimpleImputer()
# imputed_data_df = pd.DataFrame(my_imputer.fit_transform(data_df))
# imputed_data_df.columns = data_df.columns
# data_df = imputed_data_df

# Actually!!! ther is an easier way to handle missing values...
# we can just simply DROP them! :)))

data_df = data_df.drop(cols_with_missing, axis=1)
data_df
```

```
Columns with missing values:  ['ar_c', 'ar_u', 'bone', 'bser', 'o_ar_c',
'o_ar_u', 'speech']
```

[9]:

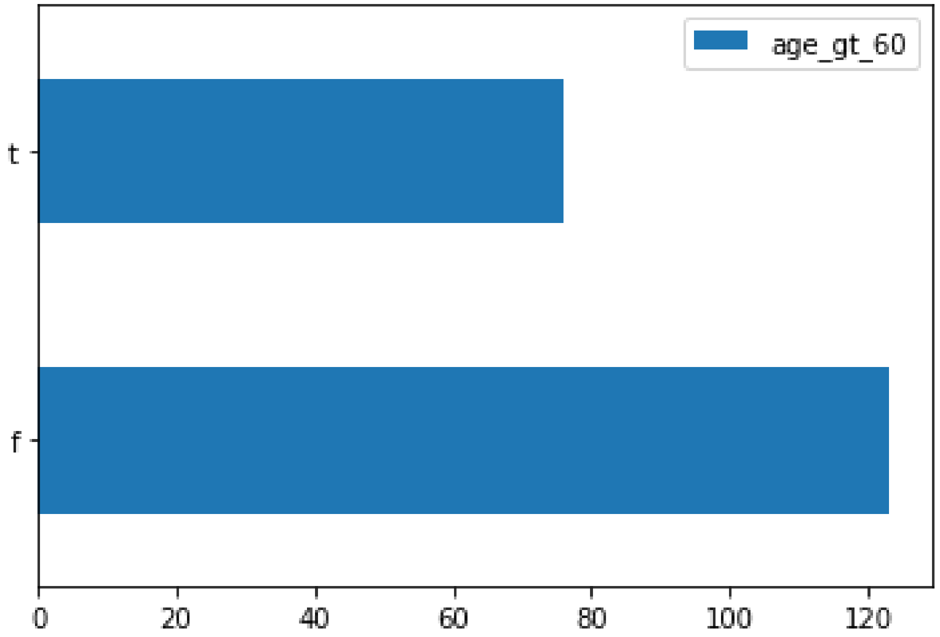
```
    age_gt_60       air airBoneGap boneAbnormal history_buzzing  \
0           f  moderate          f            t               f
1           t      mild          t            t               f
2           t      mild          t            f               f
3           t      mild          f            t               f
4           t      mild          f            t               f
..        …       …          …            …               …
194         t      mild          f            t               f
195         t      mild          f            f               f
196         f    normal          f            f               f
197         t      mild          f            f               f
198         t    normal          f            f               f

    history_dizziness history_fluctuating history_fullness history_heredity  \
0                   f                   f                f                f
1                   f                   f                f                f
2                   f                   f                f                f
3                   f                   f                f                f
4                   f                   f                f                f
..                …                 …              …              …
194                 f                   f                f                f
195                 f                   f                f                f
196                 f                   f                f                f
197                 f                   f                f                f
198                 f                   f                f                f
```

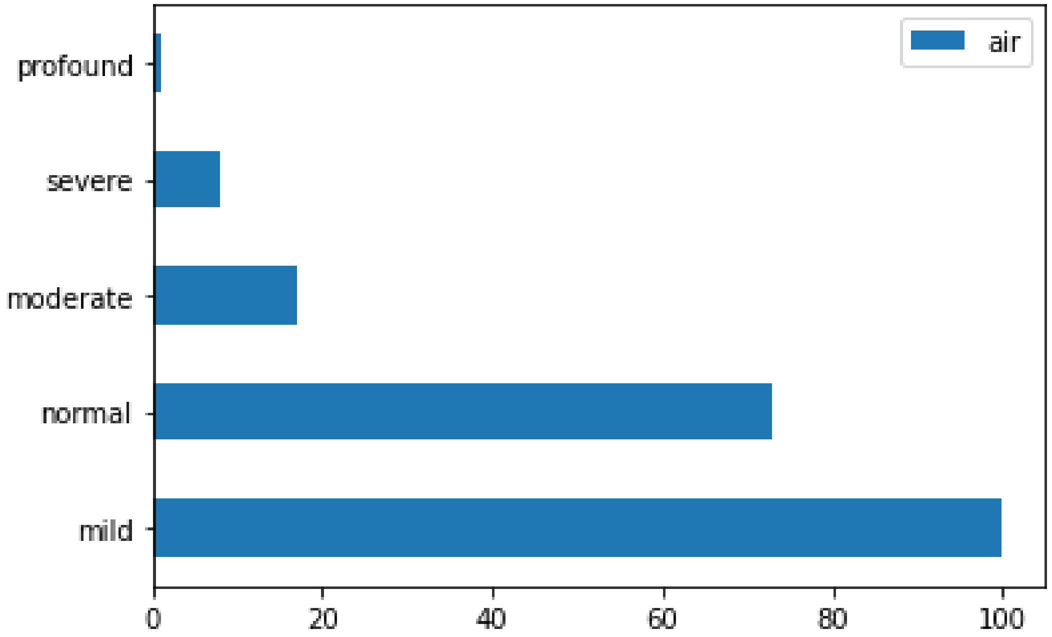
```
    history_nausea  … s_sn_gt_1k s_sn_gt_2k s_sn_gt_4k static_normal tymp  \
0                f  …          f          f          f             t    a
1                f  …          f          f          f             t   as
2                f  …          f          f          f             t    b
3                f  …          f          f          f             t    a
4                f  …          f          f          f             t    a
..             …  …        …        …        …           …  …
194              f  …          f          f          f             t    a
195              f  …          f          f          f             t    c
196              t  …          f          f          f             f    a
197              f  …          f          f          f             t    a
198              f  …          f          f          f             t    a

    viiith_nerve_signs wave_V_delayed waveform_ItoV_prolonged indentifier  \
0                    f              f                       f          p2
1                    f              f                       f          p3
2                    f              f                       f          p4
3                    f              f                       f          p5
4                    f              f                       f          p6
..                …            …                     …         …
194                  f              f                       f        p196
195                  f              f                       f        p197
196                  f              f                       f        p198
197                  f              f                       f        p199
198                  f              f                       f        p200

                               class
0                   cochlear_unknown
1        mixed_cochlear_age_fixation
2    mixed_cochlear_age_otitis_media
3                       cochlear_age
4                       cochlear_age
..                                …
194                     cochlear_age
195  mixed_cochlear_age_otitis_media
196      possible_brainstem_disorder
197                     cochlear_age
198                     cochlear_age

[199 rows x 64 columns]
```

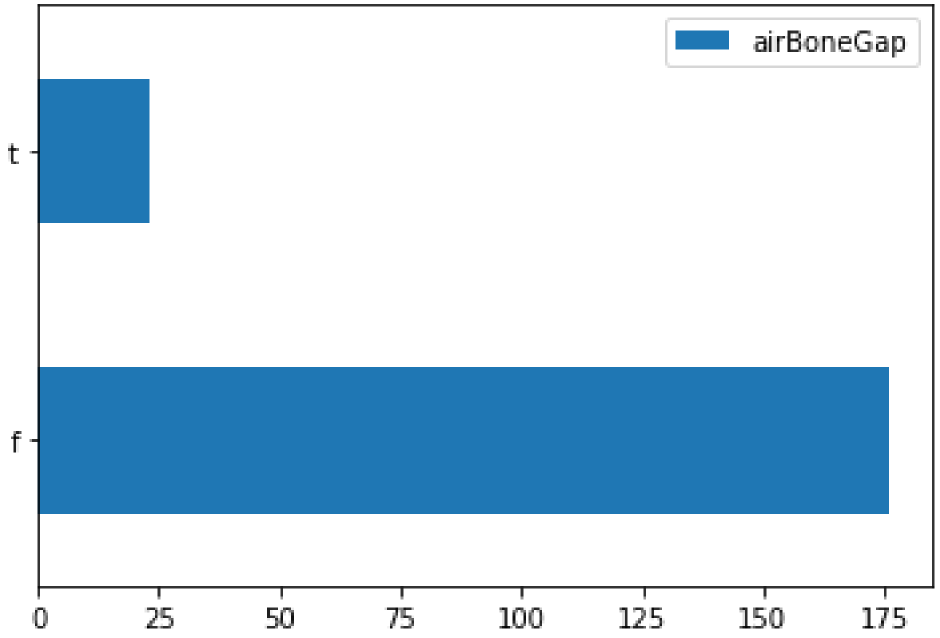
1.5 6. Graph of each values of each column

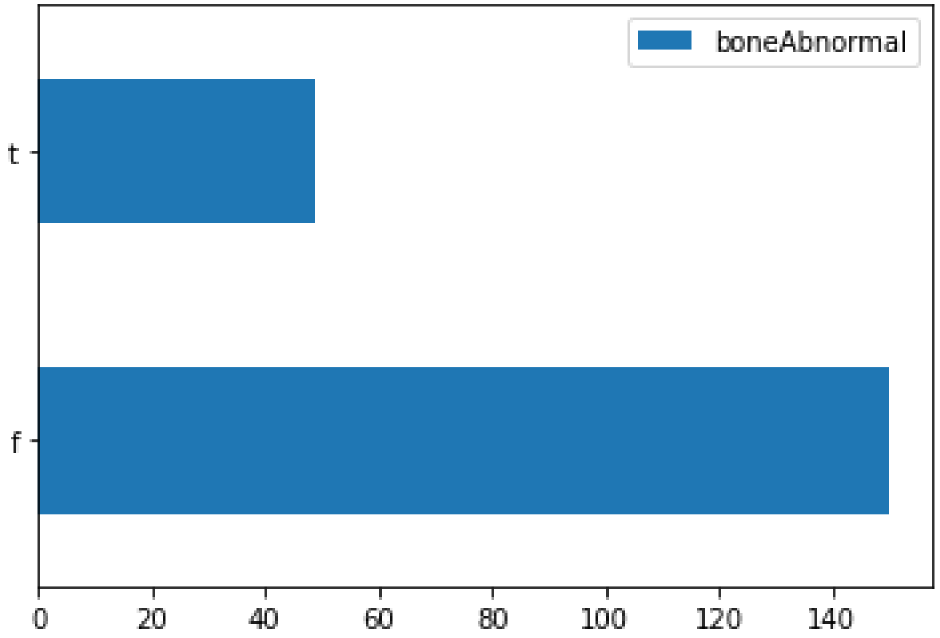
[12]:

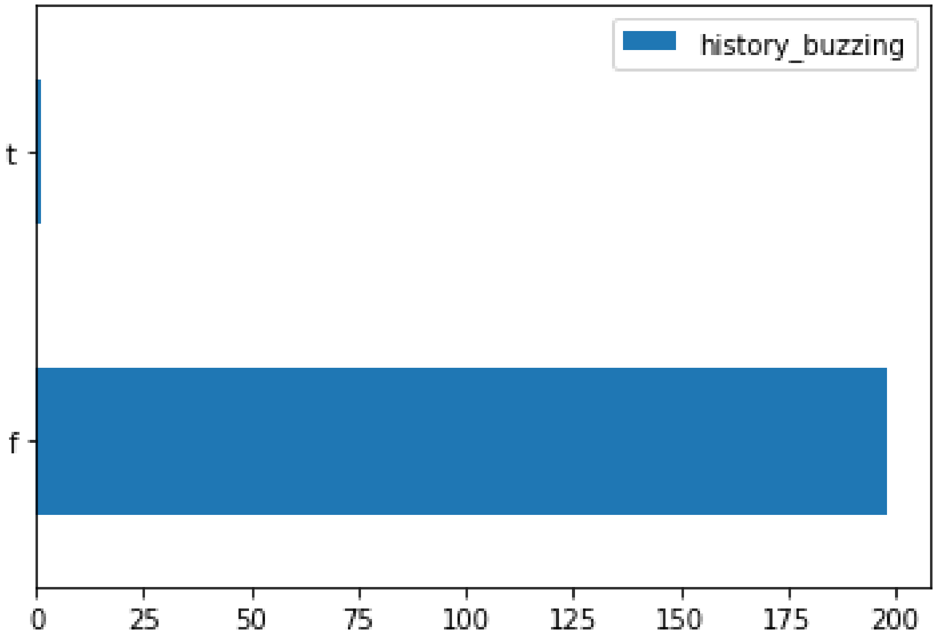
```
for col in data_df.columns:
        each_value_count = pd.DataFrame(data_df[col].value_counts())
        each_value_count.plot.barh()
```

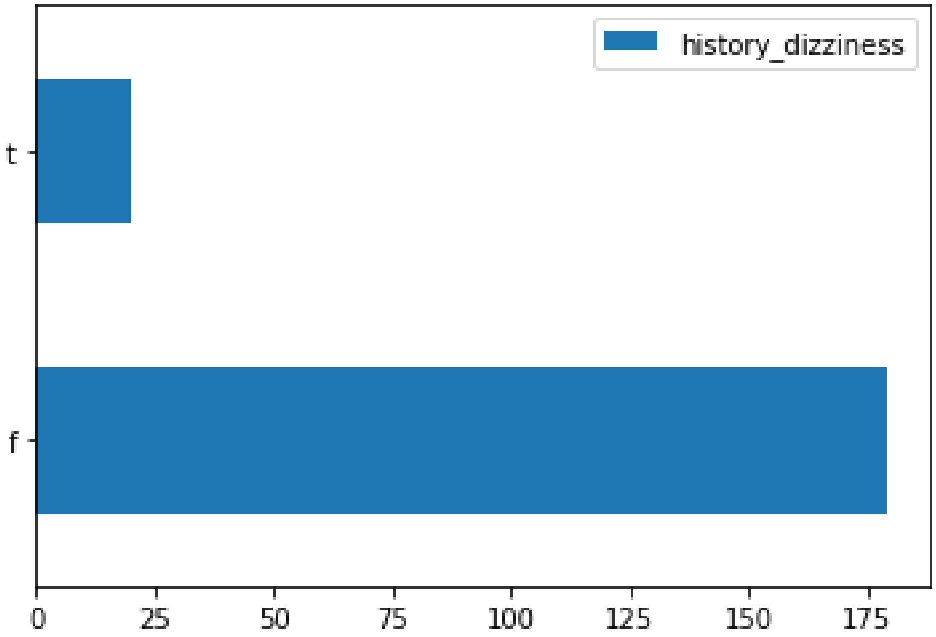


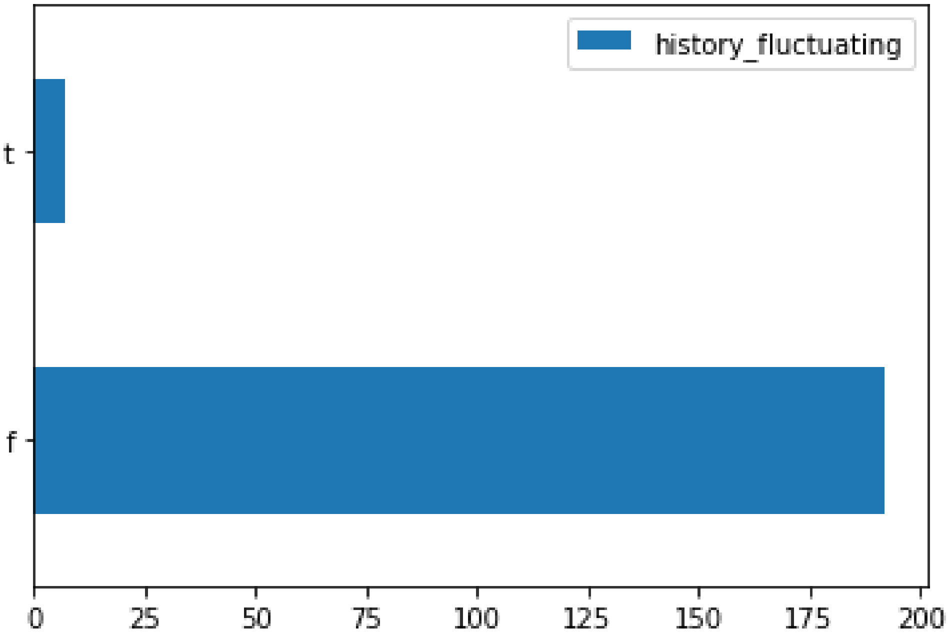


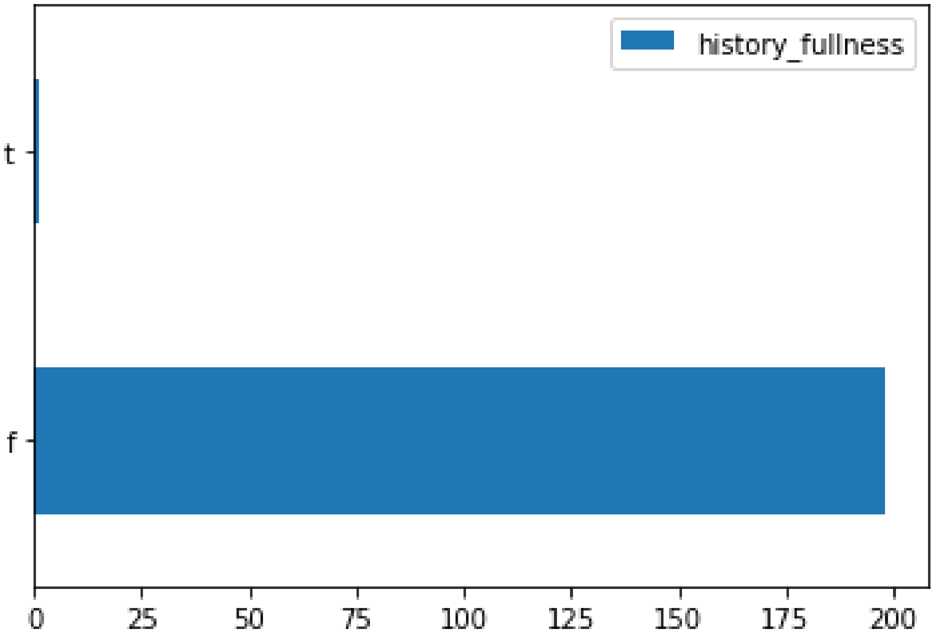

airBoneGap
t
f
0
25
50
75
100
125
150
175

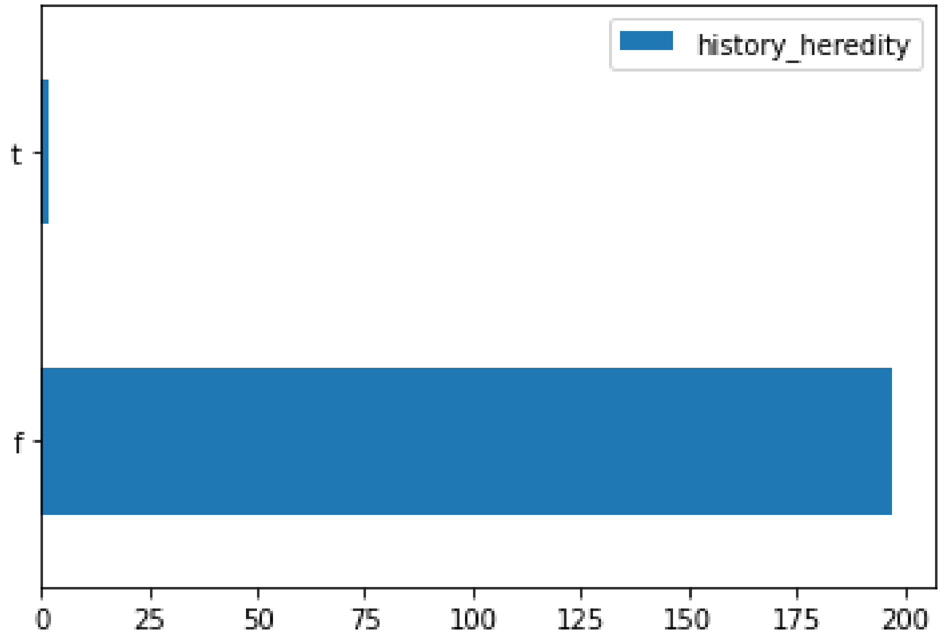

boneAbnormal
t
f
0
20
40
60
80
100
120
140

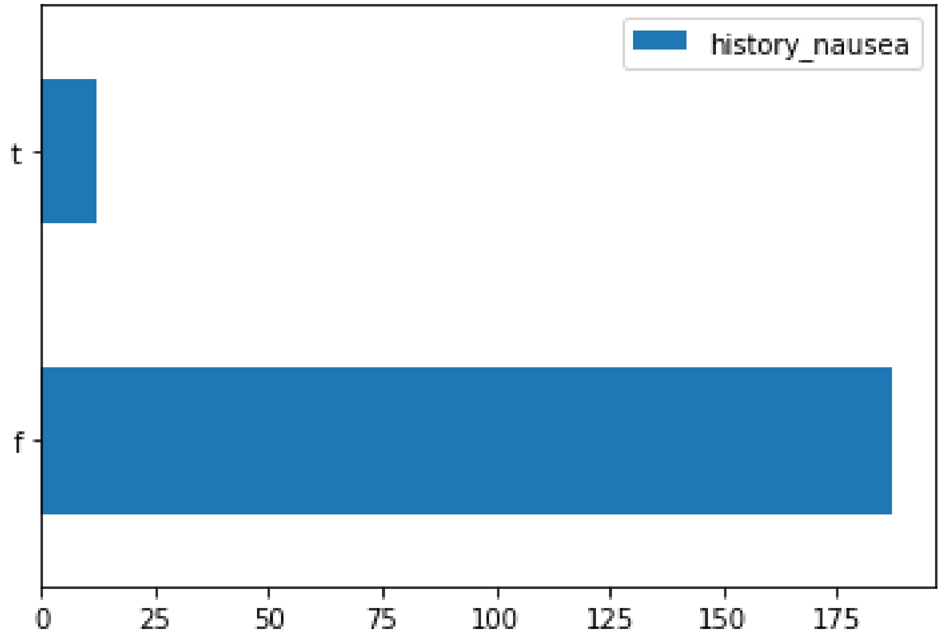

history_buzzing
t
f
0
25
50
75
100
125
150
175
200

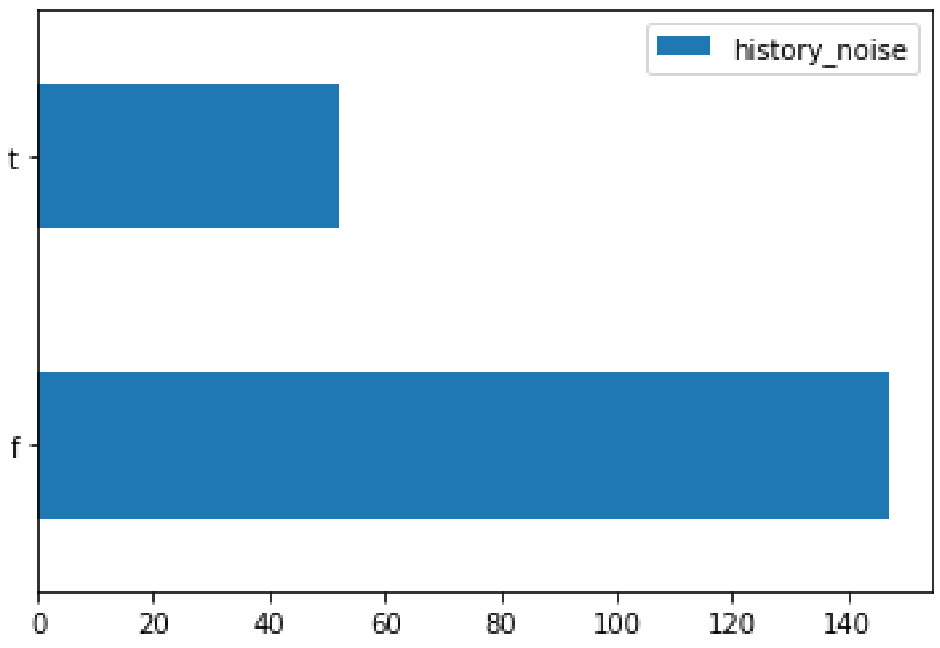

history_dizziness
t
f
0
25
50
75
100
125
150
175

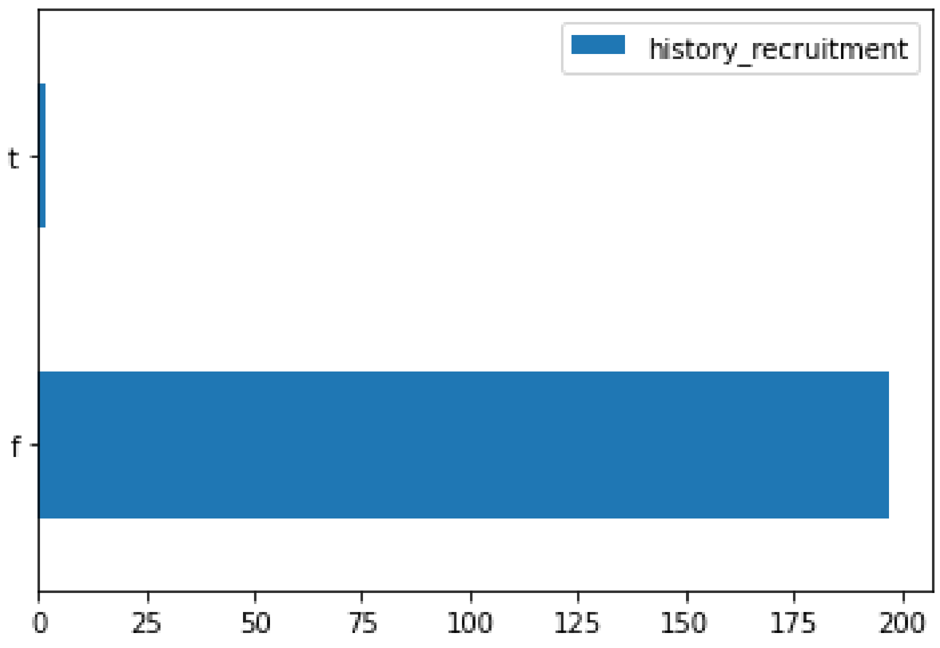

history_fluctuating
t
f
0
25
50
75
100
125
150
175
200

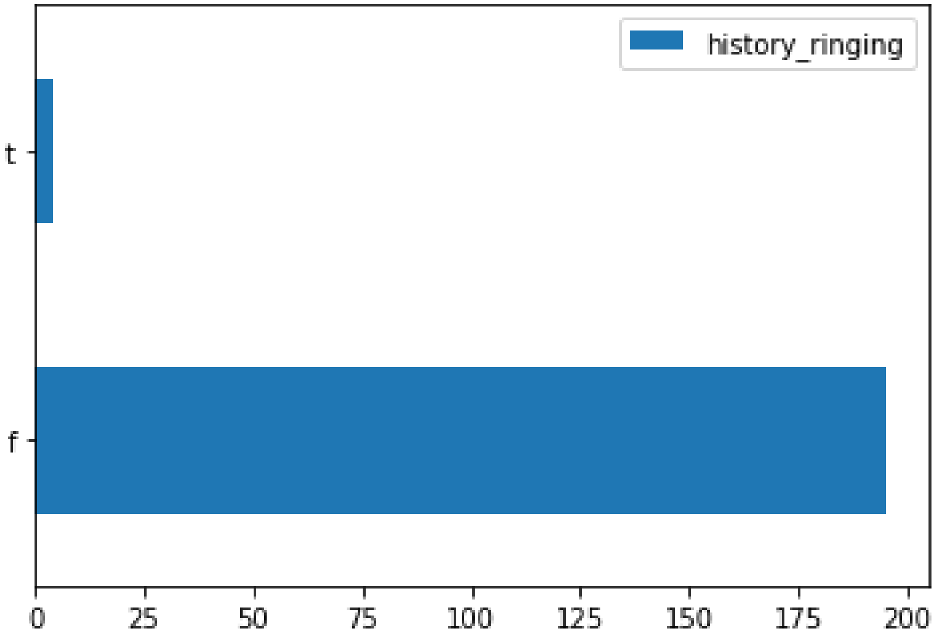

history_fullness
t
f
0
25
50
75
100
125
150
175
200

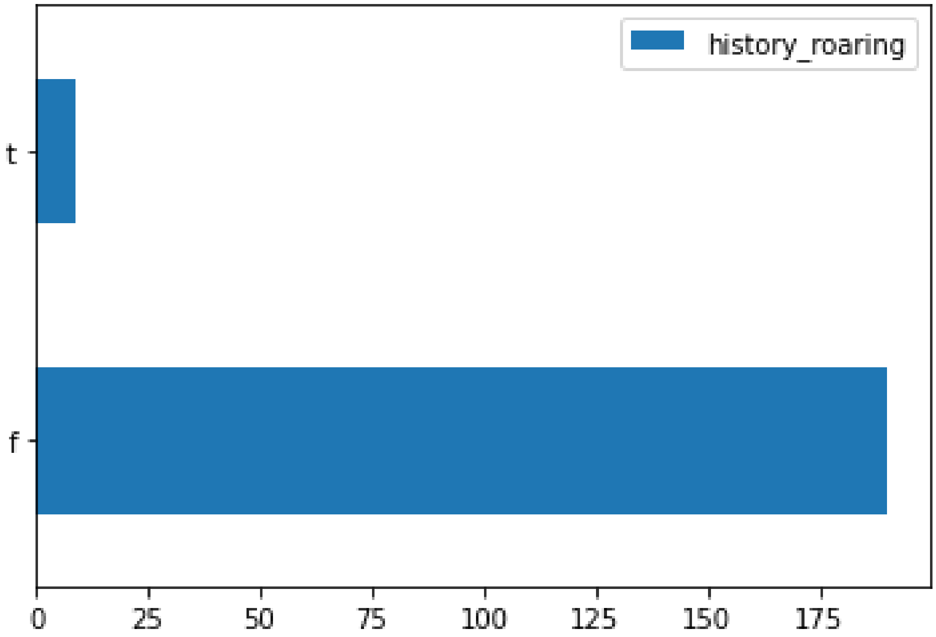

history_heredity
t
f
0
25
50
75
100
125
150
175
200

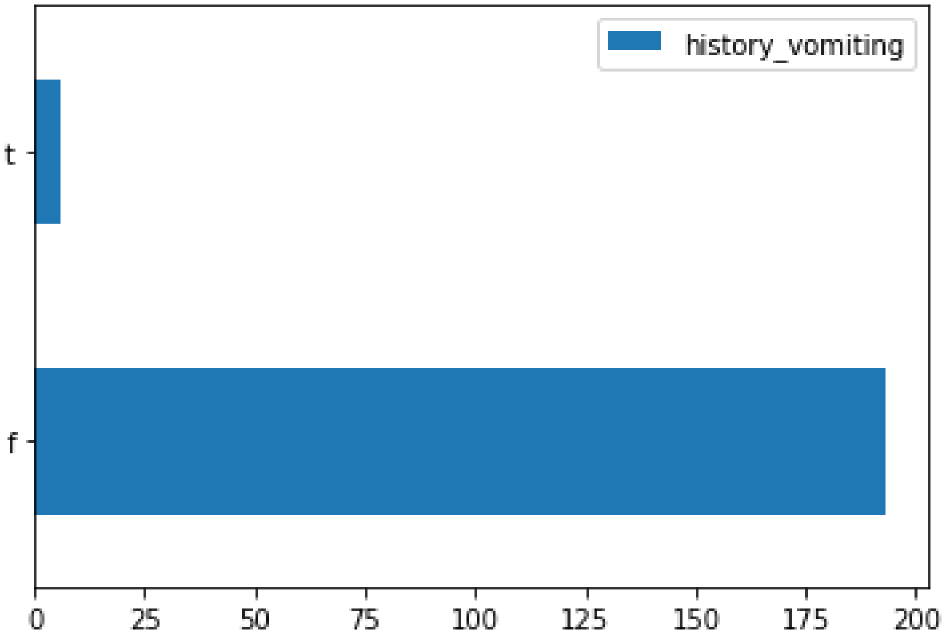

history_nausea
t
f
0
25
50
75
100
125
150
175

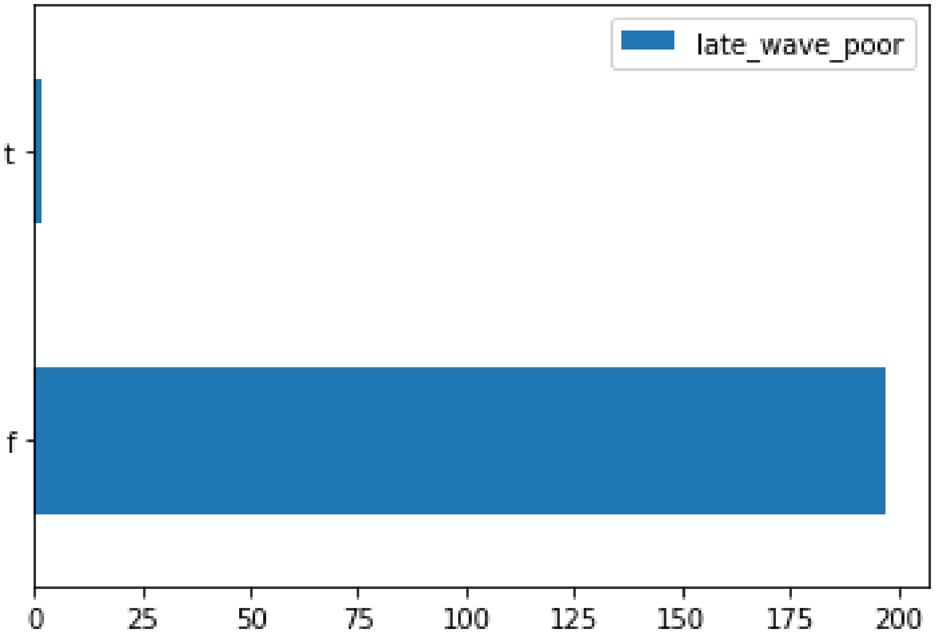

history_noise
t
f
0
20
40
60
80
100
120
140

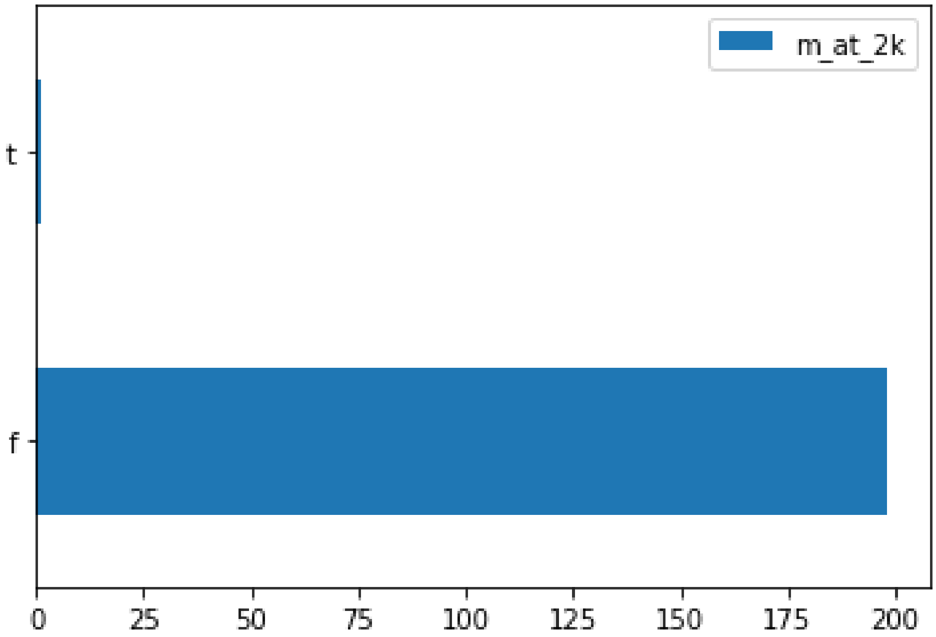

history_recruitment
t
f
0
25
50
75
100
125
150
175
200

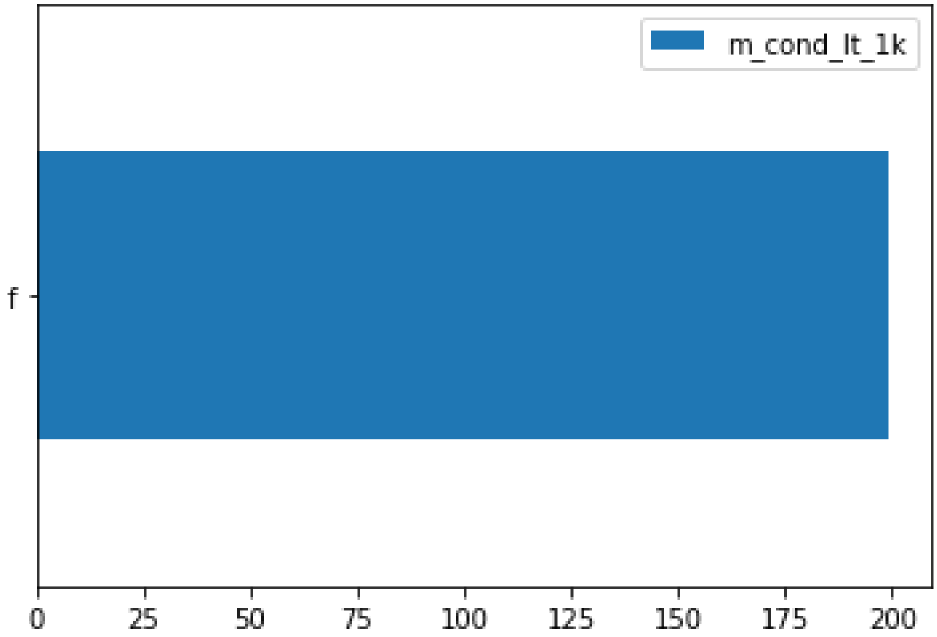

history_ringing
t
f
0
25
50
75
100
125
150
175
200

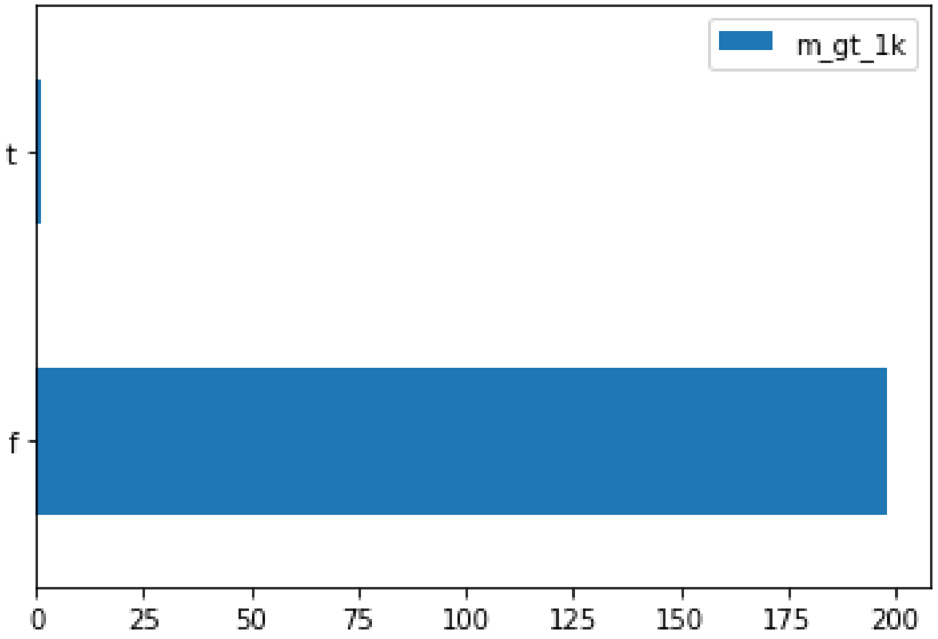

history_roaring
t
f
0
25
50
75
100
125
150
175

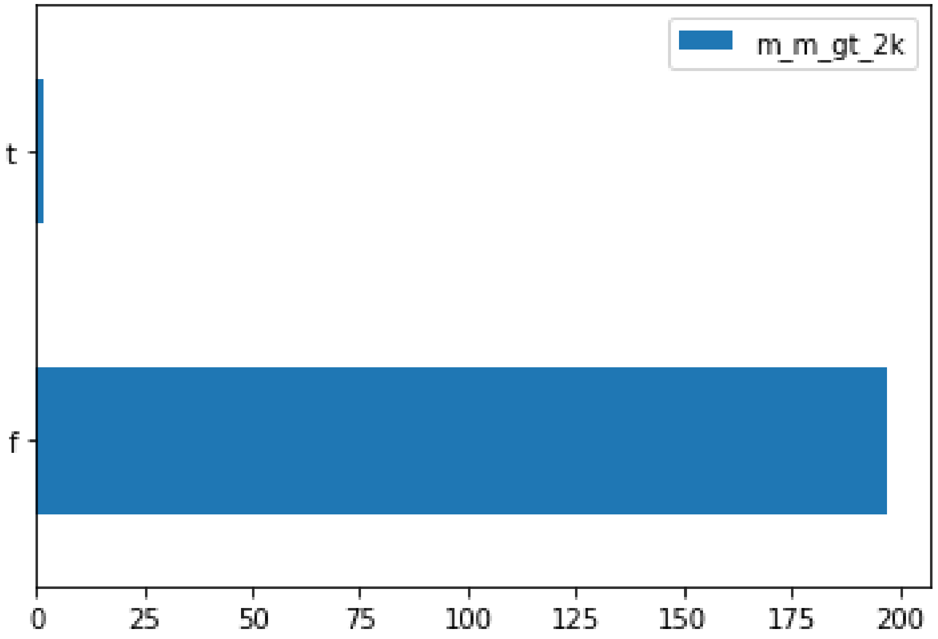

history_vomiting
t
f
0
25
50
75
100
125
150
175
200

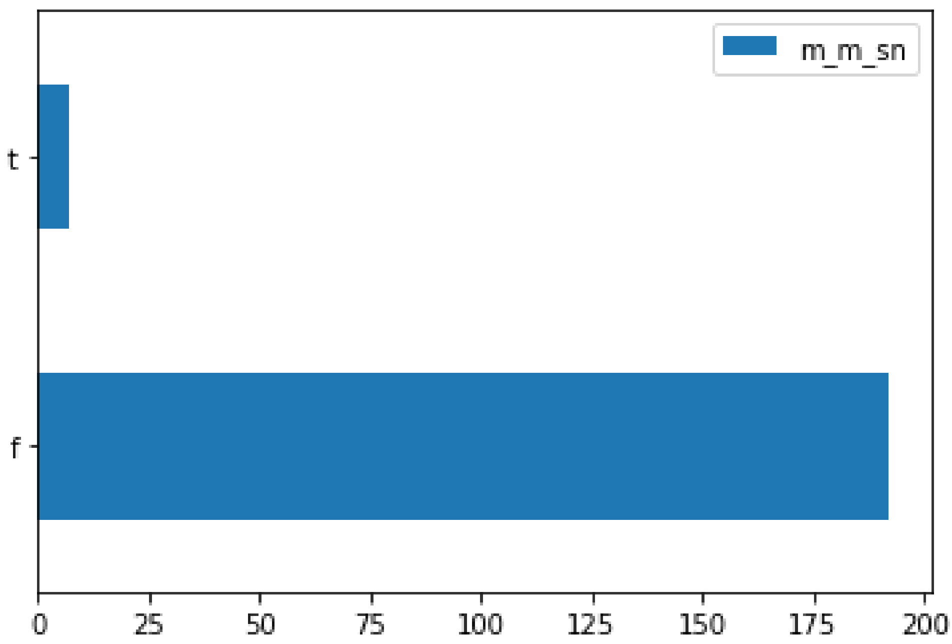

late_wave_poor
t
f
0
25
50
75
100
125
150
175
200

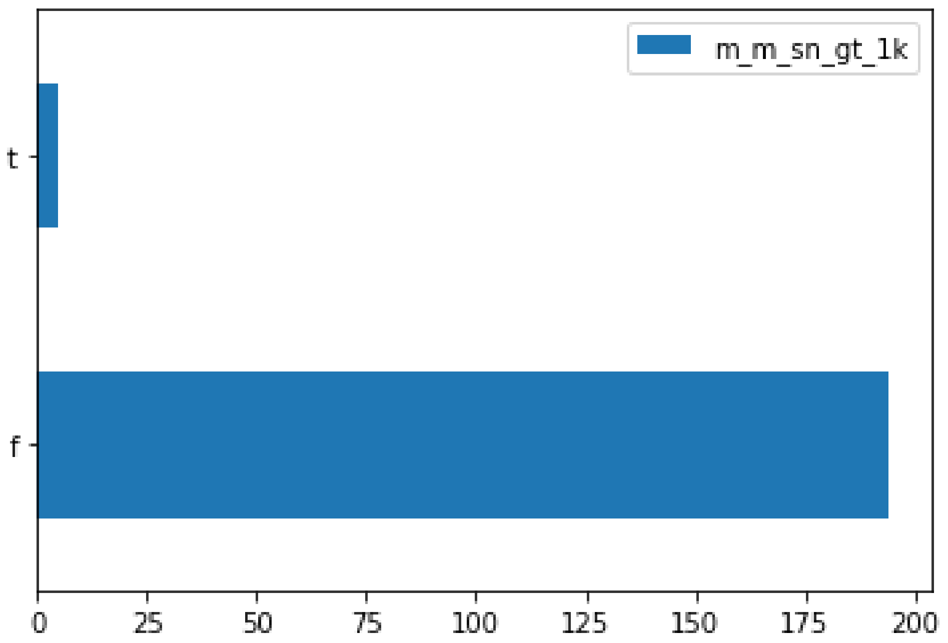

m_at_2k
t
f
0
25
50
75
100
125
150
175
200

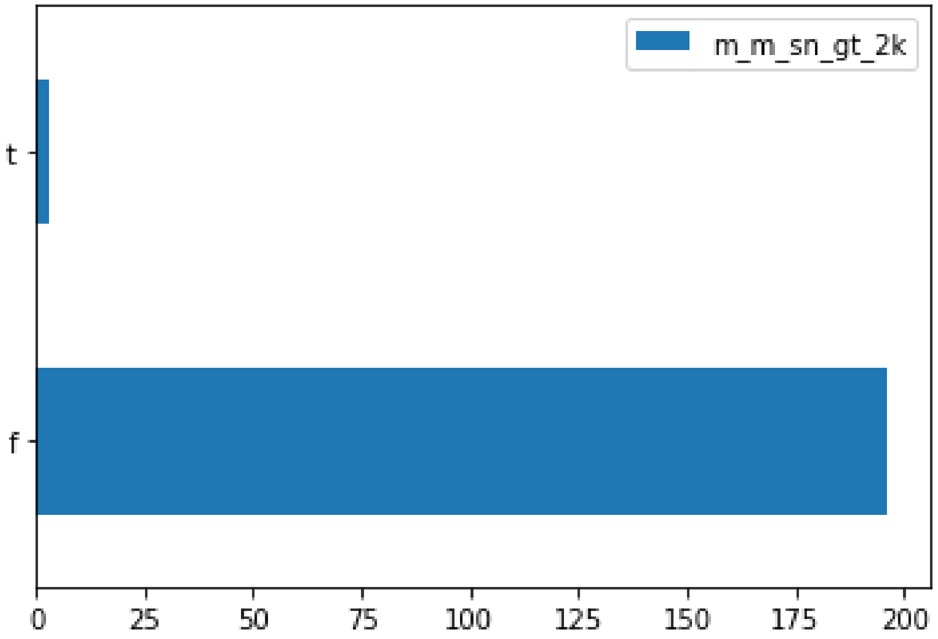

m_cond_lt_1k
f
0
25
50
75
100
125
150
175
200

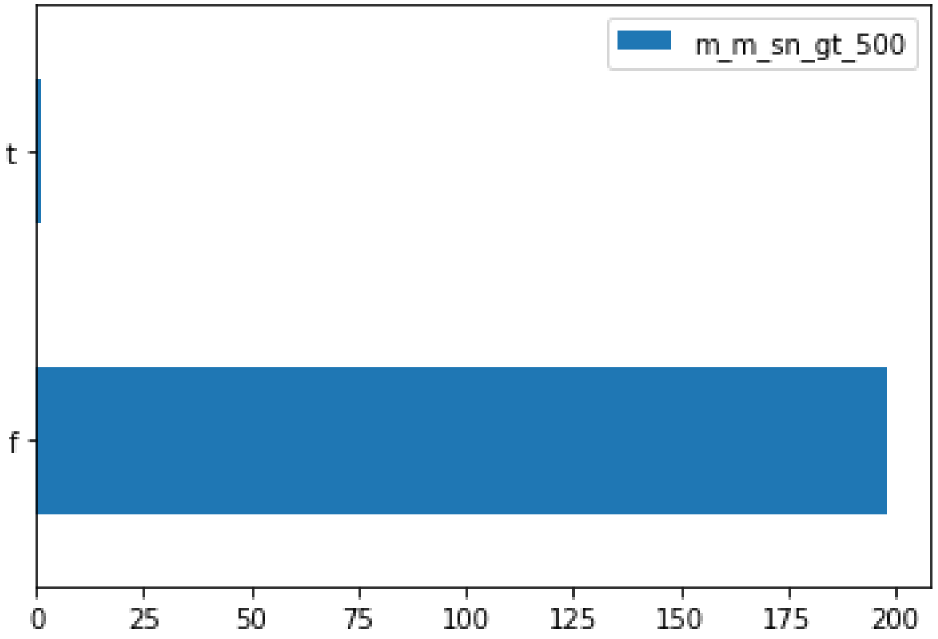

m_gt_1k
t
f
0
25
50
75
100
125
150
175
200

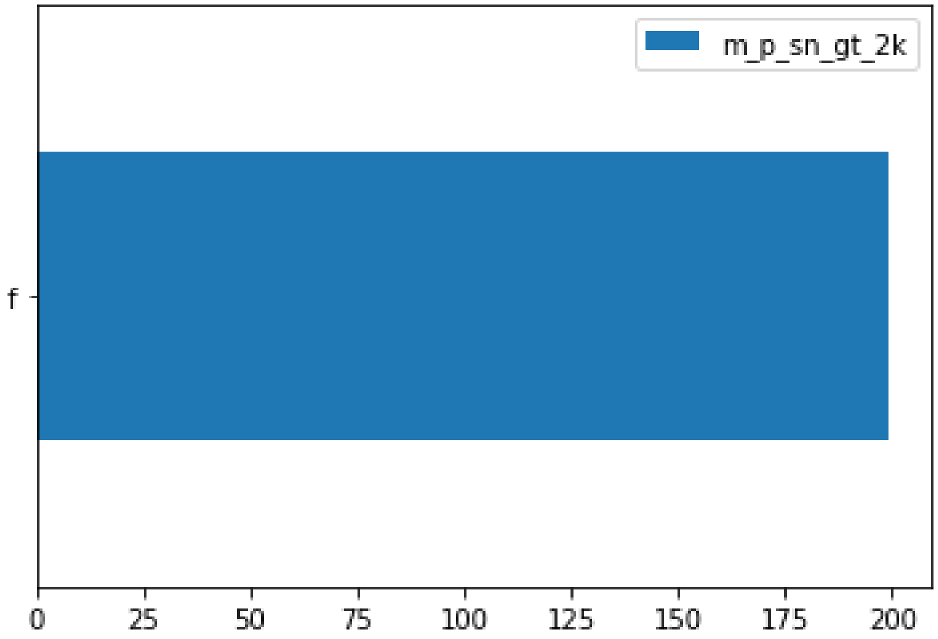

m_m_gt_2k
t
f
0
25
50
75
100
125
150
175
200

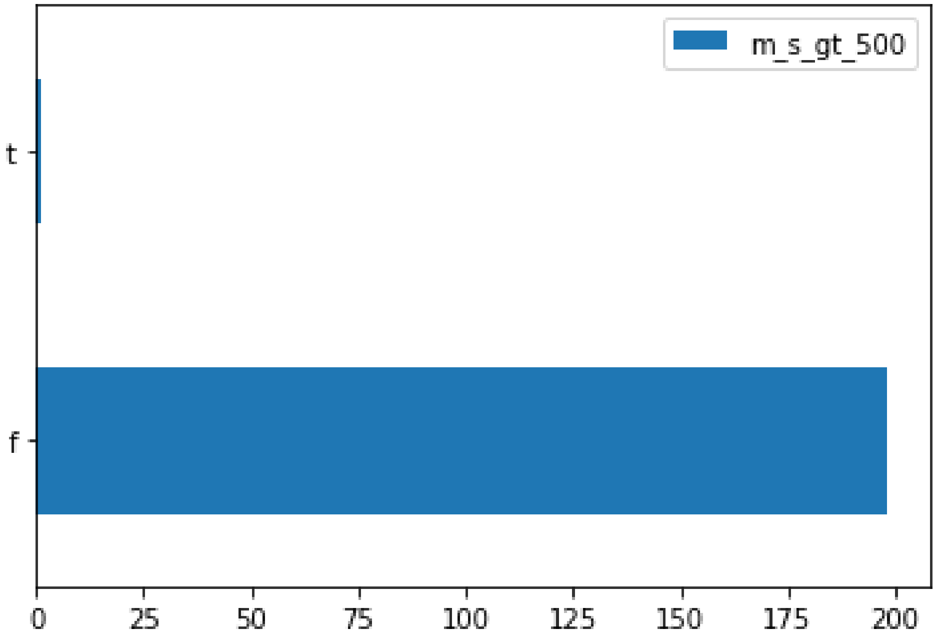

m_m_sn
t
f
0
25
50
75
100
125
150
175
200

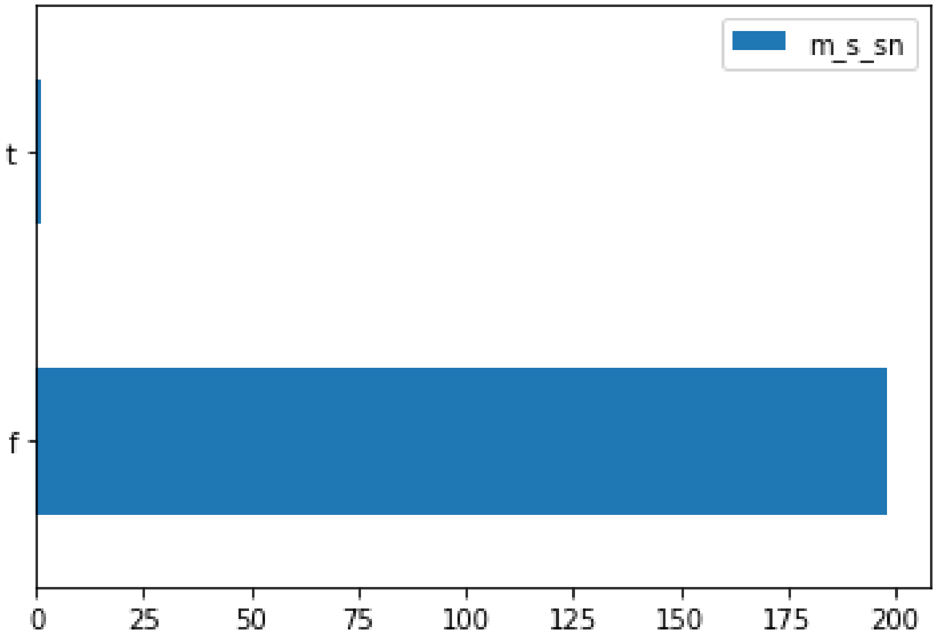

m_m_sn_gt_1k
t
f
0
25
50
75
100
125
150
175
200

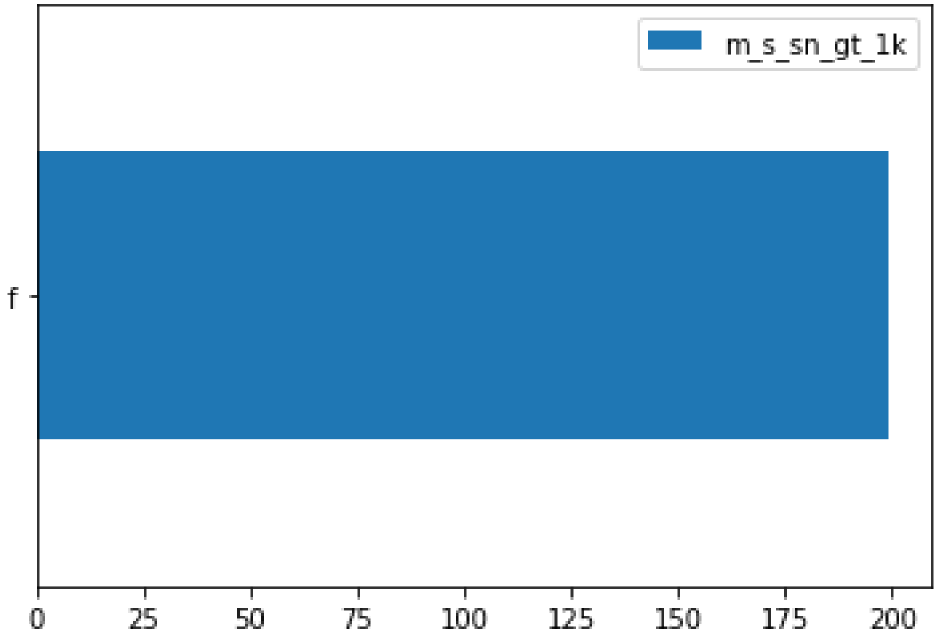

m_m_sn_gt_2k
t
f
0
25
50
75
100
125
150
175
200

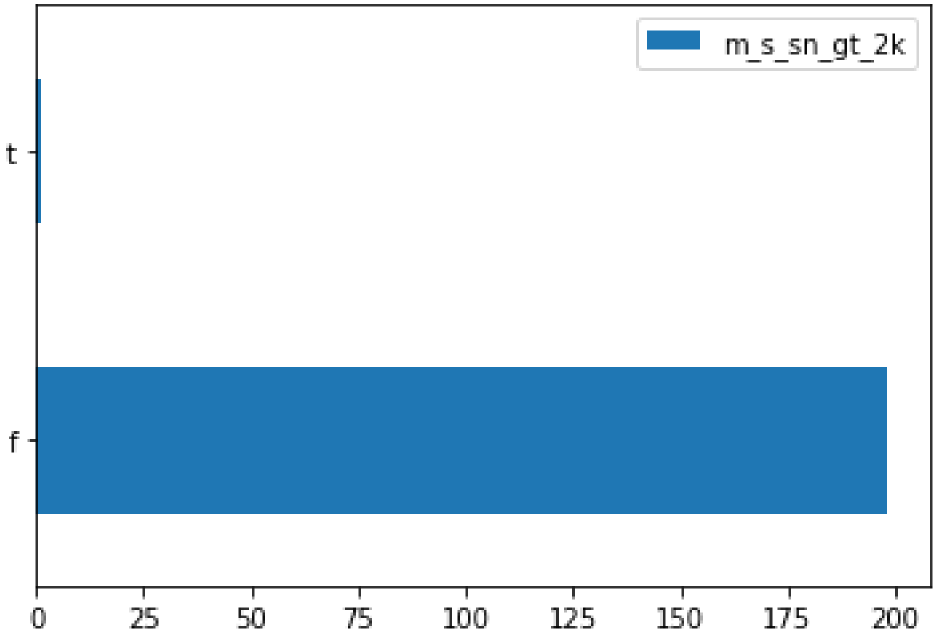

m_m_sn_gt_500
t
f
0
25
50
75
100
125
150
175
200

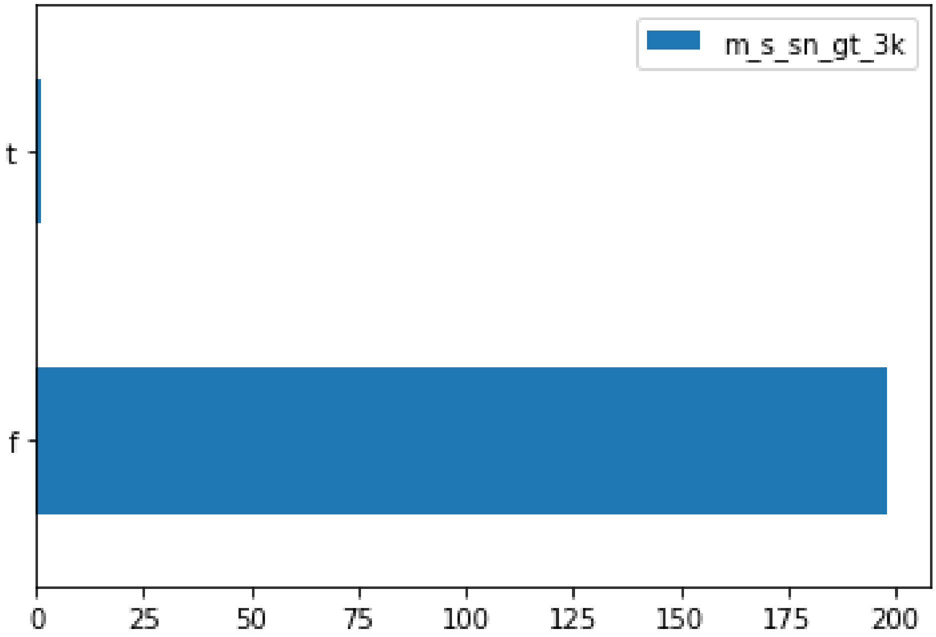

m_p_sn_gt_2k
f
0
25
50
75
100
125
150
175
200

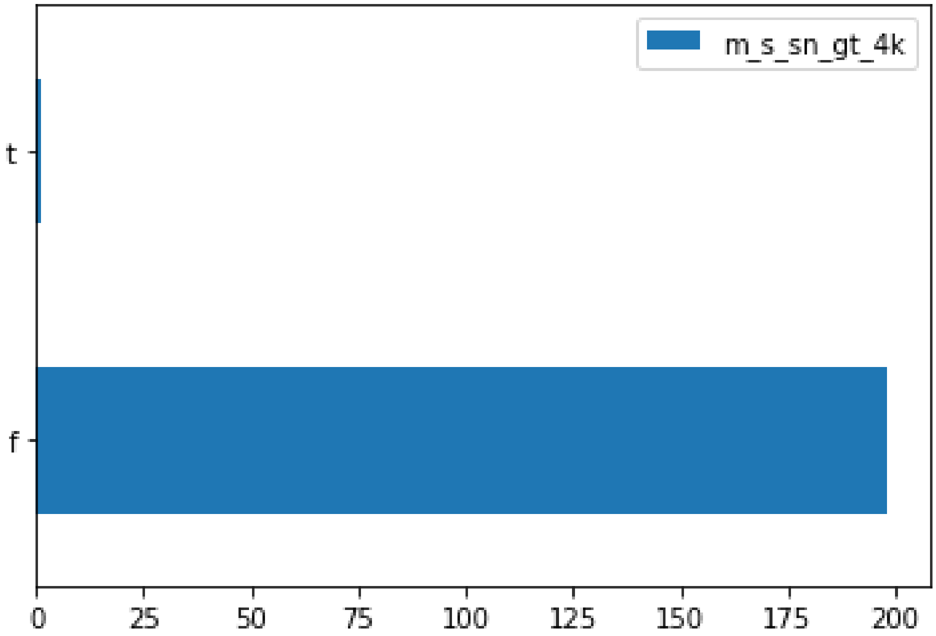

m_s_gt_500
t
f
0
25
50
75
100
125
150
175
200

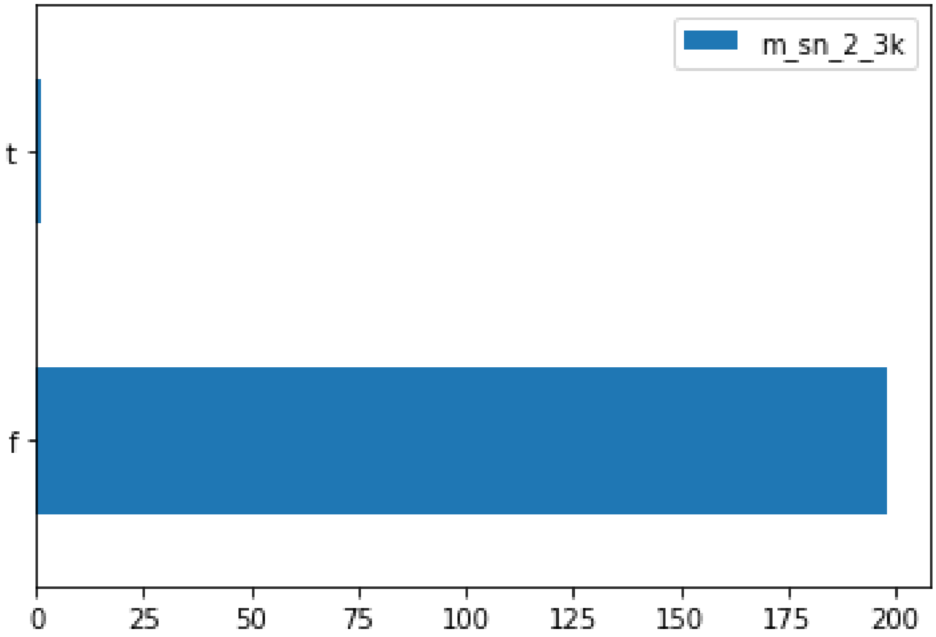

m_s_sn
t
f
0
25
50
75
100
125
150
175
200

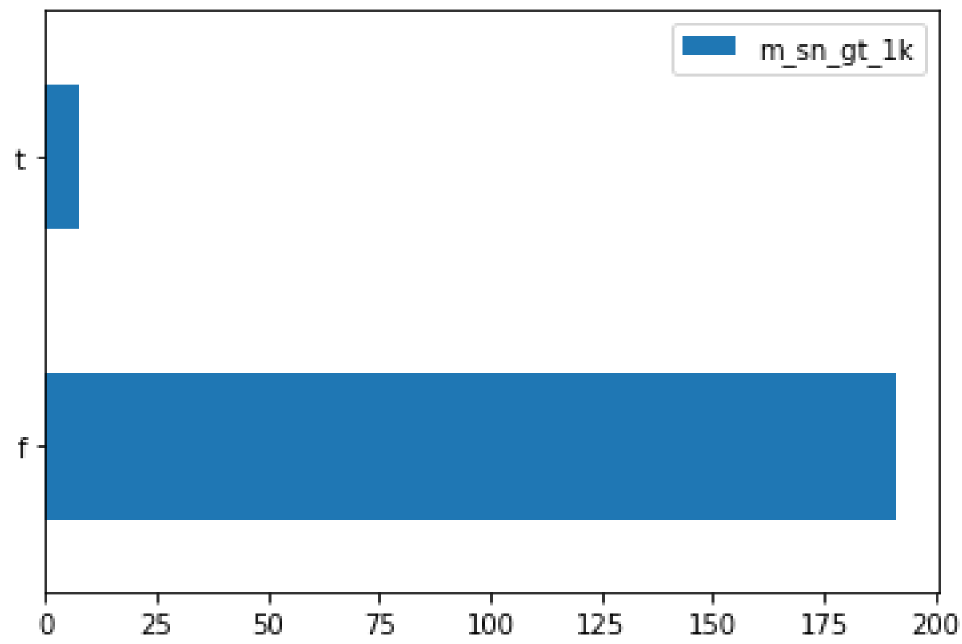

m_s_sn_gt_1k
f
0
25
50
75
100
125
150
175
200

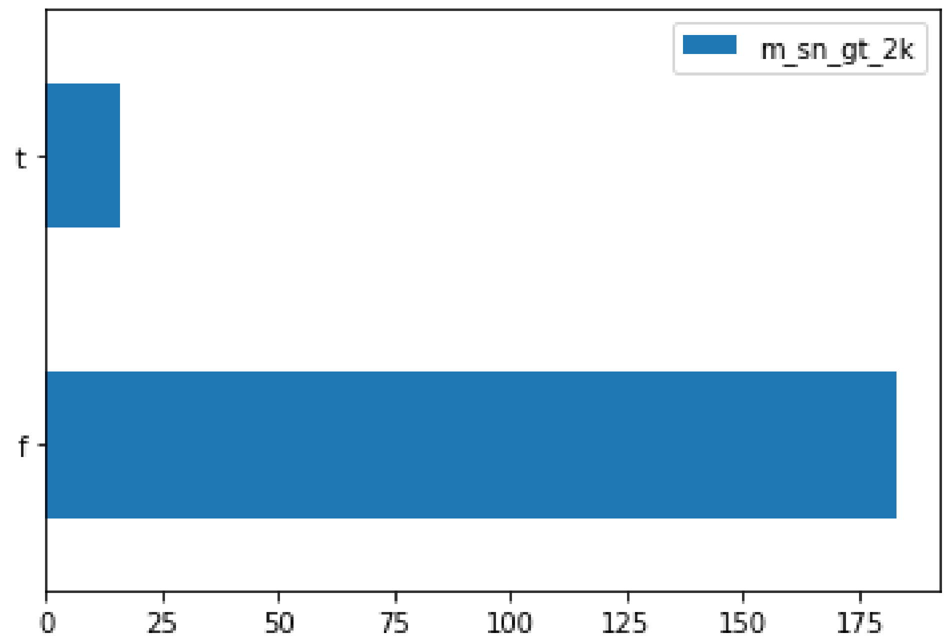

m_s_sn_gt_2k
t
f
0
25
50
75
100
125
150
175
200

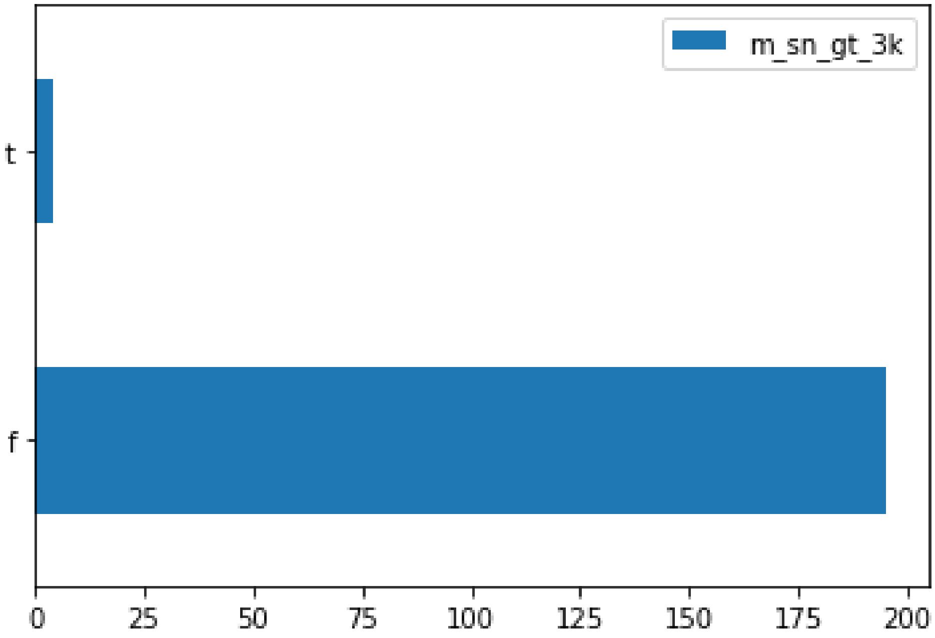

m_s_sn_gt_3k
t
f
0
25
50
75
100
125
150
175
200

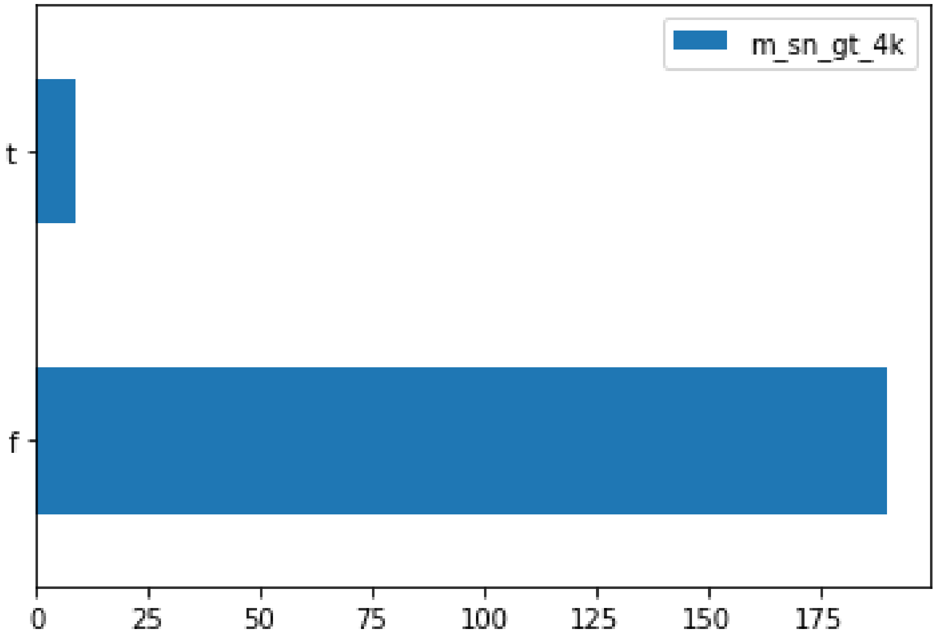

m_s_sn_gt_4k
t
f
0
25
50
75
100
125
150
175
200

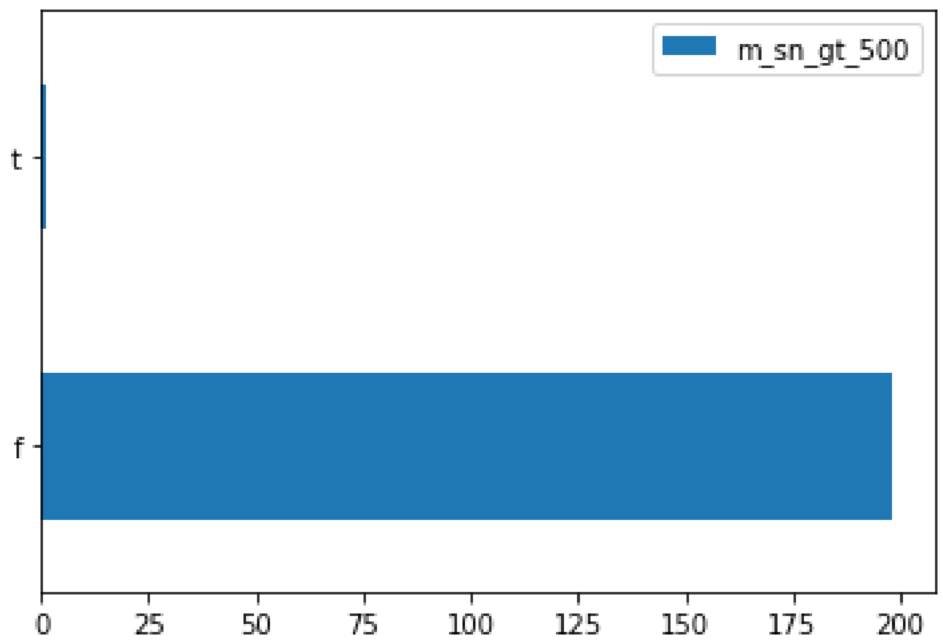

m_sn_2_3k
t
f
0
25
50
75
100
125
150
175
200

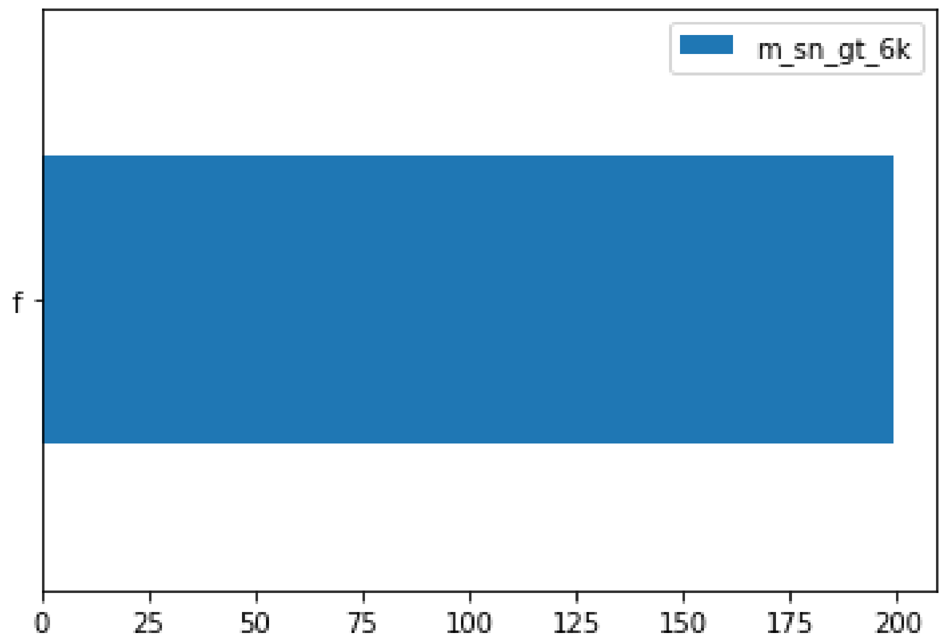

m_sn_gt_1k
t
f
0
25
50
75
100
125
150
175
200

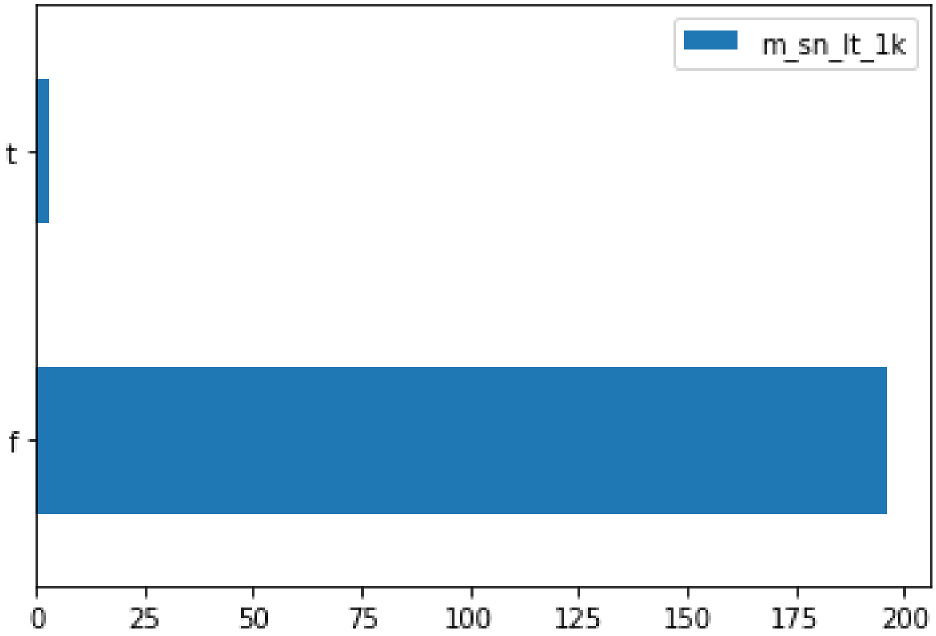

m_sn_gt_2k
t
f
0
25
50
75
100
125
150
175

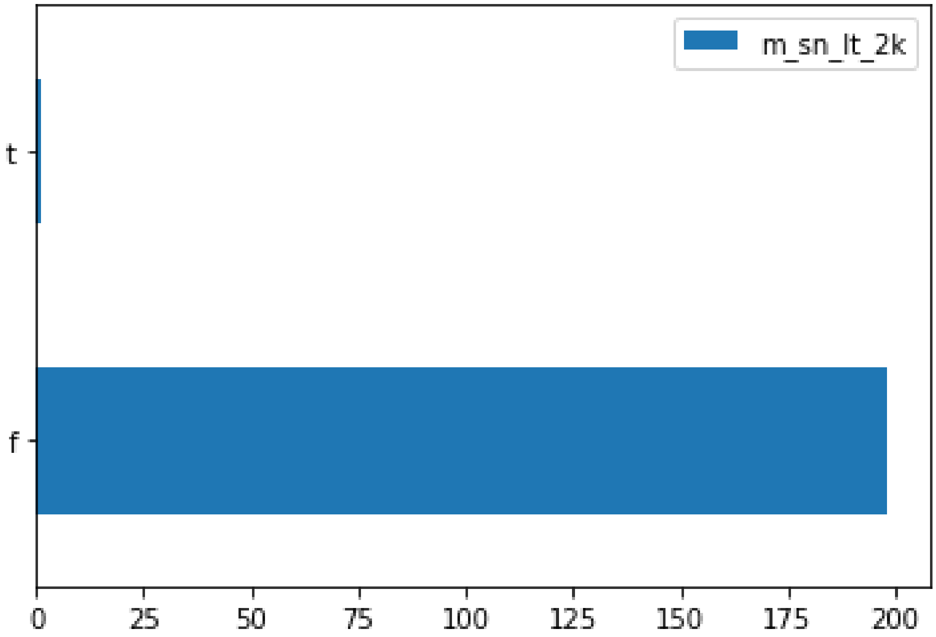

m_sn_gt_3k
t
f
0
25
50
75
100
125
150
175
200

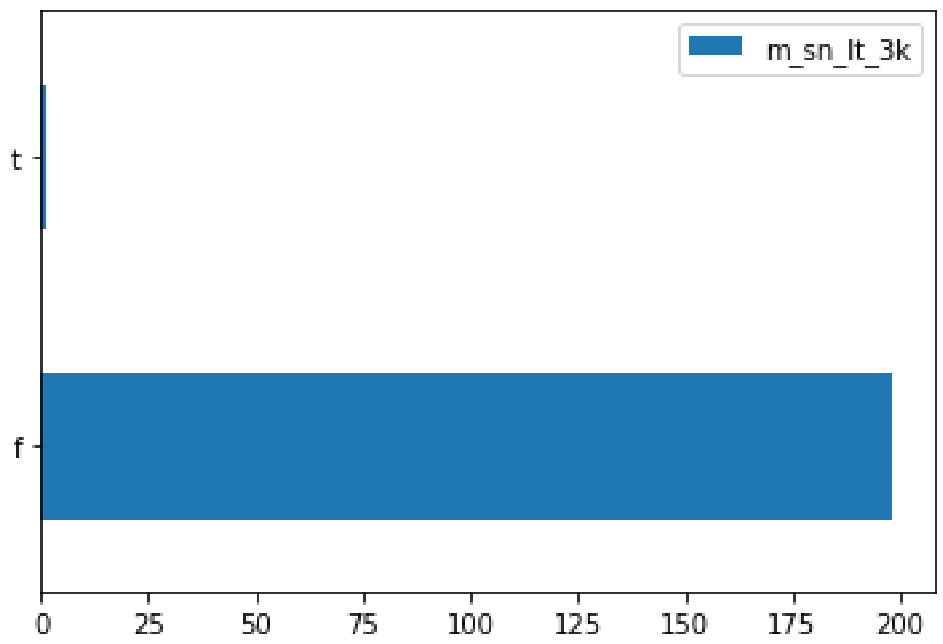

m_sn_gt_4k
t
f
0
25
50
75
100
125
150
175

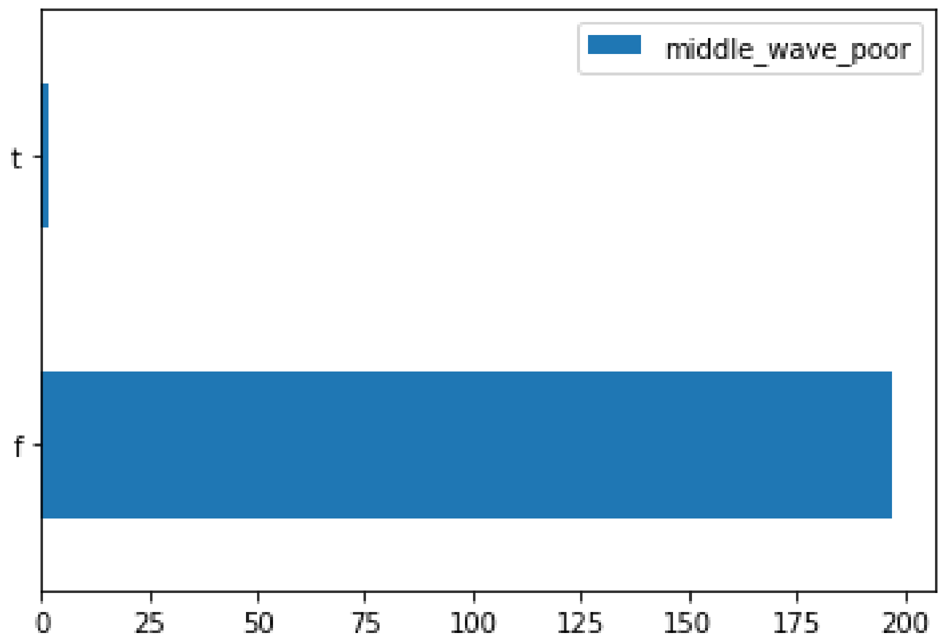

m_sn_gt_500
t
f
0
25
50
75
100
125
150
175
200

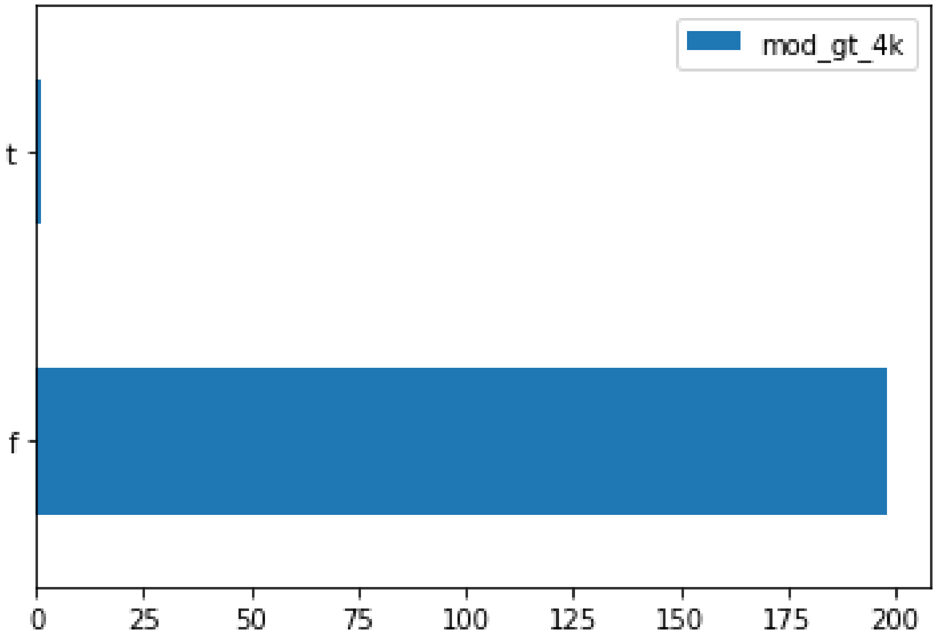

m_sn_gt_6k
f
0
25
50
75
100
125
150
175
200

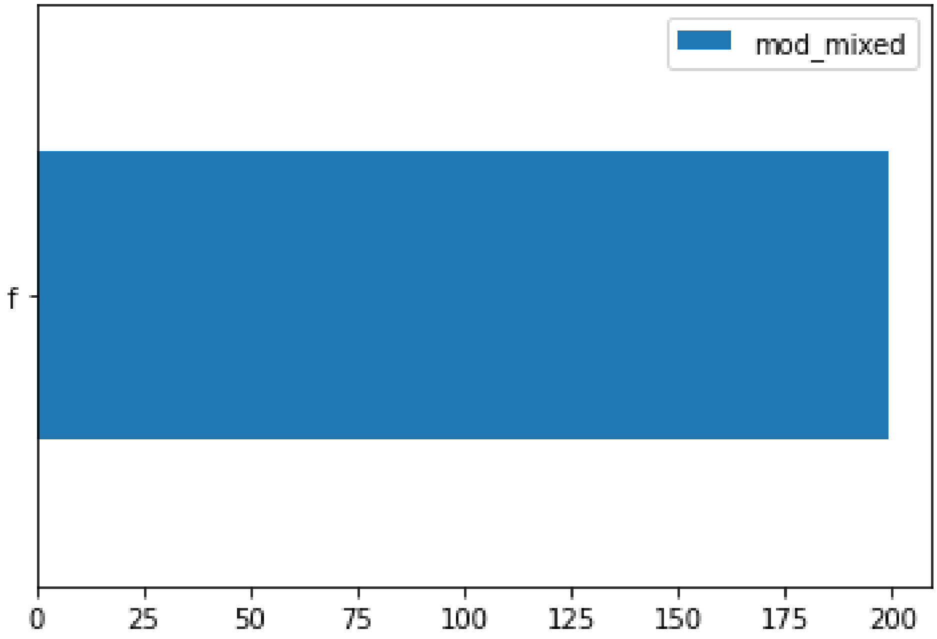

m_sn_lt_1k
t
f
0
25
50
75
100
125
150
175
200

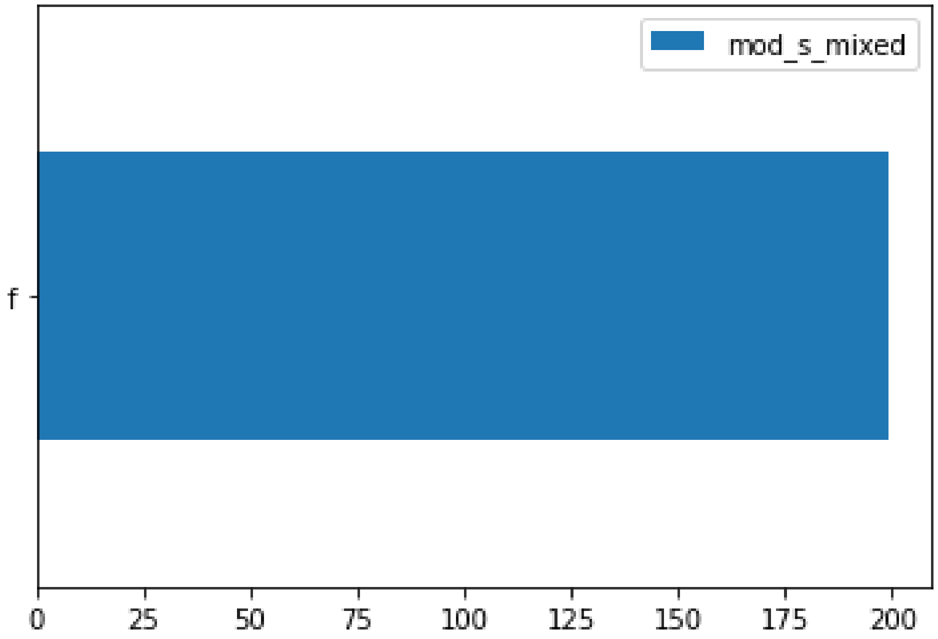

m_sn_lt_2k
t
f
0
25
50
75
100
125
150
175
200

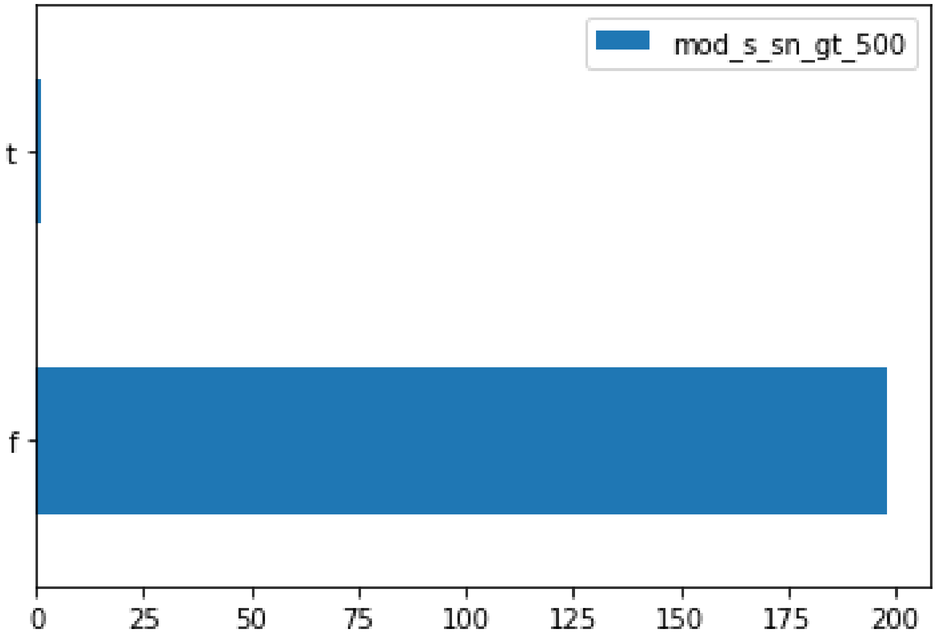

m_sn_lt_3k
t
f
0
25
50
75
100
125
150
175
200

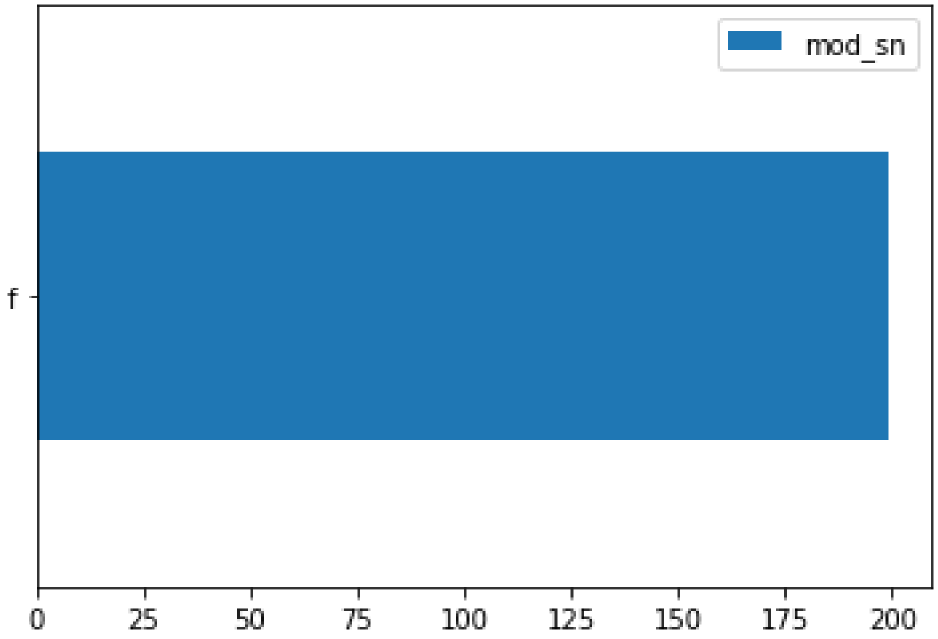

middle_wave_poor
t
f
0
25
50
75
100
125
150
175
200

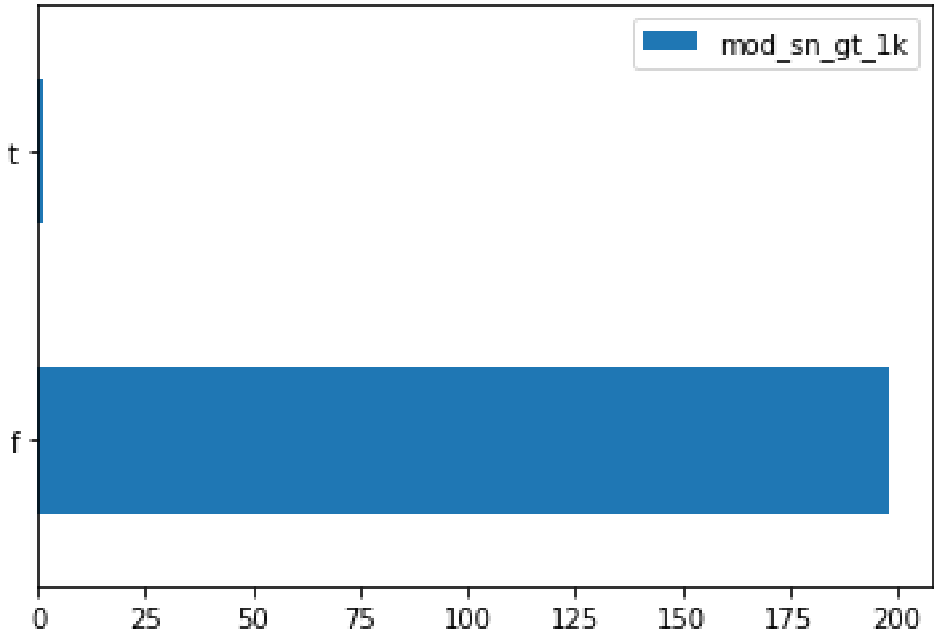

mod_gt_4k
t
f
0
25
50
75
100
125
150
175
200

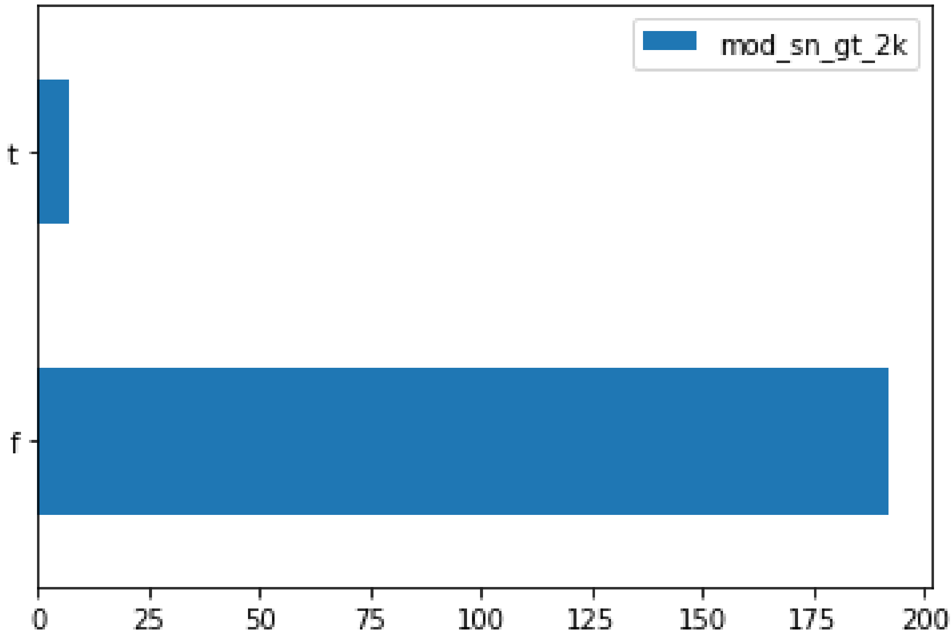

mod_mixed
f
0
25
50
75
100
125
150
175
200

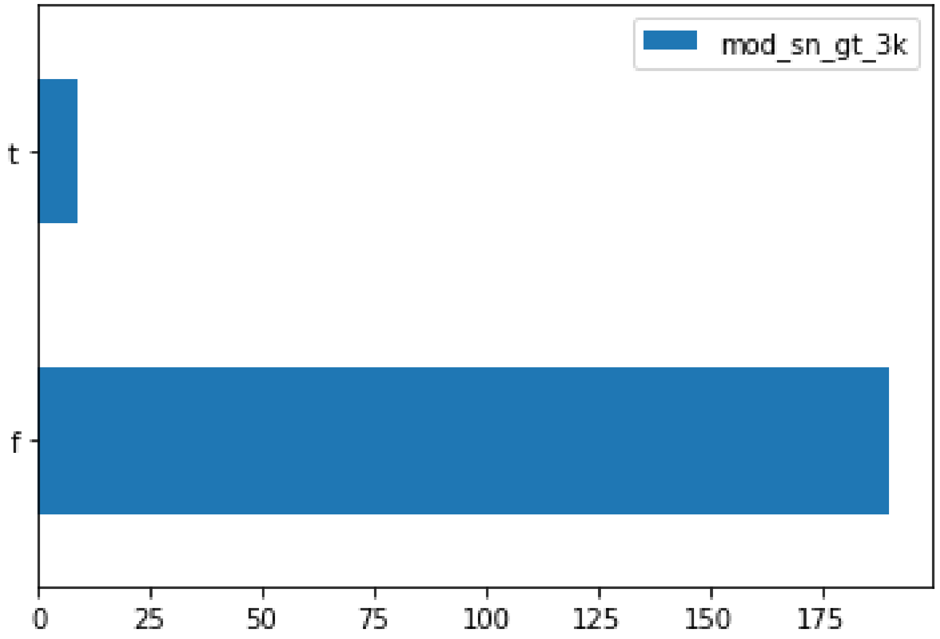

mod_s_mixed
f
0
25
50
75
100
125
150
175
200

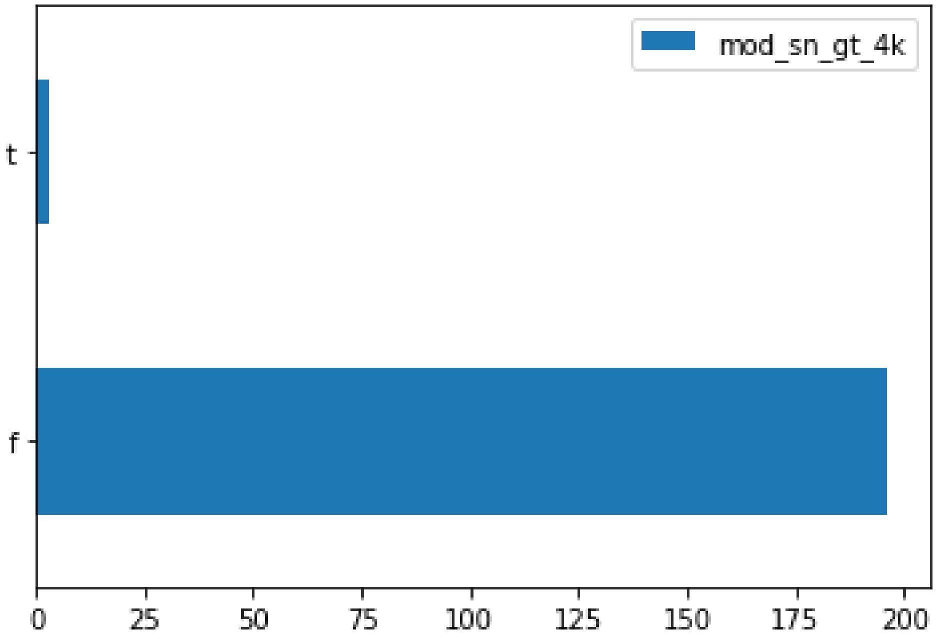

mod_s_sn_gt_500
t
f
0
25
50
75
100
125
150
175
200

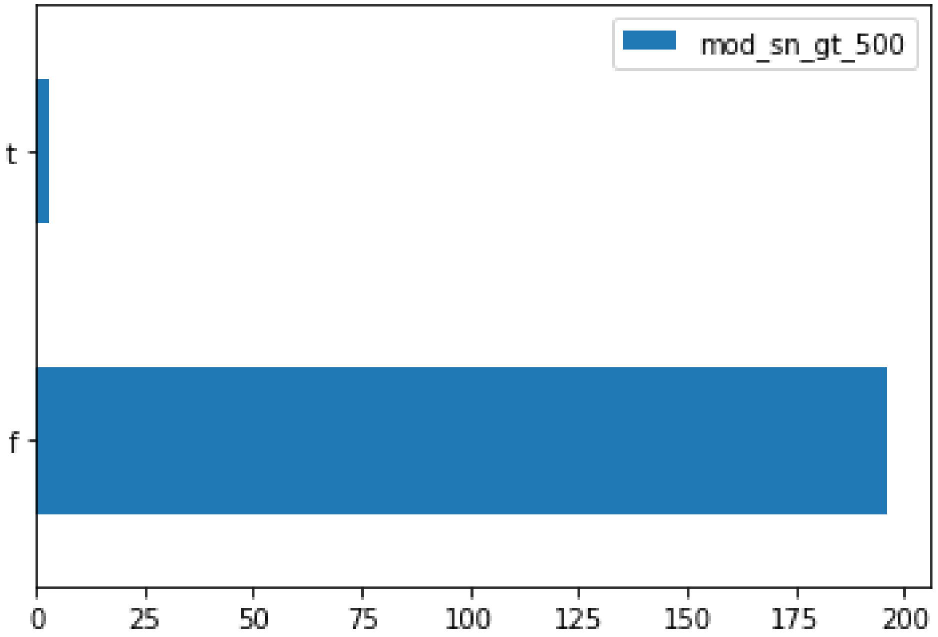

mod_sn
f
0
25
50
75
100
125
150
175
200

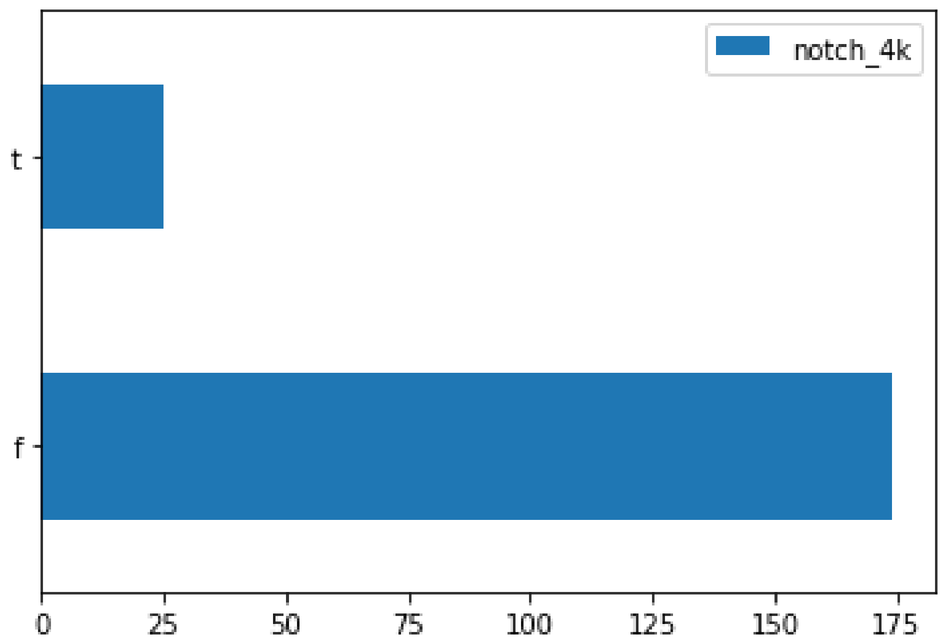

mod_sn_gt_1k
t
f
0
25
50
75
100
125
150
175
200

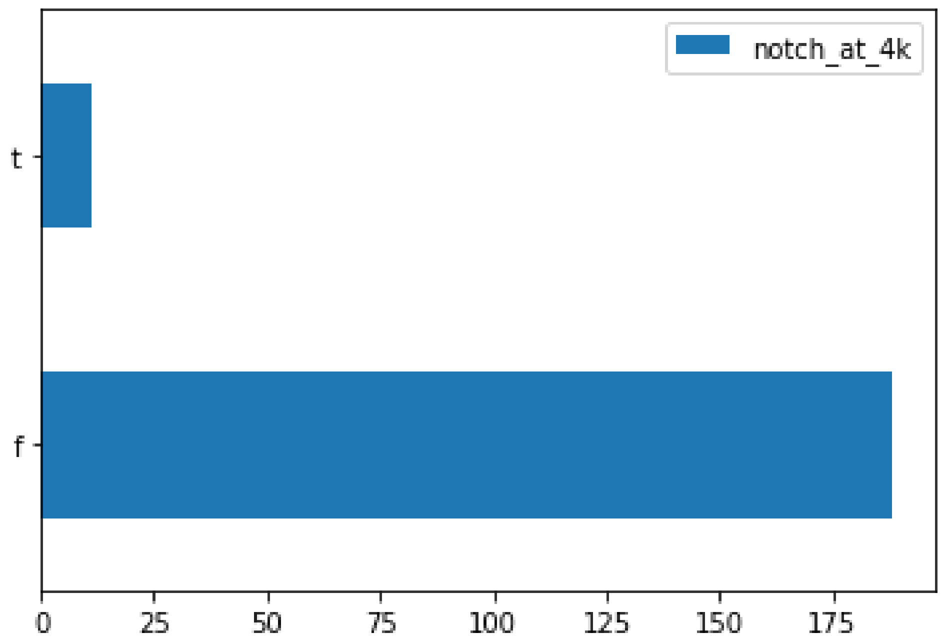

mod_sn_gt_2k
t
f
0
25
50
75
100
125
150
175
200

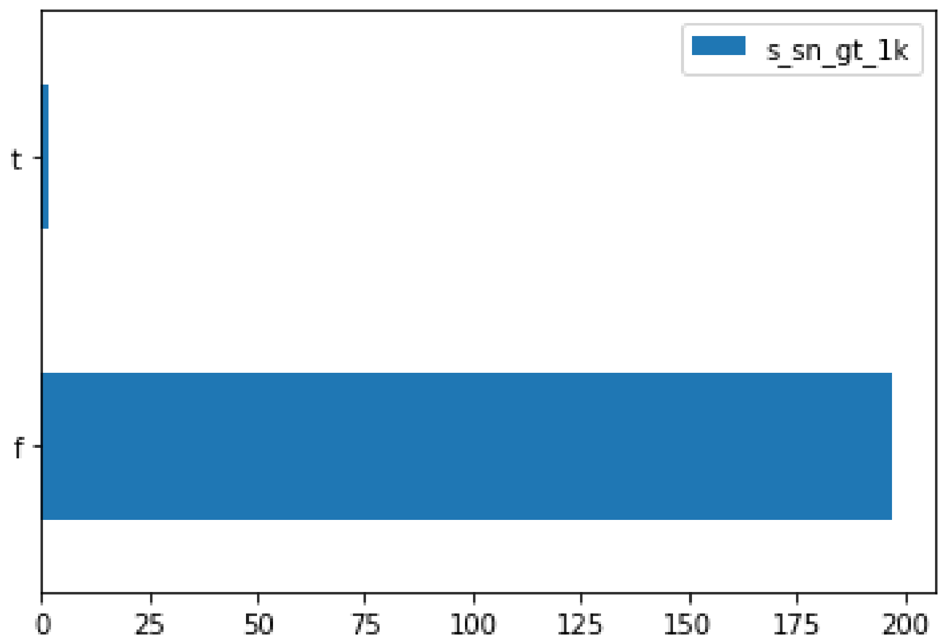

mod_sn_gt_3k
t
f
0
25
50
75
100
125
150
175

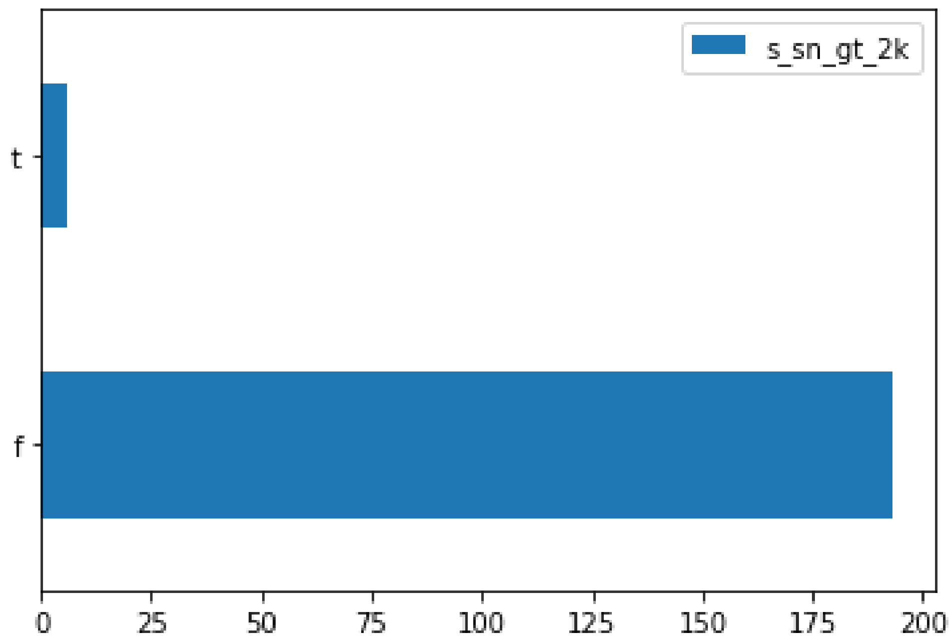

mod_sn_gt_4k
t
f
0
25
50
75
100
125
150
175
200

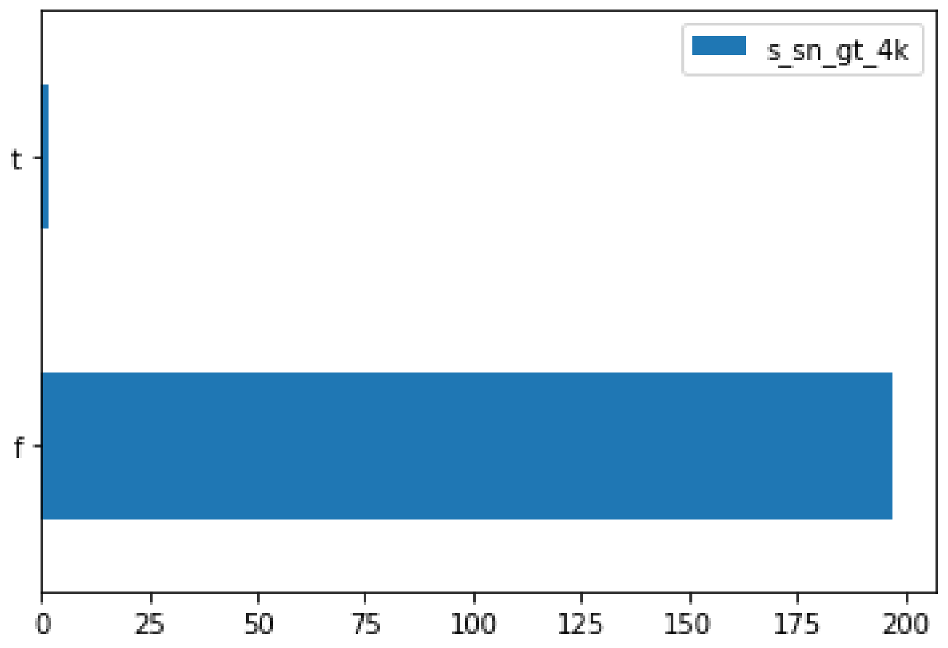

mod_sn_gt_500
t
f
0
25
50
75
100
125
150
175
200

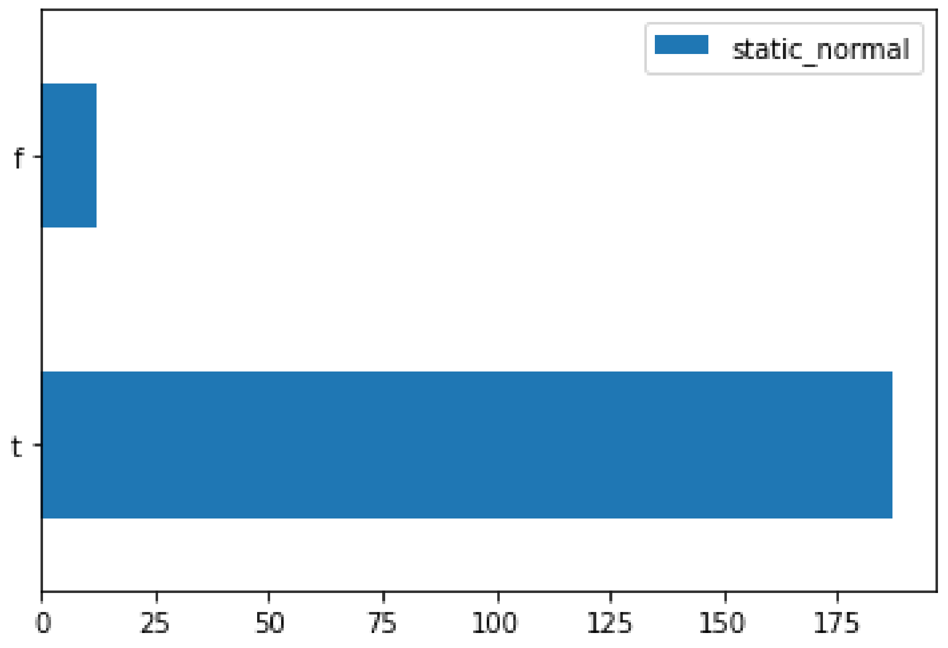

notch_4k
t
f
0
25
50
75
100
125
150
175

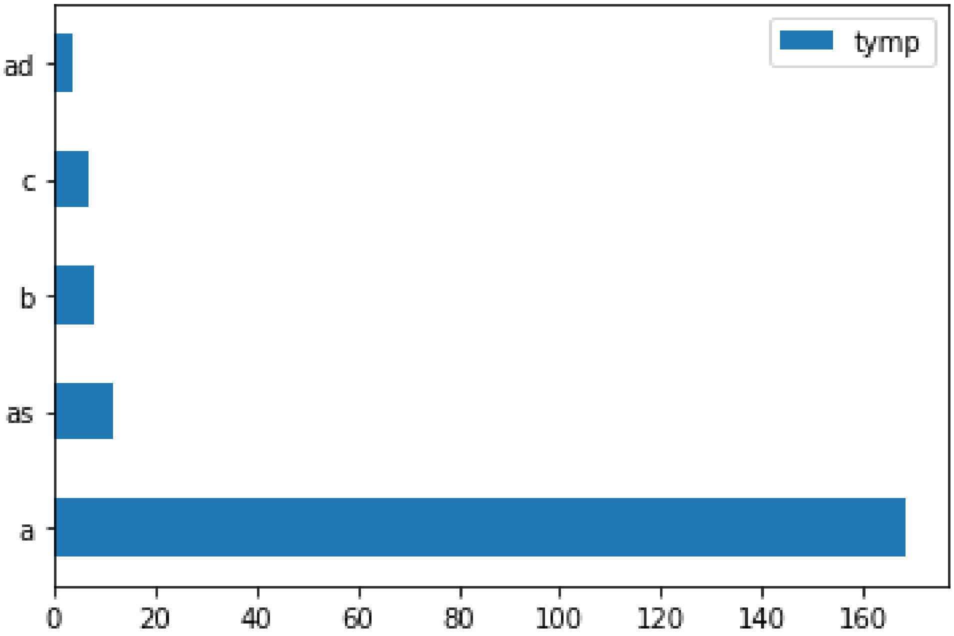

notch_at_4k
t
f
0
25
50
75
100
125
150
175

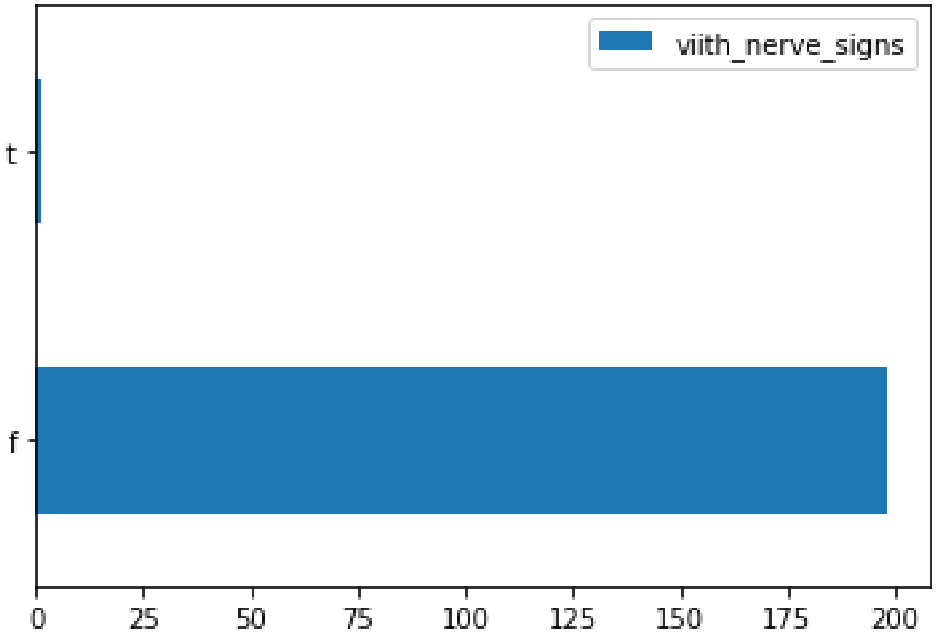

s_sn_gt_1k
t
f
0
25
50
75
100
125
150
175
200

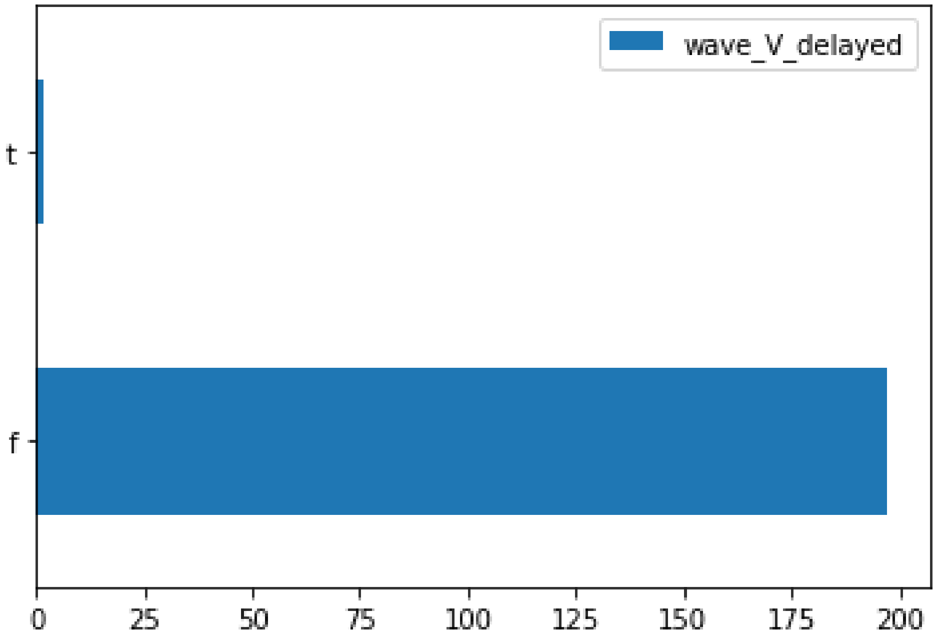

s_sn_gt_2k
t
f
0
25
50
75
100
125
150
175
200

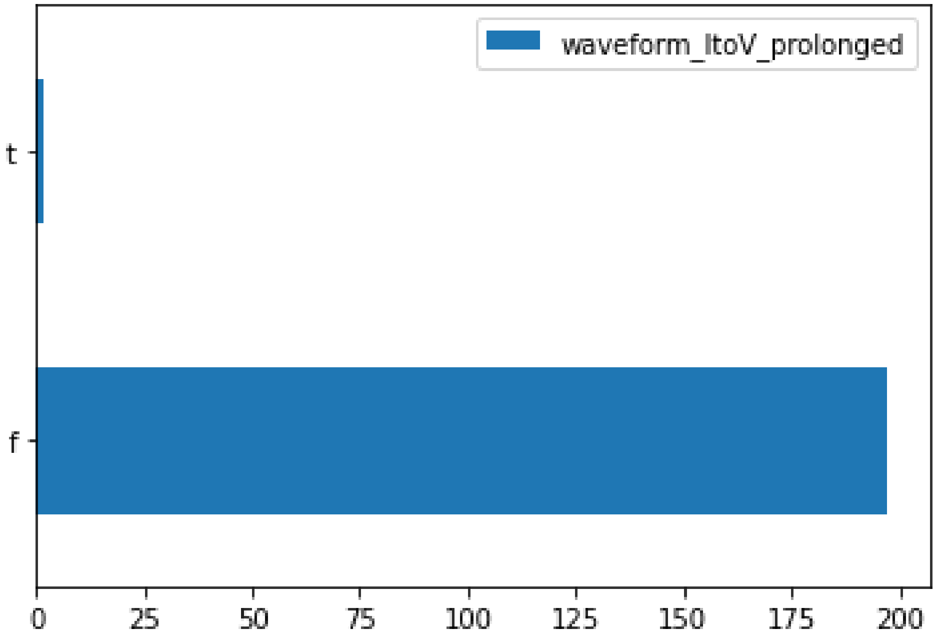

s_sn_gt_4k
t
f
0
25
50
75
100
125
150
175
200

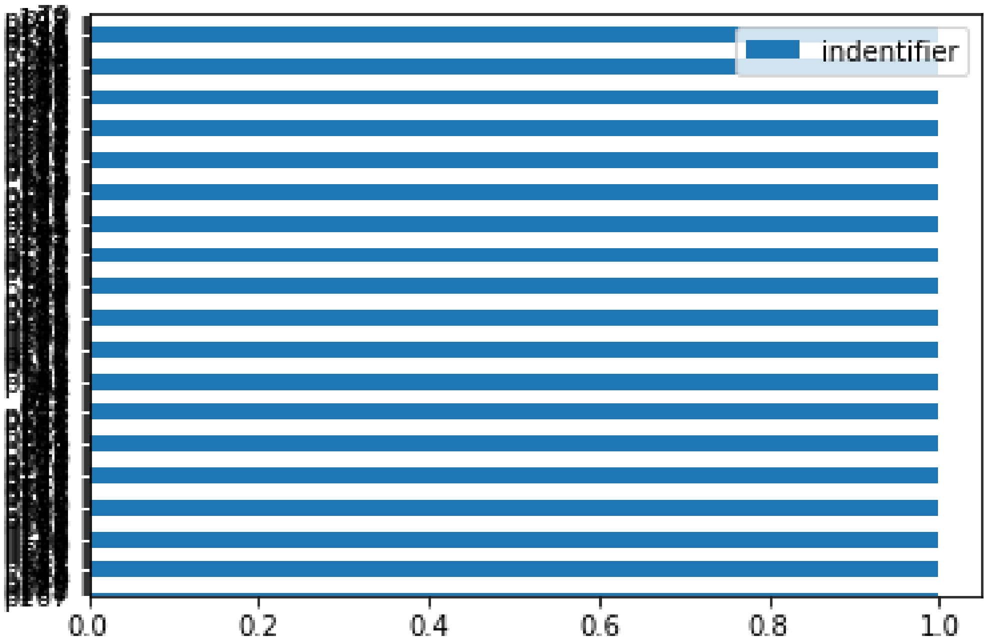

static_normal
f
t
0
25
50
75
100
125
150
175

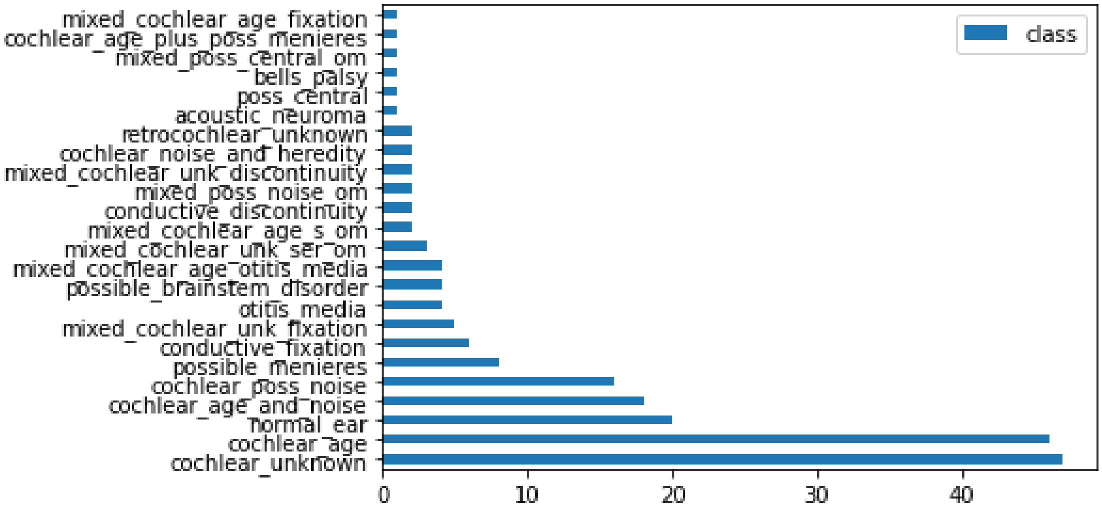

tymp
ad
c
b
as
a
0
20
40
60
80
100
120
140
160


viith_nerve_signs
t
f
0
25
50
75
100
125
150
175
200


wave_V_delayed
t
f
0
25
50
75
100
125
150
175
200


waveform_ItoV_prolonged
t
f
0
25
50
75
100
125
150
175
200

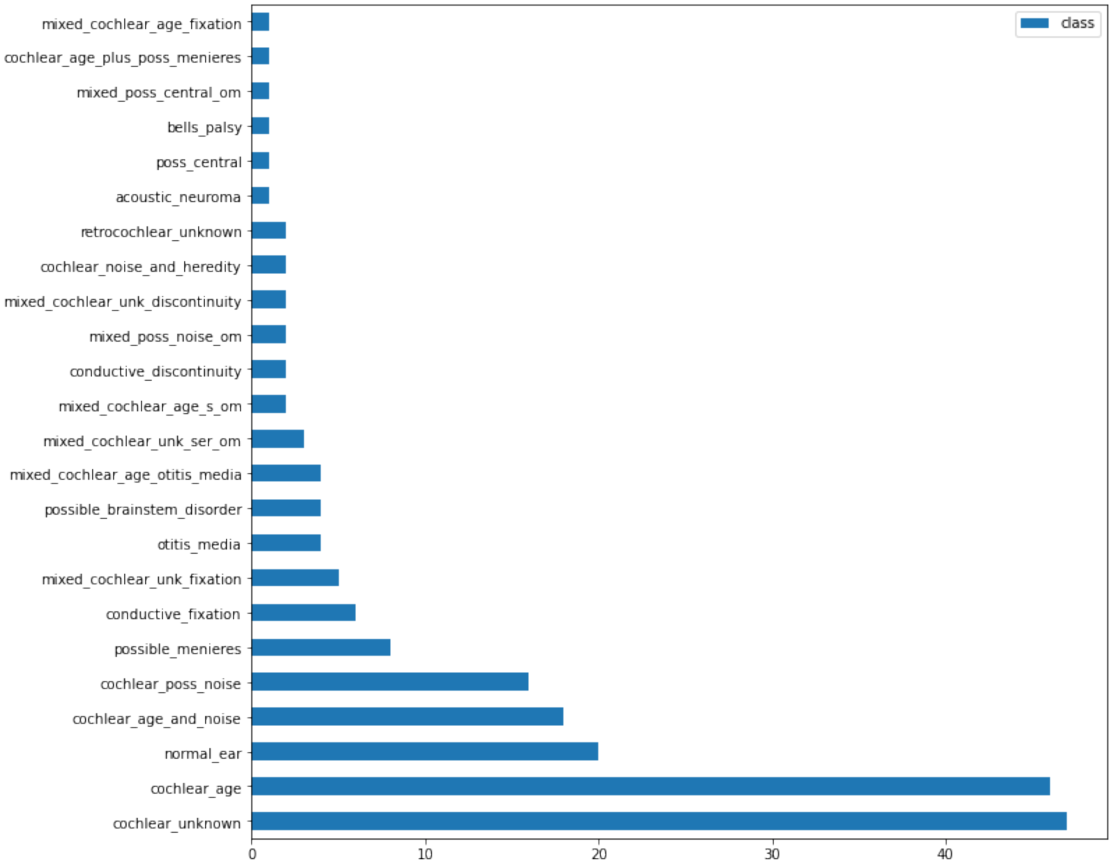




1.6 7. Draw class value counts

[11]:
```
value_count = pd.DataFrame(data_df['class'].value_counts())
value_count.plot.barh(figsize=(11,11))
```

[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd2f30bafd0>
```


class
mixed_cochlear_age_fixation
cochlear_age_plus_poss_menieres
mixed_poss_central_om
bells_palsy
poss_central
acoustic_neuroma
retrocochlear_unknown
cochlear_noise_and_heredity
mixed_cochlear_unk_discontinuity
mixed_poss_noise_om
conductive_discontinuity
mixed_cochlear_age_s_om
mixed_cochlear_unk_ser_om
mixed_cochlear_age_otitis_media
possible_brainstem_disorder
otitis_media
mixed_cochlear_unk_fixation
conductive_fixation
possible_menieres
cochlear_poss_noise
cochlear_age_and_noise
normal_ear
cochlear_age
cochlear_unknown
0
10
20
30
40