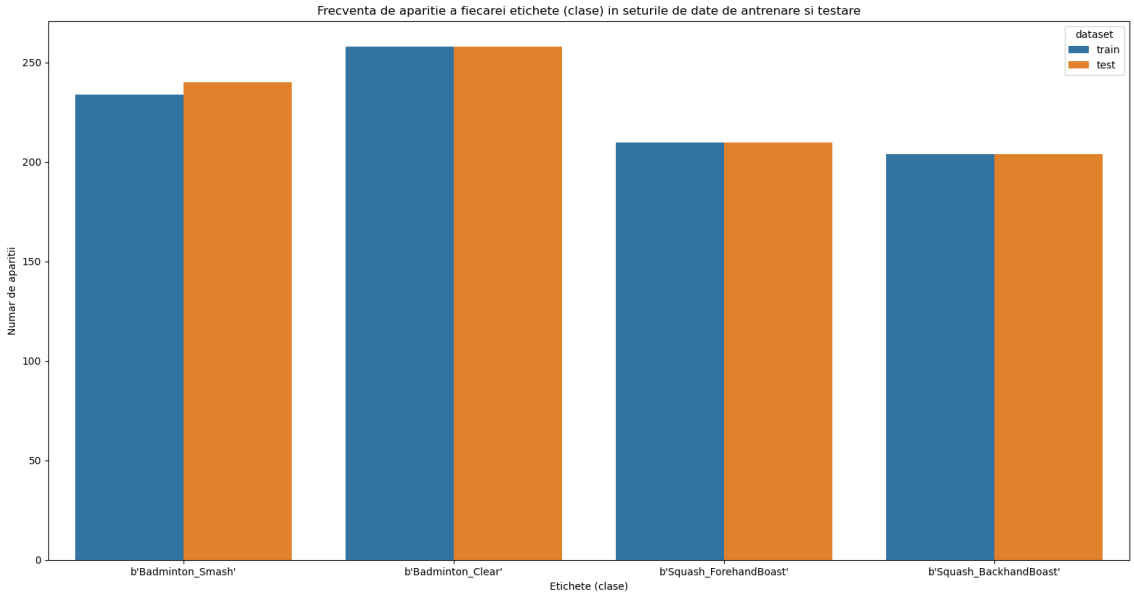
Etapa 1 ML

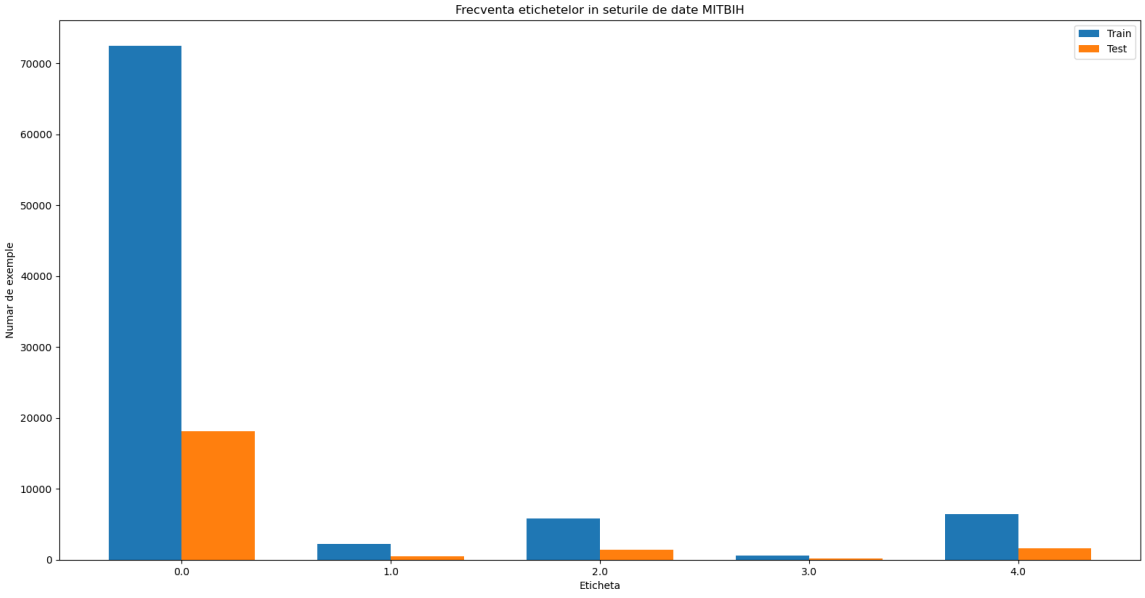
Marcu Gențiana Adeane, 343C2

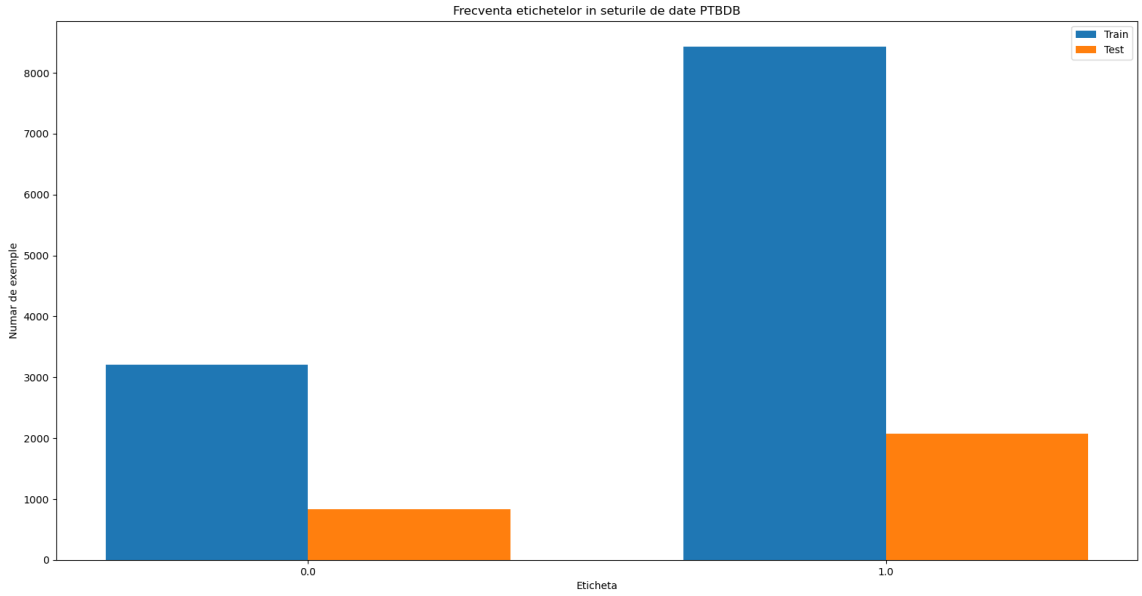
3.1 Explorarea Datelor

1. Analiza echilibrului de clase



Se poate remarca că există un echilibru între setul de date utilizat pentru antrenare și cel folosit pentru testare, deoarece pentru fiecare dintre cele patru etichete ("Badminton_Clear", "Badminton_Smash", "Squash_ForehandBoast" și "Squash_BackandBoast"), numărul de înregistrări din ambele seturi este egal, cu excepția situației în care este vorba despre "Badminton_Smash", unde numerele sunt aproape identice.



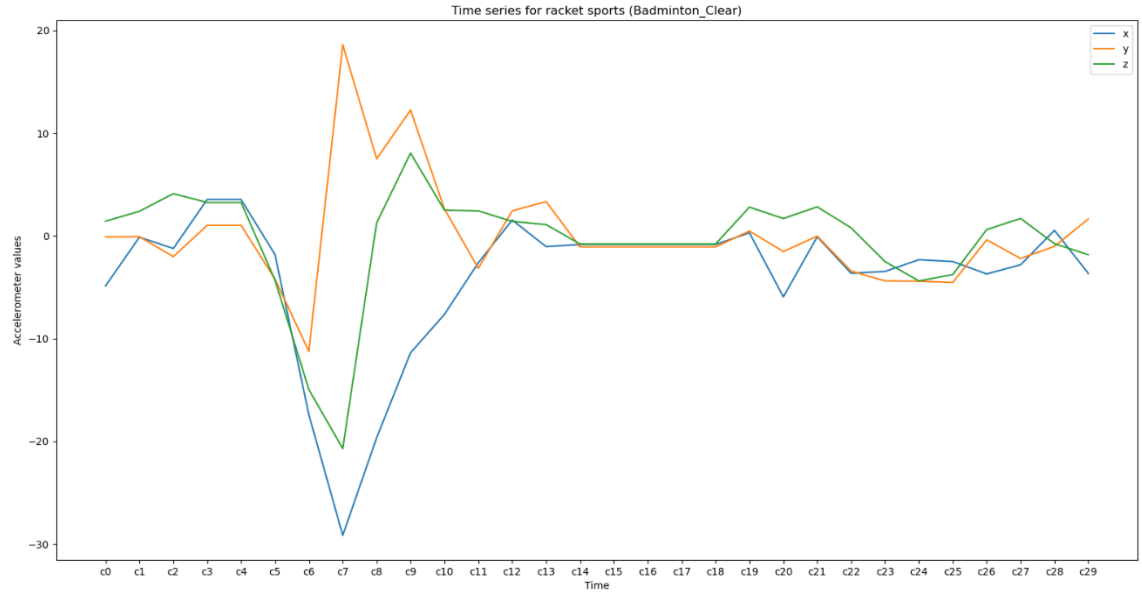


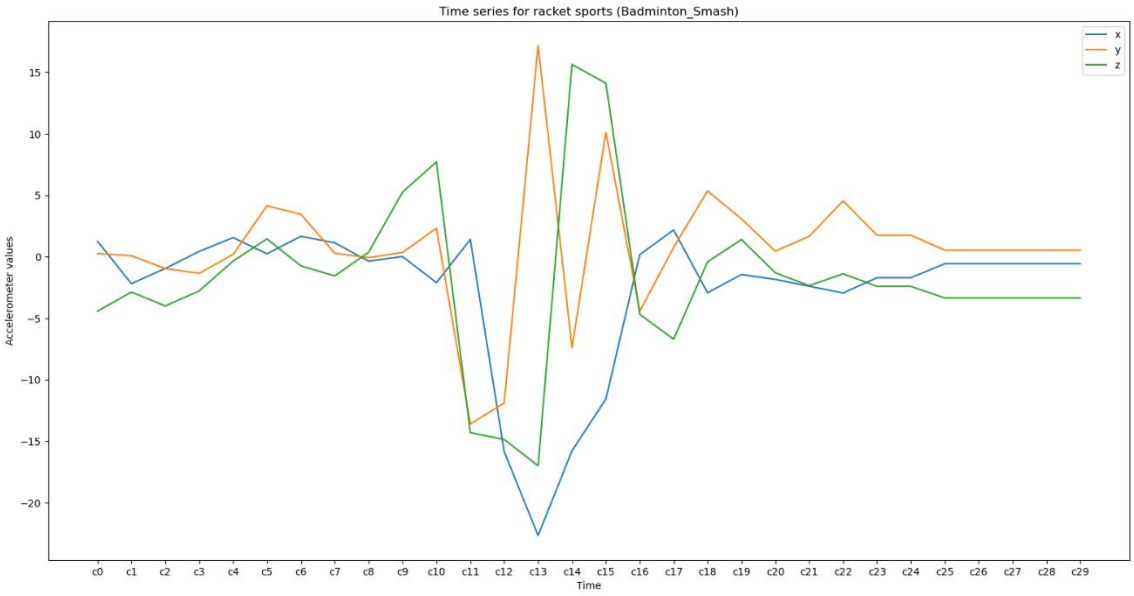
Observăm din cele două diagrame de bare de mai sus că există un dezechilibru evident între numărul de exemple din fiecare clasă, sugerând astfel avantajele standardizării datelor.

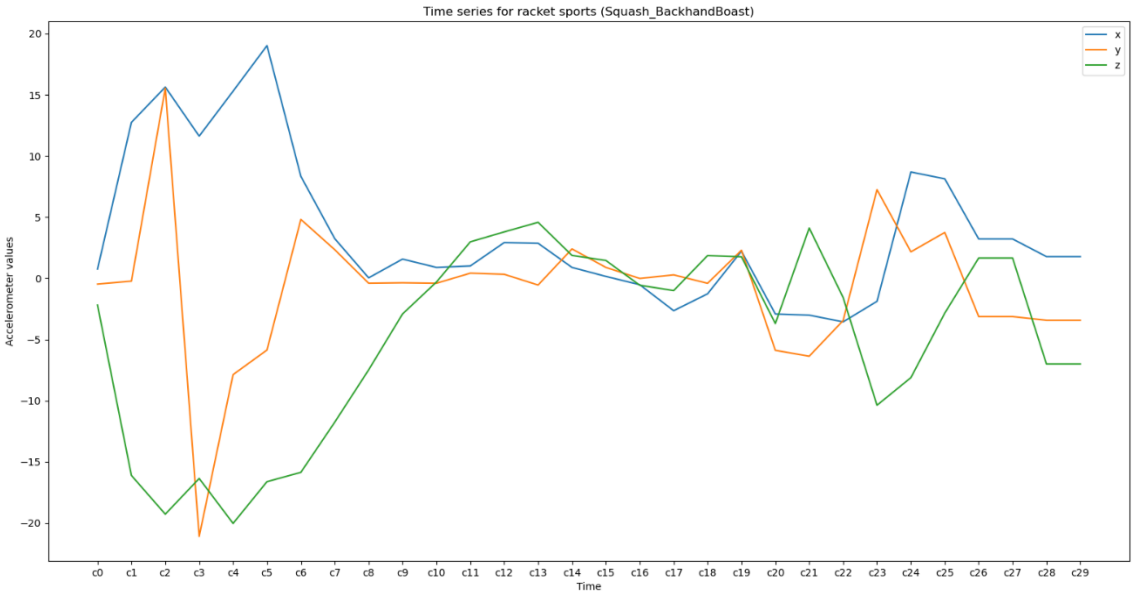
2. Vizualizarea seriilor de timp

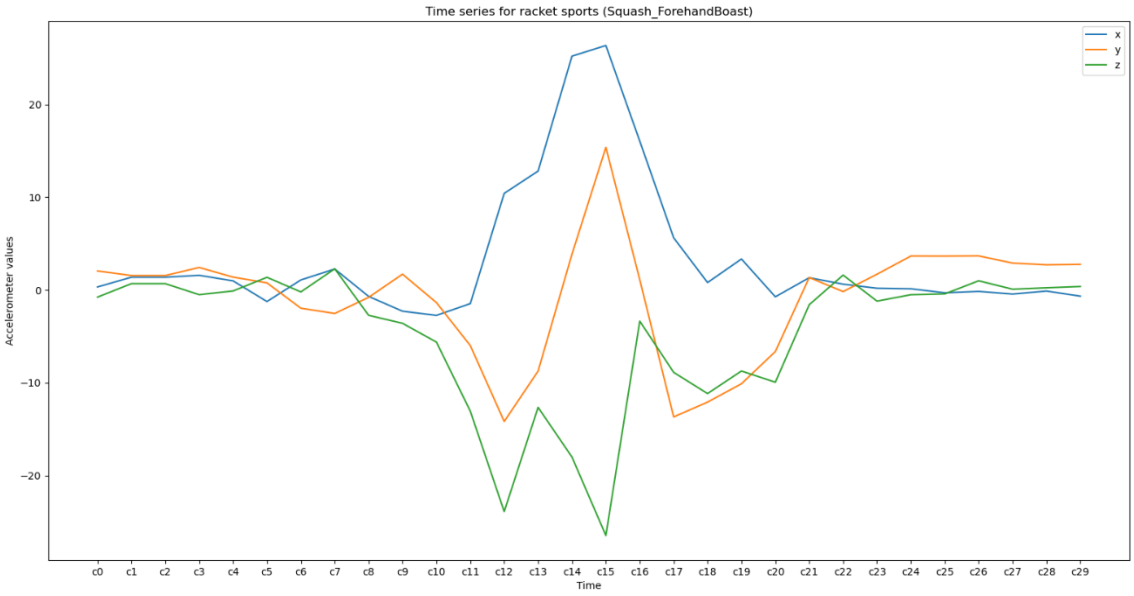
2.1. Câte un exemplu de serie pentru fiecare tip de acțiune (RacketSports)

- Accelerometru

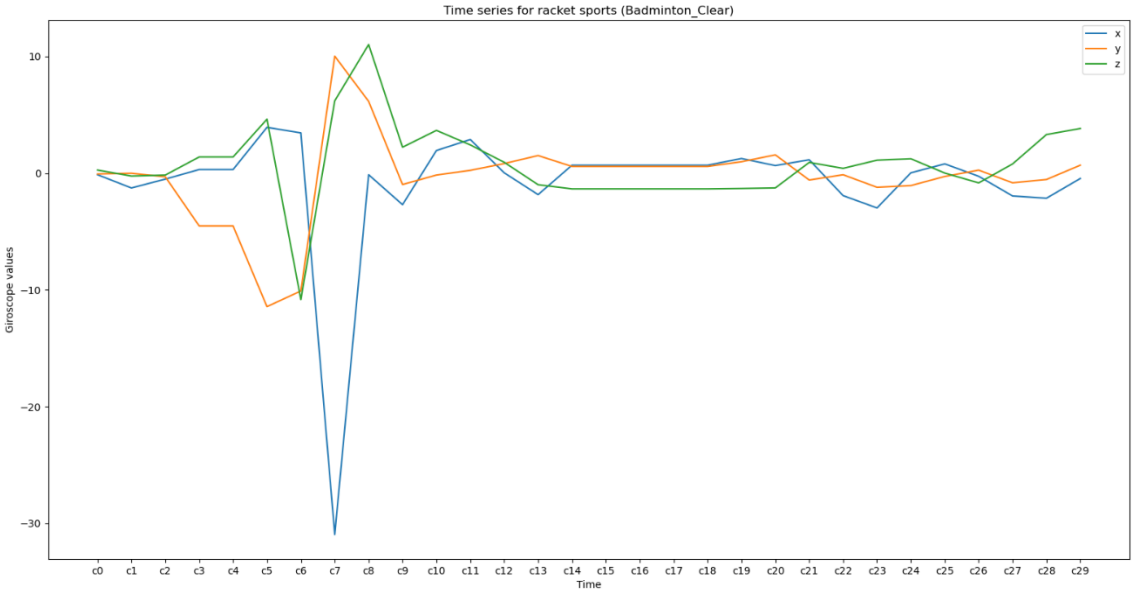


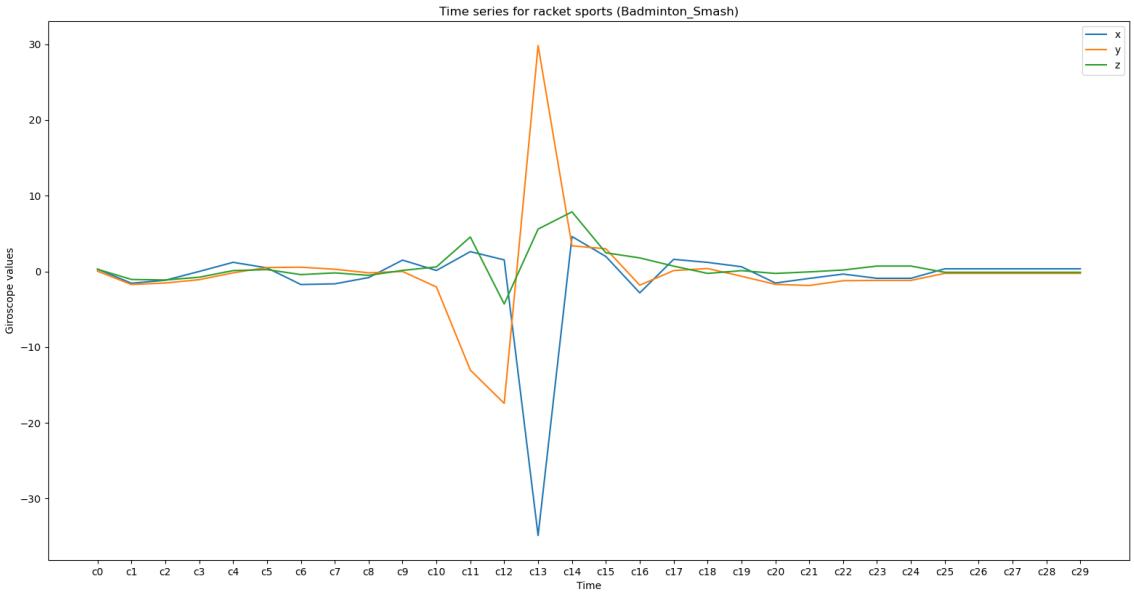

Time series for racket sports (Badminton_Smash)
x
y
z
Accelerometer values
15
10
5
0
−5
−10
−15
−20
c0 c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22 c23 c24 c25 c26 c27 c28 c29
Time

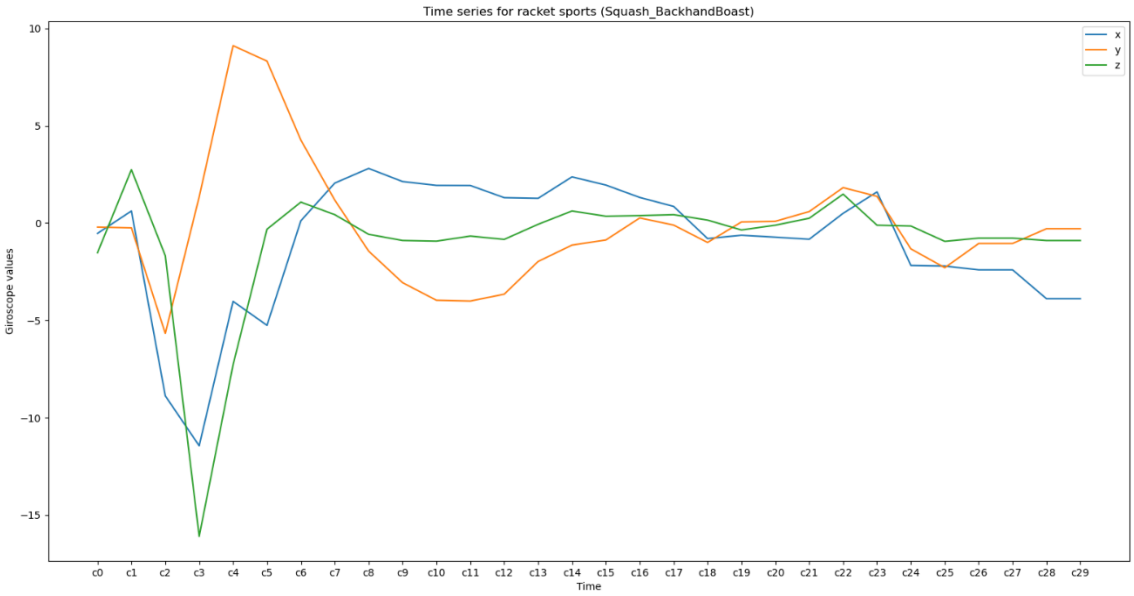

Time series for racket sports (Squash_BackhandBoast)
x
y
z
Accelerometer values
20
15
10
5
0
−5
−10
−15
−20
c0 c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22 c23 c24 c25 c26 c27 c28 c29
Time

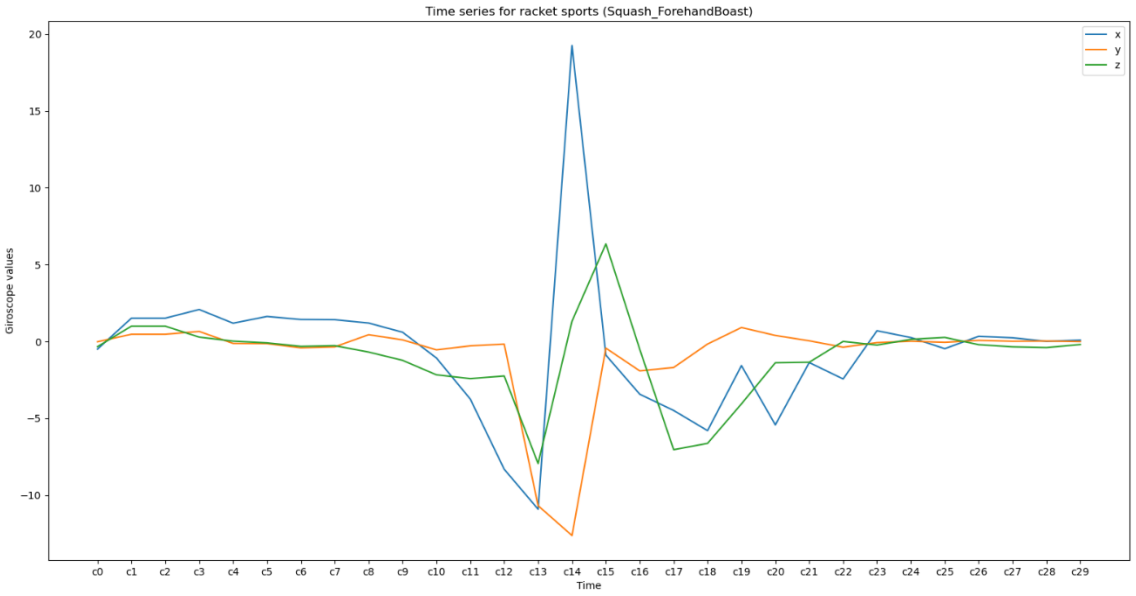

Time series for racket sports (Squash_ForehandBoast)
x
y
z
Accelerometer values
20
10
0
−10
−20
c0 c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22 c23 c24 c25 c26 c27 c28 c29
Time

- Giroscop


Time series for racket sports (Badminton_Clear)
x
y
z
Giroscope values
10
0
−10
−20
−30
c0 c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22 c23 c24 c25 c26 c27 c28 c29
Time

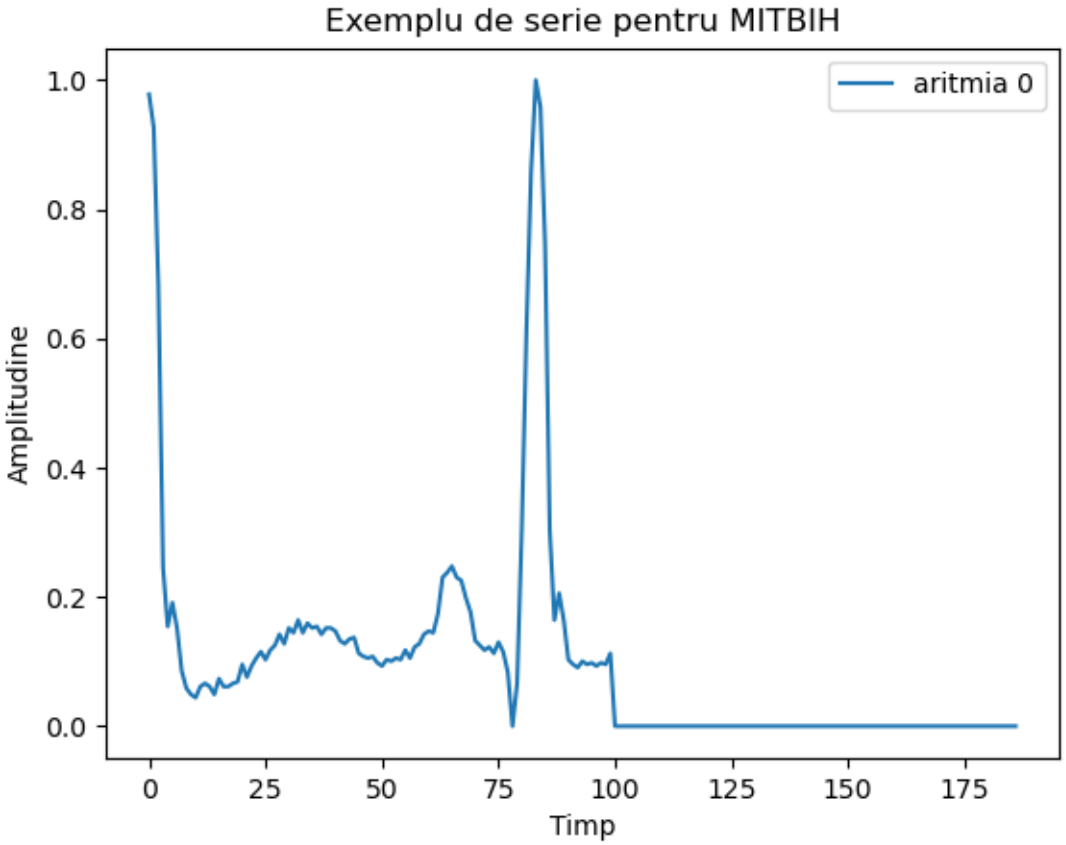

Time series for racket sports (Badminton_Smash)
x
y
z
Giroscope values
30
20
10
0
−10
−20
−30
c0 c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22 c23 c24 c25 c26 c27 c28 c29
Time

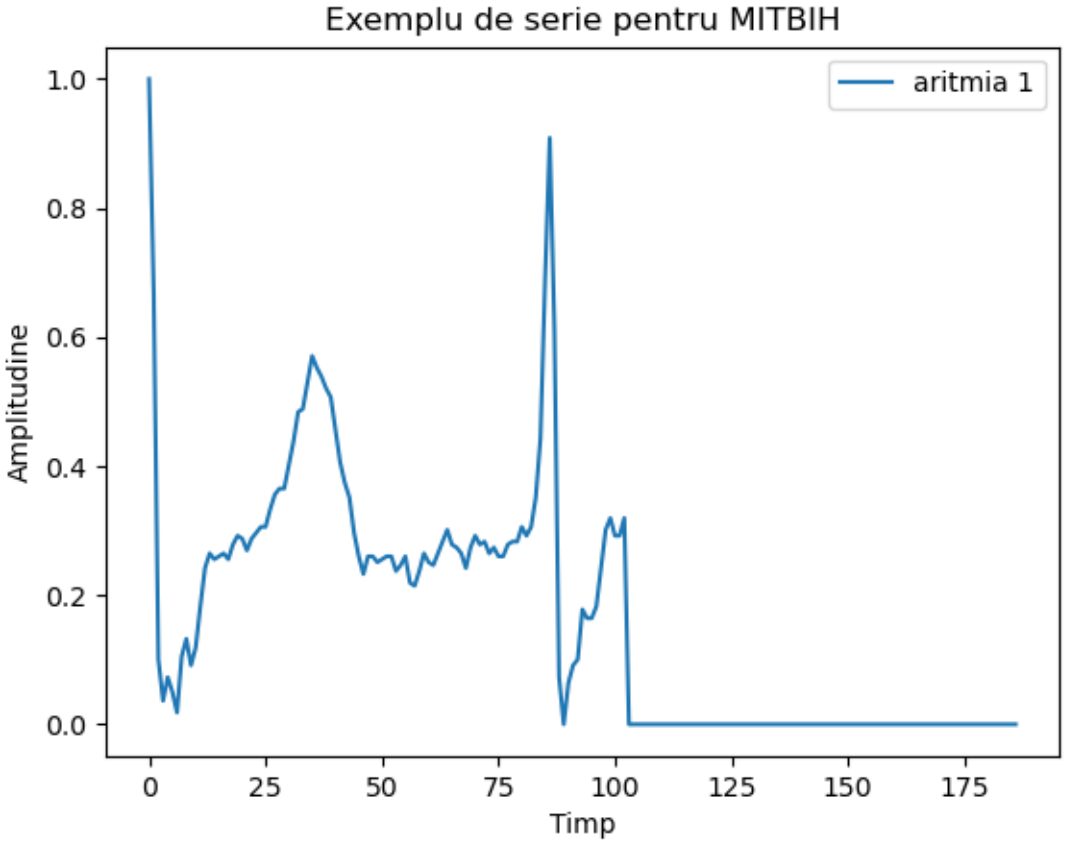

Time series for racket sports (Squash_BackhandBoast)
x
y
z
Giroscope values
10
5
0
−5
−10
−15
c0 c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22 c23 c24 c25 c26 c27 c28 c29
Time

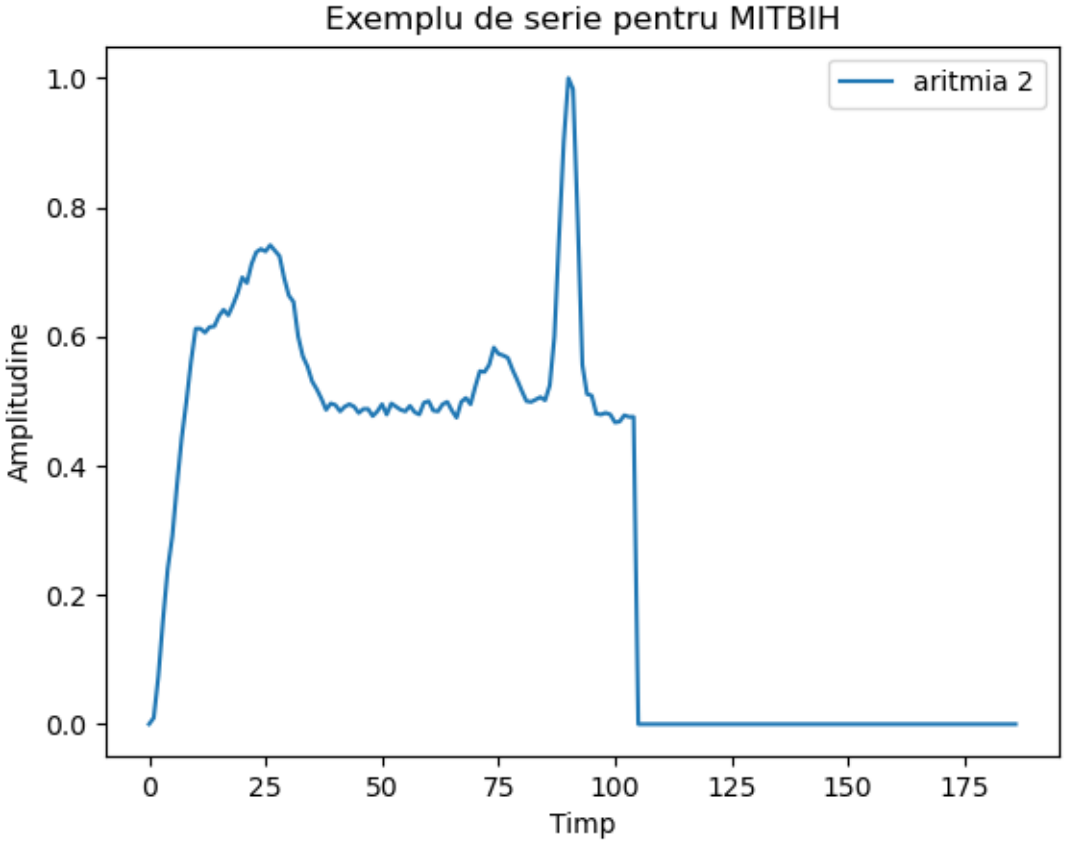


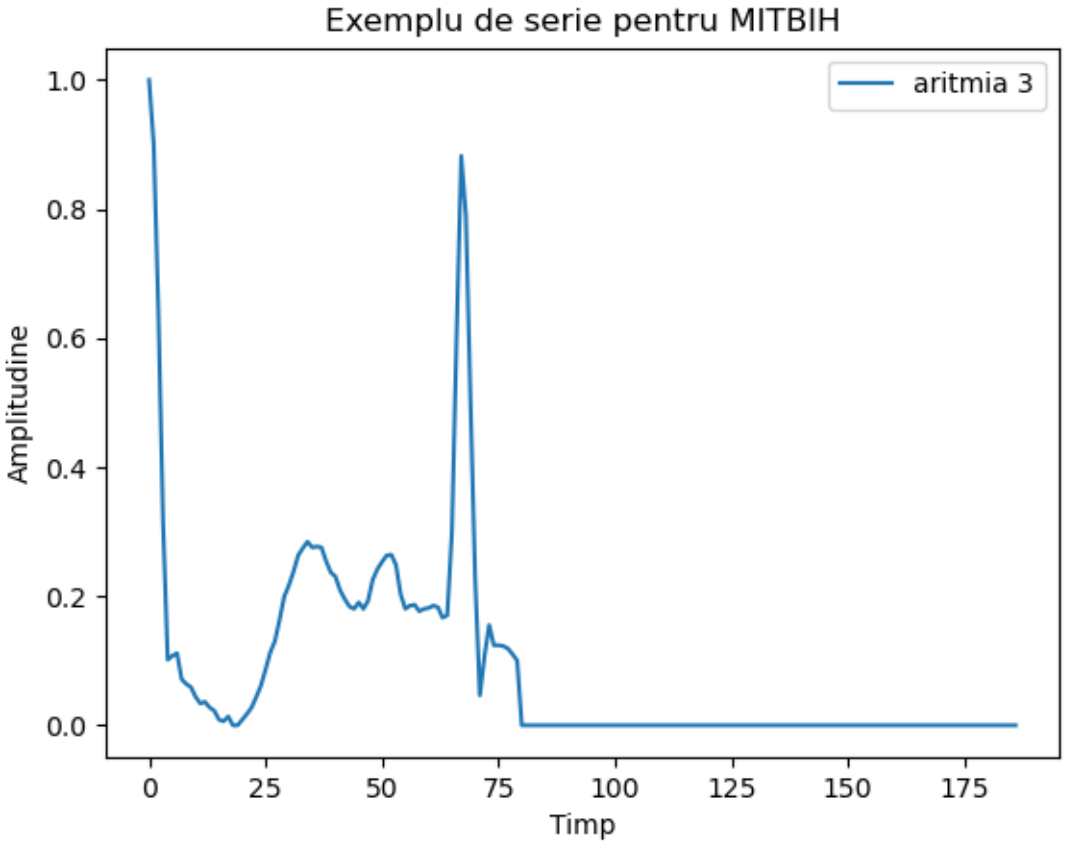
Bazându-ne pe seriile de timp pentru RacketSports, se poate observa o strânsă legătură între valorile celor trei axe de coordonate. Există mai multe zone în care curbele sunt aproape identice, ceea ce sugerează redundanța datelor, aceste zone având un impact nesemnificativ asupra predicției.

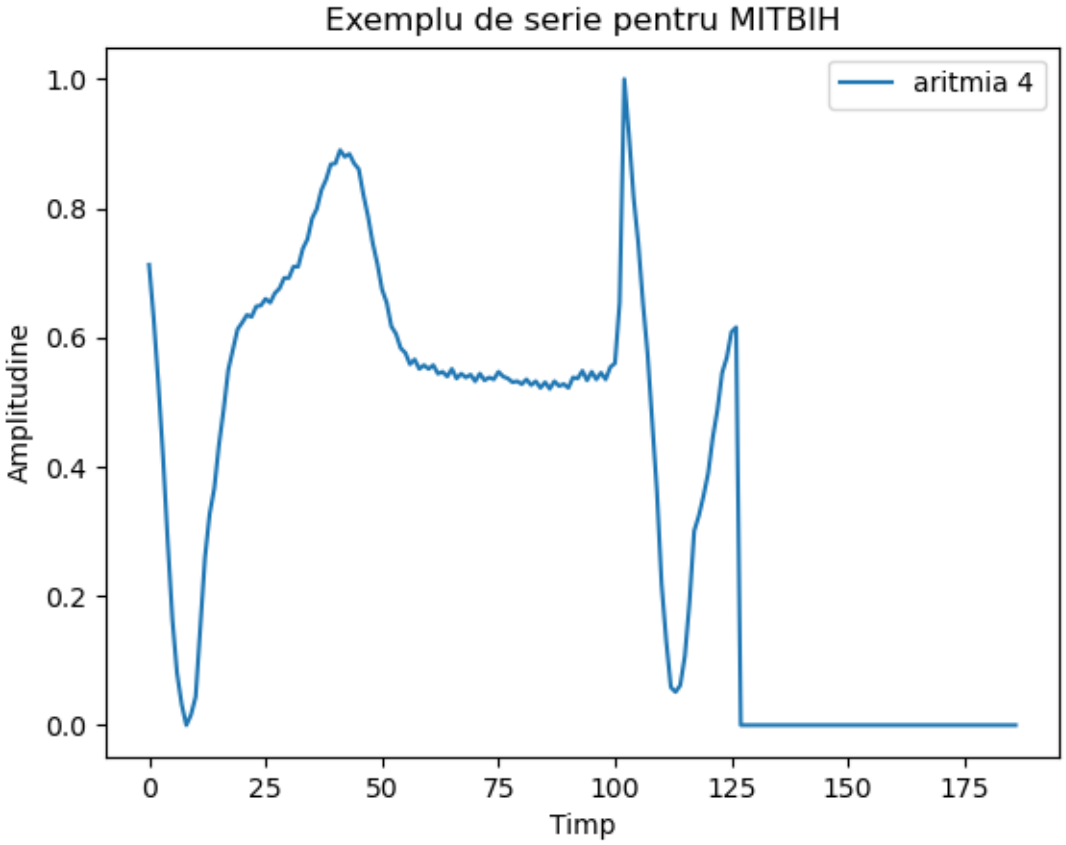
2.2. Câte un exemplu de serie pentru fiecare categorie de aritmie din seturile de date MIT-BIH / PTB

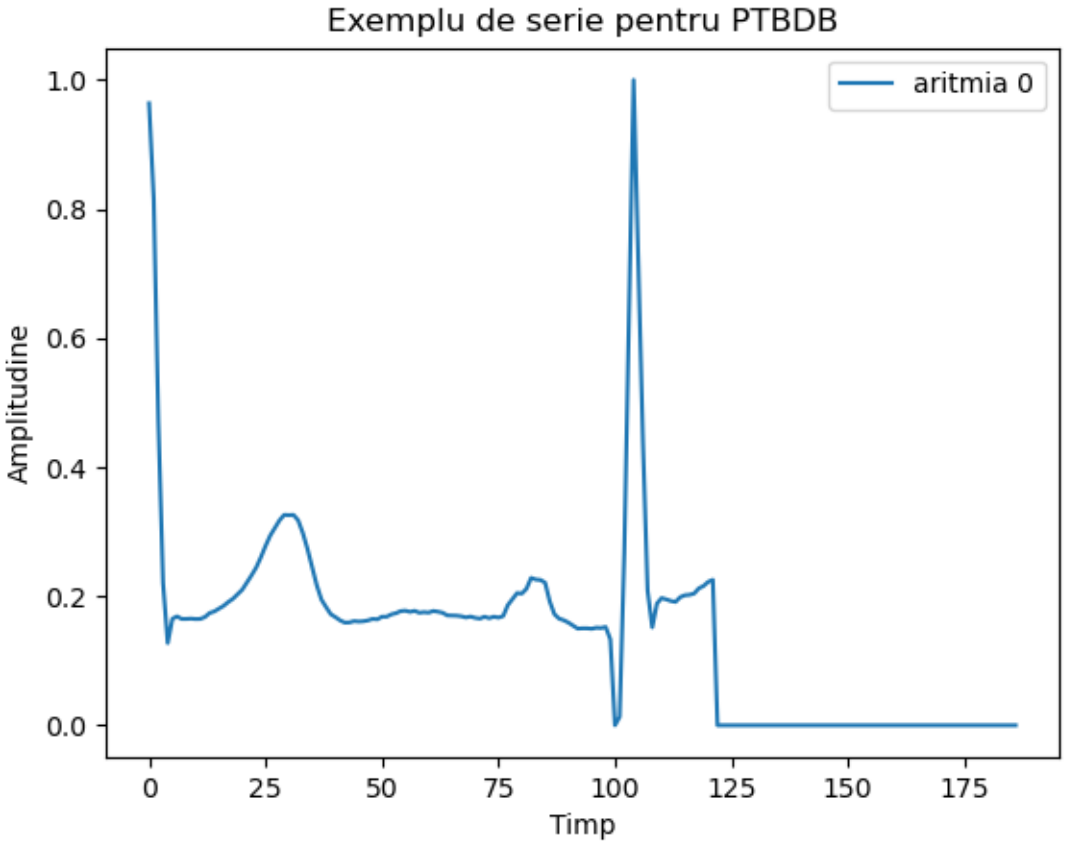


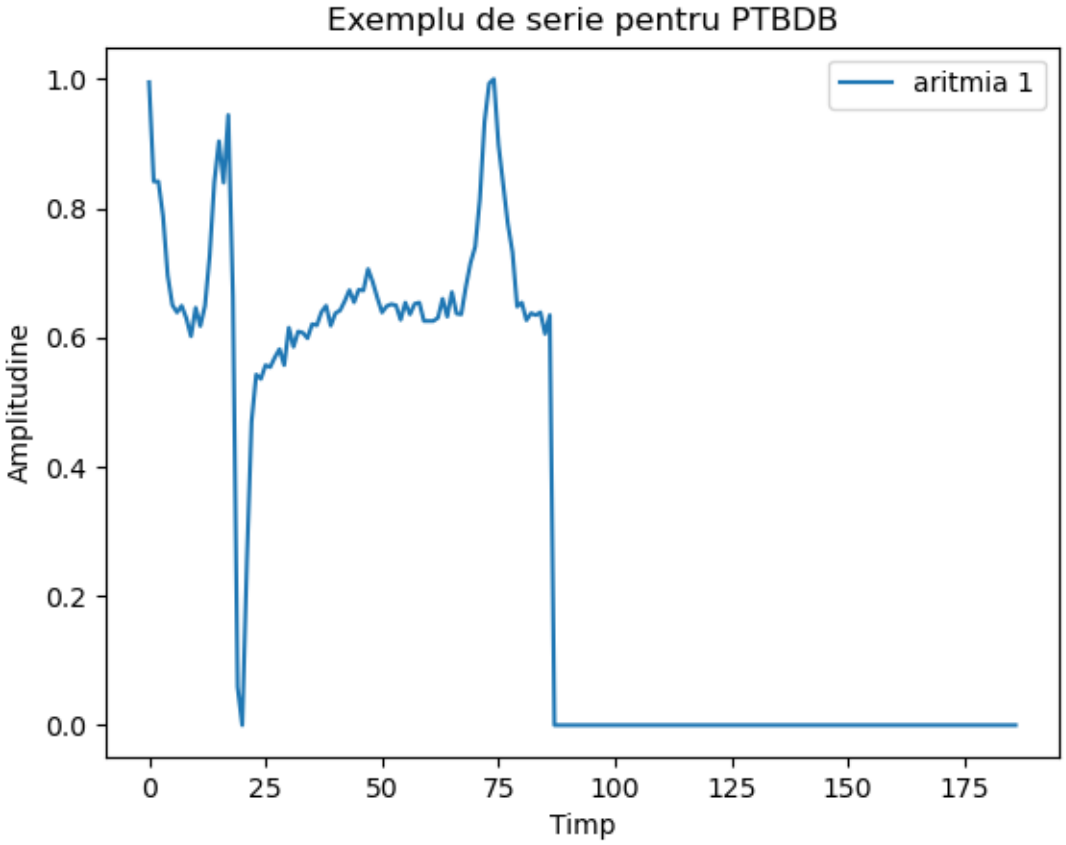

Exemplu de serie pentru MITBIH
aritmia 1
Amplitudine
Timp
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175


Exemplu de serie pentru MITBIH
aritmia 2
Amplitudine
Timp
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175

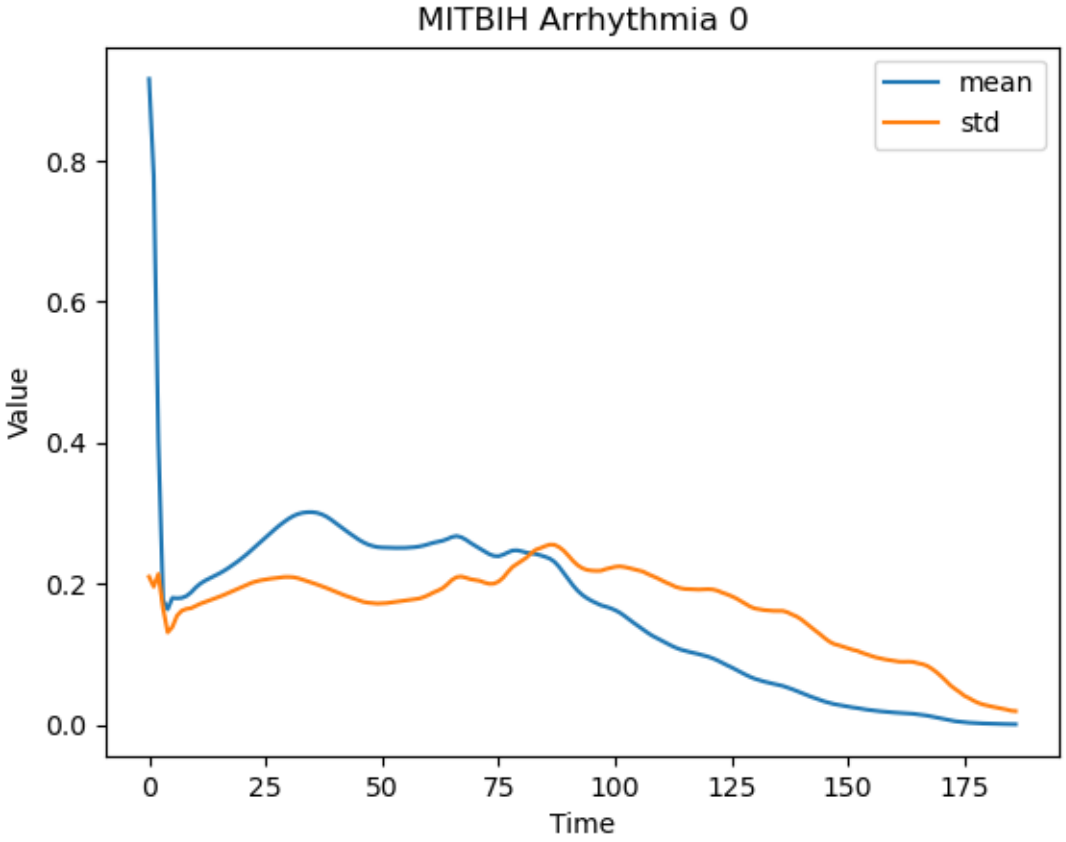

Exemplu de serie pentru MITBIH
aritmia 3
Amplitudine
Timp
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175

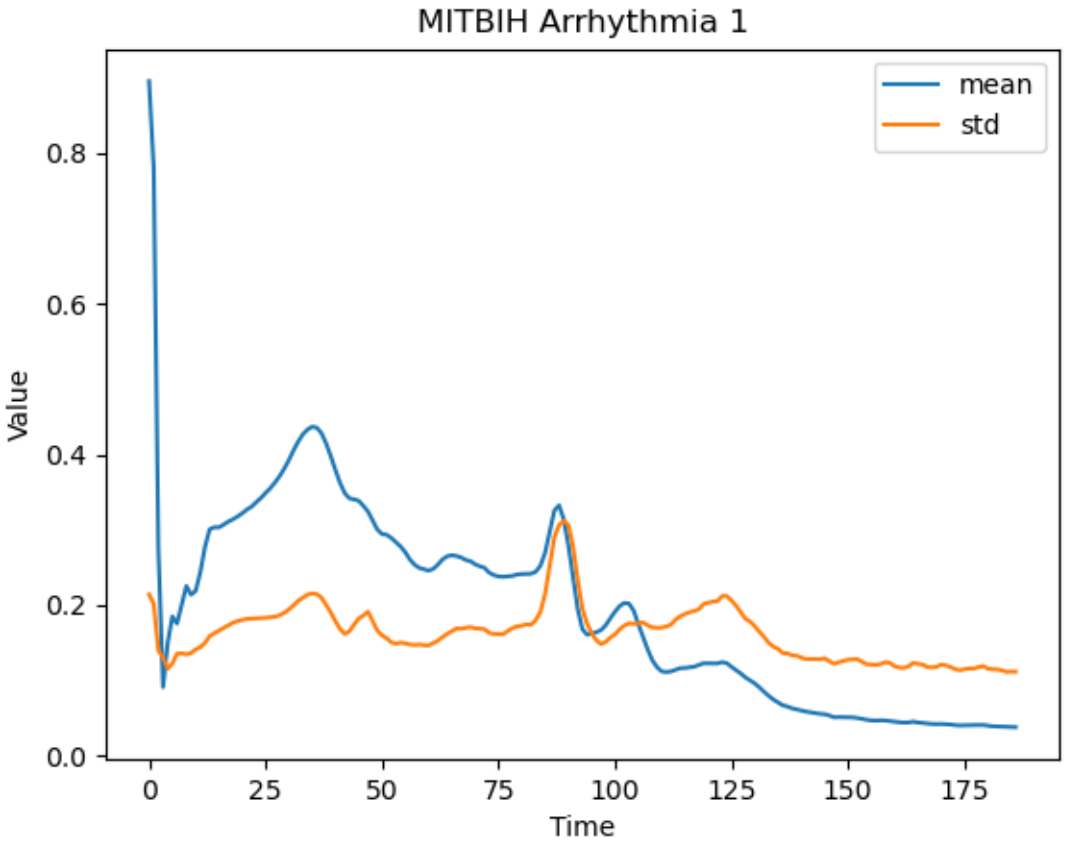

Exemplu de serie pentru MITBIH
aritmia 4
Amplitudine
Timp
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175

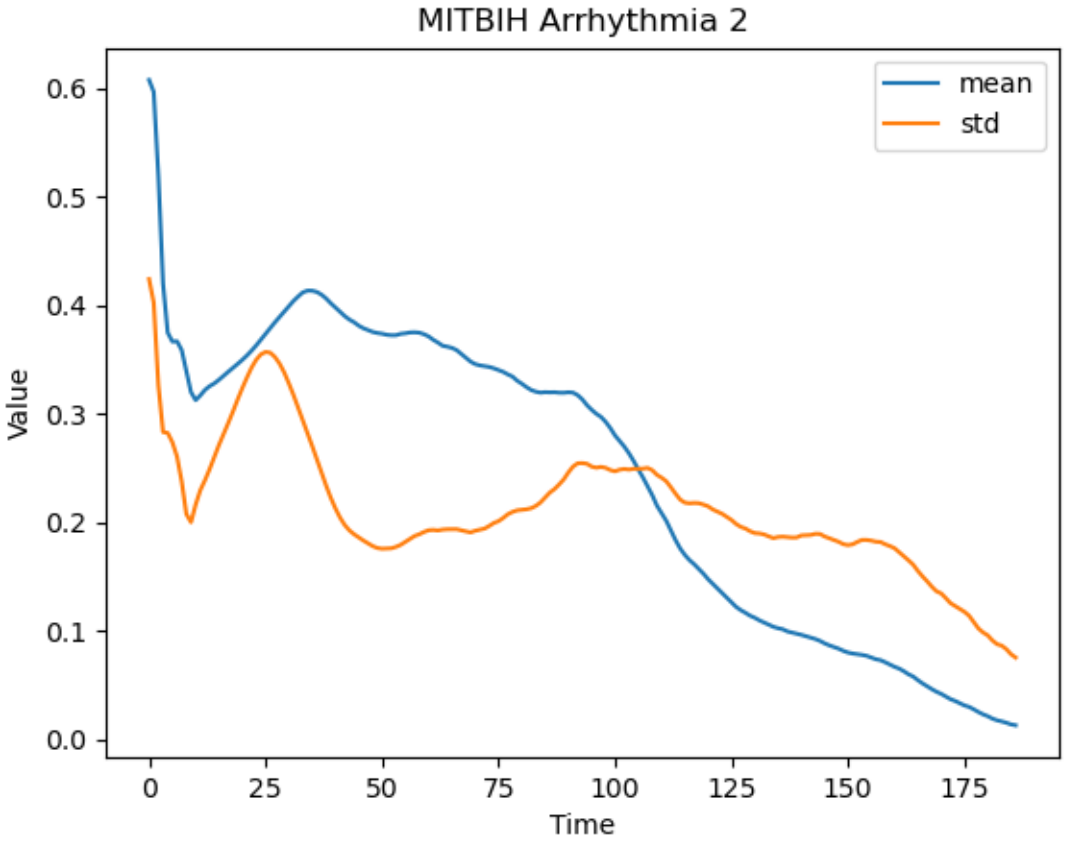

Exemplu de serie pentru PTBDB
aritmia 0
Amplitudine
Timp
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175

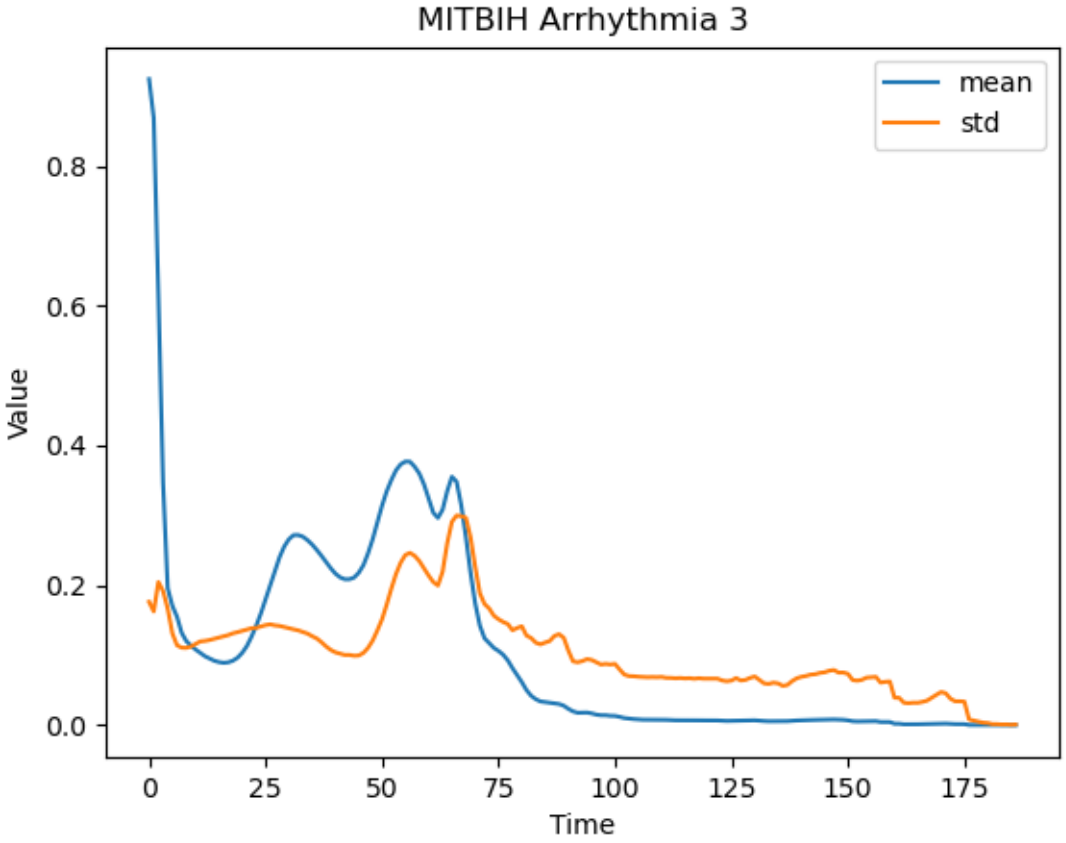

Exemplu de serie pentru PTBDB
aritmia 1
Amplitudine
Timp
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175

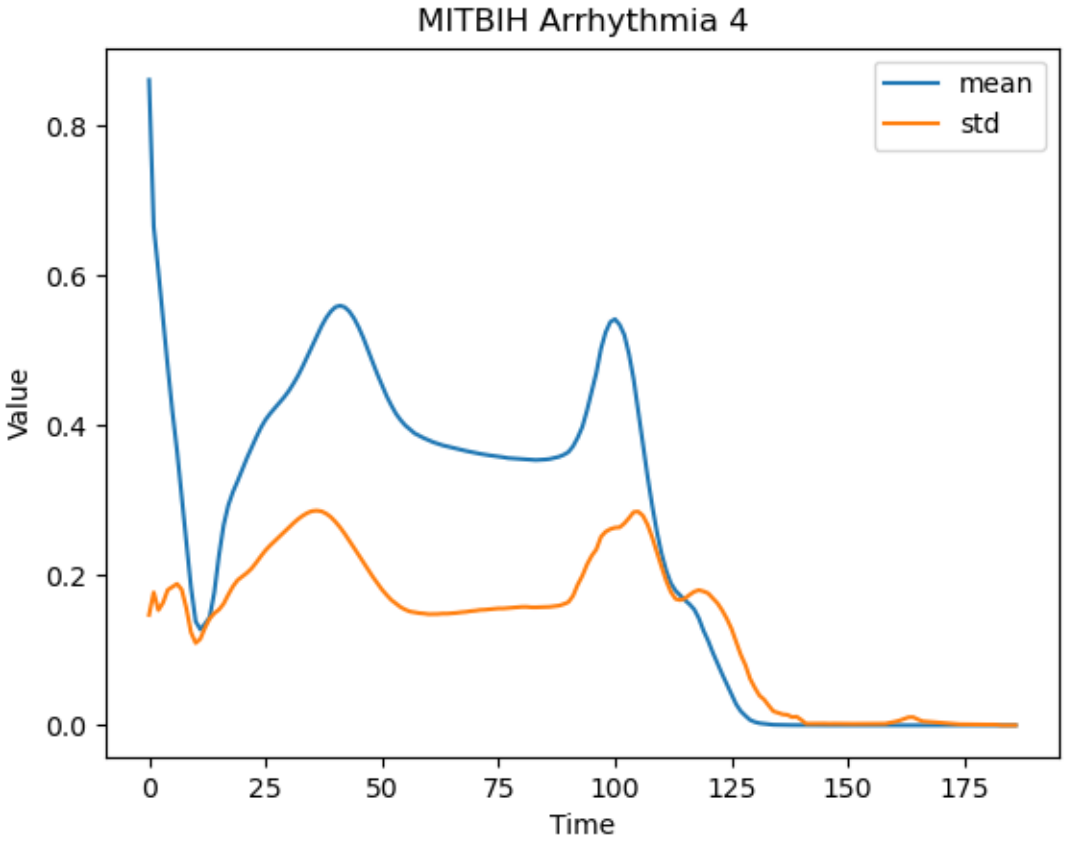
2.3. Media și deviația standard per unitate de timp, pentru fiecare clasă de aritmie (MITBIH, PTBDB)

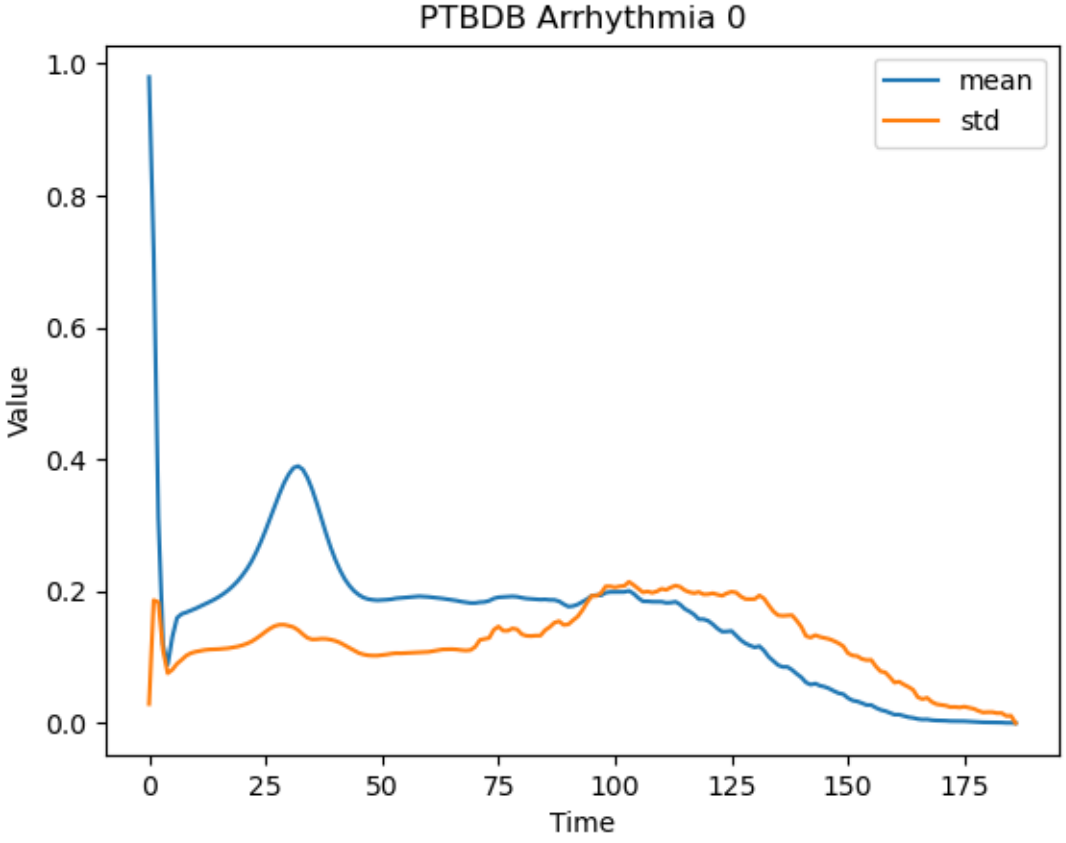


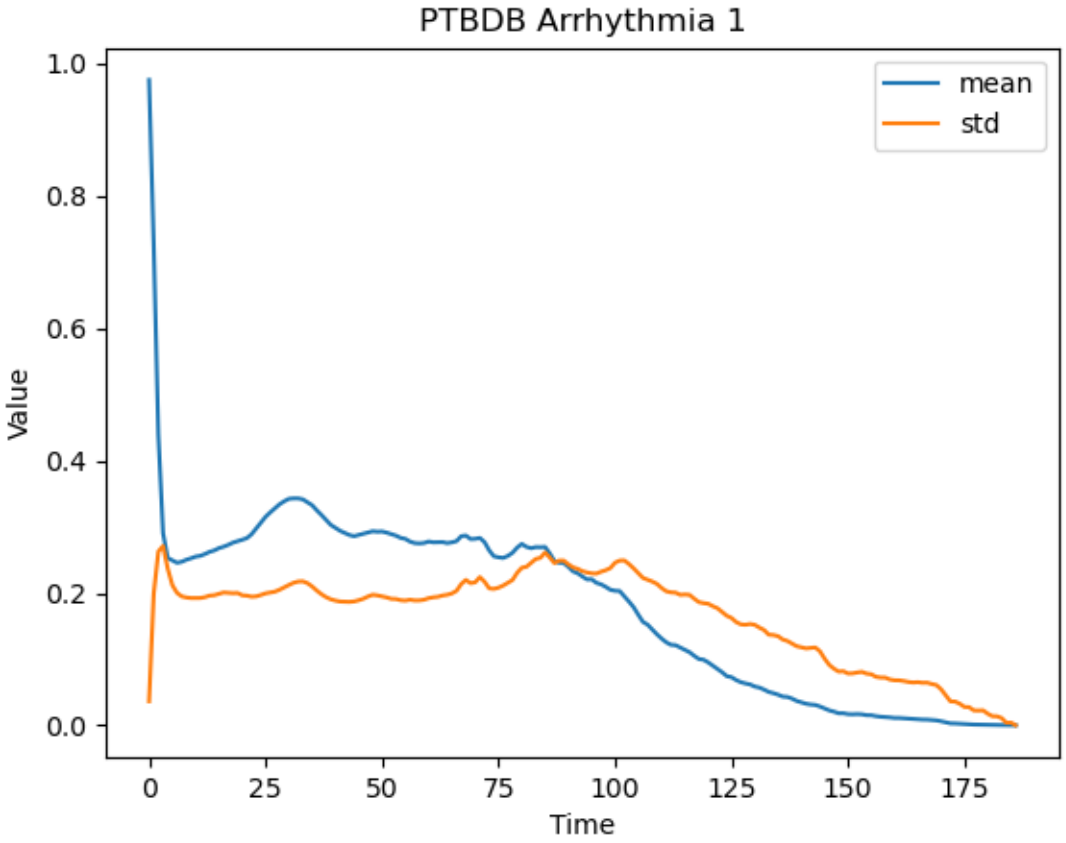



MITBIH Arrhythmia 2
mean
std
Value
Time
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
25
50
75
100
125
150
175


MITBIH Arrhythmia 3
mean
std
Value
Time
0.0
0.2
0.4
0.6
0.8
0
25
50
75
100
125
150
175


MITBIH Arrhythmia 4
mean
std
Value
Time
0.0
0.2
0.4
0.6
0.8
0
25
50
75
100
125
150
175


PTBDB Arrhythmia 0
mean
std
Value
Time
0.0
0.2
0.4
0.6
0.8
1.0
0
25
50
75
100
125
150
175

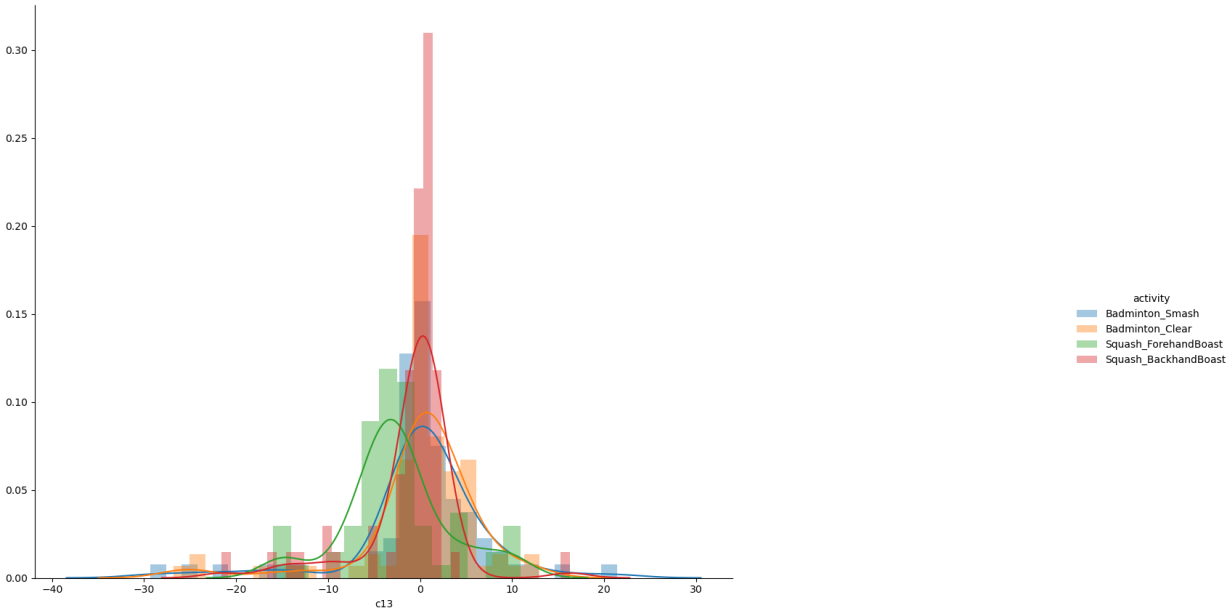


Cu privire la graficele care prezintă media și deviația standard pentru aritmii, se pot desprinde următoarele concluzii:

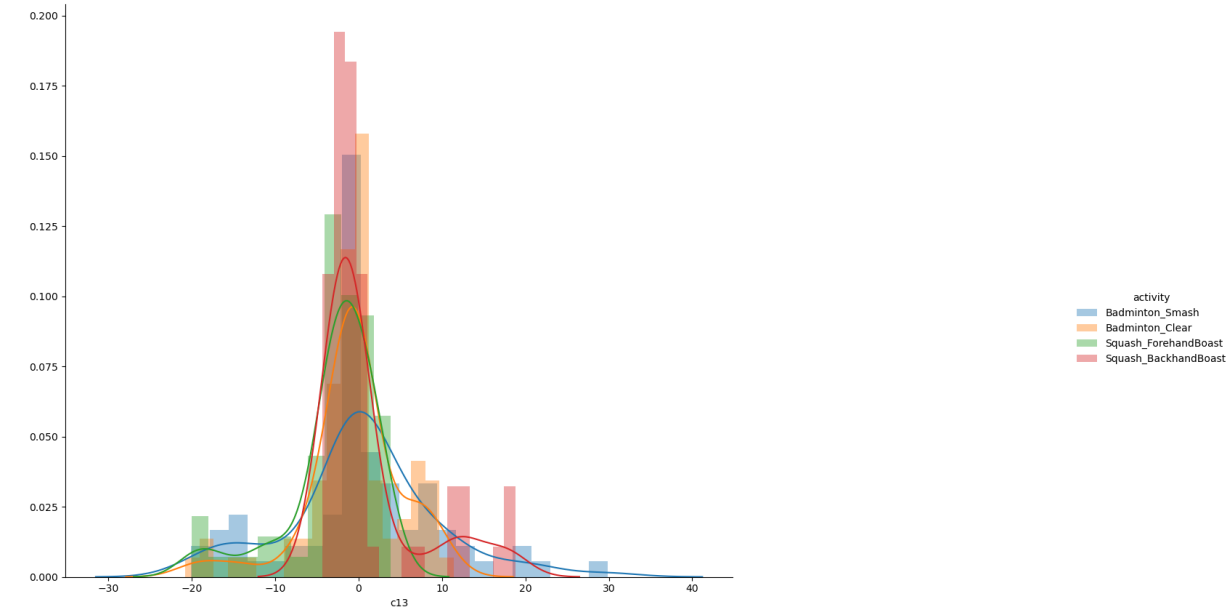
- În cazul în care media crește până la valoarea 1.0, iar deviația standard scade până la 0.0, putem presupune că seriile de date au fost completate cu zero-uri (padding) pentru a avea aceeași lungime cu celelalte serii.
- Atunci când valorile medii și de deviație standard sunt similare, putem deduce că datele sunt concentrate în jurul valorii medii, ceea ce indică o dispersie nesemnificativă. Această situație poate fi un dezavantaj pentru algoritmul de clasificare, deoarece nu va fi clar care este clasa cea mai adecvată.
- În zonele în care valorile medii și de deviație standard diferă semnificativ, datele din seturile de intrare sunt dispersate. Aceste zone pot favoriza acuratețea predicțiilor.

2.4. Distribuția valorilor per fiecare axă de accelerometru și giroscop în parte / per acțiune

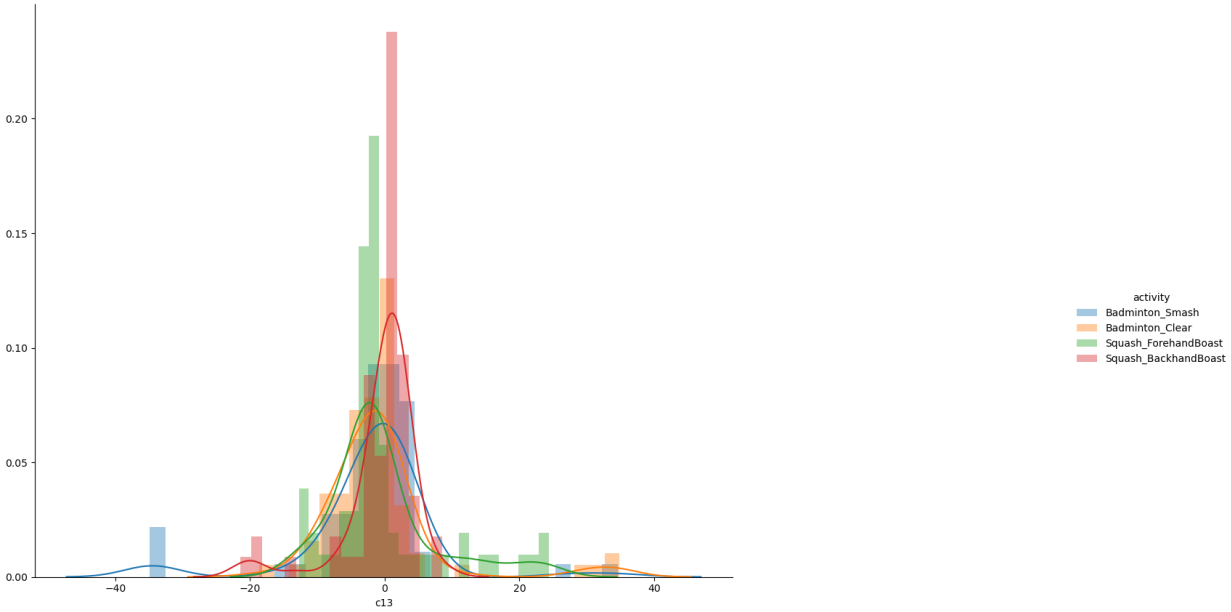
- Accelerație, axa X



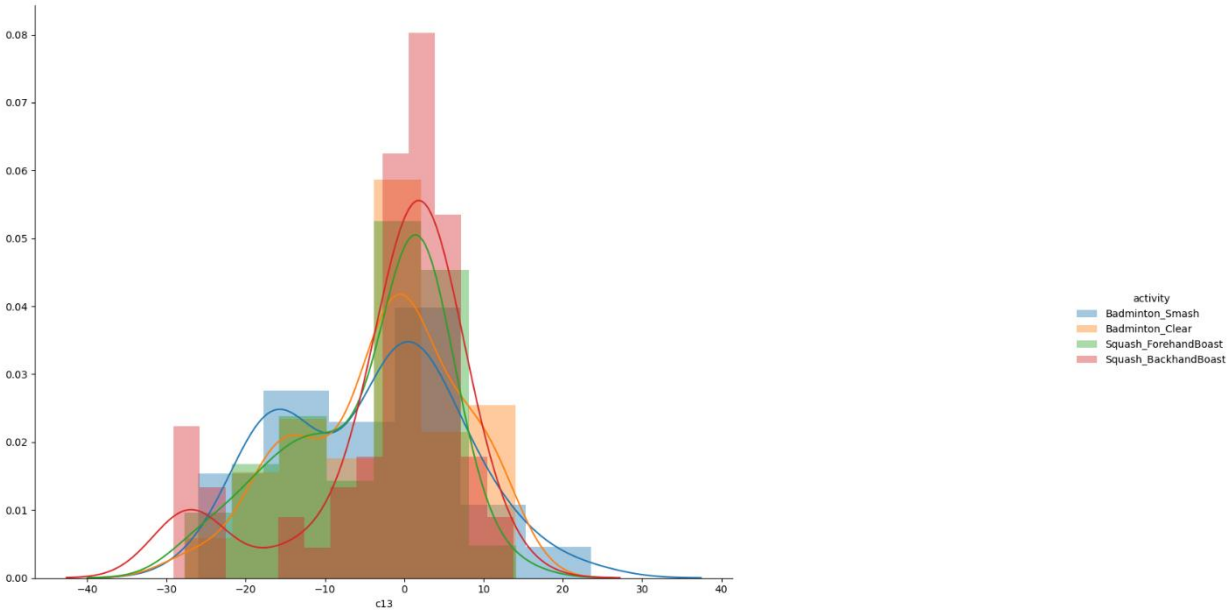
- Accelerație, axa Y



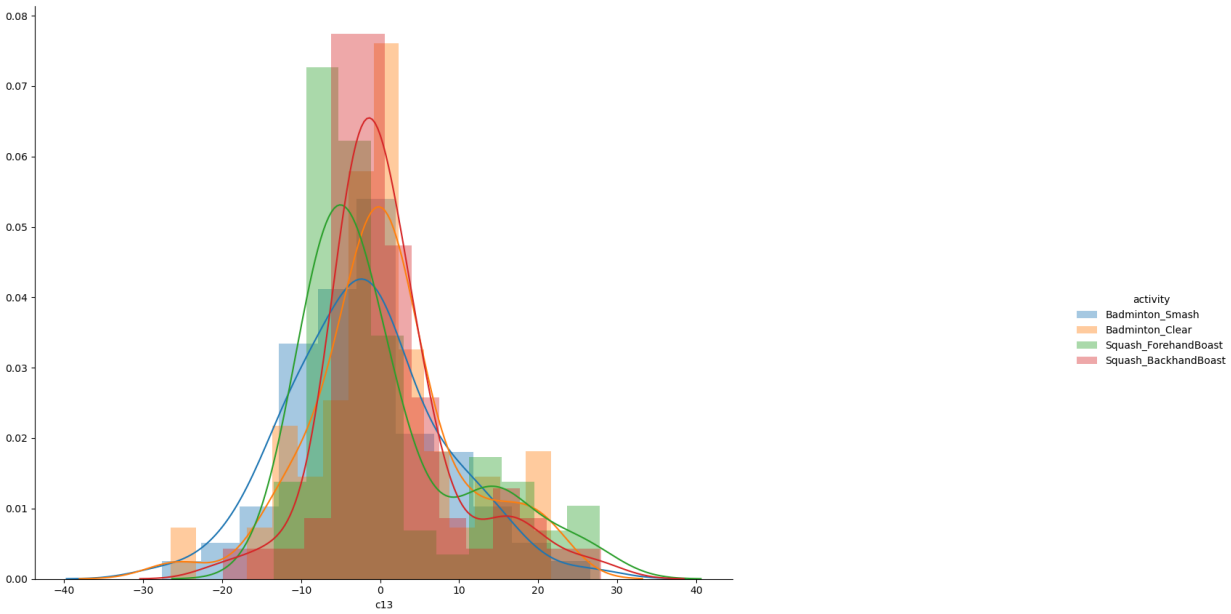
- Accelerație, axa Z



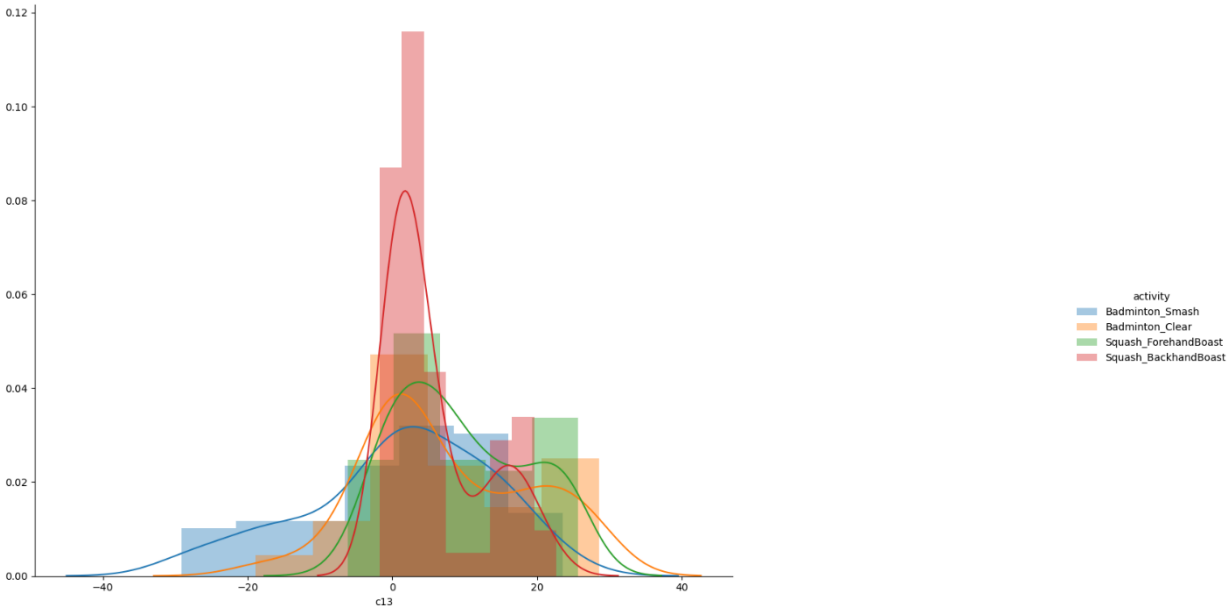
- Giroscop, axa X



- Giroscop, axa Y



- Giroscop, axa Z



Graficele de mai sus evidențiază o corelație între datele din fiecare clasă, ceea ce poate afecta acuratețea algoritmului. De asemenea, vârfurile graficelor celor patru clase se apropie, sugerând că valorile sunt în general în același interval, ceea ce indică faptul că standardizarea datelor poate nu este necesară.

3.2. Extragere manuală a atributelor și utilizarea algoritmilor clasici de Învățare Automată

1. Evaluarea algoritmilor direct pe datele de intrare

	RandomForest PTBDB		
Parametrii	Clasa / Metrica	0	1
n_estimators=50 max_depth=5 max_samples=0.5	Precision	0.818182	0.897365
	Recall	0.733813	0.934521
	F1-score	0.773704	0.915566
	mean accuracy	0.873929727	
	std accuracy	0.013609658	
n_estimators=50 max_depth=5 max_samples=0.7	Precision	0.789677	0.896067
	Recall	0.733813	0.921521
	F1-score	0.760721	0.908616
	mean accuracy	0.867060412	
	std accuracy	0.014720602	
n_estimators=50 max_depth=5 max_samples=0.9	Precision	0.783069	0.887703
	Recall	0.709832	0.92104
	F1-score	0.744654	0.904064
	mean accuracy	0.871524229	
	std accuracy	0.005835677	
n_estimators=50 max_depth=10 max_samples=0.5	Precision	0.921836	0.95677
	Recall	0.890887	0.969668
	F1-score	0.906098	0.963176
	mean accuracy	0.91721455	
	std accuracy	0.013475087	
n_estimators=50 max_depth=10 max_samples=0.7	Precision	0.92689	0.959125
	Recall	0.896882	0.971594
	F1-score	0.911639	0.965319
	mean accuracy	0.920649797	
	std accuracy	0.008813956	
n_estimators=50 max_depth=10 max_samples=0.9	Precision	0.911765	0.957041
	Recall	0.892086	0.965335
	F1-score	0.901818	0.96117
	mean accuracy	0.928205219	
	std accuracy	0.008905277	
n_estimators=50 max_depth=20 max_samples=0.5	Precision	0.959339	0.962806
	Recall	0.905276	0.984593
	F1-score	0.931524	0.973578
	mean accuracy	0.927863934	
	std accuracy	0.012457435	
n_estimators=50 max_depth=20 max_samples=0.7	Precision	0.973111	0.965258
	Recall	0.911271	0.989889
	F1-score	0.941176	0.977419
	mean accuracy	0.93232952	
	std accuracy	0.013223845	
n_estimators=50 max_depth=20 max_samples=0.9	Precision	0.970812	0.967499
	Recall	0.917266	0.988926
	F1-score	0.94328	0.978095
	mean accuracy	0.931637519	
	std accuracy	0.01016619	

n_estimators=100 max_depth=5 max_samples=0.5	Precision	0.802083	0.898273
	Recall	0.738609	0.926818
	F1-score	0.769039	0.912322
	mean accuracy	0.869119909	
	std accuracy	0.011331915	
n_estimators=100 max_depth=5 max_samples=0.7	Precision	0.805263	0.896792
	Recall	0.733813	0.928743
	F1-score	0.76788	0.912488
	mean accuracy	0.873589621	
	std accuracy	0.014698363	
n_estimators=100 max_depth=5 max_samples=0.9	Precision	0.8	0.896552
	Recall	0.733813	0.926336
	F1-score	0.765478	0.911201
	mean accuracy	0.870839891	
	std accuracy	0.014665948	
n_estimators=100 max_depth=10 max_samples=0.5	Precision	0.93	0.957366
	Recall	0.892086	0.973038
	F1-score	0.910649	0.965138
	mean accuracy	0.914809051	
	std accuracy	0.010024363	
n_estimators=100 max_depth=10 max_samples=0.7	Precision	0.915025	0.956646
	Recall	0.890887	0.966779
	F1-score	0.902795	0.961686
	mean accuracy	0.925804436	
	std accuracy	0.01435556	
n_estimators=100 max_depth=10 max_samples=0.9	Precision	0.926108	0.960934
	Recall	0.901679	0.971112
	F1-score	0.91373	0.965996
	mean accuracy	0.922366831	
	std accuracy	0.014247527	
n_estimators=100 max_depth=20 max_samples=0.5	Precision	0.967908	0.962477
	Recall	0.904077	0.987963
	F1-score	0.934904	0.975053
	mean accuracy	0.928894272	
	std accuracy	0.011807478	
n_estimators=100 max_depth=20 max_samples=0.7	Precision	0.973485	0.970269
	Recall	0.92446	0.989889
	F1-score	0.948339	0.979981
	mean accuracy	0.937138158	
	std accuracy	0.010878504	
n_estimators=100 max_depth=20 max_samples=0.9	Precision	0.968514	0.969296
	Recall	0.922062	0.987963
	F1-score	0.944717	0.978541
	mean accuracy	0.937827212	
	std accuracy	0.009267089	

n_estimators=200 max_depth=5 max_samples=0.5	Precision	0.800261	0.89697
	Recall	0.735012	0.926336
	F1-score	0.76625	0.911416
	mean accuracy	0.866027715	
	std accuracy	0.013548644	
n_estimators=200 max_depth=5 max_samples=0.7	Precision	0.797176	0.900094
	Recall	0.744604	0.923929
	F1-score	0.769994	0.911856
	mean accuracy	0.873587263	
	std accuracy	0.009875143	
n_estimators=200 max_depth=5 max_samples=0.9	Precision	0.800525	0.895765
	Recall	0.731415	0.926818
	F1-score	0.764411	0.911027
	mean accuracy	0.872556336	
	std accuracy	0.013808403	
n_estimators=200 max_depth=10 max_samples=0.5	Precision	0.917386	0.957143
	Recall	0.892086	0.967742
	F1-score	0.904559	0.962413
	mean accuracy	0.922711653	
	std accuracy	0.012962026	
n_estimators=200 max_depth=10 max_samples=0.7	Precision	0.921569	0.960859
	Recall	0.901679	0.969186
	F1-score	0.911515	0.965005
	mean accuracy	0.923398938	
	std accuracy	0.010693839	
n_estimators=200 max_depth=10 max_samples=0.9	Precision	0.922983	0.962255
	Recall	0.905276	0.969668
	F1-score	0.914044	0.965947
	mean accuracy	0.929236147	
	std accuracy	0.012057355	
n_estimators=200 max_depth=20 max_samples=0.5	Precision	0.965693	0.96516
	Recall	0.911271	0.987
	F1-score	0.937693	0.975958
	mean accuracy	0.931297413	
	std accuracy	0.008037153	
n_estimators=200 max_depth=20 max_samples=0.7	Precision	0.972046	0.967514
	Recall	0.917266	0.989408
	F1-score	0.943862	0.978338
	mean accuracy	0.936452642	
	std accuracy	0.014350315	
n_estimators=200 max_depth=20 max_samples=0.9	Precision	0.974716	0.970283
	Recall	0.92446	0.990371
	F1-score	0.948923	0.980224
	mean accuracy	0.943667368	
	std accuracy	0.012377781	

RandomForest PTBDB
pe bază de Grid Search
cu Cross Validation

Accuracy: 0.9704568876674682

Best parameters:
max_depth=20
max_samples=0.9
n_estimators=200

Metrica \ Clasa	0	1
Precision	0.973418	0.969354
Recall	0.922062	0.989889
F1-score	0.947044	0.979514
Support	834	2077
mean	209.2106	519.984689
std	416.5262	1038.010207

GradientBoosted Trees PTBDB
pe bază de Grid Search
cu Cross Validation

Accuracy: 0.9807626245276537

Best parameters:
learning_rate=0.1
max_depth=5
n_estimators=1000

Metrica \ Clasa	0	1
Precision	0.98	0.98
Recall	0.96	0.99
F1-score	0.97	0.99
mean accuracy	0.98	
std accuracy	0.00357242	

SVM PTBDB
pe bază de Grid Search
cu Cross Validation

Accuracy: 0.9587770525592579

Best parameters:
C = 10
kernel=rbf

Metrica \ Clasa	0	1
Precision	0.93	0.97
Recall	0.93	0.97
F1-score	0.93	0.97
mean accuracy	0.951980187	
std accuracy	0.003697583	

Analizând metricile obținute anterior, se poate constata că hiperparametrii au o importanță crucială, însă cel mai important este echilibrul între performanță și eficiență. Există situații în care este mai avantajos să utilizăm hiperparametrii care conduc la un efort computational mai mic, fără a afecta în mod semnificativ calitatea predicțiilor algoritmului.

2. Evaluarea algoritmilor pe atributele statistice extrase

SVM ATRIBUTE PTBDB pe bază de Grid Search cu Cross Validation		
Accuracy: 0.8474750944692545		
Best parameters: C = 10 kernel=rbf		
Clasa / Metrica	0	1
Precision	0.81	0.86
Recall	0.62	0.94
F1-score	0.7	0.9
mean accuracy	0.851473576	
std accuracy	0.005927256	

- Matrice de corelație pentru dataframe-ul format din atributele extrase din ptbdb_train:

	0	1	2	3	...	12	13	14	187
0	1.000000	0.731707	0.726746	NaN	...	0.946862	-0.893571	-0.758214	0.113723
1	0.731707	1.000000	0.981616	NaN	...	0.852746	-0.734397	-0.758698	0.248879
2	0.726746	0.981616	1.000000	NaN	...	0.826718	-0.779682	-0.799376	0.259635
3	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
5	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
6	0.815583	0.389672	0.401617	NaN	...	0.694819	-0.681796	-0.515256	0.015326
7	0.375407	0.347239	0.398489	NaN	...	0.336642	-0.414574	-0.398883	0.082552
8	0.673331	0.919582	0.960038	NaN	...	0.756770	-0.744926	-0.761920	0.243985
9	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
10	0.161986	-0.388958	-0.429006	NaN	...	-0.029356	0.087205	0.224215	-0.309546
0	0.658105	0.233019	0.301004	NaN	...	0.479998	-0.691544	-0.547908	0.014588
11	0.168943	-0.220188	-0.230975	NaN	...	0.055412	0.008998	0.136738	0.229781
12	0.946862	0.852746	0.826718	NaN	...	1.000000	-0.806882	-0.689852	0.186023
13	-0.893571	-0.734397	-0.779682	NaN	...	-0.806882	1.000000	0.949687	-0.147827
14	-0.758214	-0.758698	-0.799376	NaN	...	-0.689852	0.949687	1.000000	-0.155834
187	0.113723	0.248879	0.259635	NaN	...	0.186023	-0.147827	-0.155834	1.000000

[17 rows x 17 columns]

În cazul matricei de mai sus, putem observa că valorile de pe diagonala principală (unde se află același atribut comparat cu sine însuși) sunt egale cu 1, ceea ce este normal. De asemenea, există celule unde valorile sunt NaN (not a number), ceea ce indică faptul că nu există date pentru acele perechi de atribute.

Dacă ne uităm la celelalte valori din matrice, putem observa că unele dintre ele sunt foarte apropiate de 1 sau -1, indicând o corelație puternică între acele perechi de atribute. În general, putem spune că perechile de atribute cu valori ridicate și semnificative în matricea de corelație sunt cele mai predictive.

Asemănător se întâmplă și pentru matricea de corelație pentru dataframe-ul format din atributele extrase din ptbdb_test.