

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE  
FAKULTA BIOMEDICÍNSKÉHO INŽENÝRSTVÍ  
Katedra biomedicínské techniky



Automatická klasifikace EEG segmentů metodou DBSCAN

Automatic classification of EEG segments using DBSCAN  
algorithm

Diplomová práce

Vedoucí práce: doc. Ing. Vladimír Krajča, CSc.

Student: Bc. Marek Piorecký

květen 2016

Katedra biomedicínské techniky

Akademický rok: 2015/2016

## Z a d á n í   d i p l o m o v é   p r á c e

Student: **Bc. Marek Piorecký**  
Studijní obor: Biomedicínský inženýr  
Téma: **Automatická klasifikace EEG segmentů metodou DBSCAN**  
Téma anglicky: Automatic classification of EEG segments using DBSCAN algorithm

### Z á s a d y   p r o   v y p r a c o v á n í :

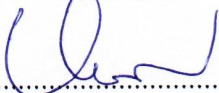
Jedním z problémů automatického zpracování EEG záznamu je klasifikace již segmentovaného záznamu. Použijte metodu DBSCAN pro klasifikaci dat na základě příznaků. V programovém prostředí MATLAB implementujte uživatelsky přívětivé prostředí pro načtení analyzovaných dat. Implementujte algoritmus DBSCAN pro automatickou klasifikaci EEG signálu na základě příznaků. Testujte algoritmus DBSCAN na simulovaných datech. Proveďte kvalitativní i kvantitativní zhodnocení použitého algoritmu. Aplikujte algoritmus na reálné EEG záznamy pacientů. Proveďte kvalitativní i kvantitativní zhodnocení použitého algoritmu. Porovnejte a statisticky vyhodnoťte rozdíl v účinnosti algoritmů DBSCAN a k-means u simulovaného i reálného EEG signálu. Analyzujte výpočetní náročnost obou algoritmů.


### Seznam odborné literatury:

- [1] Krajča V., Mohylová J., Číslicové zpracování neurofyziologických signálů, ed. Fakulta biomedicínského inženýrství, ČVUT Praha, 2011, ISBN 978-80-01-04721-7
- [2] Proakis J.G., Manolakis D.G., Digital Signal Processing, ed. 4th, Macmillan, NY, 2006, ISBN 978-0131873742
- [3] Khan, K., DBSCAN: Past, present and future, Applications of Digital Information and Web Technologies (ICADIWT), ročník 5, číslo 1, 2014, <http://80.ieeexplore.ieee.org/dialog/cvut.cz/xpl/abstractAuthors.jsp?arnumber=6814687&tag=1>

Vedoucí: doc. Ing. Vladimír Krajča, CSc.  
Konzultant: Ing. Václava Sedlmajerová

Zadání platné do: 20.08.2017

  
.....  
vedoucí katedry / pracoviště

  
.....  
děkan

V Kladně dne 20.01.2015

# Obsah

Seznam použitých symbolů a zkratek	4
Seznam tabulek	5
Seznam obrázků	7
Úvod	15
<b>1 Současný stav řešené problematiky</b>	<b>17</b>
1.1 K-means algoritmus . . . . .	17
1.2 Wave-Finder . . . . .	18
1.3 DBSCAN . . . . .	18
<b>2 Teoretická část</b>	<b>19</b>
2.1 EEG . . . . .	19
2.1.1 EEG signál . . . . .	19
2.1.2 EEG vyšetření . . . . .	23
2.2 Zpracování signálu . . . . .	25
2.2.1 Segmentace . . . . .	25
2.2.2 Příznaky . . . . .	27
2.2.3 Klasifikace . . . . .	31
2.3 K-means . . . . .	34
2.4 DBSCAN . . . . .	37
2.5 Porovnání metod k-means a DBSCAN . . . . .	41
<b>3 Metodika</b>	<b>42</b>
3.1 Data . . . . .	44
3.1.1 Třídy . . . . .	47
3.2 Načítání dat . . . . .	49
3.3 DBSCAN . . . . .	50
3.3.1 DMDBSCAN . . . . .	55

3.3.2	GRIDBSCAN . . . . .	59
3.4	Statistické zhodnocení . . . . .	64
3.4.1	Specificita . . . . .	66
3.4.2	Senzitivita . . . . .	67
3.4.3	Pozitivní prediktivní hodnota (PPV) . . . . .	67
3.5	Vyhodnocení 2D testovacích dat . . . . .	68
3.6	Vyhodnocení klasifikace EEG dat . . . . .	68
3.6.1	Vyhodnocení celých záznamů . . . . .	68
3.6.2	Výběr a vyhodnocení náhodných segmentů . . . . .	69
<b>4</b>	<b>Výsledky</b>	<b>71</b>
4.1	Klasifikace 2D - testovací data . . . . .	71
4.2	Klasifikace EEG - reálná data . . . . .	81
4.3	GUI . . . . .	92
<b>5</b>	<b>Diskuze</b>	<b>94</b>
5.1	Kvantitativní vyhodnocení . . . . .	94
5.2	Kvalitativní vyhodnocení . . . . .	95
5.2.1	2D testovací data . . . . .	96
5.2.2	Reálná EEG data . . . . .	98
<b>6</b>	<b>Závěr</b>	<b>106</b>
	<b>Reference</b>	<b>107</b>
<b>A</b>	<b>Příloha: Obsah CD</b>	<b>114</b>
<b>B</b>	<b>Příloha: Tabulky</b>	<b>115</b>
<b>C</b>	<b>Příloha: Publikovaný abstrakt</b>	<b>119</b>

## Seznam použitých symbolů a zkratek

23D	.....	23-dimensionální
EEG	.....	elektroencefalograf
WF	.....	Wave-Finder
DBSCAN	.....	Density-based spatial clustering of applications with noise
ROC	.....	receiver operating characteristic
PPV	.....	positive predictive values (pozitivní prediktivní hodnota)
FP	.....	False positive
TP	.....	True positive
FN	.....	False negative
TN	.....	True negative

## Seznam tabulek

1	Normované příznaky použité pro klasifikaci. . . . .	28
2	Porovnání vlastností DBSCAN vs. k-means. . . . .	41
3	Nastavení parametrů segmentace. . . . .	46
4	Kvantitativní zhodnocení data 1. . . . .	71
5	Kvalitativní analýza data 1. . . . .	72
6	Data 2. . . . .	73
7	Kvalitativní zhodnocení data 2. . . . .	74
8	Kvantitativní zhodnocení data 3. . . . .	75
9	Kvalitativní analýza data 3. . . . .	76
10	Kvantitativní analýza data 4. . . . .	77
11	Kvalitativní analýza data 4. . . . .	78
12	Kvantitativní analýza data 5. . . . .	78
13	Kvalitativní analýza data 5. . . . .	79
14	Celková kvalitativní analýza. . . . .	80
15	Kvantitativní vyhodnocení pro 15 pacientů. . . . .	82
16	Ukázkové výsledky analýzy pacienta 10. . . . .	83
17	Výsledná kvalitativní analýza GRIDBSCAN. . . . .	85
18	Výsledná kvalitativní analýza k-means. . . . .	85
19	Kvalitativní analýza pacient 1 k-means. . . . .	86
20	Kvalitativní analýza pacient 2 k-means. . . . .	87
21	Kvalitativní analýza pacient 3 k-means. . . . .	88
22	Kvalitativní analýza pacient 1 GRIDBSCAN. . . . .	88
23	Kvantitativní analýza pacient 2 GRIDBSCAN. . . . .	89
24	Kvalitativní analýza pacient 3 GRIDBSCAN. . . . .	90
25	Výsledná kvalitativní analýza k-means. . . . .	91
26	Výsledná analýza GRIDBSCAN. . . . .	91
27	Výsledky analýzy pacienta 1. . . . .	115
28	Výsledky analýzy pacienta 2. . . . .	115
29	Výsledky analýzy pacienta 3. . . . .	115

30	Výsledky analýzy pacienta 4. . . . .	116
31	Výsledky analýzy pacienta 5. . . . .	116
32	Výsledky analýzy pacienta 6. . . . .	116
33	Výsledky analýzy pacienta 7. . . . .	117
34	Výsledky analýzy pacienta 8. . . . .	117
35	Výsledky analýzy pacienta 9. . . . .	117
36	Výsledky analýzy pacienta 10. . . . .	118
37	Výsledky analýzy pacienta 11. . . . .	118
38	Výsledky analýzy pacienta 12. . . . .	118

# Seznam obrázků

1	Sequence diagram . . . . .	2
2	Časový průběh jednotlivých vln. . . . .	21
3	Komplex hrot-vlna. . . . .	22
4	Artefakty v EEG záznamu. . . . .	23
5	Zapojení elektrod. . . . .	25
6	Adaptivní segmentace. . . . .	26
7	Proces klasifikace pomocí k-means. . . . .	36
8	Schéma klasifikace pomocí k-means. . . . .	37
9	Proces klasifikace algoritmem DBSCAN. . . . .	39
10	Schéma klasifikace algoritmem DBSCAN. . . . .	40
11	Prostředí Wave-Finder. . . . .	43
12	Volba klinicky zajímavého úseku. . . . .	44
13	Ukázka 2D testovacích dat. . . . .	45
14	Segmenty fyziologické aktivity. . . . .	47
15	Segmenty EMG. . . . .	48
16	Segmenty epileptické aktivity. . . . .	48
17	Segmenty pomalých očních artefaktů. . . . .	48
18	Špatný kontakt elektrody. . . . .	49
19	Výběr dat. . . . .	50
20	Zjišťování ideálního počtu bodů v poloměru. . . . .	54
21	Křivka průměrů prvních $k$ hodnot. . . . .	56
22	Křivka $k$ sousedů při hodnotě parametru 3. . . . .	58
23	Křivka $k$ sousedů při hodnotě parametru 30. . . . .	58
24	Vykreslení 2 příznaků reálných EEG dat. . . . .	60
25	Posuv okna a překryv buněk. . . . .	61
26	Matice TP, TN, FP a FN. . . . .	65
27	Ukázka TP a FN segmentů. . . . .	66
28	Data 1. . . . .	72
29	Data 2. . . . .	74



30	Data 3. . . . .	76
31	Data 4. . . . .	77
32	Data 5. . . . .	79
33	Porovnání trendu rychlostí jednotlivých metod. . . . .	80
34	Porovnání výpočetní náročnosti. . . . .	82
35	FP segmenty ve třídě epileptických grafoelementů. . . . .	84
36	FP segmenty ve třídě pomalých očních artefaktů. . . . .	87
37	FP segmenty ve třídě fyziologické aktivity. . . . .	89
38	Třída EMG grafoelementů. . . . .	90
39	GUI 2D data. . . . .	92
40	GUI export. . . . .	93

# Abstrakt

Elektrickou aktivitu mozku zaznamenáváme pomocí EEG (elektroencefalografu). Nedílnou součástí vyšetření je detekce grafoelementů. Metody, které jsou založeny na matematickém principu klasifikace, je vždy nutné přizpůsobit stochastickému rázu EEG signálu. Hojně využívanou metodou je modifikovaný algoritmus k-means. Tento přístup ale skýtá omezení v prostorově prolnutých shlucích. U hustotně založeného algoritmu DBSCAN se tento problém neobjevuje. Zároveň nabízí algoritmus DBSCAN velké množství modifikací uzpůsobených pro více dimenzionální data.

Cílem diplomové práce je otestovat účinnost algoritmu DBSCAN na klasifikaci segmentů EEG signálu na základě 23 vypočtených příznaků, případně navrhnout vhodnou adaptaci algoritmu pro EEG signál.

Zpracovávaná data byla naměřena v Nemocnici Na Bulovce na 15 pacientech, kterým bylo indikováno vyšetření na základě podezření na nemoc epilepsii. Pacienti byli muži i ženy ve věku mezi 26 – 60 roky. Délka záznamu se pohybuje od jednotek do desítek minut.

Klasifikovány jsou celé záznamy (všechny segmenty z 19 kanálů). Z vyhodnocených 12 záznamů jsou vybírány náhodné segmenty, u kterých 2 nezávislí odborníci validují příslušnost k dané třídě. 3 záznamy jsou vyhodnoceny celé. V záznamu rozlišujeme třídy odpovídající epileptické, svalové a oční aktivitě, fyziologickou aktivitu a pulzní artefakty. Na základě klinického vyhodnocení byla provedena ROC analýza.

Z výsledků vyplývá, že DBSCAN není vhodný pro klasifikaci EEG záznamů, neboť není schopen správně rozdělit více jak 2 třídy grafoelementů. Modifikovaný GRIDBSCAN, který vychází z algoritmu DBSCAN, dělí 23D prostor pomocí buněk adaptivních rozměrů. Tento algoritmus má lepší senzitivitu u segmentů fyziologické aktivity, než algoritmus k-means. GRIDBSCAN zároveň klasifikuje epileptické grafoelementy do tříd s vysokou

homogenitou (nad 0,8). Nedostatky algoritmu se projevují ve špatném zařazení pomalých očních artefaktů a vysoké časové náročnosti (jednotky minut).

K zefektivnění klasifikace by přispěl optimální (individuální) výběr příznaků pro danou metodu. Modifikovaný DBSCAN lze využít k hodnocení záznamů EEG, ale stále je nutná intervence lékaře. Díky jeho vysoké výpočetní náročnosti a na vybraných příznacích prokázané téměř shodné úspěšnosti klasifikace s algoritmem k-means není vhodné v klinické praxi algoritmus k-means nahrazovat hustotním algoritmem GRIDBSCAN.

## **Klíčová slova**

EEG, klasifikace, DBSCAN, K-means, MATLAB.

# Abstract

Electroencephalography (EEG) is an electrophysiological monitoring method used for recording the electrical activity of the brain. A detection of graph elements is integral to the examination. It is essential to adapt the methods that are based on mathematical principles of classification for a stochastic character of the EEG signal. One of the most commonly used methods is modified algorithm k-means, nevertheless this method is limited by spatially entwined clusters. We do not deal with this kind of problem when using a density-based algorithm DBSCAN. Apart from that algorithm DBSCAN also provides a wide range of modifications adapted for multi-dimensional data.

The main aim of this dissertation thesis is to test the effectiveness of algorithm DBSCAN which is used for classification of the EEG signal segments. This method is based on 23 calculated symptoms. Alternatively we can propose a suitable adaptation of algorithm for the EEG signal.

The data was analyzed on the basis of taking the measurements of 15 patients who were examined in order to disprove the possible diagnosis of epilepsy. The data were measured in the hospital Nemocnice Na Bulovce. The patients were between the ages 26-60, both male and female. The length of the measurement varies from one to tens of minutes.

Records were classified whole (all segments of 19 channels). There are 12 evaluated records from which we pick random segments. The segment's competency to the specific class is validated by two independent experts. We evaluate the whole part of 3 signals. Our records are divided into separated classes corresponding with epileptic, muscular, ocular activity and with physiologic activity and pulse artifacts. We did the ROC analysis based on the clinical evaluation.

Due to the results we can see that DBSCAN is not suitable for the EEG classification of the records as DBSCAN cannot separate more than 2 classes of graph elements. Modified

GRIDBSCAN, which is based on the DBSCAN algorithm, separates 23D space using the cells with adaptable size. This algorithm is more sensible as for the segments of physiologic activity than the k-means algorithm. GRIDBSCAN also divides epileptic graph elements into classes with high homogeneity (above 0,8). The imperfection of the algorithm is visible in incorrect and time-consuming classification of slow ocular artifacts.

The classification could be more effective if we individually chose the symptoms for the specific method. Modified DBSCAN can be used for evaluation of the EEG records, nevertheless the doctor's intervention is still required. Due to the computational demand and verifiably high success of the k-means algorithm classification is not convenient to replace the k-means algorithm by a density-based algorithm in clinical practice.

## **Keywords**

EEG, classification, DBSCAN, K-means, MATLAB.

## Poděkování

Děkuji vedoucímu mé diplomové práce, doc. Ing. Vladimíru Krajčovi, CSc., a primáři MUDr. Ing. Svojmilu Petránkovi, CSc. MBA za jejich odbornou pomoc. Děkuji své konzultantce Ing. Václavě Sedlmajerové za věcné připomínky. Poděkování patří také „EEG týmu“ za společné náhledy a komentáře. V neposlední řadě děkuji také doc. Ing. Janě Vránové, CSc. za konzultace k statistickému zpracování dat.

## Prohlášení

Prohlašuji, že jsem diplomovou práci s názvem

*Automatická klasifikace EEG segmentů metodou DBSCAN*

vypracoval samostatně a použil k tomu úplný výčet citací použitých pramenů, které uvádím v seznamu přiloženém k diplomové práci.

Nemám závažný důvod proti užití tohoto školního díla ve smyslu §60 Zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

V ..... dne .....

.....

Podpis

# Úvod

Mozek můžeme chápat jako řídicí počítač lidského těla. Řídí a ovlivňuje ostatní části organismu. Proto se vědci a lékaři domnívají, že jeho pochopením se otevřou dveře k zpřesnění diagnostiky a zkvalitnění léčby. Dosud nejsou vysvětleny a popsány všechny elektrické interakce probíhající mezi neurony. Elektrický signál mozku, který jsme pomocí elektroencefalografu (EEG) schopni zaznamenat, je chaotický a působí velmi náhodně. Pro usnadnění a urychlení vyhodnocení těchto záznamů vznikla řada algoritmů sloužících k jejich segmentaci a klasifikaci. S rozvojem výpočetní techniky je možné aplikovat na signály výpočetně náročnější algoritmy. Tímto získává zpracování biosignálů více vypovídajících výsledků a je možné jeho širší využití v klinické praxi.

Cílem diplomové práce je implementovat metodiku pro automatickou klasifikaci segmentů EEG záznamu na základě příznaků charakteristických pro daných segment. Diplomová práce pojednává o možnosti využití algoritmu na bázi hustoty. Současné softwarové řešení v programu Wave-Finder (WF) využívá algoritmus k-means. Cílem práce je srovnat metodu k-means s algoritmem DBSCAN. Hustotní metoda by měla dosáhnout lepších výsledků u dat, která jsou v prostoru prolnta a nelze je oddělit přímkou. Jedním z výstupních parametrů algoritmu DBSCAN je i počet shluků. Vzhledem k tomu, že u metody k-means je počet shluků parametrem vstupním, má nově testovaná metoda přispět k zjištění ideálního počtu shluků.

Výstup je současně zapracován jako samostatně fungující modul v rámci existujícího softwarového řešení v programovacím prostředí MATLAB. Výsledky jsou exportovatelné v binárním i ASCII kódování.

Ke klasifikaci využíváme reálná data pacientů Nemocnice na Bulovce, u nichž bylo vyšetření indikováno z důvodu podezření na epilepsii. Cílem klasifikace je určení místa v záznamu, kde se vyskytují artefakty, a rozdělení celého signálu do tříd dle podobnosti jednotlivých segmentů.



Komplexní algoritmus (přípravení EEG dat, klasifikace a export) by měl v klinické praxi přispět ke zpřesnění diagnostiky, měl by být plně automatizovaný a účinnější, než dosud využívaná klasifikace pomocí metody k-means, u které je nutná interakce s uživatelem.

První část práce obsahuje popis EEG signálu, jeho genezi a měření. Jsou zde popsány základní principy vyhodnocování signálu a určení příznaků. Následující kapitoly nastiňují metody klasifikace, s bližším zaměřením na metodu DBSCAN a její varianty. V dalších kapitolách je popsána metodika zpracování a klasifikace testovacích a reálných dat. V neposlední řadě práce obsahuje výsledky použití nové metodiky v porovnání se stávajícím algoritmem a vyhodnocení poznatků.

# 1 Současný stav řešené problematiky

Algoritmy všech kategorií se snaží o co nejvíce přesné zařazení testovaných dat do patřičných shluků. Velké množství zpracovávaných dat přináší spoustu problémů. Samotný algoritmus není neúčinnější, neboť nepředstavuje komplexní řešení pro všechny druhy dat.

Technicky je za dobrý algoritmus považován takový, který splňuje tyto požadavky:

- minimální požadavky znalosti domény pro určení hodnot všech vstupních parametrů (podstatné hlavně u velkých souborů dat),
- objev libovolných tvarů shluků,
- dobrá účinnost na velkých datových souborech.

## 1.1 K-means algoritmus

V současné době se nejčastěji používá k analýze EEG záznamu metoda k-means, fuzzy c-means, popřípadě kombinace těchto metod s dalšími. [1, 2]

Metoda k-means se u EEG záznamu používá například v případech, kdy chceme detekovat choroby či analyzovat spánek. Často bývá k-means implementován spolu s neuronovými sítěmi. [3]

V klinické praxi se k-means používá k identifikování epileptických gafelementů. V rámci klasifikace bývá často k-means spojeno spolu s dalšími metodami, matematickými operacemi, aby bylo dosaženo co nejlepší klasifikace. [2]

Software Wave-Finder, jehož autorem je doc. Ing. Vladimír Krajča, CSc., metodou k-means klasifikuje segmenty EEG záznamu do 7 tříd [4]. Jsou zde rozdělovány šumové gafelementy, epileptická aktivita, pohybové artefakty a fyziologické průběhy.

## 1.2 Wave-Finder

Program Wave-Finder (WF) nabízí komplexní klinické zpracování EEG signálů. Nabízí klasifikaci pomocí neuronových sítí, kde je nutný ruční výběr grafoelementů pro naučení systému. Další možností je klasifikace metodou k-means či fuzzy c-means. [4] Software je využíván například Nemocnicí Na Bulovce, Národním ústavem duševního zdraví (NUDZ) a Psychiatrickým centrem v Bohnicích (PCP).

K-means nepřináší zcela přesné výsledky klasifikace. Neboť nevýhodou WF je nutnost odhadnutí počtu tříd uživatelem. Experimentálně bylo zjištěno, že se nejlepšího rozdělení dosáhne při vstupním parametru 7 shluků. Přesto v jednotlivých třídách nacházíme špatně zařazené segmenty. Některé třídy algoritmus mylně dělí na několik částí.

## 1.3 DBSCAN

Metoda DBSCAN se na EEG záznam v běžné praxi nepoužívá. Jedná se o starší algoritmus, který účinně odděluje prolnuté shluky a detekuje šum. Na konferenci „Knowledge Discovery and Data mining“ (KDD) 2014 byl algoritmus DBSCAN oceněn za významný přínos při extrakci dat v posledním desetiletí, neboť z něj vychází řada modifikací využívaných v současných softwarových aplikacích [5].

Jeho přínosem by mohlo být zlepšení klasifikace a získání počtu tříd jako výstupního parametru.

## 2 Teoretická část

### 2.1 EEG

Živé organismy generují biosignály, které slouží jako nosiče informací. Aktivita mozku se projevuje změnami elektrického potenciálu, které jsou měřitelné na povrchu lebky. Elektroencefalogram je projev sumace těchto neuronových potenciálů. [6]

Historie elektroencefalografu sahá do počátku 20. století, kdy byl sestrojen první přístroj. Od té doby je EEG využíváno při diagnostice a výzkumné činnosti. Při vyšetření se zjišťuje přítomnost chorob a poškození CNS, jakými jsou například: epilepsie, demence a Alzheimerova nemoc.

EEG je kvazistacionární (po částech stacionární) a stochastický signál. Jeho frekvenční i amplitudové charakteristiky se mění v čase. Charakteristiky jednotlivých segmentů (homogenní úseky záznamu) odpovídají fyziologickým dějům (pohyb, zavření očí, sluchové vjemy apod.). Amplitudová charakteristika má rozsah od 2 do 300  $\mu\text{V}$ , přičemž jako fyziologické hodnoty zdravého pacienta počítáme rozsah do 100  $\mu\text{V}$ . Frekvenční spektrum je do 100 Hz, ale obvykle se pohybujeme v rozmezí od 0,5 do 30 Hz. [7]

#### 2.1.1 EEG signál

Neurony, stavební jednotky mozku, mají elektricky nabitou membránu (50 - 70  $\mu\text{V}$ ). Neurony na rozdíl od jiných buněk nemetabolizují všechny látky, ale zpracovávají pouze glukózu a kyslík. Elektrický potenciál vzniká právě díky energii z oxydativní fosforilace glukózy. Napětí na vnitřní straně membrány je u normálního stavu negativní, na vnější pozitivní. Hodnota potenciálu neustále kolísá. Přesáhne-li hladinu 30 - 40  $\mu\text{V}$ , nastane depolarizace a neuron vyšle vzruch. Následuje inhibiční fáze (hyperpolarizace), kde dochází ke kompenzaci změny napětí membrány a návratu potenciálů do původního stavu. Neuron zde není schopen vysílat další vzruchy. [8, 9]

Neuron pomocí dendritů a těla buňky přijímá a zpracovává signál, který po neuritu (axonu) posílá k dalším neuronům. Rychlost šíření akčního potenciálu závisí na myelinizaci vláken. Více myelinizovaná vlákna vedou vzruchy rychlostí až  $120 \text{ m}\cdot\text{s}^{-1}$ . [6, 9]

Výsledný EEG signál je synchronizací elektrických nábojů membrán neuronů. Vzniká aktivitou neuronů kortexu a thalamu, který má funkci generátoru rytmů.

### **EEG elementy**

Základní EEG aktivita je rozdělena do čtyř pásem [10, 6]:

#### **Alfa rytmus**

Frekvence alfa pásma je v rozmezí 8 - 13 Hz. Jedná se o nejvíce markantní složku EEG signálu. Nejvíce je patrná v týlním laloku. Amplituda této elektrické aktivity se pohybuje v rozmezí 30 - 80  $\mu\text{V}$ . Alfa aktivita je fyziologickým příznakem, který se projevuje u vyzrálého mozku při zavřených očích.

#### **Beta rytmus**

Frekvence beta pásma se pohybuje v rozmezí 13 - 30 Hz, jedná se o nejrychlejší změnu aktivity. Nejlépe měřitelná aktivita je ve frontální oblasti. Amplituda měřeného signálu se nachází mezi 10 - 30  $\mu\text{V}$ . Beta aktivita se nemění otevřením očí ani vědomými pohyby. Naopak změní se při změně myšlenkových stavů (náročnější logické operace).

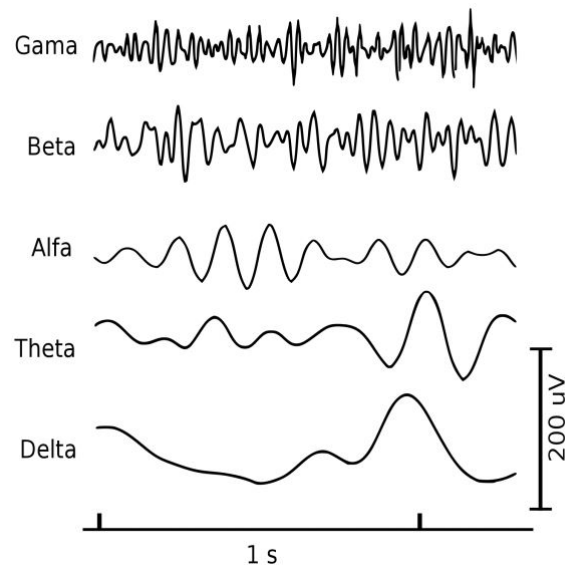
#### **Theta rytmus**

Frekvenci theta pásma nacházíme mezi 4 - 8 Hz. Nejdominantnější je v temporální oblasti. Amplituda je při normálním fyziologickém stavu do 30  $\mu\text{V}$ . Tato aktivita se vyskytuje ve spojení s učením a paměťovými procesy.

## Delta rytmus

Do delta pásma řadíme frekvence mezi 0,5 - 4 Hz. U dospělých osob se jedná o patologický nález v EEG záznamu. Často indikuje nádory. U dětí se fyziologicky vyskytuje v hlubokém spánku.

Na obrázku č. 2 jsou zobrazeny časové průběhy alfa, beta, gama, delta a theta rytmů.



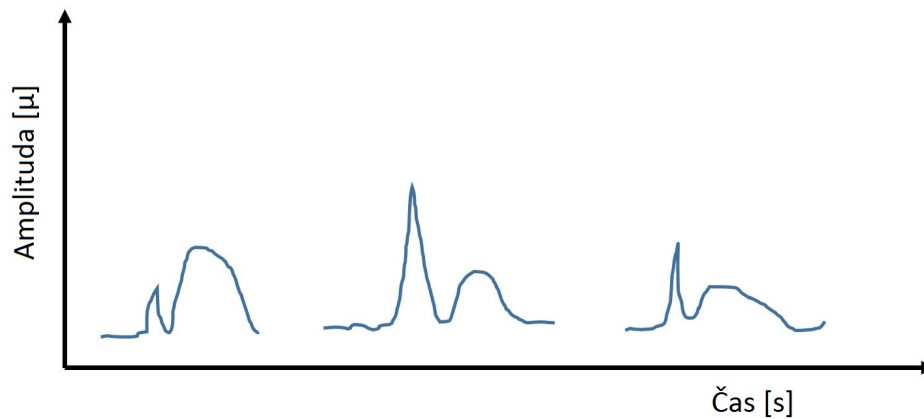
Obrázek 1: Časový průběh jednotlivých vln, převzato z: [6, 11], upraveno: Martin Jílek.

Mezi další (již patologické) vlny signálu patří například:

### Hrot a komplex hrot-vlna

Hrot neboli spike je ostrá trojúhelníkovitá vlna s vysokou amplitudou (i nad 100  $\mu\text{V}$ ). Tento grafoelement se často vyskytuje v komplexech tzv. „multiple spike komplex“, což je těsné spojení několika ostrých vln za sebou. Ve spojení hrot-vlna, pokud se tyto komplexy nacházejí napříč kanály, indikují epileptickou aktivitu. Amplituda se pohybuje nad 75  $\mu\text{V}$ . Epileptické záchvaty se vyskytují v určitou denní dobu, nemusejí se tedy zobrazit

při krátkém EEG vyšetření. Pro vyvolání epileptické aktivity můžeme použít například fotostimulací. [12, 13]



Obrázek 2: Komplex hrot-vlna v různých podobách, převzato z: [12], upraveno.

### Artefakty

Artefakty mohou mít fyzikální nebo biologický původ. Mezi fyzikální se řadí síťový brum, impulzní rušení nebo elektrostatické potenciály. Nedostatečně vodivě spojené elektrody se v záznamu projevují fluktuacemi, které se pomalu vrací k základní linii. Podobné artefakty mohou způsobovat i polámané dráty. [6]

Biologické artefakty mají pestřejší a zároveň proměnlivější charakter. Jsou to například změny aktivity v důsledku pohybu očí, elektrická změna potenciálu následkem srdeční aktivity či pohybem svalů. EEG křivka je výrazně rušena třesem víček, proto si je pacient často přidržuje. Pocení způsobuje změnu kožního potenciálu. Artefakt v tomto případě má vysokou amplitudu s pomalým průběhem. Některé, více vzácné křivky, mohou způsobit kardiostimulátory či různé druhy kovů (zubní plomby). [6]



Obrázek 3: Artefakty, které můžeme v EEG záznamu pozorovat. Převzato z [10], upraveno.

### 2.1.2 EEG vyšetření

V klinické praxi se jedná o neinvazivní vyšetření, které obvykle trvá půl hodiny, ale u některých chorob je třeba dlouhodobějšího pozorování (např. u epilepsie). Přístroj pro záznam elektrické aktivity se nazývá elektroencefalograf. Jeho důležitým úkolem je odfiltrování šumu a zesílení elektropotenciálů.

Vyšetření je mnohdy doplněno záznamem dalších signálů: EKG, EMG, EOG a další.

Elektrická aktivita je snímána pomocí elektrod. Počet snímacích elektrod koreluje s počtem kanálů. Jedná se tedy o záznam více kanálů (v běžné praxi okolo 18 - 20 kanálů, maximálně je možno sledovat až 256 kanálů). Aktivita má velmi nízkou amplitudu, proto následuje mnohonásobné zesílení a poté převedení do digitální podoby a vykreslení výsledné křivky. Tvar a charakter křivky je odrazem aktuální aktivity mozku. EEG signál má tedy nepravidelný průběh, některé biologické procesy mají ale typickou odezvu (charakteristické znaky v záznamu). [14, 15]



Jednotlivé artefakty lze v EEG záznamu detekovat vizuálně (lékař), při 24 hodinovém záznamu je takovýto postup pro každého pacienta zvláště časově extrémně náročný.

### Elektrody

Elektrody pro snímání EEG jsou z AgCl. Na hlavě jsou rozmístěny v systému 10-20 (desetdvacet). Obvod lebky je rozdělen na úseky po 10 % a 20 %. Rozměření ve dvou zbývajících kolmých rovinách je analogické. U EEG se využívá unipolárního i bipolárního zapojení. Unipolární zapojení znamená, že snímaná aktivita je rovna napětí mezi aktivní a referenční elektrodou (či svorkou). Bipolární zapojení znamená měření napětí mezi dvěma aktivními elektrodami.

Na obrázku níže jsou označeny elektrody, které jsou ve svých pozicích kotveny pomocí gumové čepice, jejíž jsou součástí. Elektrody jsou označeny písmeny a pořadovými čísly.

A – ear lobe (ušní – referenční)

C – central (centrální – na vrcholu lebky)

P – parietal (parietální – temenní)

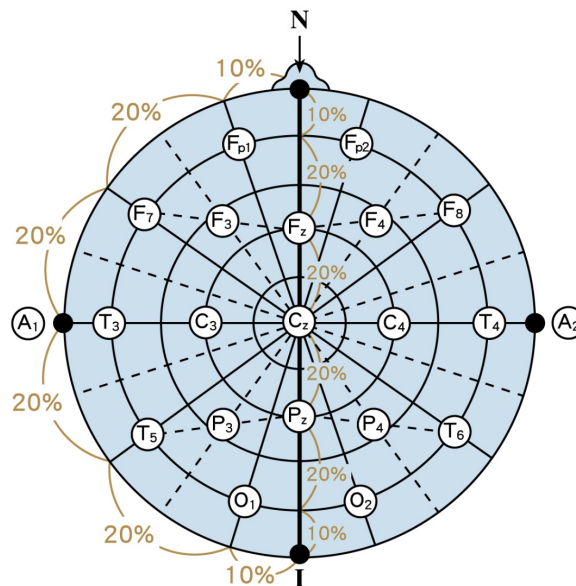
Fp – frontopolar (frontopolární – přední kolem pomyslného pólu)

F – frontal (frontální – přední)

O – occipital (okcipitální – týlní)

T – temporal (temporální – spánková)

Lichá čísla označují snímaná místa nad levou hemisférou, sudá čísla nad hemisférou pravou.



Obrázek 4: Zapojení elektrod v systému 10-20 [16].

## 2.2 Zpracování signálu

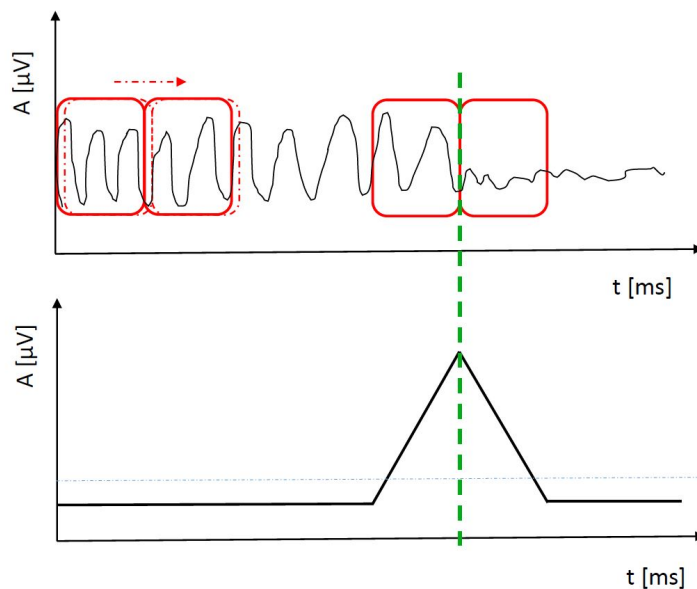
Celé předzpracování EEG signálu probíhá analogově (zesílení, filtrace). V digitální podobě v počítači se zpracovává již samotný záznam EEG. Záznam působí velmi chaoticky a je v něm zahrnuto i biologické rušení. Před samotnou klasifikací je potřeba provést segmentaci a výpočet příznaků. Oba kroky zásadně ovlivňují klasifikaci, ať už je prováděna jakoukoli klasifikační metodou.

### 2.2.1 Segmentace

Segmentace rozdělí záznam na úseky, ve kterých má křivka podobný charakter. Jsou dvě možnosti segmentace EEG záznamu:

**Konstantní** – fixní délka segmentu, každý úsek záznamu obsahuje stejný počet vzorků. Nevýhoda této segmentace tkví v tom, že mnohdy rozdělíme artefakt, vlnu, v polovině.

**Adaptivní** – různé délky segmentů, kdy je záznam rozdělen na díly, které mají průběh záznamu s podobným charakterem, jak je naznačeno na obrázku č. 6 níže. Tento druh segmentace je výrazně lepší pro následnou klasifikaci, neboť do výpočtu příznaků není zanesena chyba - nedochází k přerušení záznamu, který patří do jedné skupiny (například rozdělení hrotu a vlny). [17]



Obrázek 5: Adaptivní segmentace EEG záznamu, popis pod obrázkem. Převzato z: [7], upraveno.

Jak je patrné z obrázku č. 6, záznam je projížděn pomocí dvou spojených plovoucích oken (červeně). V každém okně v daném okamžiku (v dané pozici) jsou zjištěny charakteristiky záznamu a udělán rozdíl mezi oběma okny (spodní graf). Pokud okna projedou celý záznam, získáme křivku, která charakterizuje stacionaritu/nestacionaritu. V této křivce hledáme lokální maxima, která se nacházejí nad stanovenou mezí pro segmentaci (bledě modrá). Lokální maxima nám udávají pozice hranic segmentů (zeleně).

### 2.2.2 Příznaky

Příznak je číselná hodnota charakterizující vlastnosti záznamu v daném segmentu. Více využívané jsou spektrální vlastnosti, oproti dynamickým vlastnostem signálu. Mezi hlavní příznaky používané při klasifikaci v časové oblasti patří například absolutní hodnota amplitudy signálu. Ve frekvenční oblasti jsou hlavními příznaky převážně výkony frekvenčních pásem. [18]

K často využívaným příznakům patří například spektrální výkony pro frekvenční pásma. Pro každý algoritmus, postup klasifikace, jsou vybírány příznaky tak, aby maximalizovaly úspěšnost klasifikace. Příznaky jsou voleny tak, aby mezi sebou nekorelovaly a měly vypovídající váhu. [19, 20]

Pro svou klasifikaci používám příznaky popsané ve skriptech Krajča, Mohylová, z roku 2011 [7], jejichž výpočet je implementován ve Wave-Finderu. Příznaky jsou normovány v intervalu (0,1), jsou tedy bez jednotkové veličiny. Program WF nabízí možnost výpočtu až 24 příznaků, které jsou uspořádány ve vektoru. V obecné rovině existuje velké množství příznaků (já zvolil 23 příznaků, které jsou uvedeny v tabulce č. 1 níže). Příznaky tak vytváří n-rozměrný vektor, který je vypočítáván pro každý segment EEG záznamu. [7, 4]

Tabulka 1: Normované příznaky použité pro klasifikaci [7].

Příznak	Popis
SIGM	variabilita signálu
APOS	maximální pozitivní hodnota v daném segmentu
ANEG	maximální negativní hodnota v daném segmentu
DELT1	hodnoty signálu v části delta frekvenčního pásma (0,5 Hz – 1,5 Hz)
DELT2	hodnoty signálu v části delta frekvenčního pásma (2,0 Hz – 3,5 Hz)
THET1	hodnoty signálu v části theta frekvenčního pásma (4,0 Hz – 5,5 Hz)
THET2	hodnoty signálu v části theta frekvenčního pásma (6,0 Hz – 7,5 Hz)
ALPH1	hodnoty signálu v části alfa frekvenčního pásma (8,0 Hz – 10,0 Hz)
ALPH2	hodnoty signálu v části alfa frekvenčního pásma (10,5 Hz – 12,5 Hz)
SIGMA	hodnoty signálu v části beta frekvenčního pásma (18,0 Hz – 29,0 Hz)
BETA	hodnoty signálu v beta frekvenčním pásmu (13,5 Hz – 29,0 Hz)
MAX1D	maximální hodnota první derivace
MAX2D	maximální hodnota druhé derivace
mf	střední frekvence
MD1	střední hodnota první derivace
MD2	střední hodnota druhé derivace
mob	Hjorthův parametr (Hjorth, 1975), mobility
comp	Hjorthův parametr (Hjorth, 1975), complexity
act	Hjorthův parametr (Hjorth, 1975), activity
LOfC	length of curve, délka křivky
NLinE	nonlinear energy (Automatic EEG analysis during long-term monitoring in the ICU, Agarwal, Gotman)
ZC	počet průběhů nulou
Peaks	frekvence maximálního vrcholu ve spektru

V tabulce č. 1 jsou zjednodušeně popsány příznaky vybrané ke klasifikaci EEG záznamů. Jednotlivá pásma (alfa, delta, theta) jsou rozdělena vždy do dvou oblastí v rámci svého frekvenčního pásma. V každé oblasti jsou pak určovány charakteristiky signálu ve frekvenčním spektru.

Maximální negativní *ANEG* (v negativním směru) a pozitivní hodnota *APOS* (v pozitivním směru) jsou extrémy amplitudy pro daný segment. Udávají extrémní hodnotu napětí daného segmentu po odečtení stejnosměrné složky napětí ( $A_{DC}$ ) uvedenou ve vzorci č. 1. Negativní směr v signálu znamená kladnou hodnotu amplitudy (směr nahoru) a naopak kladný směr znamená zápornou amplitudu (směr dolů). [7]

Stejnoseměrnou složku vypočítáme podle vzorce [7]:

$$A_{DC} = \frac{\sum_{i=1}^L y_i}{L}, \quad (1)$$

kde  $L$  je délka segmentu a  $y_i$  je amplituda  $i$ -tého vzorku v segmentu.

Maximální hodnota první derivace  $MAX1D$  (rovnice č. 2) a střední hodnota první derivace  $MD1$  (rovnice č. 3) určují sklona průměrný sklon křivky signálu.[7]

$$MAX1D = \max(y_{i+1} - y_i), \quad (2)$$

$$MD1 = \frac{\sum_{i=1}^n (y_{i+1} - y_i)}{n} \quad (3)$$

kde  $y_i$  je amplituda  $i$ -tého vzorku v segmentu a  $n$  je počet vzorků.

Maximální hodnota druhé derivace  $MAX2D$  (rovnice č. 4) a střední hodnota druhé derivace  $MD2$  (rovnice č. 5) určují špičatost křivky. [7]

$$MAX2D = \max(y_{i+4} - 2y_{i+2} + y_i), \quad (4)$$

$$MD2 = \frac{\sum_{i=1}^n y_{i+4} - 2y_{i+2} + y_i}{n}, \quad (5)$$

kde  $y_i$  je amplituda  $i$ -tého vzorku v segmentu a  $n$  je počet vzorků.

Hjortovy parametry jsou ukazatele statistických vlastností používaných při zpracování signálů z časové oblasti. Parametry jsou nazývány aktivita, mobilita a komplexita. Jsou běžně využívány v klinické praxi pro analýzu EEG záznamů. [21]

Aktivita představuje sílu signálu, rozptyl časové funkce (čtverec standardních odchylek). Hodnotí se výkonové spektrum ve frekvenční oblasti.

$$Activity = var(y(t)), \quad (6)$$

kde  $(y(t))$  reprezentuje signál.

Mobilita (*Mobility*) představuje střední frekvenci - podíl standardní odchylky výkonového spektra. Toto je definováno jako podíl druhé mocniny z rozptylu první derivace signálu a rozptylem signálu samotným. [22]

$$Mobility = \sqrt{\frac{var\left(y(t) \cdot \frac{dy}{dt}\right)}{var(y(t))}}, \quad (7)$$

kde  $var(y(t))$  je Hjorthův parametr aktivita,  $\frac{dy}{dt}$  je derivace amplitudy daného vzorku segmentu podle času a  $y(t)$  je velikost amplitudy vzorku v segmentu.

Komplexita (*Complexity*) je změna frekvence. Parametr je porovnáním signálu s čistě sinusovou vlnou. [22]

$$Complexity = \frac{Mobility\left(y(t) \cdot \frac{dy}{dt}\right)}{Mobility(y(t))}, \quad (8)$$

kde *Mobility* je Hjorthův parametr mobilita,  $\frac{dy(t)}{dt}$  je derivace amplitudy vzorku segmentu podle času a  $y(t)$  je amplituda vzorku v daném segmentu.

Další parametr délku křivky (*LOfC*) si můžeme představit jako délku signálu, který roztáhneme jako provázek. Počítá se jako součet absolutních hodnot rozdílů amplitud jednotlivých vzorků v daném segmentu. [7]

$$LOfC = \sum_{i=1}^{N_s} abs[y(i) - y(i + 1)], \quad (9)$$

kde  $N_s$  je počet vzorků v segmentu a  $y(i)$  je amplituda  $i$ -tého vzorku v segmentu.

Počet průběhů nulou je číslo, které značí, kolikrát křivka prošla z kladných hodnot do záporných a naopak. Hranice přechodu je definována hodnotou elektrického napětí  $0,01 \mu\text{V}$ . Hodnota se opět počítá po odstranění stejnosměrné složky amplitudy (viz rovnice č. 1).

Parametr *Peaks* nám udává počet špiček v daném segmentu. Parametr je počítán ze spektra signálu daného segmentu pomocí rychlé fourierovy transformace (FFT). Parametr *Peaks* odpovídá frekvenci dominantního vrcholu spektra signálu. [7]

NLinE je parametr charakterizující signál z hlediska energie. Udává průměrný výkon v hlavní energetické zóně. Hlavní energetická zóna je pásmo, které je soustředěné na průměru frekvence z 80,% celkové energie spektra. Průměrný výkon v hlavní energetické zóně se získá dělením výkonu v tomto pásmu a jeho šířkou. [23, 7]

$$NLinE(i) = y^2(i) - y(i-1)y(i+1), \quad (10)$$

kde  $y(i)$  je amplituda v  $i$ -tém vzorku segmentu. Používá se proto, že odráží koncentraci energie ve spektru. V případě, že je výkon ve spektru soustředěn v jedné oblasti, hlavní energetická zóna je úzká a průměrný výkon v ní je tedy velmi velký.

### 2.2.3 Klasifikace

Klasifikace slouží k označení segmentů EEG, které mají podobnou charakteristiku (typickým příkladem mohou být segmenty s epileptickými grafoelementy). Hlavním cílem klasifikace je blížit se vizuálnímu hodnocení lékaře. V dlouhodobých záznamech EEG signálu má upozornit na diagnosticky zajímavé úseky, které může individuálně lékař vyhodnotit.

V dnešní době se již těžko obejdeme bez výpočetní techniky. Díky pokroku s křemíkovými materiály mohou počítače v současnosti provádět rychle složité výpočetní operace. EEG signál je zpracováván především ve frekvenční oblasti [14]. Úspěšnost metod



používaných ke klasifikaci se odvíjí od správně extrahovaných příznaků, které popisují vlastnosti klasifikovaných objektů. “Každá metoda automatické klasifikace je jenom tak dobrá, jak kvalitní jsou použité příznaky. [7]“ [24]

Záznamy aktivity mozku mají různou délku. U dlouhodobých záznamů se běžně nepřístupuje k jeho hodnocení bez předchozí klasifikace. Je potřeba digitálně zpracovat data a roztrždit je na oblasti zájmu, které má lékař zkontrolovat, a oblasti s normální aktivitou. S nástupem moderních technologií, elektronického zpracování biomedicínských dat, je možné rychle vyhodnotit takto velké soubory. Existuje několik druhů metod určených pro klasifikaci. Klasifikační metody se dělí do dvou kategorií: učení s učitelem a učení bez učitele.

### **Učení s učitelem**

Metody učení s učitelem reprezentují proces, při kterém program tzv. „učíme“. Učení představuje počáteční vstup uživatele v podobě definice správných vzorů (etalonů). Pomocí těchto etalonů se algoritmus učí a do stejných tříd přiřazuje data podobná vstupním vzorům. Patří sem například neuronové sítě a genetické algoritmy. Neuronové sítě se využívají například v kombinaci s k-means při zkoumání spánkových stavů v EEG záznamu. U těchto metod je nutná intervence lékaře, který v EEG záznamu musí najít příznaky typické pro jednotlivé třídy, do kterých chce záznam rozdělit. [25]

### **Učení bez učitele**

Pokud není potřeba vzorů, a podobnost mezi jednotlivými body je dána matematickým vztahem, hovoříme o klasifikaci bez učitele. Mezi neznámější algoritmy patří metoda k-means, která je také v současnosti využívána ke klasifikaci EEG záznamu. Tyto algoritmy můžeme dále dělit podle matematických vztahů, které je definují.

### **Prototypově(pouze vzdálenostně) založené**

#### **K-means**

K-means je často používaný algoritmus shlukování s jedním vstupním parametrem - počtem shluků. Body jsou přiřazovány ke shlukům (klastrům) následkem přepočítávání polohy centra a vzdálenosti bodů k němu. Tato metoda je vždy konečná a rychle konverguje k řešení. V současnosti se tento algoritmus používá při klasifikaci segmentů EEG, proto se mu budu níže věnovat podrobněji. [7]

#### **Fuzzy c-means**

V k-means jsou data rozdělena do shluků, ve kterých každý bod přísluší právě k jednomu shluku. Při použití metody fuzzy c-means mohou jednotlivé datové prvky patřit do více než jedné skupiny. Každý bod je pak definován parametrem příslušnosti k jednotlivým třídám. Bod tedy může patřit do jednoho, ale i do více shluků. [26]

#### **PAM (Partitioning Around Medoids)**

PAM, někdy také nazývaný k-medoids, je algoritmus velmi podobný algoritmu k-means. Těžištěm shluku je jeden z bodů datového souboru. Algoritmus je více robustní. Medoid je přiřazen do shluku, pokud jeho průměrná odlišnost u všech objektů v shluku je minimální – tj. že je nejvíce centrálně umístěný bod v shluku. [24]

### **Hustotně založené**

#### **DBSCAN**

DBSCAN je hustotně založená metoda, která rozděluje segmenty podle množství bodů (sousedů) v jejich blízkém okolí. Odděluje tedy od sebe místa řidších a místa hustších bodů. Vstupním parametrem je poloměr – radius. Tento parametr je možné spočítat automaticky ze vstupních dat. [27]

#### **DENCLUE**

DENCLUE je nejmladší metoda vycházející z algoritmu DBSCAN. Je realizována pomocí

výpočtu histogramu. Metoda dobře klasifikuje nerovnoměrně hustotně rozložená data, ale časově a paměťově se řadí mezi náročnější. [27]

### **Grafově založené**

#### **Hierarchické shlukování**

Jedná se o časově náročnou metodu. Výsledkem je sada vnořených uskupení, vytváří se hierarchický strom – dendrogram. Existují dva možné postupy. Aglomerativní – začneme s jednotlivými body jako shluky a postupně slučujeme nejbližší dvojice shluků, dokud nedosáhneme jednoho shluku. Opačný – divizní postup spočívá v dělení z jednoho počátečního shluku. Chceme-li konkrétní počet shluků, provedeme řez v dané rovině – vrstvě stromu. [24]

#### **Chameleon**

Mladá, výpočetně náročná metoda Chameleon, která slouží k dynamickému modelování shluků, je účinná ve 2D prostoru. Z matice podobnosti jsou vytvářeny hrany charakterizované vzdálenostmi bodů. Postupně v hierarchické posloupnosti jsou rozrušovány nejdelší hrany a tím dochází ke vzniku nových shluků z jednoho původního. Tento algoritmus vychází z hierarchické metody SNN-DBSCAN. [28]

#### **HMM Markovo shlukování**

HMM Markovo shlukování je pravděpodobnostní metoda vhodná pro klasifikaci časových řad. Nepatří ale mezi často využívané metody, ačkoli již byla použita i na klasifikaci surového EEG. [29]

## **2.3 K-means**

Nejjednodušší používanou metodou je v současnosti metoda k-means, která je také součástí programu na zpracování signálů, Wave-Finderu. Jedná se o výpočetně méně náročnou metodu. Vztahy mezi body jsou počítány pomocí vzdálenosti, převážně Euklidovské,

případně Mahalanobisovi či Hammingovi. Cílem je vytvoření shluků s co nejpodobnějšími body, které jsou zároveň od ostatních shluků dost vzdálené. [30]

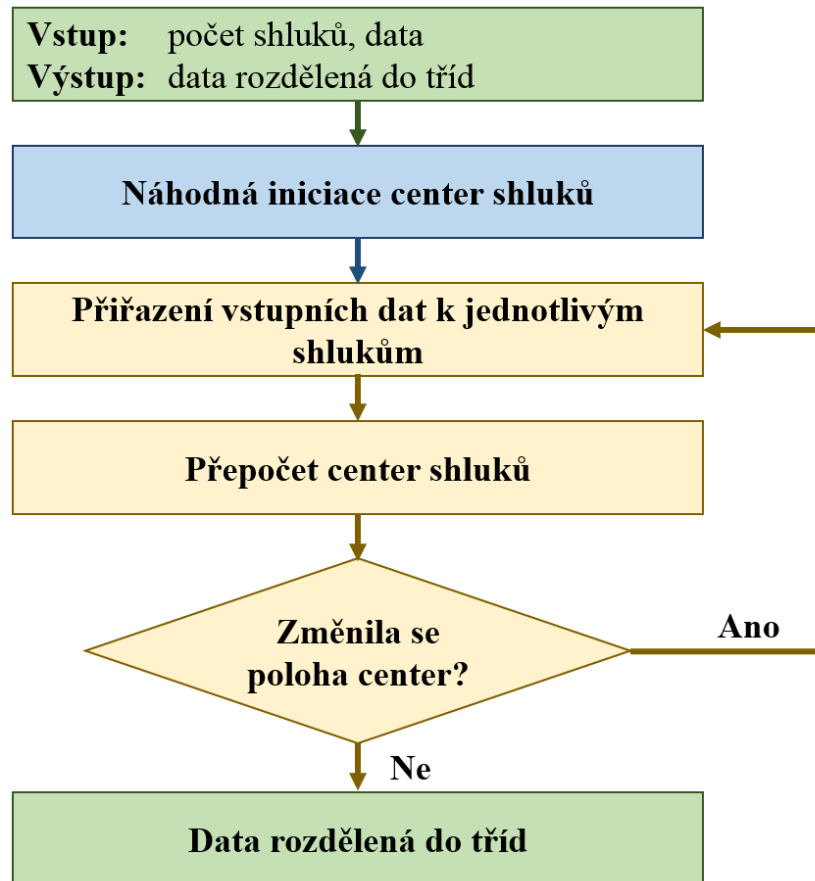
Uživatel zadává jako vstupní parametr počet shluků, do kterých mají být body rozřazeny. V programu WF je na základě empirické zkušenosti nastavena hodnota na 7 shluků [4]. Tento počet shluků se u epileptických pacientů osvědčil a klasifikace epileptické aktivity byla ze subjektivního pohledu odborníka úspěšnější více jak v 50 %.

U metody k-means lze vypočítat pomocí metody siluet ideální počet shluků a není tak potřeba vstupního parametru. Metoda ověřuje konzistentnost dat uvnitř shluků. Tato technika stručně graficky znázorňuje, jak moc je objekt podobný třídě, ve které je zařazen. Hodnota siluety je tedy měřítkem podobnosti objektu vlastnímu shluku v porovnání s ostatními shluky. Rozmezí hodnot se pohybuje od -1 do 1, kde 1 indikuje správné zařazení bodu do třídy. Pokud většina bodů má hodnotu blížíci se 1, pak byla data správně klasifikována. Čím více bodů má hodnotu siluet blížíci se -1, tím je vyšší chybovost klasifikace. [31, 32]

Časově a paměťově se jedná o velmi náročnou metodu, neboť jako vstupní parametr do výpočtu siluet je třeba klasifikace záznamu pro všechny varianty počtu shluků, ze kterých chceme vybírat. Pro záznam EEG se nám nepodařilo (do 1 hodiny) pro 30 minutový záznam s 23 příznaky spočítat automatickou klasifikaci pro 7 shluků. Při snížení velikosti souboru a počtu příznaků program počítal 2 hodiny. Realizace automatické klasifikace EEG pomocí metody k-means v programu MATLAB se tedy nezdařila.

## K-means postup

Vstupním parametrem k-means je počet shluků, do kterých mají být vstupní data rozdělena. Výstupním parametrem jsou body rozřazené do tříd. Na obrázku č. 7 níže je načrtnut proces klasifikace prostřednictvím k-means metody.



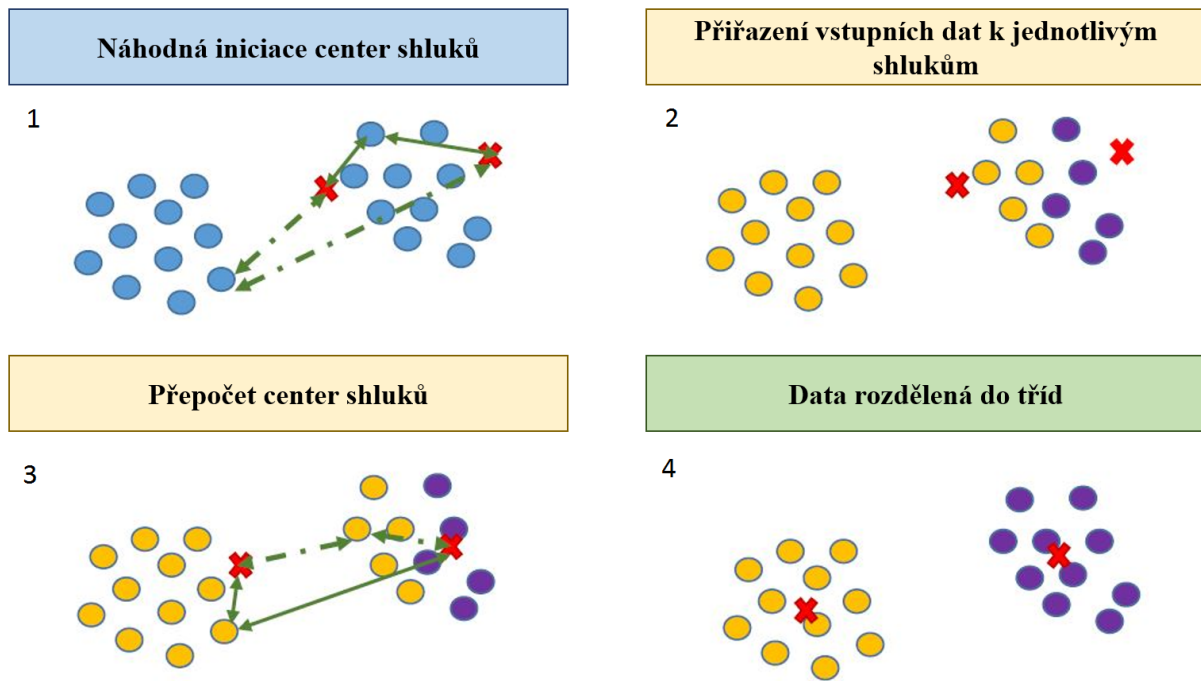
Obrázek 6: Proces klasifikace do tříd prostřednictvím algoritmu k-means.

**1. část:** V prvním kroku dochází k náhodné inicializaci center shluků. Vstupní data jsou normována od 0 do 1, tudíž i středy se nacházejí v normovaném prostoru  $(0,1)$ .

**2. část:** Pro všechny body vstupních dat se počítá jejich vzdálenost od všech center shluků. Body jsou pak přiřazeny ke shluku, k jehož centru jsou nejbližší.

**3. část:** Dojde k přepočítání center shluků.

**4. část:** Nová centra shluků odpovídají těžištím shluků. Proces přiřazování bodů vstupních dat ke shlukům a přepočítání center probíhá tak dlouho, dokud se mění centra jednotlivých shluků. Jakmile těžiště zůstávají na svém místě, již se nemění vzdálenost a všechny body v daném shluku jsou nejbližší k těžišti, které k shluku patří. [7, 33]



Obrázek 7: Vizualizace procesu klasifikace pomocí metody k-means, čísla odpovídají popisu nad obrázkem.

## 2.4 DBSCAN

Shlukovací techniky využíváme pro extrakci skrytých a zajímavých vzorů z velkých souborů dat. DBSCAN se řadí mezi hustotně založené prostorové algoritmy. Je průkopníkem algoritmů založených na hustotě. Z jeho původní myšlenky vychází např. metody OPTIC a DENCLUE. Tato metoda je vhodná pro aplikaci na data s odlehlými body (šumem). [34]

Cílem je rozřazení bodů do shluků na základě hustoty, kdy je oblast bodů s podobnou hustotou klasifikována jako jedna třída. DBSCAN popisuje shluk jako oblast s vysokou hustotou bodů, která je oddělená místem s nízkou hustotou. [35]

Vstupními parametry jsou poloměr  $Eps$  (radius) a počet bodů v radiusu  $k$ . Některé varianty DBSCANu umožňují automatický výpočet vstupních parametrů, tudíž je poté algoritmus zcela soběstačný a není nutná žádná intervence uživatele. Bohužel stejně jako u metody k-means je tento postup značně výpočetně náročný.

Jednoduchý DBSCAN má stejně jako k-means řadu problémů. Jsou to:

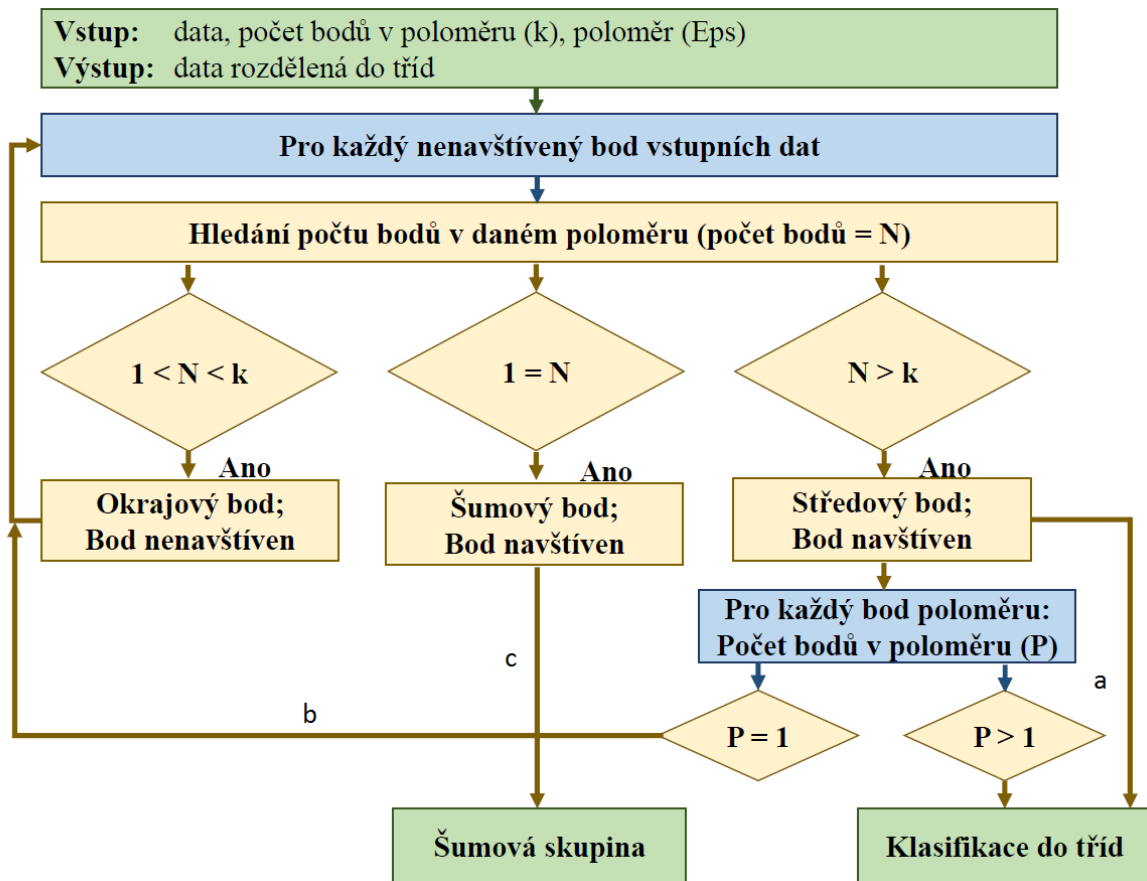
1. nutný vstup uživatele – zadání parametrů
2. náchylnost k chybě pokud se mění hustotní rozložení bodů v rámci shluků
3. vysoká výpočetní náročnost

Mnozí odborníci se pokusili překonat tyto nevýhody a prvotní algoritmus DBSCAN modifikovali či slučovali s jinými technikami. Vznikl tak například VDBSCAN, FDBSCAN, DMDBSCAN, IDBSCAN a další.

## DBSCAN postup

V metodě DBSCAN nezávisí na iniciačním místě, neboť nepočítáme s náhodným středem shluků. Počet shluků je výchozí nikoli vstupní parametr. Začne se jakýmkoli bodem dat, která klasifikujeme. V blízkém okolí bodu (zadaném poloměru) hledáme počet bodů, které do něj spadají. DBSCAN body dělí do 3 skupin:

- okrajový bod - počet bodů v poloměru shodný se vstupním parametrem  $k$ , označuje okraje shluku
- středový - počet bodů v poloměru vyšší než vstupní parametr  $k$
- šumový bod - počet bodů v poloměru menší než vstupní parametr  $k$ , označuje odlehlé body

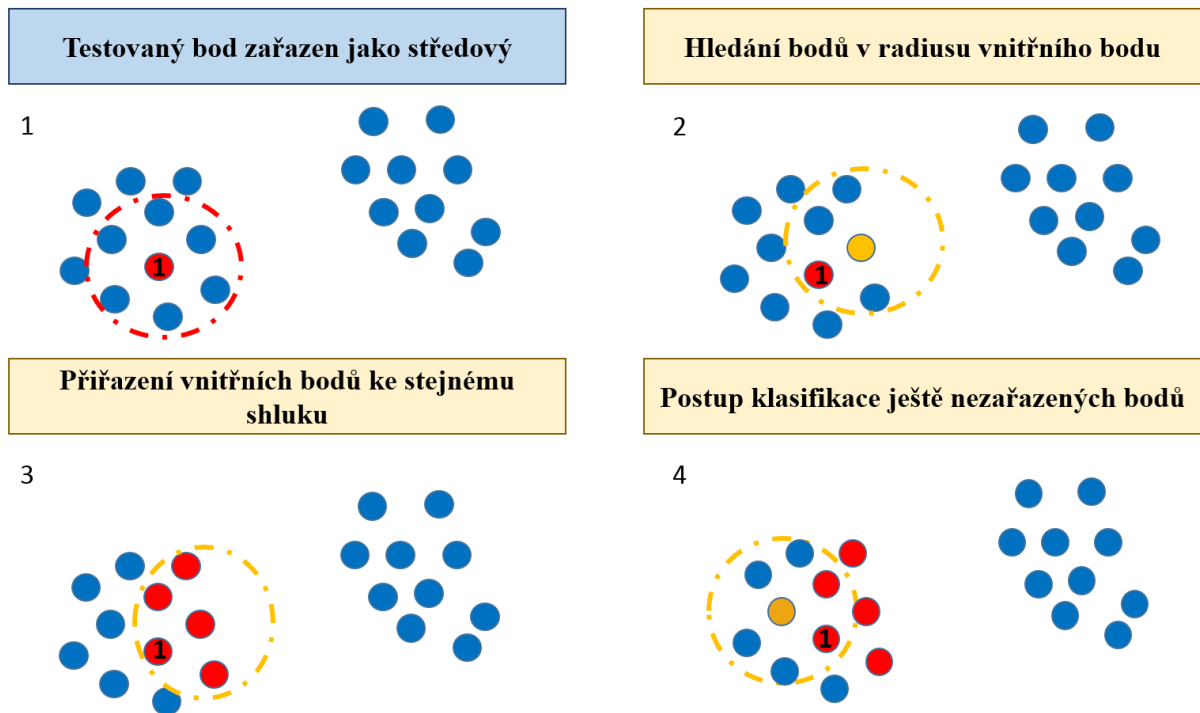


Obrázek 8: Proces klasifikace do tříd prostřednictvím algoritmu DBSCAN.

Algoritmus DBSCAN klasifikuje na základě rozhodovacího kritéria počet bodů v poloměru. Pokud je počet bodů dostatečný je aktuální testovaný bod zařazen do shluku (trasa a). Čísla shluků vzrůstají od 1 do  $n$ . Následně jsou testovány body jeho poloměru. Ty jsou na základě stejného kritéria řazeny do stejného shluku, nebo jsou postoupeny opětovnému testování (trasa b). Šumové body jsou vždy klasifikovány rovnou do speciální třídy (trasa c).



V následujícím obrázku č. 10 je zobrazen postup klasifikace DBSCANem na 2D datech.



Obrázek 9: Proces klasifikace do tříd prostřednictvím algoritmu DBSCAN, čísla odpovídají popisu pod obrázkem.

**1. část:** Pro zvolený bod zjišťujeme podle počtu bodů ( $k$ ) v jeho blízkém okolí ( $Eps$ ), zda se jedná o středový, okrajový nebo šumový bod.

**2. část:** Z bodů patřících do poloměru iniciačního bodu (červený bod s číslem 1) vybereme jeden bod (oranžový).

**3. část:** U oranžového bodu také hledáme jeho počet bodů spadajících do jeho blízkého okolí.

**4. část:** Podle počtu bodů v okolí oranžového bodu označíme tento bod jako středový nebo jako okrajový a zařadíme jej do shluku k původnímu (červenému s číslem 1) bodu. Pokud tento bod byl označen jako středový, tak všechny body v jeho poloměru přiřadíme do stejného shluku. Předchozí část (3.) opakujeme pro všechny body, které patřily do poloměru červeného iniciačního bodu (označený číslem 1).

Pokud jsme vyčerpali všechny body, které spadaly do blízkého okolí iniciačního bodu (červený s číslem 1), tak v algoritmu automaticky přejdeme na další bod, který ještě nebyl zařazen do žádné třídy. Celý proces se opakuje, dokud nejsou všechny body navštíveny a zařazeny do třídy, nebo hodnoceny jako šum. Tím, že se počet tříd zvyšuje od 1 do  $n$ , tak  $n$  je výstupní hodnota algoritmu, kdy získáme počet tříd pro klasifikovaná data. Počet tříd jako výstupní parametr je hlavní výhodou metody DBSCAN.

## 2.5 Porovnání metod k-means a DBSCAN

V tabulce č. 2 je porovnání výhod i nevýhod obou algoritmů. V současnosti na klasifikaci EEG záznamů využívaný k-means byl nastaven dle testovaných subjektů a počet shluků neodpovídá reálnému počtu tříd v záznamu. Na klasifikaci EEG nově aplikovaný algoritmus DBSCAN má oproti k-means velkou výpočetní náročnost.

Tabulka 2: Porovnání vlastností algoritmů DBSCAN vs. k-means [25, 27, 33, 7].

<b>K-means</b>	<b>DBSCAN</b>
+ rychlost	+ detekce prolnutých shluků
+ komplexnost - různá vstupní data	+ separace šumu
+ variabilita změnou výpočtu vzdáleností	+ velký počet dostupných modifikací
+ jednoduchost	+ automaticky dán počet shluků
– vstupní parametr počet shluků	– vysoká výpočetní náročnost
– neseparuje prolnuté shluky	– chyba při nerovnoměrné hustotě

## 3 Metodika

Metod klasifikace je nepřehledné množství, většina známých metod není univerzální na všechny typy dat. Hojně využívanou metodou na klasifikaci EEG signálu je metoda k-means z kategorie neučících se metod, z kategorie učících se jsou to neuronové sítě. Metodou k-means nedocílíme správného rozdělení všech grafoelementů signálu.

DBSCAN, průkopnický algoritmus pracující na základě hustoty, se využívá při klasifikaci satelitních snímků a své uplatnění se pokouší najít i v radiografii [36]. Tato metoda umí rozlišit prolnuté shluky, jejichž výskyt v 2D prostoru příznaků předpokládám a se kterými si algoritmus k-means neumí poradit. Zároveň tento algoritmus poskytuje jako svůj výstupní parametr hodnotu počtu shluků. Na základě těchto vlastností jsem zvolil DBSCAN jako algoritmus pro pilotní studii na hustotně založenou klasifikaci EEG dat.

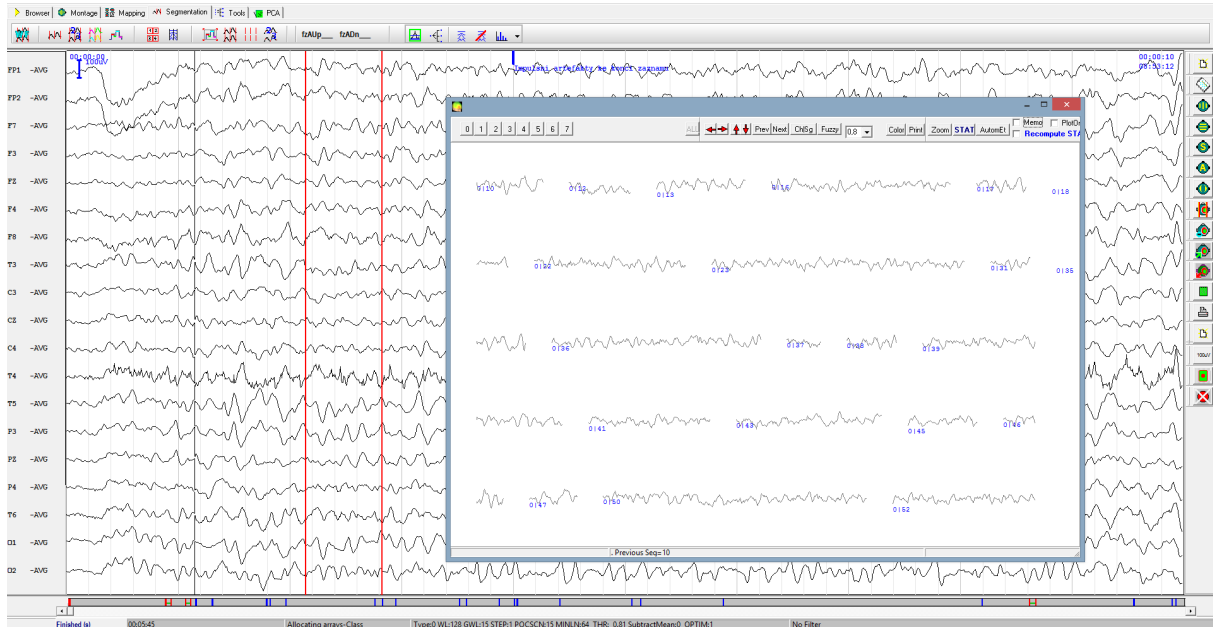
V odstavcích níže bude popsán postup implementace algoritmu DBSCAN, jeho testování a implementace modifikací za účelem vylepšení klasifikace EEG signálů.

### Programové prostředí

Metodika pro klasifikaci EEG segmentů je zpracovávána v programovém prostředí MATLAB 2015a, ve kterém vzniká i nový, komplexní program pro práci s EEG signály. Vizualizace EEG dat, jejich segmentace a výpočet příznaků jsou realizovány pomocí programu WF, který je využíván v klinické praxi.

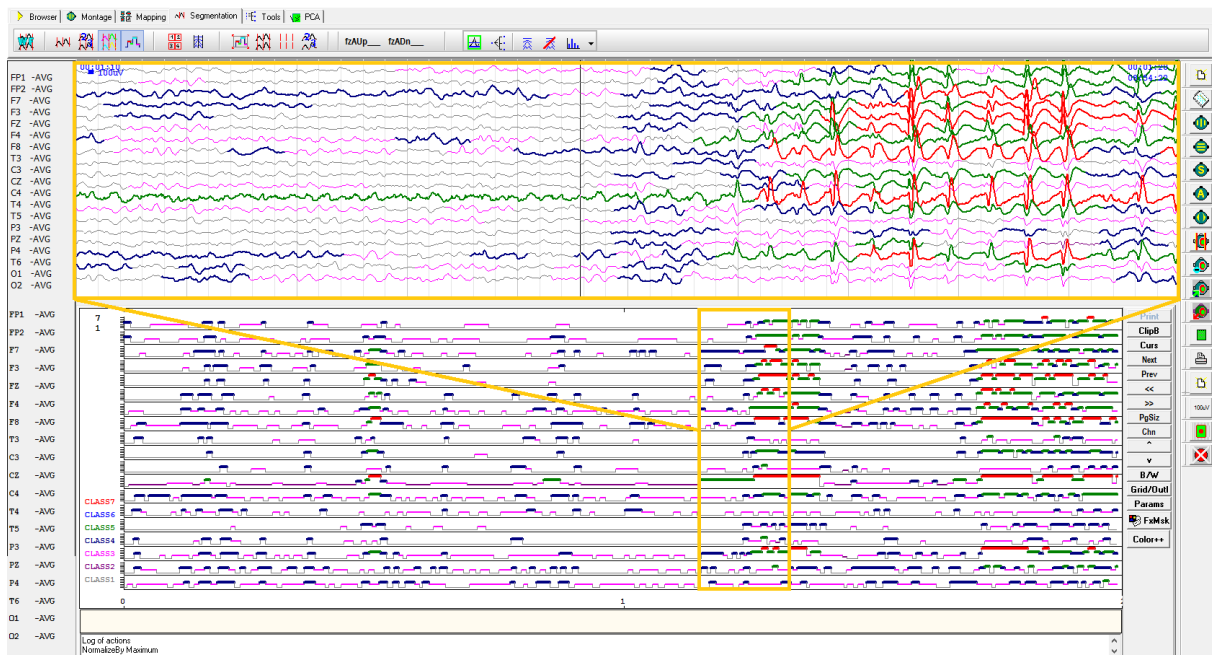
**MATLAB** (zkratka pro matrix laboratory) je programovací prostředí s rozsáhlou knihovnou funkcí primárně určený pro matematické operace; práce s daty probíhá formou matic. MATLAB umožňuje vykreslování grafů, implementaci vlastních algoritmů i vytváření vlastního uživatelského rozhraní. Jedná se o rozšířený program ve vědecké, akademické i technicko-vývojové sféře.

**Wave-Finder** (zkratka WF) je program, jehož autorem je pan doc. Ing. Vladimír Krajča, CSc. Využívají ho lékařské zařízení i výzkumné ústavy pro zpracování EEG záznamů. Umožňuje vizuální i matematické hodnocení digitálních EEG záznamů. Uživatel si volí montáž, výběr počtu příznaků, nastavení segmentace a dalších parametrů ovlivňujících charakter příznaků využívaných pro klasifikace. [4]



Obrázek 10: Vizualizace EEG záznamu a zvolené třídy v prostředí programu Wave-Finder [37].

Na obrázku č. 12 níže vidíme v dolní části okno (temporální profil) s časově zhuštěným signálem a barevně vyznačenými úseky. Významné části signálu jsou podle tříd různě obarvené. Lékař si výběrem (kliknutím myši) může zvolit zvýrazněnou oblast a tu si v horní části okna detailně prohlédnout, aniž by musel vyhodnocovat celý signál.



Obrázek 11: Klasifikace celého signálu - volba klinicky zajímavého úseku, oranžově zobrazená část s barevně oddělenými třídami v oblasti zájmu [37].

V následujících dvou kapitolách budou popsány společné části - příprava dat a jejich načítání. Následovat budou jednotlivé metody s dílčími výsledky.

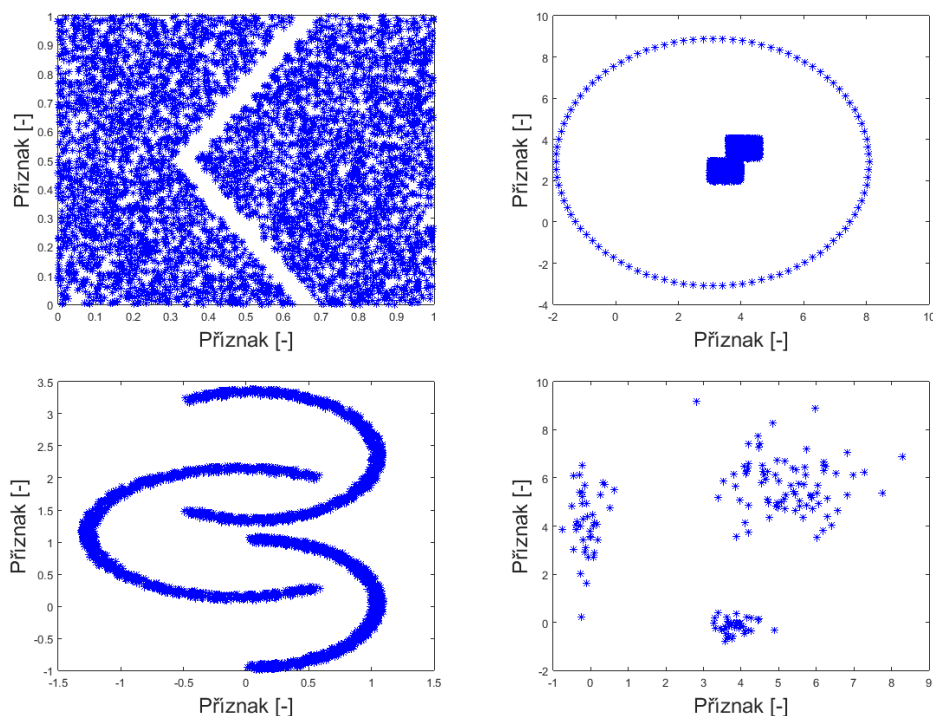
### 3.1 Data

Našimi daty jsou příznaky - číselné hodnoty, vlastnosti charakterizující signál. Počet příznaků není pro většinu klasifikačních metod omezen, ale s jeho velikostí stoupá výpočetní náročnost a někdy dochází i k snížení kvality klasifikace.

#### Testovací 2D data

Pomocí funkcí programu MATLAB jsem si vytvořil testovací 2D data (2 příznaky vykreslené v závislosti na sobě). Data tvoří prolnuté shluky, obrazce hustotně odlišné a data s šumovými body, tak aby bylo možné dobře vizualizovat rozdíly mezi algoritmem k-means

a metodou DBSCAN. Zároveň se jedná o různě velká data (různé počty segmentů viz obrázek č. 13), tak aby bylo možné vyhodnotit kvantitativní rozdíl mezi metodami. Uložena jsou ve formátu .txt v ASCII kódování.



Obrázek 12: Ukázka vytvořených testovacích 2D dat.

### Reálná EEG data

Data pocházejí od pacientů z Nemocnice Na Bulovce, byla získána na základě návrhu projektu "Analýza mikrostruktury a makrostruktury dlouhodobých EEG záznamů algoritmy umělé inteligence a číslicového zpracování signálu", který byl schválený etickou komisí Nemocnice Na Bulovce dne 28. 06. 2011.

Signály byly zaznamenávány v 19 kanálech při unipolárním zapojení na přístroji Brainquick. Byl použit analogový filtr pásmové propusti 0 - 70 Hz. Data byla vzorkována frekvencí 128 Hz a převedena za použití 12 bitového převodníku. Testovací data byla naměřena na 15 pacientech, kterým bylo indikováno vyšetření na základě podezření na nemoc epilepsii. Pacienti byli muži i ženy ve věku mezi 26 - 60 roky. Délka záznamu

se pohybuje od jednotek po desítky minut (standardní klinické vyšetření se pohybuje v rozmezí 15 - 30 min.). Počet segmentů v jednom záznamu se pohyboval od jednotek tisíc do desítek tisíc segmentů. Pomocí programu WF provádíme segmentaci a výpočet příznaků pro jednotlivé segmenty (viz kapitola EEG signál). Segmentace má nastavení parametrů udáno v následující tabulce č. 3.

Tabulka 3: Nastavení parametrů segmentace

Parametr	Nastavení
Window Length	128 vzorků
G Window Length	15 vzorků
STEP	1 vzorek
Optim	1 [-]
MINLENGTH	64 vzorků
Number of Scan pts	15 vzorků
Treshold	81 [-]
Max Segm Length	1024 vzorků

Parametr Window Length udává délku dvojitého okna. Při adaptivní segmentaci signál projíždějí dvě spojená okna (viz obrázek č. 6 v sekci 2.2.1), každé má délku 64 bodů, dohromady tedy 128.

Parametr G Window Length je délka okna, ve kterém se hledá přesná pozice maxima. Pro hledání maxima jsou potřeba minimálně 3 body. Na každou stranu od maxima musí být hodnota nižší.

Parametr STEP je krok. V signálu se pohybujeme s krokem 1.

Parametr Optim zapíná a vypíná funkci optimalizace hranice segmentu. Jedná se o okno, v jehož středu je původně detekovaná hranice, která může mít špatnou polohu. Na každou stranu od bodu původně detekovaného jako hranice hledáme, zda signál neklesá. Do nejnižšího bodu umisťujeme hranici segmentu. Počet bodů, o které se díváme na každou stranu, je definován parametrem Number of Scan pts.

Parametr Threshold udává hranici, nad kterou je rozdíl dvou oken vyhodnocen jako dostatečný pro segmentaci (rozdělení signálu na dvě části).

Parametr MINLENGTH udává nejmenší velikost segmentu a Max Segm Length udává nejdelší možnou velikost segmentu. Velikost segmentů je primárně nastavována kvůli přehlednosti. Často se v signálu vyskytují dlouhé úseky stejné aktivity. Ty jsou segmentovány pouze pro snazší optické vyhodnocení.

Výstupem, ve kterém jsou uloženy příznaky pro všechny segmenty signálu, je soubor s příponou .nra.

### 3.1.1 Třídy

Ve zkoumaných záznamech se vyskytují třídy: fyziologické aktivity, EMG, epileptické a oční aktivity, pulzní artefakty a rovné čáry. Kromě rovných čar a pulzních artefaktů jsou všechny třídy popisovány a klasifikovány i ve studii „Classification of transient events in eeg recording“, zabývající se klasifikací EEG záznamů [38].

#### Fyziologická aktivita mozku

Mezi fyziologickou aktivitu řadíme sinusové vlny a veškerou aktivitu, která nepatří do žádné třídy grafoelementů.

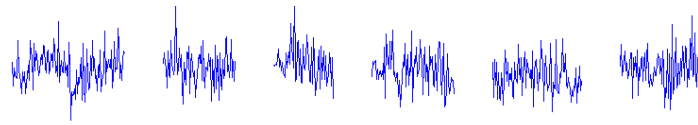


Obrázek 13: Ukázka segmentů fyziologické aktivity, zobrazeno ve WF [37].

#### EMG svalová aktivita a šumové grafoelementy

Svalová aktivita je ve třídě, do které řadíme i zašuměné segmenty se síťovým brumem, kdy šum zcela zkreslí původní signál.

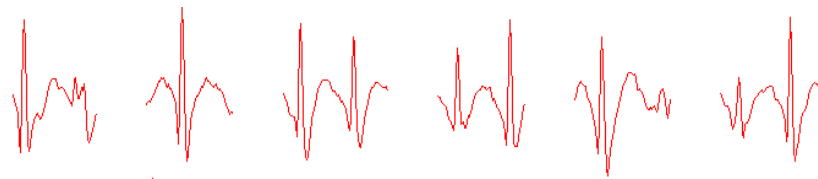




Obrázek 14: Ukázka segmentů EMG (svalových) grafoelementů, zobrazeno ve WF [37].

### Epileptická aktivita

Epileptická aktivita tvoří nejvíce klinicky zajímavou třídu. Epileptické grafolementy mají proměnlivý charakter, co se týká amplitudy. Algoritmy mají tedy tendenci epileptickou aktivitu dělit do 2 tříd podle velikosti amplitudy vln.



Obrázek 15: Ukázka segmentů epileptické aktivity s vysokou amplitudou, zobrazeno ve WF [37].

### Pomalé oční artefakty

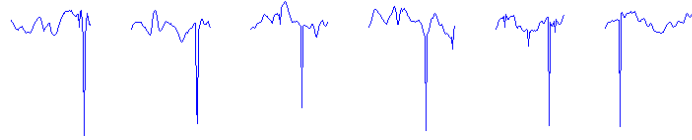
Jedná se například o pomalé vlny způsobené oční aktivitou. Může se jednat i o artefakt způsobený elektrodou. Jako v předchozím případě mají algoritmy tendenci dělit skupinu do dvou tříd podle velikosti amplitudy.



Obrázek 16: Ukázka segmentů pomalých očních artefaktů - pomalých vln, zobrazeno ve WF [37].

### Pulzní artefakty

Pulzní artefakty se projevují úzkým kladným hrotem s vysokou amplitudou. Tyto artefakty se vyskytují většinou pouze v jednom kanálu - byly způsobeny jednou elektrodou.



Obrázek 17: Ukázka segmentů artefaktů ze špatného kontaktu elektrody, zobrazeno ve WF [37].

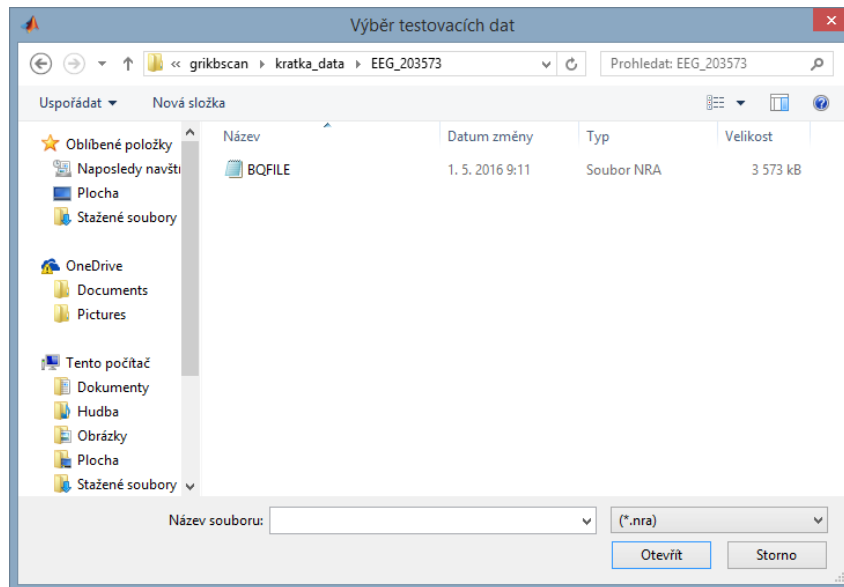
### Rovné čáry

V signálu se vyskytují i rovné úseky, kdy nebyl zaznamenán žádný EEG signál. Jedná se o segmenty s konstantní amplitudou. Těchto segmentů, vyskytují-li se v signálu, bývá zlomkové množství (např. 14 segmentů z celkového počtu 40000 segmentů). Nejedná se o projev EEG aktivity. Nejsou tedy klinicky významným artefaktem, který bychom potřebovali klasifikovat.

Výše popsané grafoelementy vyskytující se v segmentech jsou uloženy v souboru s příponou .nra. Jedná se o soubor s normalizovanými příznaky extrahovanými z jednotlivých segmentů. Normované příznaky jsou v binárním kódování uloženy v souboru .nra, který je potřeba načíst do programu MATLAB, ve kterém jsem implementoval algoritmy ke klasifikaci.

## 3.2 Načítání dat

Data (soubory .nra) jsou načítána pomocí grafického okna (obrázek č. 19) i s možností filtru formátu. U 2D testovacích dat není potřeba další úpravy, data jsou uložena ve formě matice  $(n,2)$ , kde  $n$  je počet segmentů.



Obrázek 18: Plovoucí okno pro výběr dat [39].

Data exportovaná z WF mají formát jednoho sloupce - uložení pod sebou, včetně technických údajů. Z hlavičky vyčteme počet kanálů, jejich délky a počet příznaků. Pomocí délek kanálu si spočítám začátky jednotlivých kanálů. Na základě těchto dvou informací (délky kanálů a počtu příznaků) si data uložím do matice  $(n, m)$ , kde  $n$  je počet segmentů v celém signálu a  $m$  je počet příznaků pro jednotlivé segmenty.

Výsledná funkce je uložena pod názvem "nacteni".

### 3.3 DBSCAN

Metoda DBSCAN je popsána pomocí definic [34, 40, 27]:

1. definice: Aby byl bod středový, musí mít alespoň jeden bod blíže, než je hodnota poloměru  $Eps$ .
2. definice: Existují dva druhy bodů, které patří do shluku - okrajové a středové. Středový bod má ve svém okolí minimálně definovaný počet bodů. Okrajový je takový, který není středovým bodem, ale leží v blízkosti jiného středového bodu.

3. definice: Bod je hustotně dosažitelný, pokud v jeho blízkosti leží bod, který patří zároveň do poloměru jiného středového bodu.
4. definice: Může se stát, že dva okrajové body nebudou hustotně dosažitelné, jak je popsáno v předchozí definici. Musí proto existovat jiný bod, který bude splňovat předchozí definici s každým z nich zvlášť.
5. definice: Je-li bod  $p$  náležící do shluku  $A$  hustotně dosažitelný (viz definice č. 3 a 4) s bodem  $q$ , pak i bod  $q$  patří do shluku  $A$ .
6. definice: Šum je množina bodů, které nebyly zařazeny do žádného shluku.

Algoritmus DBSCAN jsem implementoval na základě popisu metody DBSCAN v článkách [34, 40]. DBSCAN jako prvotní verze algoritmu vyžaduje 2 vstupní hodnoty - poloměr  $Eps$  (radius) a počet bodů v poloměru  $k$ . Níže je popsán pseudokód implementované verze DBSCANu. Veškeré proměnné v pseudokódech jsou uváděny bez diakritiky, aby odpovídaly proměnným použitým v reálném algoritmu.

**Data:**  $k$ ,  $Eps$

**Result:** Klasifikace segmentů

```

if bod ještě nebyl navštíven then
    vzdalenost = vzdálenost všech bodů od bodu zájmu
    body_polomeru = počet bodů v poloměru  $Eps$ 
    if počet(body_polomeru) > 1 a počet(body_polomeru) <  $k$  then
        bod_zajmu = okrajový
        shluk = zatím neklasifikujeme
    end
    if počet(Body poloměru) = 1 then
        bod_zajmu = šumový
        shluk = nezařazujeme do žádného shluku
    end
    if počet(Body poloměru) ≥  $k$  then
        while Nenavštíví všechny Body poloměru do
            vzdalenost2 = vzdálenost všech bodů od bodu zájmu
            body_polomeru2 = počet bodů v zadaném poloměru
            if počet(Body poloměru) > 1 then
                shluk = číslo sluku
                if počet(Body poloměru) ≥  $k+1$  then
                    | bod_zajmu = středový
                else
                    | bod_zajmu = okrajový
                end
            end
        end
    end
end

```

### Pseudokód 1: Algoritmus DBSCAN

V algoritmu jsou definovány stavy, kdy je bod klasifikován jako šumový, okrajový a středový pomocí počtu bodů v poloměru okolo klasifikovaného bodu. U středového bodu

jsou klasifikovány body v jeho poloměru na základě stejných pravidel a na základě hustotní dostupnosti jsou do stejných shluků přidávány i body jejich poloměru. Šumové body nejsou přiřazovány do žádné třídy; na konci algoritmu tvoří jednu skupinu nezařazených bodů.

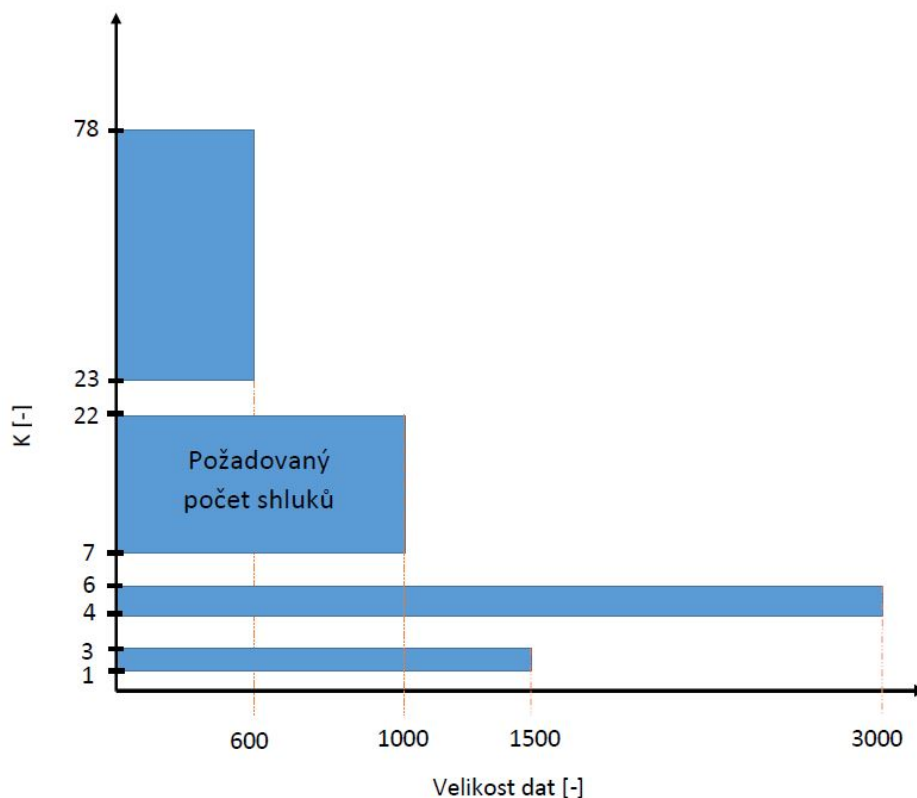
V článku [41] je uveden vzorec pro automatický výpočet hodnoty  $Eps$  - poloměru, který je funkční pro dvoudimenzionální data.

$$Eps = \left( \frac{\left( \prod_{i=1}^{max(x)-min(x)_i} \right) \cdot k \cdot \gamma \cdot (0,5 \cdot n + 1)}{m \cdot \sqrt{\pi^n}} \right)^{\frac{1}{n}} \quad (11)$$

V rovnici č. 11 představuje  $x$  data zapsaná formou matice  $(m, n)$  a  $m$  vyjadřuje počet segmentů a  $n$  počet příznaků.  $k$  je počet bodů v poloměru a  $\gamma$  je interpolační koeficient. [41]

Nyní vstupní hodnotou DBSCANu zůstává pouze počet bodů v poloměru. Podle návrhu od Aliho Touky [42] získáme hodnotu počtu bodů v poloměru, která bude univerzální pro použití napříč testovanými daty.

U každého souboru dat testujeme postupně hodnoty  $k$  (od 1 do  $N$ ). Z hodnot sestavíme grafy. Každý graf vypovídá o jedné sadě testovacích dat. Na ose  $x$  je uvedena průměrná hodnota počtu bodů zařazených v jednom shluku. Na ose  $y$  se nachází hodnota  $k$ , pro kterou získáme daný počet shluků. Z každého grafu získáme rozpětí  $k$ , pro které získáme správný počet tříd a správné zařazení bodů do nich. Ze všech těchto rozpětí vybereme hodnotu, která se vyskytuje ve výběrech napříč všemi testovanými daty.



Obrázek 19: Graf zjišťování ideálního počtu bodů  $K$  v poloměru v závislosti na velikosti vstupních dat, podrobnější popis uveden pod obrázkem.

Graf na obrázku č. 20 ukazuje rozdělení do tříd pro různé hodnoty  $K$  pro jeden soubor testovacích dat o velikosti 3000 bodů. Pro  $k = 1, 2$  a  $3$  byla data rozdělena do 2 skupin. Pro  $K = 4, 5$  a  $6$  všechny body náležely do jedné třídy (všech 3000 bodů je v jedné skupině). Pro rozpětí  $K$  od 7 do 22 dostáváme požadovaný počet tříd 3 a správné rozřazení dat. Pro hodnoty  $K$  vyšší jak 22 opět získáváme špatný počet shluků. Na základě výsledků ze všech takto vytvořených grafů z testovaných dat jsem pro klasifikaci zvolil hodnotu  $k = 15$ .

Simulovaná 2D data jsou metodou DBSCAN klasifikována dle požadavků, pokud se jedná o větší soubory dat (stovky a více bodů). Ale při aplikaci na EEG signál DBSCAN nedosahuje požadované klasifikace. Vizuální vyhodnocení odborníkem určilo úspěšnost klasifikace menší, jak 50 %. Díky široké škále modifikací DBSCANu na různé typy dat se nabízí několik možností úpravy algoritmu (například DMDBSCAN a GRIDBSCAN).

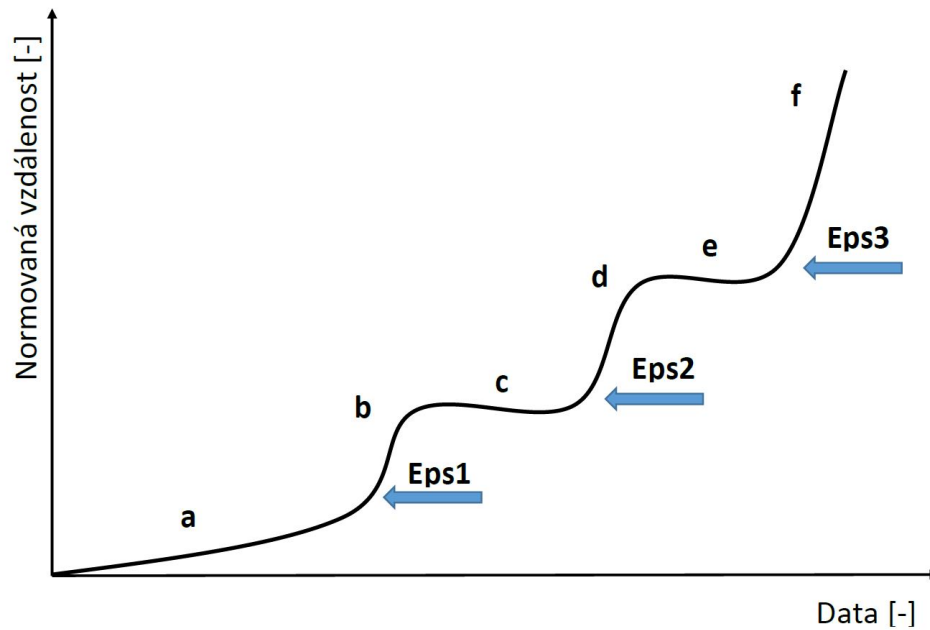
### 3.3.1 DMDBSCAN

DMDBSCAN počítá automaticky parametr  $Eps$  i pro vícerozměrná data, která jsou nerovnoměrně hustotně rozložená. Jelikož algoritmus DBSCAN špatně klasifikoval EEG data, předpokládám, že chyba mohla být způsobena právě nerovnoměrně hustotně rozloženými daty. Pokud klasifikujeme nerovnoměrnou hustotu, DMDBSCAN by měl nalézt vhodný poloměr pro každou úroveň hustoty.

Poloměr je vypočítáván z křivky nejbližších sousedů. Každý bod má od sebe vzdálené ostatní body - jednoznačná hodnota vzdálenosti. Počítáme vzdálenost všech bodů ke každému bodu v souboru. Tyto hodnoty seřadíme vzestupně. Pro každý bod vybereme jeho první tři sousedy a uděláme z nich průměrnou hodnotu [35]. Takto získané průměrné hodnoty prvních třech nejbližších sousedů seřadíme vzestupně podle velikosti. Ze získaného souboru hodnot vytvoříme křivku. V jejích kolenech je hodnota vhodného poloměru pro daný soubor dat. [43]

Existuje i druhá varianta, kdy nebereme průměr prvních třech nejbližších sousedů, ale vezmeme přímo hodnotu třetího nejbližšího souseda. [35]





Obrázek 20: Křivka průměrů prvních třech nejbližších hodnot  $k$ , písmena odlišují intervaly jednotlivých hustotních úrovní,  $Eps$  označují pozici ideální hodnoty poloměru pro danou hustotní úroveň. [35].

Na obrázku č. 21 je ukázána křivka pro hustotně tříúrovňová data. Mezi  $a$  a  $b$  se nacházejí data jedné hustotní úrovně, pro která je vypočtený vhodný poloměr  $Eps1$ . Mezi  $c$  a  $d$ ,  $e$  a  $f$  jsou analogicky další dvě hustotní úrovně s adekvátními poloměry  $Eps2$  a  $Eps3$ .

Počet nejbližších sousedů si můžeme zvolit. Studie prokázala, že od 4. souseda dále se výsledky prakticky neliší a křivka zůstává stejná [44]. Proto ponechám první 3 nejbližší sousedy, jak je uvedeno ve studii [35].

**Data:** *data*

**Result:** křivka 3. nejbližších bodů

**for** 1:*všechny body* **do**

**for** 1:*počet příznaků* **do**

        | vzdalenost = Euklidovská vzdálenost

**end**

    suma vzdalenosti = průměrná vzdálenost přes všechny dimenze

    serazene = vzestupně suma vzdáleností

    B1 = výběr prvních „k“ nejbližších sousedů

    B2 = kumulativní součet z B1

    B3 = součet všech hodnot vektoru B1

    prumer\_ksousedu = B3/k

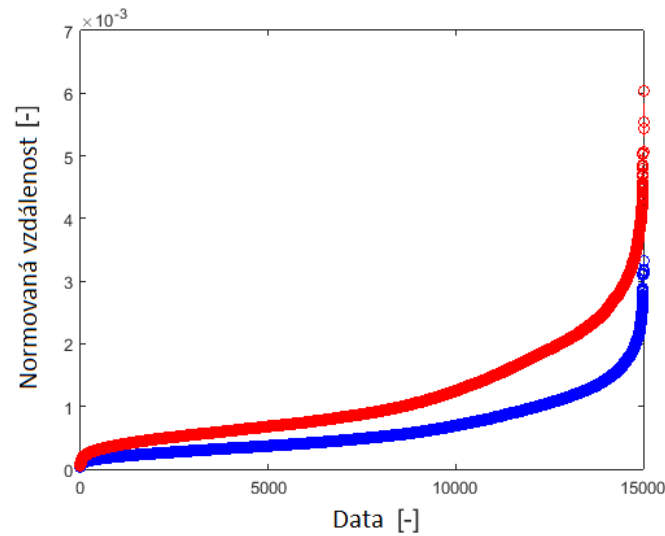
**end**

vzdalkdist = seřazené vzdálenosti pro vykreslení křivky

**Pseudokód 2:** Automatický výpočet poloměru u DMDBSCANu

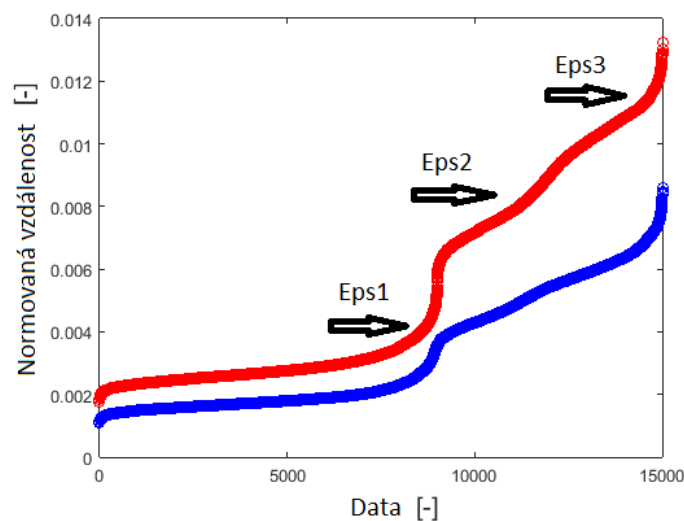
Pokud by křivka ukázala více poloměrů - hustotně různorodě rozložená data, pro každý poloměr by se počítal algoritmus DBSCAN zvlášť. V našem případě se ani jednou (při hodnotě parametru  $k = 3$ ) na reálných EEG datech nestalo, že by křivka měla víc jak jedno "koleno" (viz obrázek č. 22). Ani další studie na reálných datech neměla více hodnot poloměrů [43]. Inflexní body na křivce získáváme až při hodnotách nejbližších sousedů nad 50 při využití prostoru 23 příznaků.

Na obrázku č. 22 a č. 23 je modrou křivkou naznačen průměr prvních  $k$  sousedů. Červená křivka představuje hodnoty  $k$ -tého souseda.



Obrázek 21: Křivka prvních  $k$  sousedů při hodnotě parametru  $k = 3$  [39].

Pokud snížíme počet dimenzí, získáme při hodnotě  $k = 30$  rozdělení prostoru na tři různé hustotní úrovně (viz obrázek č. 23). S vyšším počtem nejbližších bodů (hodnotou  $k$ ) vzrůstá výpočetní náročnost algoritmu. Právě vzhledem ke zvýšení času výpočtu není vhodné metodu využívat pro výpočet poloměrů pro každý signál EEG zvlášť.



Obrázek 22: Křivka prvních  $k$  sousedů při hodnotě parametru  $k = 30$  s rozlišením tří úrovní hustot bodů [39].

Metoda DMDBSCAN se díky své výpočetní náročnosti ukázala pro praxi jako nevhodné řešení. Na testovacích datech byl výrazný kvantitativní rozdíl. DMDBSCAN byl 10 násobně pomalejší než algoritmus k-means. Navíc s počtem dimenzí výpočetní náročnost výrazně stoupá. Další možností, která by mohla řešit nerovnoměrnost rozložení dat prostoru, je algoritmus GRIDBSCAN.

### 3.3.2 GRIDBSCAN

GRIDBSCAN je fúzí myšlenek několika algoritmů. Základní z nich říká, že data mohou být součástí buněk, které rovnoměrně rozdělují prostor. Obsah buňky s nízkou hustotou (malým počtem bodů) je rovnou klasifikován jako šum. [45] Vstupním parametrem je počet buněk, na které je prostor rozdělen.

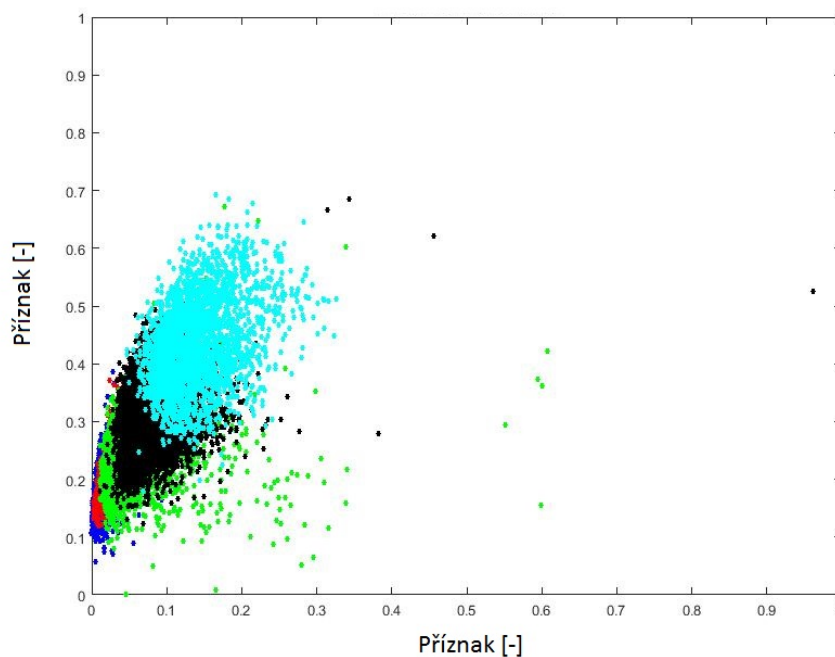
Ve studii [46] používají princip DBSCANu a mřížkového shlukování k filtraci 2D dat. Zachovány jsou jen buňky, které obsahují hustotně nadprahová data. Zároveň náležitost ke třídě je definována podle sousedství. Pokud buňka hustotně nadlimitní sousedí s již klasifikovanou, je přiřazena do stejné třídy. Pokud se postupným posuvem přes všechny buňky dostanu k buňce, která nesousedí s žádnou již klasifikovanou, je vytvořena nová třída. [46]

Obdobného postupu využívají i ve studii [47]. Zde je každá buňka definovaná parametrem  $\varepsilon$ , který matematicky vyjadřuje hustotu bodů. Tento parametr charakterizuje variaci vzdáleností bodů, ale i rovnoměrnost rozložení bodů v rámci buňky.

Je možné také po odstranění šumových buněk pro zbylé (hustotně nadlimitní) buňky počítat koeficient podobnosti. Pokud spolu buňky sousedí, je pomocí vzorců vypočítaná podobnost buněk. Pokud podobnost detekujeme jako nadprahovou (vstupním parametrem je prahová hodnota podobnosti pro sloučení buněk), obsah buněk řadíme do stejné třídy. Díky výpočetní náročnosti ale není postup vhodný pro více dimenzionální data.

U všech těchto postupů algoritmu GRIDBSCAN vzniká problém v námi klasifikovaném 23 dimenzionálním prostoru, neboť po rozdělení buňky obsahují velmi malé počty bodů a těžko se odlišují šumové body a jednotlivé třídy.

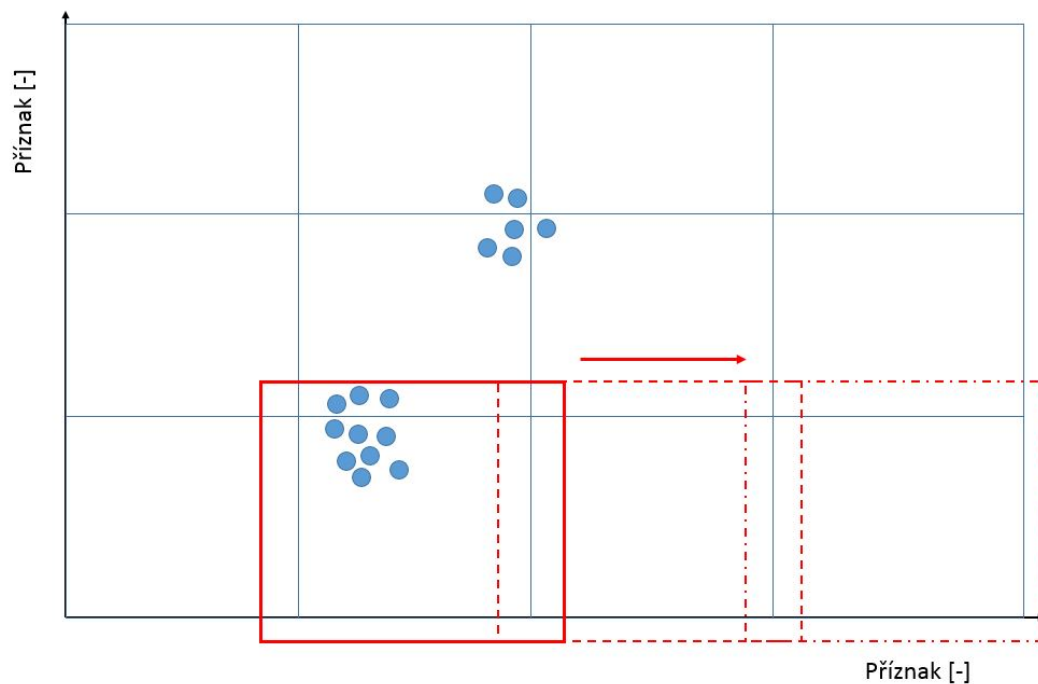
Další možnosti nastínil článek, který využívá mřížku ke zrychlení DBSCANu [48]. Mřížky jsou zde propojeny překryvy. DBSCAN probíhá v každé buňce zvlášť. Díky překryvu jsou některé body klasifikovány vícekrát. K propojení shluků (v každé buňce má třída jiné číslo) dochází právě přes body v oblasti překryvu (viz obrázek č. 25).



Obrázek 23: Vykreslení dvou příznaků z reálných EEG dat [39].

Na obrázku č. 24 jsou v závislosti na sobě vykresleny normované amplituda a frekvence. Barevně jsou odděleny tříd po již proběhlé klasifikaci. EEG data, vykreslíme-li 2 různé příznaky v závislosti na sobě, jsou zhuštěna v jedné části, jak je vidět na obrázku č. 24. Předpokládáme tedy i v 23D prostoru nerovnoměrně rozloženou hustotu dat. GRIDBSCAN se nabízí jako vhodná modifikace, pomocí které proběhne klasifikace EEG segmentů. Problém vzniká ve velkém počtu dimenzí a tím malému počtu bodů v jednotlivých buňkách.

Na základě těchto poznatků jsem zvolil algoritmus GRIDBSCAN ze studie [48], kdy prostor bude dělen na určitý počet buněk s větším překryvem mezi sebou.



Obrázek 24: Naznačení posuvu okna (červeně) v příznakovém prostoru pro místní výpočet DBSCANu v jednotlivých buňkách a překryv těchto buněk v prostoru.

**Data:** *hranice*

**Result:** Klasifikace segmentů

```

for 1:počet buněk do
  hranice = pevně stanovené hranice
  body = body patřící do buňky
  if Buňka obsahuje nějaké body then
    [co,trida,typ] = dbscan(body)
    cislo_bunky = vždy se posouvá s každou buňkou
    matice = [data, cislo bunky, trida, typ]
  end
end

for 1:počet opakujících se bodů do
  if Bod je středový then
    cisla_shluku = ve kterých třídách se bod vyskytuje
    FN = nejmenší číslo třídy
    cislo_bunky = vždy se posouvá s každou buňkou
    for 1:počet tříd, ve kterých se vyskytoval do
      | matice2 = všechny body tříd, kde se vyskytoval přepíšu na hodnotu FN
    end
  end
end

for 1:počet řádků matice2 do
  opakuj = body, které se opakují
  vysledek = uloží jen jeden bod (s nejnižším číslem buňky)
end

```

### Pseudokód 3: Algoritmus GRIDDBSCAN

Pro reálná data s více příznaky (více jak 4D prostor) GRIDBSCAN nedosahuje požadovaných výsledků. Díky velkému prostoru se v jednotlivých buňkách vyskytuje malé množství bodů. Proto jsem se rozhodl pro výběr modifikace GRIDBSCANu, kdy rozměry buněk automaticky přizpůsobím datům. Počet buněk bude nastavován jako fixní para-

metr, nebude tedy prostor dělen do velkého množství buněk (u původního GRIDBSCANu běžně 50 buněk). Současně díky adaptivním rozměrům budou buňky obsahovat dostatečně velký počet bodů a prázdná místa příznakového prostoru nebudou počítána.

Počet buněk je tedy zadáván jako fixní parametr. Čím vyšší počet buněk, tím náročnější je algoritmus na výpočet. Při příliš vysokém počtu buněk se opět objeví problém jako u původního GRIDBSCANu - nedostatečný počet bodů v buňce. Při testování vyšší počet buněk znamenal vyšší výpočetní náročnost - nutný vyšší počet cyklů při slučování. Nižší počet buněk snižoval kvalitu klasifikace, neboť se blížíme klasifikaci obyčejného DBSCANu (bez dělení buněk). Proto jsem počet buněk nastavil na 8, tj. před slučováním budou body rozděleny do 8 částí.

Cílem je vytvořit buňky s homogenní hustotou. Hranice buněk se tedy vytvářejí na základě přepočtu vzdáleností navzájem mezi jednotlivými body [27]. Do stejné buňky jsou přiřazeny body, které jsou si nejbližší, počet bodů v jednotlivých buňkách je tedy odlišný. Jakmile jsou vytvořeny buňky, které obsahují rovnoměrně rozložené body (mezi rovnoměrně rozložené body se v této fázi berou i prázdné buňky), odstraní se prostor bez dat (prázdné buňky). Na základě koeficientu podobnosti (vztahován ke středu buňky) vypočítaného z charakteristik ve frekvenčním spektru jsou slučovány buňky podle uživatelem zvoleného prahu. Práh jsem volil tak, abych klasifikací získal počet tříd, který odpovídá třídám vyskytujícím se v EEG signálu (viz. kapitola 3.1.1).

Účinnosti všech algoritmů jsou porovnány na 2D testovacích datech. Zvolený algoritmus GRIDBSCAN je na reálných EEG datech porovnáván se současnou metodou k-means.

## Hodnocení algoritmů

U algoritmů provádím kvalitativní i kvantitativní zhodnocení. Kvantitativní porovnání provádím pomocí vyhodnocení časové náročnosti jednotlivých algoritmů. Algoritmus 5x po sobě klasifikuje totožná data. Ze zjištěných časů mediánem zvolím hodnotu, která bude



reprezentovat čas výpočtu algoritmu pro příslušný objem dat. Kvalitativní vyhodnocení realizují pomocí statistické metody, kdy srovnávám účinnost algoritmů na testovaná data. EEG signál od jednoho pacienta je klasifikován jako jeden celek. Klasifikace neprobíhá po jednotlivých kanálech, ale napříč celým signálem. Klasifikace EEG dat byla hodnocena dvěma experty: doc. Ing. Vladimírem Krajčou, CSc. a primářem MUDr. Ing. Svojmílem Petránkem, CSc. MBA.

### 3.4 Statistické zhodnocení

Účinnost algoritmů jsem testoval pomocí ROC analýzy. ROC analýza je statistický postup pro binární vyhodnocení dat. Hodnotí se správná a falešná pozitivita a správná a falešná negativita. Jedná se o statistickou analýzu využívanou v diagnostických testech, při rozhodování lékařů. Byla využita i při hodnocení EEG signálů [49]. Po konzultaci s doc. Ing. Janou Vránovou, CSc z Ústavu lékařské biofyziky a lékařské informatiky a díky již aplikovanému postupu na EEG data jsem se rozhodl pro statistické vyhodnocení klasifikace EEG signálu pomocí ROC analýzy.

Data můžeme rozdělit podle toho, zda do shluku patří, nebo jsou do něj zařazena mylně a náleží k jinému shluku. Pro každý segment máme tedy možnost binárního stavu 1 - správně zařazen, nebo 0 - špatně zařazen.

Na obrázku č. 26 je zobrazena tzv. konfuzní matice (Confusion matrix). Ta obsahuje informace u skutečné náležitosti do tříd (vyhodnocení experty) a předpokládané rozdělení (klasifikace pomocí algoritmů). Hodnotí se zde výkonnost algoritmů v rámci jedné třídy. [50]

		Předpovídaný stav	
		Pozitivní	Negativní
Skutečný stav	Přítomný	TP 10409	FP 8
	Nepřítomný	FN 24378	TN 4951

Obrázek 25: Konfuzní matice zobrazující hodnoty TP, TN, FP a FN získané z klasifikace jednoho celého záznamu pro třídu fyziologické aktivity.

Obrázek č. 26 ukazuje reálná data z 1 celého záznamu pro vyhodnocení třídy fyziologické aktivity. Body (popis v rozích matice v obrázku č. 26) s ohledem na jejich zařazení ve shluku a vztahu k datům jsou popsány pomocí 4 vztahů [51]:

#### TP (True positive)

Takto se označují správně zařazená data ve shluku. Pokud daný shluk představuje například skupinu špatně připojených elektrod, tak jako TP jsou označeny právě segmenty s projevem pulzních artefaktů.

#### FP (False positive)

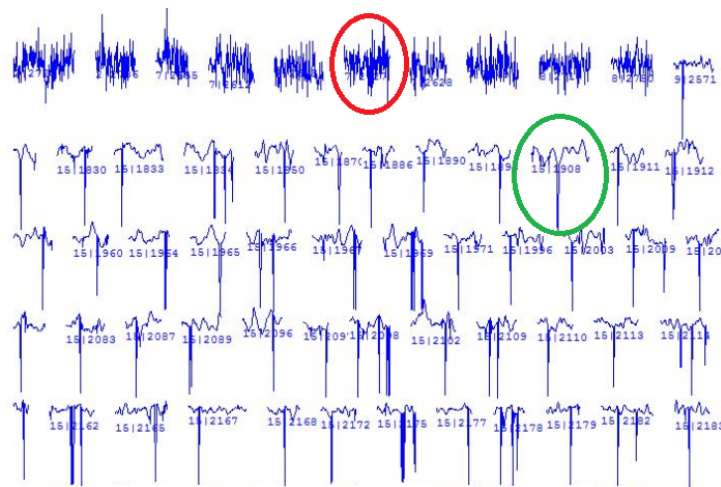
Body, které jsou do shluku zařazeny špatně, jsou označeny jako falešně pozitivní. Pokud tedy ve skupině segmentů představující pulzní artefakty nalezneme například segment s EMG aktivitou, jedná se o špatně - falešně zařazený segment.

### FN (False negative)

Jako falešně negativní jsou označeny body, které do shluku patří, ale jsou přiřazeny ke shluku jinému. Tak by tedy byly označeny segmenty s pulzními artefakty, které nebyly algoritmem zařazeny do skupiny těchto grafoelementů.

### TN (True negative)

Správně negativní jsou takové segmenty, které nepatří do zkoumané třídy a nejsou do ní zařazeny. Podle příkladu se jedná o všechny zbylé klasifikované segmenty, které nejsou ve třídě pulzních artefaktů a nepatří do ní.



Obrázek 26: Třída s vyznačenými TP (zeleně) a FN (červeně) segmenty, zobrazeno ve WF [37].

Z hodnot TP, FN, TN a FP můžeme vypočítat další charakteristiky vypovídající o algoritmu a jeho klasifikaci.

#### 3.4.1 Specificita

Diagnostická specificita udává pravděpodobnost negativního výsledku u zdravých osob (nemají zkoumanou chorobu). Ukazuje pravděpodobnost s jakou jsou segmenty, které nepatří do zkoumaného shluku, zařazeny do jiného shluku. Specificita je tedy poměr počtu

skutečně negativních ku součtu skutečně negativních a falešně pozitivních segmentů. [51, 52]

$$\textit{Specificita} = \frac{TN}{TN + FP} \quad (12)$$

### 3.4.2 Senzitivita

Diagnostická senzitivita udává pravděpodobnost úspěšné detekce choroby. Udává citlivost s jakou jsou ve shluku zařazeny všechny segmenty do něj patřící. Ukazuje poměr počtu detekovaných (například epileptických grafoelementů) segmentů ve třídě a celkového počtu segmentů s epileptickou aktivitou v zkoumaném signálu. Senzitivita je poměr skutečně pozitivních ku součtu skutečně pozitivních a falešně negativních segmentů. [51, 52]

$$\textit{Senzitivita} = \frac{TP}{TP + FN} \quad (13)$$

### 3.4.3 Pozitivní prediktivní hodnota (PPV)

Prediktivní pozitivní hodnota popisuje výkon diagnostického testu. PPV v našem případě můžeme přirovnat k homogenitě třídy. Jedná se o podíl správně klasifikovaných dat v celé třídě. PPV je poměr skutečně pozitivních hodnot ku všem hodnotám zařazeným do zkoumaného shluku. [53]

$$PPV = \frac{TP}{TP + FP} \quad (14)$$

PPV má, spolu se senzitivitou, pro náš soubor dat nejvíce vypovídající hodnotu pro vyhodnocení účinnosti klasifikace v rámci tříd.

### 3.5 Vyhodnocení 2D testovacích dat

Osy u 2D testovacích  $x$  a  $y$  reprezentují příznaky, tak jako kdybychom z 23 prostoru v EEG datech vykreslili pouze 2 z 23 příznaků v závislosti na sobě. Podrobnější popis viz kapitola 3.1. U 2D dat jsou hodnoceny jednotlivé parametry v rámci shluků a následně i napříč klasifikovanými daty.

Kapitola s výsledky z 2D testovacích dat obsahuje kvantitativní i kvalitativní vyhodnocení pro každý soubor dat dohromady. Tabulky s časy klasifikace vyjadřují výpočetní náročnost algoritmu vztaženou na počet klasifikovaných segmentů. Kvalitativní analýza popisuje účinnost algoritmu na danou sadu testovacích dat. Každá sada dat reprezentuje specifický problém (prolnuté shluky, nerovnoměrně rozložená data, velký a malý počet segmentů, apod.).

### 3.6 Vyhodnocení klasifikace EEG dat

Vzhledem k velkému počtu dat (jeden pacient = desítky tisíc segmentů) jsem klasifikaci rozdělil na 2 části:

- vyhodnocení celých záznamů
- výběr a vyhodnocení náhodných segmentů.

#### 3.6.1 Vyhodnocení celých záznamů

U třech pacientů byly expertem vyhodnocovány výsledky klasifikace všech segmentů. Každý pacient představuje jednu skupinu: EEG signál s výraznou epileptickou aktivitou, signál s pulzními artefakty a výraznou EMG aktivitou a signál s nízkou amplitudou epileptických gaoelementů a vysokým procentem segmentů fyziologických vln.

V každé třídě byly označeny segmenty, které do třídy patří (TP). U segmentů, které do třídy byly zařazeny mylně (FP), byla udána příslušnost k jiné třídě. Tak byl určen reálný počet jednotlivých grafoelementů, nezávisle na počtu detekovaném algoritmem. Pro jednotlivé třídy v každém ze tří záznamů byly vypočteny hodnoty PPV a určena úspěšnost klasifikace na základě konzistentnosti tříd.

### 3.6.2 Výběr a vyhodnocení náhodných segmentů

K potvrzení funkčnosti algoritmu je zapotřebí testování napříč širším spektrem probandů. Provedl jsem klasifikaci na dalších 12 pacientech (celkem tedy bylo klasifikováno 15 záznamů od 15 pacientů).

Abych zmenšil objem dat, od každého pacienta jsem náhodně vybral 50 segmentů (postup popsán níže), u kterých expert stanovil příslušnost k dané třídě.

Klasifikované segmenty vizualizované pomocí WF jsou uspořádány po stovkách na jednotlivých stránkách. Počet stran je dán velikostí třídy a délkou segmentů. Pro každého pacienta jsou tyto hodnoty jedinečné. Ze souboru .qua získáme počty segmentů v jednotlivých kanálech pro každou třídu. Jejich součtem (napříč kanály) získáme počet segmentů v každé třídě. Počet stran spočítáme podle počtu stran v programu WF. Tato hodnota je vstupem do funkce generující náhodnou pozici segmentů. Skript pomocí funkce *rand* generuje náhodná čísla stránek a pozic na stránce. Čísla stran jsou v rozmezí 1 až maximální počet stran pro daného pacienta. Maximální hodnota počtu segmentů na stránce je nastavena na 200 (při krátkých úsecích se počet segmentů pohyboval okolo 150 segmentů na stránku). Pokud je počet segmentů na stránce nižší, pokračujeme v číslování od začátku stránky dále až do konečné hodnoty. Pokud bychom tedy měli najít na stránce 146. segment a strana by měla 130 segmentů, vybereme 16. segment.

Odborník vizuálně zkontroloval celou klasifikaci (všechny segmenty signálu pacienta) a určil, která třída (číselně) odpovídá funkční třídě (třída epileptických grafoelementů,

EMG a šumových grafoelementů, fyziologické aktivity apod.). Následně vyhodnotil náhodně vybraných 50 segmentů.

Pro náhodně vybrané segmenty (tvořily třídy stejně jako v celé klasifikaci) probíhala klasifikace stejně jako u vyhodnocení celých záznamů. Byly určeny hodnoty TP, FP, TN, FN ve třídách a parametry senzitivita, specificita a PPV.

## 4 Výsledky

### 4.1 Klasifikace 2D - testovací data

Postupně byla klasifikována jednotlivá 2D testovací data všemi algoritmy. Nad každým obrázkem testovaných dat jsou uvedeny tabulky s rychlostmi klasifikace reprezentující kvantitativní zhodnocení algoritmů. V další tabulce pod obrázkem jsou uvedeny hodnoty senzitivity (SENZ), specifity (SPEC) a PPV pro jednotlivé algoritmy. Hodnoty těchto parametrů jsou zaokrouhleny na 3 desetinná místa.

#### Data 1

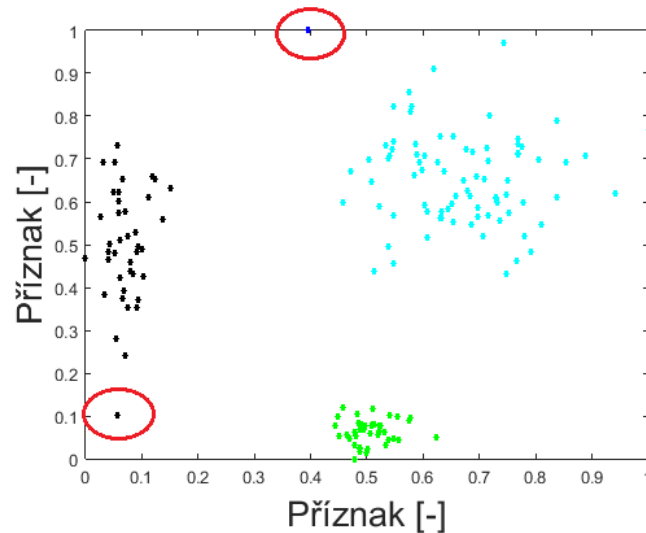
První data (163 bodů) obsahují velmi malý počet bodů, který obecně není ideální pro hustotní algoritmy. Shluky jsou v prostoru dobře separované, což zpřesňuje klasifikaci metodou k-means. Zároveň ale data obsahují i 2 šumové body. Algoritmus k-means šum na rozdíl od hustotních algoritmů neodděluje.

Tabulka 4: Kvantitativní porovnání jednotlivých metod klasifikace na 2D datech - 163 bodů (obrázek č. 28).

Algoritmus	Čas klasifikace [s]
k-means	0,017
k-means + siluety	0,080
DBSCAN	0,027
DMDBSCAN	0,076
GRIDBSCAN	0,059

Z tabulky č. 4 vyplývá, že algoritmus k-means, u kterého je potřeba zadat vstupní parametr, je nejrychlejší. DMDBSCAN je díky automatickému výpočtu všech parametrů z původní sady dat časově nejnáročnějším z hustotních algoritmů. Automatizované k-means (výpočet ideálního počtu shluků pomocí siluet) je nejpomalejším algoritmem.





Obrázek 27: Data 1 s oddělenými shluky a 2 šumovými body.

Všechny testované algoritmy dokáží detekovat takto separované shluky. Červené kroužky na obrázku č. 28 označují šumové body. Jak je vidět na obrázku, hustotní algoritmus zde odhalil pouze jeden (horní) z nich. Všechny metody klasifikovaly až na šumové body správně všechny 3 třídy.

Tabulka 5: PPV, senzitivita a specificita na 2D testovacích datech 1 (obrázek č. 28)

Parametry		K-means	DBSCAN	DMDBSCAN	GRIDBSCAN
Shluk 1	SENZ [-]	1,000	1,000	1,000	1,000
	SPEC [-]	0,988	1,000	1,000	1,000
	PPV [-]	0,9875	1,000	1,000	1,000
Shluk 2	SENZ [-]	1,000	1,000	1,000	1,000
	SPEC [-]	1,000	1,000	1,000	1,000
	PPV [-]	1,000	1,000	1,000	1,000
Shluk 3	SENZ [-]	1,000	1,000	1,000	1,000
	SPEC [-]	0,991	0,991	0,991	0,991
	PPV [-]	0,976	0,976	0,976	0,976
Shluk 4	SENZ [-]	0,000	0,500	0,500	0,500
	SPEC [-]	-	1,000	1,000	1,000
	PPV [-]	-	1,000	1,000	1,000

Z tabulky č. 5 vyplývá, že hustotní algoritmy detekovaly jeden šumový bod, měly tedy lepší citlivost (senzitivitu). Ale ani jeden algoritmus nedokázal data zcela správně

klasifikovat. K-means neumí extrahovat z dat šum a hustotní algoritmy zde mají problém s malou velikostí souboru dat.

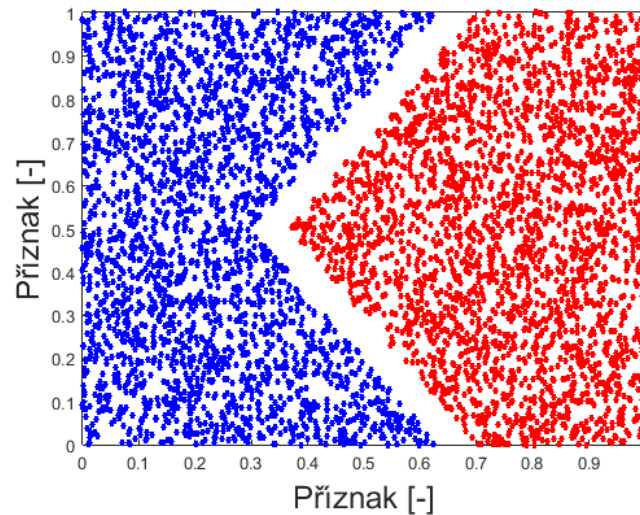
## Data 2

Druhá data (5000 bodů) obsahují oproti prvním velký počet bodů, který je ideální pro hustotní algoritmy. Zároveň je naznačeno částečné prolnutí shluků. Prolnuté shluky se v 2D prostoru myslí shluky, které nemůžeme oddělit přímkou. U takto drobného prolnutí záleží na iniciačním těžišti při výpočtu metodou k-means. Podle jeho polohy zvládne/nezvládne tento algoritmus data správně oddělit.

Tabulka 6: Kvantitativní porovnání jednotlivých metod klasifikace na 2D testovacích datech - 5000 bodů (obrázek č. 29).

Algoritmus	Čas klasifikace [s]
k-means	1,163
k-means + siluety	7,317
DBSCAN	2,586
DMDBSCAN	5,774
GRIDBSCAN	1,465

V tabulce č. 6 je patrné, že s vysokým počtem klasifikovaných bodů výrazně stoupá výpočetní náročnost algoritmu DMDBSCAN. Délka výpočtu algoritmu DBSCAN a k-means je téměř vyrovnaná.



Obrázek 28: Data vhodná ke klasifikaci hustotními algoritmy.

Při 3 pokusech k-means 2x klasifikovalo data do správných shluků. Proto u všech testovaných algoritmů získáváme stejný obrázek č. 29. S vyšším počtem bodů ke klasifikaci výrazně narůstá čas výpočtu u všech automatických algoritmů (u kterých není nutné zadávat ručně vstupní parametr).

Tabulka 7: PPV, senzitivita a specificita na 2D testovacích datech 2 (obrázek č. 29)

Parametry		<b>K-means</b>	<b>DBSCAN</b>	<b>DMDBSCAN</b>	<b>GRIDBSCAN</b>
<b>Shluk 1</b>	<b>SENZ [-]</b>	1,000	1,000	1,000	1,000
	<b>SPEC [-]</b>	1,000	1,000	1,000	1,000
	<b>PPV [-]</b>	1,000	1,000	1,000	1,000
<b>Shluk 2</b>	<b>SENZ [-]</b>	1,000	1,000	1,000	1,000
	<b>SPEC [-]</b>	1,000	1,000	1,000	1,000
	<b>PPV [-]</b>	1,000	1,000	1,000	1,000

V tabulce č. 7 vidíme, že všechny algoritmy detekovaly oba shluky správně, ani jeden bod nebyl zařazen do špatného shluku.

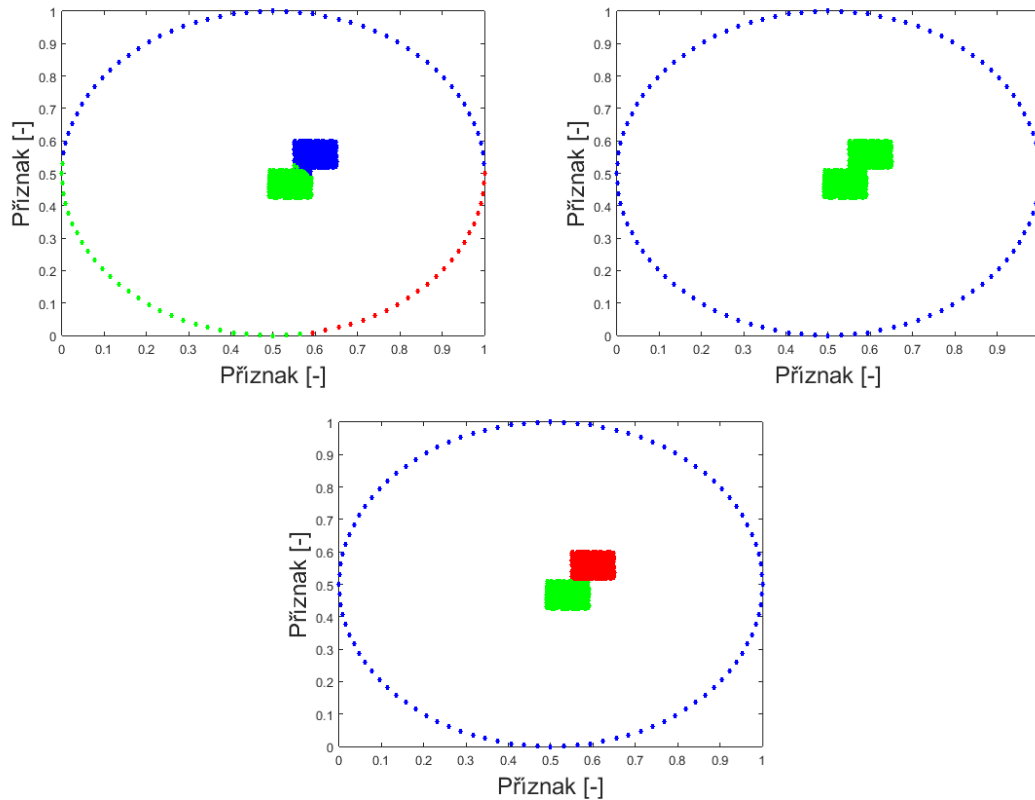
**Data 3**

Třetí data (2101 bodů) obsahují hranou spojené shluky vnořené do kruhového shluku. K-means si nedokáže s takovýmito daty poradit, neoddělí spojené obdélníky ani jim opsanou kružnici. DBSCAN má problém v odlišení dvou na sebe navazujících shluků. Jediný algoritmus GRIDBSCAN klasifikoval tato data zcela správně.

Tabulka 8: Kvantitativní porovnání jednotlivých metod klasifikace na 2D testovacích datech - 2101 bodů (obrázek č. 30).

Algoritmus	Čas klasifikace [s]
k-means	0,207
k-means + siluety	1,100
DBSCAN	0,491
DMDBSCAN	1,158
GRIDBSCAN	0,449

Z tabulky č. 8 je opět patrná dominance algoritmu k-means. DBSCAN klasifikuje rychleji, než GRIDBSCAN. DMDBSCAN i k-means se siluetami jsou algoritmy časově velmi náročné i při malém souboru dat.



Obrázek 29: Porovnání klasifikace dat 3 pomocí metod, zleva k-means, DBSCAN, dole GRIDBSCAN.

Na obrázku č. 30 vidíme vlevo nahoře klasifikaci pomocí k-means. Klasifikace je zcela nepřesná, algoritmus k-means neklasifikoval ani jeden shluk správně. DBSCAN (vpravo nahoře) neoddelil na sebe těsně nasedající shluky. Jediný algoritmus GRIDBSCAN klasifikoval data správně (v obrázku č. 30 dole).

Tabulka 9: PPV, senzitivita a specificita na 2D testovacích datech 3 (obrázek č. 30)

Parametry		<b>K-means</b>	<b>DBSCAN</b>	<b>DMDBSCAN</b>	<b>GRIDBSCAN</b>
<b>Shluk 1</b>	<b>SENZ [-]</b>	0,323	1,000	1,000	1,000
	<b>SPEC [-]</b>	1,000	1,000	1,000	1,000
	<b>PPV [-]</b>	1,000	1,000	1,000	1,000
<b>Shluk 2</b>	<b>SENZ [-]</b>	0,973	1,000	1,000	1,000
	<b>SPEC [-]</b>	1,000	0,087	0,087	1,000
	<b>PPV [-]</b>	1,000	0,500	0,500	1,000
<b>Shluk 3</b>	<b>SENZ [-]</b>	0,985	0,000	0,000	1,000
	<b>SPEC [-]</b>	0,973	-	-	1,000
	<b>PPV [-]</b>	0,976	-	-	1,000

Jediný algoritmus GRIDBSCAN klasifikoval data 3 správně (viz tabulka č. 9). DBSCAN i DMDBSCAN detekovaly pouze 2 shluky.

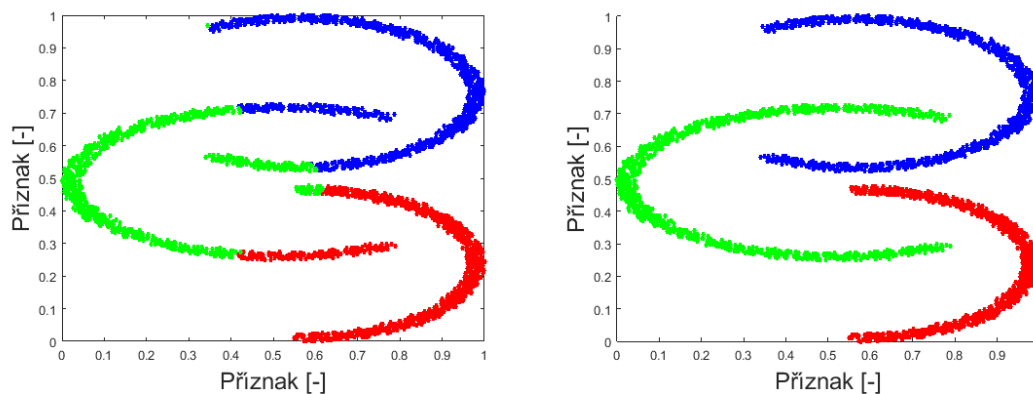
#### Data 4

Data 4 (3000 bodů) obsahují prolnuté shluky. K-means nedokáže odlišit takto prostorově prolnutá data. DBSCAN i jeho další modifikace tato data klasifikují správně.

Tabulka 10: Kvantitativní porovnání jednotlivých metod klasifikace na 2D testovacích datech - 3000 bodů (obrázek č. 31).

Algoritmus	Čas klasifikace [s]
k-means	0,508
k-means + siluety	2,758
DBSCAN	0,914
DMDBSCAN	2,002
GRIDBSCAN	0,767

V tabulce 10 je opět patrná nejmenší rychlost výpočtu algoritmu k-means. DBSCAN a GRIDBSCAN mají téměř podobné časy. DMDBSCAN a k-means se siluetami jsou několikanásobně pomalejší vůči ostatním algoritmům.



Obrázek 30: Porovnání klasifikace metod, první obrázek klasifikace k-means, druhý klasifikace všech modifikací DBSCANu.

Na obrázku č. 31 vidíme 3 prolnuté shluky. V levé části je vidět nepřesná klasifikace metodou k-means. Ani jeden shluk není klasifikován zcela správně. Hustotní algoritmy si s takovými daty poradí a klasifikují správně.

Tabulka 11: PPV, senzitivita a specificita na 2D testovacích datech 4 (obrázek č. 31)

Parametry		<b>K-means</b>	<b>DBSCAN</b>	<b>DMDBSCAN</b>	<b>GRIDBSCAN</b>
<b>Shluk 1</b>	<b>SENZ [-]</b>	0,798	1,000	1,000	1,000
	<b>SPEC [-]</b>	0,885	1,000	1,000	1,000
	<b>PPV [-]</b>	0,776	1,000	1,000	1,000
<b>Shluk 2</b>	<b>SENZ [-]</b>	0,801	1,000	1,000	1,000
	<b>SPEC [-]</b>	0,972	1,000	1,000	1,000
	<b>PPV [-]</b>	0,935	1,000	1,000	1,000
<b>Shluk 3</b>	<b>SENZ [-]</b>	0,897	1,000	1,000	1,000
	<b>SPEC [-]</b>	0,886	1,000	1,000	1,000
	<b>PPV [-]</b>	0,797	1,000	1,000	1,000

Tabulka č. 11 ukazuje, že senzitivita i specificita při detekování prolnutých shluků algoritmem k-means nedosahuje 100 %, jako tomu je u hustotních algoritmů.

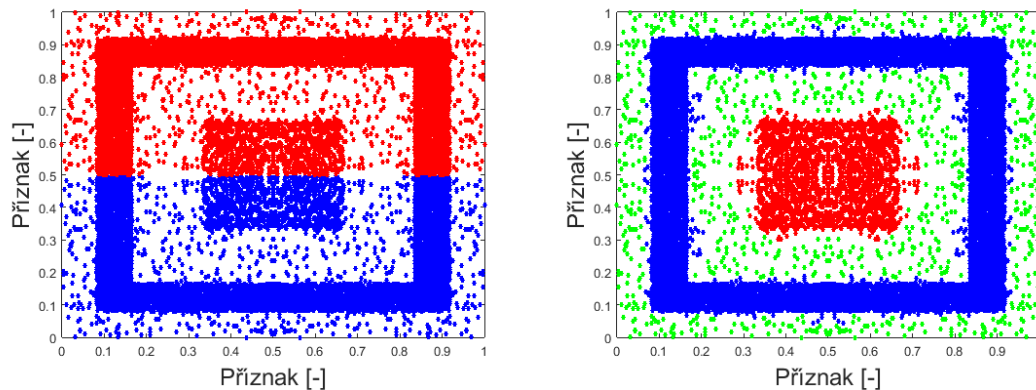
## Data 5

Poslední data (18224 bodů) obsahují 2 úrovně, které k-means neklasifikuje správně. Jedná se o prolnuté shluky a šum. DBSCAN a všechny jeho testované modifikace detekují šum a správně rozliší prolnuté shluky. V reálných EEG datech předpokládáme výskyt šumu i prostorově prolnutých shluků. Tato data by tedy hustotní algoritmy měly oproti algoritmu k-means klasifikovat správně.

Tabulka 12: Kvantitativní porovnání jednotlivých metod klasifikace na 2D testovacích datech - 18224 (obrázek č. 32).

<b>Algoritmus</b>	<b>Čas klasifikace [s]</b>
k-means	4,892
k-means + siluety	22,234
DBSCAN	13,097
DMDBSCAN	19,023
GRIDBSCAN	8,238

Z tabulky č. 12 je patrné, že u dat 5 algoritmus k-means a GRIDBSCAN klasifikovaly stejnou rychlostí.



Obrázek 31: Porovnání klasifikace metod. Obrázek vlevo představuje klasifikaci prostřednictvím algoritmu k-means, obrázek vpravo klasifikaci prostřednictvím všech variant DBSCANu.

Na obrázku č. 32 vpravo je vidět červený obdélník, který je vnořen do modrého obdélníku. Zeleně jsou označeny šumové body. V obrázku z klasifikace metodou k-means je viditelné, že šumové body jsou řazeny rovnoměrně mezi shluky. Body byly algoritmem k-means rozděleny na 2 téměř shodné poloviny.

Tabulka 13: PPV, senzitivita a specificita na 2D testovacích datech 5 (obrázek č. 32)

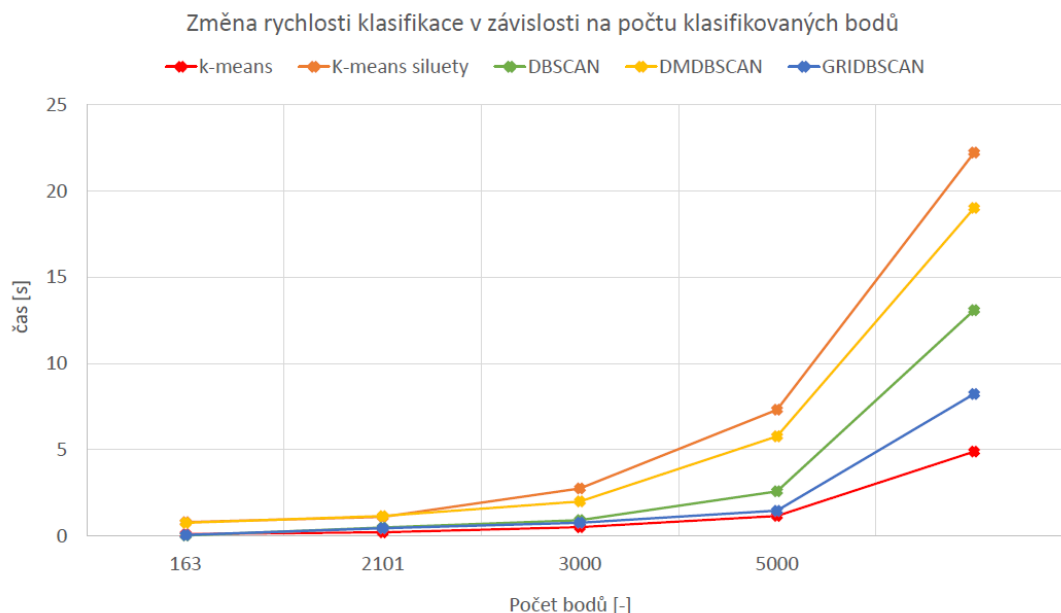
Parametry		<b>K-means</b>	<b>DBSCAN</b>	<b>DMDBSCAN</b>	<b>GRIDBSCAN</b>
<b>Shluk 1</b>	<b>SENZ [-]</b>	0,798	1,000	1,000	1,000
	<b>SPEC [-]</b>	0,987	0,989	0,989	0,988
	<b>PPV [-]</b>	0,776	0,990	0,990	0,990
<b>Shluk 2</b>	<b>SENZ [-]</b>	0,801	1,000	1,000	1,000
	<b>SPEC [-]</b>	0,997	0,996	0,996	0,995
	<b>PPV [-]</b>	0,934	0,991	0,991	0,990
<b>Shluk 3</b>	<b>SENZ [-]</b>	0,897	0,932	0,932	0,925
	<b>SPEC [-]</b>	0,897	1,000	1,000	1,000
	<b>PPV [-]</b>	0,987	0,931	0,931	1,000

Tabulka č. 13 potvrdila, že algoritmus k-means není schopný tato data správně klasifikovat. Oproti tomu hustotně založené algoritmy správně oddělily shluky od šumu.



## Shrnutí klasifikace 2D testovacích dat

V následujícím grafu na obrázku č. 33 je viditelná změna rychlosti klasifikace pro jednotlivé algoritmy v závislosti na velikosti klasifikovaných dat.



Obrázek 32: Porovnání trendu rychlosti klasifikace jednotlivých metod v závislosti na velikosti klasifikovaných dat.

K-means je výpočetně nejméně náročný, klasifikace probíhala vždy v nejkratším čase. DBSCAN a GRIDBSCAN klasifikují s podobnou asymptotickou složitostí  $n^2$ . DMDBSCAN a k-means se siluetami jsou časově velmi náročné. Z grafu je patrné, že se zvyšujícím se počtem dat se zvyšuje poměr časové náročnosti k-means a GRIDBSCANu.

Pomocí mediánu vypočtená celková senzitivita s specificita napříč testovanými 2D daty je uvedena v tabulce č. 14.

Tabulka 14: Celková senzitivita a specificita na 2D testovacích datech.

Parametr	K-means	DBSCAN	DMDBSCAN	GRIDBSCAN
SENZ	0,950	0,994	0,994	1,000
SPEC	0,993	0,998	0,998	1,000

Z tabulky č. 14 je patrné, že hustotně založené algoritmy klasifikovaly s větší přesností než algoritmus k-means. Z hustotních algoritmů si s těsně navazujícími, prolnutými a zároveň zašuměnými daty nejlépe poradil algoritmus GRIDBSCAN. Měl by tedy být nejvhodnější modifikací pro aplikaci na EEG data.

## 4.2 Klasifikace EEG - reálná data

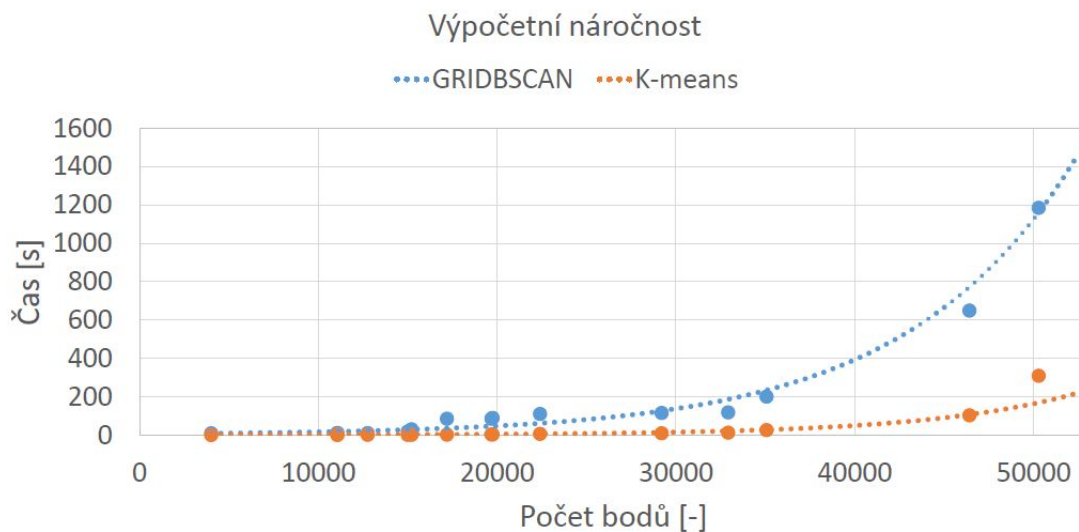
Celkem bylo klasifikováno 15 pacientů algoritmy k-means a GRIDBSCAN. První část obsahuje kvantitativní vyhodnocení, tedy hodnoty časů a počty segmentů v 15 klasifikovaných signálech. Další část ukazuje výsledky z 50 náhodných segmentů z každého z 12 pacientů. V třetí části jsou výsledky klasifikace na celých 3 záznamech ze třech pacientů. V tabulkách reprezentujících kvalitativní část analýzy jsou opět uvedeny hodnoty senzitivity (SENZ), specifity (SPEC) a PPV pro jednotlivé algoritmy (k-means, GRIDBSCAN).

### Kvantitativní analýza 15 pacientů

U každého algoritmu byly změřeny 5x časy klasifikace pro všechny sady dat. V tabulce č. 15 níže je medián zjištěných časů pro každou velikost datového souboru.

Tabulka 15: Kvantitativní vyhodnocení výpočetní náročnosti algoritmů na reálných datech.

Počet segmentů [-]	Čas GRIDDBSCAN [s]	Čas k-means [s]
50292	1186,003	310,098
46413	650,276	103,876
35062	201,068	26,789
32920	119,615	13,761
29194	115,923	10,012
22391	110,831	7,538
19736	90,665	3,920
19668	86,038	3,769
17183	85,947	2,897
15204	30,533	2,562
14992	19,796	2,105
12754	11,780	1,979
11058	11,085	1,890
11051	10,982	1,890
3995	9,816	1,126



Obrázek 33: Porovnání výpočetní náročnosti algoritmů k-means a GRIDDBSCAN na reálných datech, rovnice křivky pro GRIDDBSCAN je  $y = 5,8599 \cdot e^{0,0001 \cdot x}$ ,  $R^2 = 0,8933$ ; pro algoritmus k-means  $y = 0,4384 \cdot e^{0,0001 \cdot x}$ ,  $R^2 = 0,9695$ .

V grafu na obrázku č. 34 je vidět, že výpočetní náročnost algoritmu GRIDBSCAN výrazně stoupá s velikostí klasifikovaných dat. Algoritmus k-means je méně výpočetně, tudíž i časově náročný.

### Kvalitativní analýza 12 pacientů

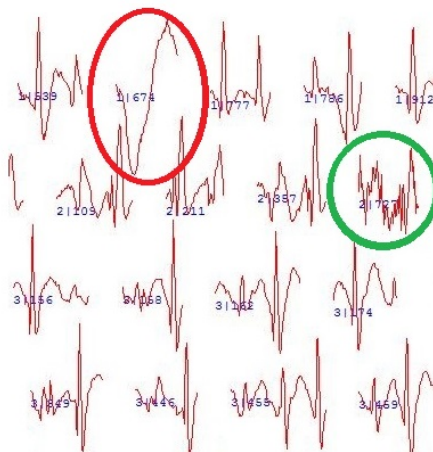
U 12 pacientů bylo dvěma experty hodnoceno 50 náhodných segmentů s různými počty tříd. Následující tabulka č. 16 ukazuje výsledky analýzy jednoho pacienta. Takto zpracované jsou výsledky ze všech 12 pacientů (viz příloha). Zkrácené FYZ znamená shluk fyziologické aktivity, OČNÍ znamená třídu pomalých očních artefaktů, EMG značí třídu svalových a zašuměných segmentů a EPIL značí shluk epileptických grafoelementů.

Tabulka 16: Ukázkové výsledky analýzy pacienta 10 (ostatní tabulky viz příloha na stránce č. 115).

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,400	0,500
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000
OČNÍ	SENZ [-]	1,000	0,036
	SPEC [-]	0,979	1,000
	PPV [-]	0,750	1,000
EMG	SENZ [-]	1,000	1,000
	SPEC [-]	1,000	0,978
	PPV [-]	1,000	0,833
EPIL	SENZ [-]	1,000	1,000
	SPEC [-]	0,896	0,896
	PPV [-]	0,286	0,286

Z tabulky č. 16 vyplývá vyšší senzitivita algoritmu GRIDBSCAN pro segmenty fyziologické aktivity. Naopak pro pomalou oční aktivitu má GRIDBSCAN citlivost nižší. Epileptickou aktivitu rozdělují oba algoritmy se stejnou účinností. EMG aktivitu lépe klasifikuje algoritmus k-means.

V následujícím obrázku č. 35 je ukázka mylně zařazených segmentů (FP) do shluku epileptických grafoelementů pacienta 4. Červeně vyznačený je segment pomalé oční aktivity. Zeleně zakroužkovaný segment je na hranici, kdy záleží na odborníkovi, který klasifikaci vyhodnocuje, do jaké třídy jej zařadí. Segment může náležet do třídy EMG a šumových grafoelementů, nebo tento segment i přes zašumění znatelně reprezentuje třídu epileptických grafoelementů.



Obrázek 34: FP segmenty ve třídě epileptických grafoelementů, zobrazeno ve WF [37].

Pokud bychom náhodným výběrem zvolili červeně označený segment v této třídě epileptických grafoelementů, získáme nepřesně vypovídající hodnoty o této skupině, která je jinak až na dva zakroužkované segmenty zcela homogenní (složena z jednoho druhu grafoelementů).

Z 12 pacientů se jen u některých vyskytovaly epileptické grafoelementy. Některé záznamy obsahují pouze fyziologickou aktivitu a pomalé oční artefakty. V náhodném výběru nemusejí být vždy obsaženy segmenty reprezentující všechny třídy obsažené v signálu. Neboť například třída EMG a šumových grafoelementů byla pouze na jedné stránce a 50 segmentů bylo náhodně vybíráno z celkového počtu 240 stran.

Další dvě tabulky č. 17 a č. 18 ukazují výslednou analýzu pro jednotlivé třídy ze všech 12 pacientů. Jedná se o medián ze všech 12 hodnot specifity, senzitivity a PPV u jednotlivých tříd pro každého z 12 pacientů.

Tabulka 17: Výsledná senzitivita, specifita a PPV jednotlivých tříd pro GRIDBSCAN

Parametry	FYZ	OČNÍ	EMG	EPIL
<b>SENZ</b> [-]	0,935	0,591	0,375	1,000
<b>SPEC</b> [-]	0,550	1,000	0,990	0,896
<b>PPV</b> [-]	0,921	1,000	0,833	0,400

Tabulka 18: Výsledná senzitivita, specifita a PPV jednotlivých tříd pro k-means

Parametry	FYZ	OČNÍ	EMG	EPIL
<b>SENZ</b> [-]	0,400	0,917	1,000	1,000
<b>SPEC</b> [-]	1,000	0,990	1,000	1,000
<b>PPV</b> [-]	1,000	0,9375	1,000	1,000

Z obou tabulek je patrné, že označí-li algoritmus k-means nějaký segment jako patřící do daného shluku reprezentujícího určitý grafoelement, pak s vysokou pravděpodobností (PPV) je tento segment daným grafoelementem. K-means má nižší senzitivitu pro segmenty s fyziologickou aktivitou, ostatní hodnoty vypovídají o jeho lepší účinnosti oproti algoritmu GRIDBSCAN.

### Kvalitativní analýza 3 pacientů

Na 3 pacientech byla provedena analýza všech segmentů klasifikace celého záznamu. Vyhodnocení celého záznamu má vyšší vypovídající hodnotu, než analýza 50 náhodných segmentů záznamu, proto jsou níže uvedeny tabulky všech hodnot zjištěných kvalitativní analýzou na 3 pacientech, včetně následně dopočítaných ukazatelů kvality klasifikace.

V tabulkách jsou popsány grafoelementy, které se v daném signálu vyskytovaly. Zkrácené FYZ znamená shluk fyziologické aktivity, OČNÍ znamená třídu pomalých očních

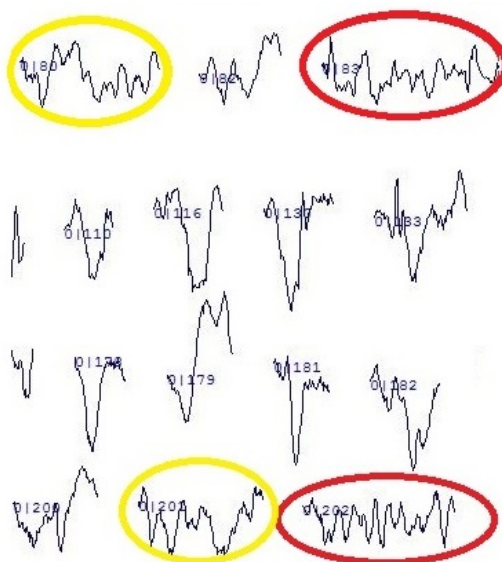
artefaktů, EMG značí třídu svalových a zašuměných segmentů, ROVN jsou rovné čáry, PULZ značí pulzní artefakty a EPIL značí shluk epileptických grafoelementů.

### K-means

Tabulka 19: Výsledná senzitivita, specifita a PPV jednotlivých tříd pro algoritmus k-means pro pacienta 1

Parametry	FYZ	OČNÍ	EMG	ROVN	EPIL
<b>TP</b> [-]	10409	1200	1000	14	1902
<b>FN</b> [-]	24378	657	170	13	0
<b>FP</b> [-]	8	7278	4	3203	4696
<b>TN</b> [-]	4951	30611	38572	36516	33148
<b>SENZ</b> [-]	0,299	0,646	0,855	0,519	1,000
<b>SPEC</b> [-]	0,998	0,808	1,000	0,919	0,876
<b>PPV</b> [-]	0,999	0,142	0,996	0,004	0,288

Signál hodnocený v tabulce č. 19 obsahoval 5 tříd: fyziologickou aktivitu, pomalé oční artefakty, svalovou EMG aktivitu, rovné čáry a třídu epileptické aktivity. K-means dělí segmenty signálu do 7 tříd, některé třídy byly tedy mylně rozděleny do více tříd. Jak vyplývá z tabulky č. 19 fyziologická aktivita ve vysokém počtu je obsažena ve více třídách. Pomalé oční artefakty jsou spojeny s třídou jiných segmentů. Stejně tak rovné čáry k-means neodděluje a jsou součástí třídy, která obsahuje část segmentů fyziologické aktivity.



Obrázek 35: FP segmenty ve třídě pomalých očních grafoelementů, zobrazeno ve WF [37].

V obrázku č. 36 jsou vyznačeny (červeně) segmenty, které náleží do třídy fyziologické aktivity a byly mylně přiřazeny do třídy pomalých očních artefaktů. Oranžově zvýrazněné segmenty jsou hraniční segmenty, které by mohly být zařazeny jak do jedné, tak do druhé třídy.

Tabulka 20: Výsledná senzitivita, specificita a PPV jednotlivých tříd pro algoritmus k-means pro pacienta 2

Parametry	FYZ	OČNÍ	EMG	PULZ	EPIL
TP [-]	21074	244	1457	499	4199
FN [-]	19509	27	13	38	3209
FP [-]	32	91	68	10	35
TN [-]	9662	49915	48739	49730	42834
SENZ [-]	0,519	0,900	0,991	0,929	0,567
SPEC [-]	0,997	0,998	0,997	1,000	0,999
PPV [-]	0,998	0,7284	0,955	0,980	0,992

Signál druhého pacienta obsahoval také 5 tříd, konkrétně fyziologickou aktivitu, pomalé oční artefakty, svalovou EMG aktivitu, pulzní artefakty a třídu epileptické aktivity. K-means dělí segmenty signálu do 7 tříd, některé třídy byly tedy mylně rozděleny do více tříd. Jak vyplývá z tabulky č. 20, fyziologická aktivita byla klasifikována pouze se spe-



cificitou 0,519. Pomalé oční artefakty byly v tomto signálu detekovány s vyšší přesností. Pulzní artefakty k-means klasifikuje s vysokou pravděpodobností  $PPV = 0,999$ .

Tabulka 21: Výsledná senzitivita, specificita a PPV jednotlivých tříd pro algoritmus k-means pro pacienta 3

Parametry	FYZ	OČNÍ	EMG	EPIL
<b>TP</b> [-]	6701	0	238	1404
<b>FN</b> [-]	13285	1300	6	693
<b>FP</b> [-]	0	0	0	1366
<b>TN</b> [-]	2377	21063	22119	18900
<b>SENZ</b> [-]	0,335	0,000	0,975	0,670
<b>SPEC</b> [-]	1,000	-	1,000	0,933
<b>PPV</b> [-]	1,000	-	1,000	0,507

V případě očních artefaktů v signálu z pacienta č. 3 byla senzitivita nulová, shluk nebyl algoritmem vůbec detekován. Pomalé artefakty v tomto případě byly součástí třídy fyziologické aktivity, neměly tak svou vlastní třídu. EMG svalová aktivita a fyziologická aktivita jsou klasifikovány s vysokou specificitou, což znamená, že segmenty jiných artefaktů (nepatřících do testované třídy) jsou správně klasifikovány do jiných shluků. Epileptická aktivita u tohoto pacienta byla klasifikována s menší přesností.

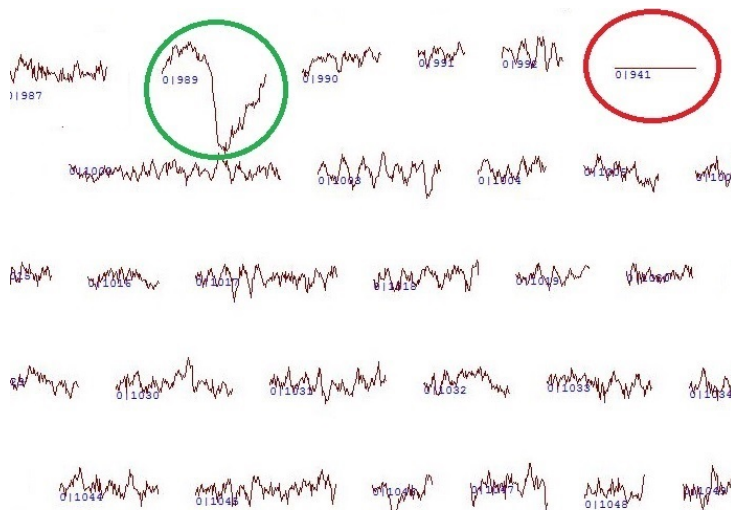
## GRIDBSCAN

Tabulka 22: Výsledná senzitivita, specificita a PPV jednotlivých tříd pro algoritmus GRIDBSCAN pro pacienta 1

Parametry	FYZ	OČNÍ	EMG	ROVN	EPIL
<b>TP</b> [-]	28175	26	826	0	840
<b>FN</b> [-]	6511	1831	448	27	1062
<b>FP</b> [-]	2308	6955	0	0	56
<b>TN</b> [-]	2690	30934	38472	39539	37788
<b>SENZ</b> [-]	0,812	0,014	0,648	0,000	0,442
<b>SPEC</b> [-]	0,538	0,816	1,000	-	0,999
<b>PPV</b> [-]	0,924	0,004	1,000	-	0,938

Jak je patrné z tabulky č. 22, algoritmus GRIDDBSCAN nedokáže detekovat shluk obsahující rovné čáry. Vysokou citlivost má v tomto případě algoritmus na segmenty fyzi-

ologické aktivity. Vysoká specifita zde byla při klasifikaci EMG aktivity a epileptických grafoelementů.



Obrázek 36: FP segmenty ve třídě fyziologické aktivity, zobrazeno ve WF [37].

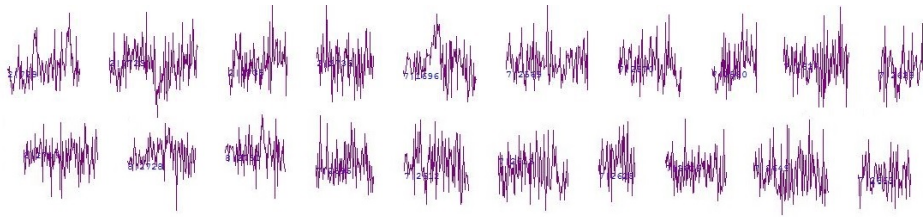
V obrázku č. 37 jsou zvýrazněny falešně pozitivní segmenty, které do shluku fyziologické aktivity nepatří. Jedná se o pomalý oční artefakt (zeleně) a rovnou čáru (červeně). Oba zvýrazněné artefakty jsou do třídy fyziologické aktivity mylně řazeny algoritmem GRIDBSCAN i k-means.

Tabulka 23: Výsledná senzitivita, specifita a PPV jednotlivých tříd pro algoritmus GRIDBSCAN pro pacienta 2

Parametry	FYZ	OČNÍ	EMG	PULZ	EPIL
TP [-]	40383	0	20	5361	2706
FN [-]	0	271	1450	1	4702
FP [-]	5358	0	0	15	8
TN [-]	4536	50006	48807	44900	42861
SENZ [-]	1,000	0,000	0,014	1,000	0,365
SPEC [-]	0,458	-	1,000	1,000	1,000
PPV [-]	0,883	-	1,000	1,000	0,997

U pacienta 2 algoritmus GRIDBSCAN nenalezl třídu pomalých očních artefaktů. Fyziologická aktivita byla zařazena všechna do jednoho shluku. V tomto shluku se ale vy-

skytovaly i další segmenty, které sem nepatřily. Výbornou účinnost zde měl algoritmus na pulzní artefakty.



Obrázek 37: Třída EMG a šumových grafoelementů, zobrazeno ve WF [37].

V obrázku č. 38 jsou segmenty EMG grafoelementů. Tato třída má PPV rovné 1, stejně tak specificitu. Senzitivita je ale velmi nízká, neboť třída neobsahuje všechny segmenty, které do ní náleží (viz tabulka č. 23). V Záznamu se vyskytovalo větší množství segmentů s EMG a šumovými grafoelementy.

Tabulka 24: Výsledná senzitivita, specificita a PPV jednotlivých tříd pro GRIDBSCAN pro pacienta 3

Parametry	FYZ	OČNÍ	EMG	EPIL
<b>TP</b> [-]	19947	0	205	1279
<b>FN</b> [-]	39	1300	39	818
<b>FP</b> [-]	89	0	0	39
<b>TN</b> [-]	2288	21063	22119	20227
<b>SENZ</b> [-]	0,998	0,000	0,840	0,610
<b>SPEC</b> [-]	0,963	-	1,000	0,998
<b>PPV</b> [-]	0,996	-	1,000	0,970

Při klasifikaci záznamu pacienta 3 GRIDBSCAN opět nenalezl třídu pomalých očních artefaktů. Vysokou účinnost zaznamenal na třídách fyziologické aktivity a EMG svalové aktivity.

## Shrnutí vyhodnocení klasifikace kompletních 3 záznamů

Další dvě tabulky č. 25 a č. 26 ukazují výslednou analýzu pro jednotlivé třídy ze 3 kompletních záznamů pro oba algoritmy.

Tabulka 25: Výsledná senzitivita, specificita a PPV jednotlivých tříd pro algoritmus k-means

Parametry	FYZ	OČNÍ	EMG	ROV	PULZ	EPIL
<b>SENZ</b> [-]	0,335	0,446	0,975	0,519	0,929	0,670
<b>SPEC</b> [-]	0,998	0,608	1,000	0,919	1,000	0,933
<b>PPV</b> [-]	0,999	0,040	0,996	0,004	0,980	0,507

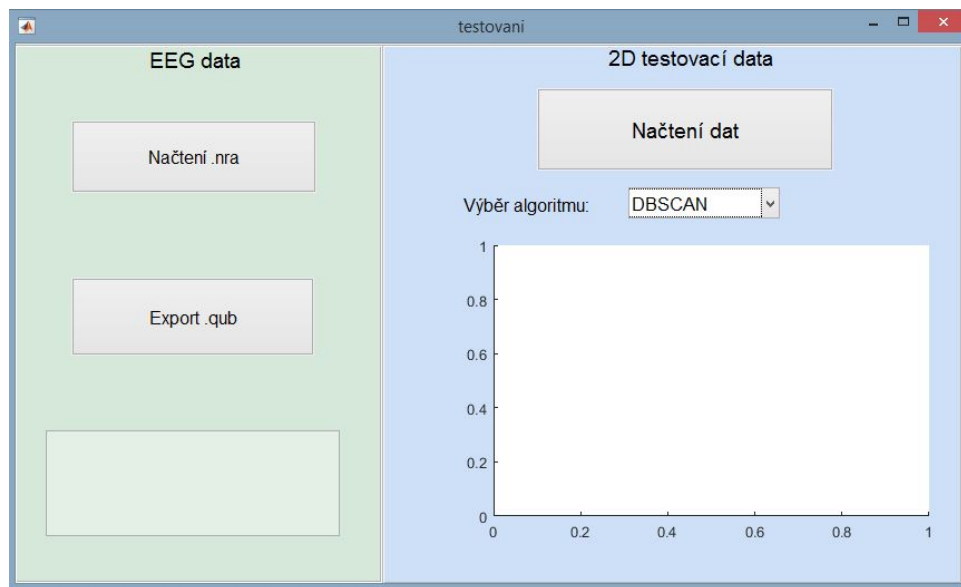
Tabulka 26: Výsledná senzitivita, specificita a PPV jednotlivých tříd pro algoritmus GRIDBSCAN

Parametry	FYZ	OČNÍ	EMG	ROV	PULZ	EPIL
<b>SENZ</b> [-]	0,998	0,000	0,648	0,000	1,000	0,442
<b>SPEC</b> [-]	0,538	-	1,000	-	1,000	0,999
<b>PPV</b> [-]	0,924	-	1,000	-	0,997	0,970

Oční artefakty algoritmus GRIDBSCAN není schopen vůbec oddělit. K-means algoritmus klasifikuje tyto artefakty s nízkou úspěšností (PPV = 0,04, senzitivita pod 0,5). Fyziologická aktivita je lépe rozřazena algoritmem GRIDBSCAN (PPV = 0,924, senzitivita = 0,998). Velmi dobrých výsledků algoritmus také dosahuje u EMG a pulzní aktivity. K-means má nízkou hodnotu PPV pro epileptickou aktivitu (PPV = 0,507), což znamená, že pokud algoritmus zařadí segment do třídy epileptických grafoelementů, jen s 51% pravděpodobností náhodně vybraný segment bude reprezentovat epileptickou aktivitu.

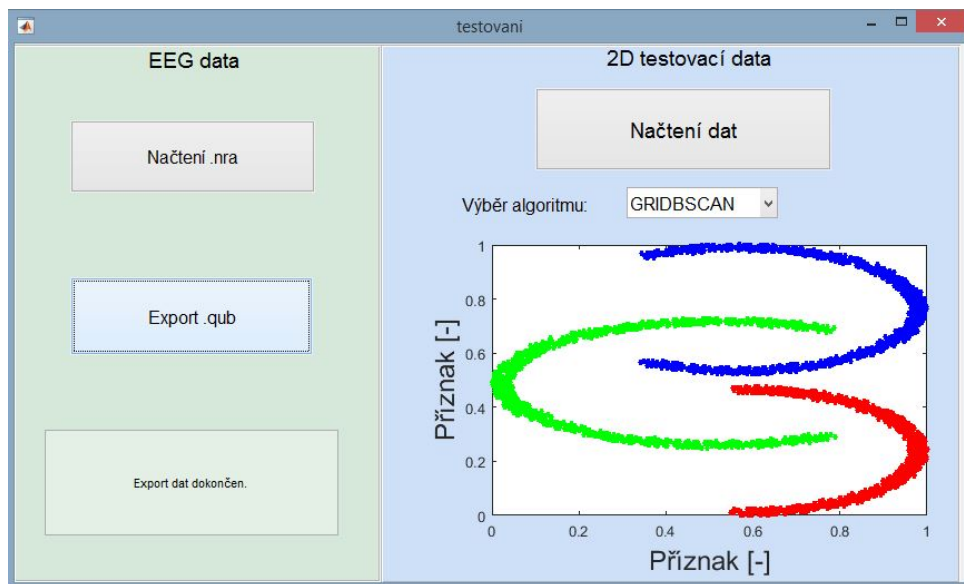
### 4.3 GUI

Vytvořil jsem jednoduché grafické uživatelské rozhraní (GUI) pro snadné načtení dat a export výsledků klasifikace. Barevně jsou v okně odděleny části načtení a klasifikace reálných dat (zeleně) a klasifikace a zobrazení 2D testovacích dat (modrá sekce). Na obrázku č. 39 jsou v levé části okna tlačítka pro načtení reálných EEG segmentů v souboru s příponou .nra. Po načtení zmáčkne tlačítko Export, v okně pod tímto tlačítkem se objeví hláška, abychom vyčkali, až proběhne výpočet. Jakmile klasifikace proběhne a data jsou vyexportována do složky, objeví se hláška, že export byl dokončen, viz obrázek č. 40.



Obrázek 38: Grafické prostředí pro načtení dat a export klasifikovaných dat.

V pravé části obrázku č. 39 je tlačítko pro načtení 2D testovacích dat. Pod tímto tlačítkem se nachází popupmenu, kterým vybíráme algoritmus ke klasifikaci načtených dat. V grafu v dolní části se po výběr algoritmu zobrazí výsledek klasifikace 2D dat, viz obrázek č. 40.



Obrázek 39: GUI pro načtení dat a export klasifikovaných dat po načtení a spuštění klasifikace.

## 5 Diskuze

### 5.1 Kvantitativní vyhodnocení

Testovací 2D data č. 1 a 2 byla sestrojena tak, aby poukázala na rychlost klasifikace při zachování přesnosti klasifikace i algoritmem k-means. Data 1 a 2 byla klasifikována s vysokou přesností pomocí všech testovaných algoritmů, ale algoritmus k-means byl výrazně nejrychlejším z nich. Rychlost je dána i potřebou vstupu uživatele v podobě zadání parametru počtu shluků. Algoritmus k-means spojený s výpočtem siluet, reprezentující automatickou klasifikaci metodou k-means, byl ze všech algoritmů nejpomalejším. Výpočetní náročnost hustotních algoritmů výrazně roste se zvyšujícím se objemem klasifikovaných dat. Tento trend můžeme pozorovat na obrázku č. 33, ze kterého je patrné, že u vyššího počtu dat narůstá čas výpočtu klasifikace u hustotních algoritmů strměji než u algoritmu k-means.

U reálných dat z EEG záznamů klasifikace hustotními algoritmy probíhala v řádu minut. Nejdelsí klasifikace probíhala téměř 20 minut viz tabulka č. 15. U hustotního algoritmu záleží, zda počítáme všechny parametry automaticky ze vstupních dat, nebo zadávali parametry uživatel. Hodnoty uvedené v tabulce č. 15 jsou pro algoritmus s automatickým výpočtem parametrů. U algoritmu GRIDBSCAN se zadáním vstupních parametrů se čas klasifikace reálných EEG dat snížil o jednotky minut. Dle rozměru jednotlivých dat se klasifikace pomocí algoritmu k-means počítala na jednotky až desítky sekund (nejdelší výpočet probíhal 5 minut, data obsahovala více jak 50 tisíc segmentů a každý segment byl popsán 23 příznaky). Časový rozdíl mezi jednotlivými typy algoritmů (GRIDBSCAN, k-means) byl v jednotkách minut.

U algoritmu GRIDBSCAN je časová náročnost daná tím, že jeden bod je navštíven vícekrát a pokaždé jsou přepočítávány vzdálenosti k ostatním bodům při hledání bodů v poloměru. Asymptotická složitost, která ukazuje horní mez náročnosti pro data o „n“ bodech, odpovídá u DBSCANu n-násobku logaritmu o základu n ( $n \cdot \log(n)$ ) [54, 44]. Při

špatně zvolených parametrech může být pro mnou naprogramovaný DBSCAN a GRID-BSCAN asymptotická složitost až  $n^2$ . Časovou náročnost lze snížit změnou počtu bodů v poloměru případně velikostí poloměru. Upravený algoritmus GRIDBSCAN automaticky počítá vstupní parametry. Snížit u něj výpočetní náročnost můžeme tedy jen zvýšením počtu buněk. V jednotlivých buňkách se bude nacházet poté menší počet bodů. DBSCAN probíhající v jednotlivých buňkách bude klasifikovat pouze  $n$  bodů, což bude počet v jednotlivých buňkách, který je několikanásobně menší, než velikost vstupních dat. Pokud bychom vypočítali vzdálenosti všech bodů mezi sebou, uložili je a přistupovali k nim postupně s probíhající klasifikací po bodech, jednalo by se o rychlejší variantu, která by ale byla paměťově náročná.

## 5.2 Kvalitativní vyhodnocení

Úspěšnost klasifikace algoritmů je vyhodnocena na základě třech parametrů (senzitivita, specificita a PPV). Parametr PPV říká, s jakou pravděpodobností segment zařazený do třídy X opravdu reprezentuje třídu X. Mohli bychom ho tedy přirovnat k homogenitě třídy. Jedná se tedy například o pravděpodobnost, že segment označen jako EMG aktivita, je svalovou (EMG) aktivitou. Proto má tento parametr pro klinické hodnocení nejvíce vypovídající hodnotu v rámci analýzy po jednotlivých třídách. Není vhodné ho ale paušalizovat na celý algoritmus. Například u metody k-means, jak vyplývá z tabulek č. 19, 20 a 21, by výrazně snížil její hodnotu PPV. Ta je u fyziologické aktivity nízká a pokud bychom chtěli PPV globalizovat na celý signál, zastínila by jedna nízká hodnota u velké třídy (vysoký počet segmentů) správně klasifikované třídy s menším počtem segmentů. Při klasifikaci EEG signálu je důležité správně detekovat grafoelementy jako jsou například epileptická aktivita. Rovné čáry nejsou projevem aktivity mozku. A fyziologická aktivita má pro lékaře až druhotný charakter. Jejich mylné rozdělení do více skupin nepůsobí výrazné klinické problémy.

Parametr senzitivita říká, kolik segmentů je zařazeno do shluku správně a kolik segmentů do shluku patřících je přiřazeno k jiným shlukům. Říká tedy, jak je algoritmus



dobry při detekci třídy. Bude-li se jednat například o třídu epileptické aktivity, získáme hodnotu, která napoví, kolik segmentů epileptické aktivity bylo zařazeno mylně do jiné třídy. Jedná se tedy o klinicky významný ukazatel.

Specifická ukazuje na to, jak je algoritmus schopný detekovat segmenty, které do vybraného shluku nepatří. Jedná se tedy o parametr, jehož váha ze tří zkoumaných charakteristik je nejnižší. Tento parametr má ve výsledcích vysoké hodnoty, neboť u malých shluků (např. EMG či pulzní aktivity) jsou správně oddělené několikanásobně větší shluky fyziologické aktivity. Pro jeho vyšší vypovídající hodnotu by bylo nutné klasifikovat data se stejným počtem segmentů v jednotlivých třídách, což u reálného EEG záznamu není možné.

### 5.2.1 2D testovací data

Hustotní algoritmy klasifikují hůře malé datové soubory (reprezentovány daty na obrázku č. 28). Nízký počet klasifikovaných bodů způsobuje nepřesnost z důvodu malého počtu bodů v poloměru potřebných pro určení středového bodu. U těchto dat je nutné zadat větší poloměr a menší počet bodů v něm. Na základě toho dochází k zařazení šumového bodu do shluku (viz tabulka č. 5), neboť šumový bod náleží do poloměru (díky jeho velikosti) posledního bodu ve shluku. V datech 2 dochází tedy ke špatné separaci šumu od řídkých dat (malý počet dat ve velkém příznakovém prostoru).

Data 2 na obrázku č. 29 reprezentují mírně prolnuté shluky (nedochází k velkému překryvu shluků). Prolnuté shluky nelze v prostoru rozdělit přímkou. V tomto případě jsou shluky i se svými těžišti rovnoměrně rozloženy v prostoru. K-means takovéto shluky rozdělí podle polohy iniciačních středů správně, nebo s několika FN a FP klasifikovanými body. Hustotní algoritmy správně jako svůj výstupní parametr určují počet shluků 2.

Data 3 reprezentují dva velmi blízké shluky s okolními šumovými body, které oba shluky obklopují. Algoritmy DBSCAN a DMDBSCAN nedokáží odlišit dva velmi blízké shluky (viz obrázek č. 30 vpravo nahoře), které řadí do jedné třídy. K-means neklasi-

fikuje ani jeden shluk zcela správně, jak je patrné z tabulky č. 9. Šumové body dělí mylně do tří shluků. Ani vnitřní obdélníky nejsou rozděleny správně, právě díky těžišti, které je ovlivněno špatně klasifikovanými šumovými body. GRIDBSCAN rozdělí velmi blízké shluky a šumové řadí do jedné třídy. Všechny zkoumané parametry (senzitivita, specificita, PPV) jsou pro GRIDBSCAN rovny 1. Tato data ukazují, že GRIDBSCAN je nejuniverzálnější verzí DBSCANu.

Prolnuté shluky reprezentují také data 4 na obrázku č. 31. Algoritmus k-means je nedokáže oddělit, ani pokud bychom iniciační centra uměle umístili do středů shluků požadovaného výsledku klasifikace. Hustotní algoritmy prolnuté shluky oddělují bez jediného chybně zařazeného bodu, jak jsem předpokládal z výsledků uváděných v odborných studiích. Pokud by se tedy v příznakovém prostoru EEG signálu vyskytovaly prolnuté shluky, u hustotních metod odvozených z algoritmu DBSCAN předpokládám správnou klasifikaci (viz tabulka č. 11).

Testovací data 5 obsahují kombinaci prolnutých shluků a šumu. K-means šum nedokáže detekovat, neboť třídy vznikají na základě vzdáleností mezi body. Každý bod je přiřazen k nejbližšímu těžišti shluku, tudíž jsou šumové body řazeny mezi shluky, nikoli do zvláštní skupiny. Díky přiřazení šumu ke shluku se mění poloha centra shluku. Tím může nastat chyba při klasifikaci ostatních bodů, které by do shluku měly patřit, nebo naopak jsou v něm zařazeny mylně. Hustotní algoritmy klasifikují shluky s velmi malou chybou (viz tabulka č. 13). Několik bodů šumu bylo díky své těsné blízkosti k shluku do shluku zařazeno (viz obrázek č. 32). U algoritmu DBSCAN a jeho modifikací nevzniká problém falešně zařazeného bodu shluku do šumové třídy. Jinými slovy, bod není mylně vyhodnocen jako šumový, ačkoli náleží některému ze shluků.

Z tabulky č. 14 je patrné, že jediný algoritmus GRIDBSCAN měl všechny parametry rovné 1, tedy stoprocentní účinnost klasifikace. Třídy rozdělené algoritmem GRIDBSCAN byly homogenní. Specificitu měly zbylé tři algoritmy shodnou 0,998. Senzitivita je shodná u algoritmů DBSCAN a DMDBSCAN 0,994. Nejnižší senzitivitu měl algoritmus k-means 0,950. Výsledek analýzy odpovídá předpokládanému stavu. Data byla vytvářena tak, aby

reprezentovala předpokládaný výskyt rozložení v příznakovém prostoru reálných dat. Testovací data měla dokázat, že k-means nedokáže prostorově prolnuté shluky správně klasifikovat. Jeho nejnižší klasifikační účinnost ze všech testovaných algoritmů tento předpoklad potvrdila.

### 5.2.2 Reálná EEG data

Reálné signály obsahovaly různé počty tříd podle výskytu různých typů grafoelementů. K-means má obecně nastavený počet 7 tříd. Pokud se v klasifikovaném signálu vyskytuje menší počet tříd (testovaná data obsahovala nejvíce 5 tříd), jsou třídy mylně děleny na více částí. Při nastavení parametru 7 shluků klasifikuje k-means dobře signály obsahující výraznou epileptickou aktivitu, EMG aktivitu a pulzní artefakty. Pokud signál obsahuje pouze fyziologickou aktivitu, pomalé oční artefakty a epileptickou aktivitu s nízkou a vysokou amplitudou, dochází k chybnému rozdělení fyziologické a epileptické aktivity do více tříd, neboť algoritmus je zadaným parametrem nucen dosáhnout konkrétního počtu tříd. Oproti tomu hustotní algoritmy nejsou vázány počtem tříd. Počet tříd naopak vychází z charakteru klasifikovaných dat.

V jednom ze 3 celých zkoumaných signálů se vyskytovaly rovné čáry. Hustotní algoritmy i algoritmus k-means je klasifikovaly do shluku fyziologické aktivity. Počet čar v signálu byl 14, což vůči celkovému počtu segmentů (39746) je zanedbatelné množství. Rovné čáry zároveň nerepresentují mozkovou aktivitu, nejedná se o EEG signál. Na základě toho jsem se rozhodl nezahrnout výsledky jejich klasifikace do rozhodování o lepší účinnosti některého z klasifikačních algoritmů.

### 12 pacientů

Pro testování algoritmů napříč pacienty jsem díky velkému počtu dat (vyhodnocení celých záznamů 15 pacientů by znamenalo milióny segmentů), zvolil náhodný výběr 50 segmentů z každého z 12 pacientů. Analyzoval jsem úspěšnosti klasifikace těchto segmentů. Seg-

menty tříd fyziologické aktivity a pomalých očních grafoelementů se vyskytovaly u všech 12 pacientů. U 9 pacientů se vyskytovaly segmenty svalové EMG aktivity a síťového šumu, z čehož záznam pacienta 9 byl velmi zašuměný. U čtyřech pacientů byla v 50 vybraných segmentech i epileptická aktivita. Díky náhodnému výběru se v 50 segmentech nemusely vždy promítnout všechny třídy obsažené v klasifikaci celého signálu. Zároveň mezi vybranými segmenty nemusely být reprezentativní vzorky třídy (viz obrázek č. 35). Náhodným výběrem mohl být zvolen například pouze jediný segment, který do dané třídy nepatří. To by se mohlo stát, pokud by tedy třída měla 100 segmentů a reálné PPV třídy by bylo 0,99. Ale při výběru onoho jednoho chybně zařazeného segmentu by PPV bylo nulové.

Algoritmus GRIDBSCAN nenalezl třídu EMG a šumových artefaktů reprezentovanou náhodně zvolenými segmenty. Tyto segmenty byly mylně přiřazovány k fyziologické aktivitě. Většinou se jednalo o segmenty mírně zašuměné, s nízkou amplitudou (cca 2  $\mu$ ). Při pohledu na celková data GRIDBSCAN třídu EMG grafoelementů, vyskytovala-li se v záznamu, našel a měla téměř vždy vysokou hodnotu PPV.

U většiny pacientů GRIDBSCAN detekoval fyziologickou aktivitu s vysokými hodnotami parametrů senzitivity a PPV (nad 0.8). V signálu od pacienta 11 nebyla detekována ani samostatná třída pomalých očních artefaktů (viz tabulka č. 37 v příloze). U tohoto pacienta bylo všech 50 vybraných segmentů zařazeno v třídě fyziologické aktivity. Tento pacient ukazuje chyby spojené s vysokou senzitivitou algoritmu GRIDBSCAN pro fyziologickou aktivitu. K-means dělí fyziologickou aktivitu do několika skupin. Algoritmus GRIDBSCAN vytváří ve většině případů klasifikace jednu velkou třídu fyziologické aktivity. Do té jsou ale mnohdy mylně zahrnuty i segmenty náležící k jiným třídám. Parametr senzitivity ale tyto mylně zařazené segmenty nevyhodnocuje, proto jsou jeho hodnoty u algoritmu GRIDBSCAN tak vysoké.

Algoritmus GRIDBSCAN u 12 pacientů (50 segmentů) měl u 3 pacientů velmi nízké hodnoty PPV pro epileptickou aktivitu (pod 0,5), viz tabulka č. 17. Ačkoli při celkovém vizuálním vyhodnocení klasifikace byla epileptická aktivita tříděna s vysokou účinností.

Dva náhodně vybrané epileptické grafoelementy byly na hranici zařazení do třídy epileptických grafoelementů.

Algoritmus k-means u 12 pacientů (50 segmentů) měl dle předpokladů velmi nízké hodnoty senzitivity pro třídu fyziologických segmentů. Výsledné hodnoty jsou vidět v tabulce č. 18. To je dáno rozdělením těchto segmentů do většího počtu tříd a jejich zastoupením (více jak polovina segmentů v záznamu) v analyzovaných datech. Pomalé oční artefakty i EMG aktivita je oproti algoritmu GRIDBSCAN klasifikována s vyšší přesností. Oproti předpokladům měl u těchto pacientů algoritmus k-means nízké hodnoty PPV pro epileptickou aktivitu. Pravděpodobně náhodným výběrem byl zvolen segment, který byl v celkové klasifikaci mylně zařazen do jiného shluku (tzv. FN segment). Nejednalo se tedy o segment z třídy epileptických grafoelementů, ačkoli tato třída v celkově roztríděném signálu byla zastoupena s vysokou přesností klasifikace.

Celkově na 12 pacientech měl algoritmus k-means lepší výsledky při klasifikaci epileptické aktivity, EMG aktivity a pomalých očních artefaktů. Pouze fyziologická aktivita byla lépe klasifikována algoritmem GRIDBSCAN. Při optickém vyhodnocení nebyla pozorována tak nízká úspěšnost klasifikace metodou GRIDBSCAN, jako ukazují výsledky z 12 pacientů. Nepřesnosti tohoto vyhodnocení jsou způsobeny náhodným výběrem 50 segmentů z několika desítek tisíc segmentů. V takto malém poměru vybraných segmentů není možné obsáhnout reprezentativní rozložení vzorků v jednotlivých třídách. Dochází tak ke zkreslení účinnosti obou algoritmů, jak v pozitivním tak v negativním slova smyslu. Vyhodnocení klasifikace celých záznamů u 15 pacientů z časových důvodů nebylo možné. Experti by museli vyhodnotit milióny segmentů.

### 3 pacienti

U 3 testovaných pacientů byla analyzována klasifikace všech segmentů jednotlivých záznamů. Jeden pacient obsahoval třídy: fyziologické aktivity, pomalé oční aktivity, EMG a epileptické aktivity a rovné čáry. Druhý pacient namísto třídy rovných čar obsahoval třídu

pulzních artefaktů. V záznamu posledního pacienta byly obsaženy segmenty fyziologické, pomalé oční, EMG a epileptické aktivity. Každá klasifikace představovala tedy různé stavy testovaných EEG dat.

### **pacient 1**

U prvního pacienta metoda k-means měla opět malou hodnotu senzitivity (0,299) pro fyziologickou aktivitu, viz tabulka č. 19. Parametry PPV a senzitivity pro pomalou oční aktivitu byly pod hranicí úspěšné klasifikace (0,75). To je způsobeno tím, že algoritmus klasifikoval několik smíšených tříd. Pomalé oční artefakty tvořily 4 třídy spolu s fyziologickou aktivitou. Fyziologická aktivita je v signálu více zastoupena, proto hodnota PPV byla pouze 0,142. Třída rovných čar v reálné klasifikaci nebyla vůbec nalezena, rovné čáry byly klasifikovány do stejné třídy s fyziologickou aktivitou. Díky nadbytečnému počtu tříd (7) u k-means je možné tvrdit, že třída rovných čar byla detekována, ale má velkou chybovost, neboť většinu (3203 z 3217 segmentů) tvořila fyziologická a pomalá oční aktivita. Epileptická aktivita byla klasifikována do jednoho shluku. Žádný segment nebyl v jiném shluku, parametr senzitivity byl tedy nejvyšší - 1. Zároveň ale shluk obsahoval i další segmenty, které do něj nepatřily (FP segmenty). Díky tomu parametr PPV dosáhl pouze hodnoty 0,288.

Algoritmu GRIDBSCAN rozdělil signál pacienta 1 do 5 tříd. Pomalá aktivita byla stejně jako u algoritmu k-means sloučena s fyziologickou aktivitou. Podle amplitudy byla tato třída rozdělena na 2 části. Pomalá oční aktivita má nízké hodnoty parametrů PPV a senzitivity (0,004 a 0,014) díky vysokému počtu segmentů fyziologické aktivity. Veškerá EMG aktivita vyskytující se v signálu nebyla zařazena do jedné třídy, ale třídu tvořily pouze segmenty EMG aktivity, proto PPV třídy bylo rovno 1. Epileptická aktivita byla algoritmem GRIDBSCAN rozdělena do dvou tříd podle velikosti amplitudy. Proto velikost senzitivity byla pouze 0,442, ale PPV 0,938 ukazuje, že pouze malý zlomek segmentů byl k epileptické aktivitě zařazen špatně. Rovné čáry byly přiřazeny k fyziologické aktivitě a jejich třída nebyla nalezena.

**pacient 2**

U pacienta 2 byla fyziologická aktivita algoritmem k-means rozdělena do 2 tříd. Na rozdíl od signálu pacienta 1 ale nebyla hodnota PPV snížena počtem pomalých očních artefaktů. Ty byly v tomto signále metodou k-means klasifikovány do vlastního shluku s hodnotou parametrů  $PPV = 0,728$  a senzitivitou  $= 0,900$ . EMG aktivita byla klasifikována s velmi malým počtem FP bodů (segmenty s jinou aktivitou zařazené ve třídě mylně) a pouze s 13 segmenty EMG aktivity zařazenými v jiných shlucích. Pulzní artefakty algoritmus k-means také správně odděluje do samostatné třídy s vysokou hodnotou PPV ( $0,980$ ), jedná se tedy o třídu s většinovým zastoupením jednoho druhu segmentů (reprezentuje jeden druh grafoelementů či aktivity). Senzitivita pro pulzní artefakty pro algoritmus k-means byla  $0,929$ , takže je malá část pulzní artefaktů byla přiřazena k jiné třídě. Epileptická aktivita byla rozdělena do dvou tříd. Parametr PPV zůstal vysoký díky homogenitě obou tříd, ale senzitivita vyšla  $0,567$  právě díky rozdělení třídy do dvou.

Algoritmus GRIDBSCAN správně detekoval 5 tříd. Opět ale sloučil pomalou oční aktivitu a fyziologickou aktivitu dohromady. Pomalé oční artefakty tedy neměly svou vlastní třídu. Všechny segmenty fyziologické aktivity byly zařazené do jednoho shluku, proto byl parametr senzitivity rovný  $1,000$ . EMG aktivita byla chybně řazena do shluku fyziologické aktivity. Ve třídě EMG aktivity byly zařazené pouze segmenty s vysokou amplitudou. Pulzní artefakty byly do své třídy řazeny s vysokou přesností ( $PPV = 0,997$ , senzitivita  $= 1,000$ ). Epileptická aktivita byla stejně jako u algoritmu k-means rozdělena do dvou tříd podle velikosti amplitudy. Parametr PPV byl ještě o  $0,005$  vyšší, než u algoritmu k-means ( $PPV = 0,992$ ). PPV třídy epileptické aktivity bylo vysoké, ale senzitivita byla u algoritmu GRIDBSCAN velmi nízká ( $0,365$ ), neboť segmentů s nižší amplitudou, které byly mylně přiřazeny do jiných shluků, bylo více než epileptických grafoelementů s vysokou amplitudou. Epileptické grafoelementy s vysokou amplitudou tvořily homogenní třídu.

**pacient 3**

Algoritmus k-means u pacienta 3 klasifikoval do stejných tříd pomalé oční artefakty a epileptické grafoelementy s nízkou amplitudou. Třída pomalých očních artefaktů tedy nebyla vůbec vytvořena. Fyziologická aktivita byla rozdělena do čtyř tříd, kdy jedna třída obsahovala pouze fyziologickou aktivitu, další tři třídy obsahovaly i jiné grafoelementy. Proto u třídy fyziologické aktivity je opět nízká senzitivita (0,335). Hodnota PPV byla u třídy fyziologické aktivity rovna 1, třída tedy neobsahovala segmenty s jinými grafoelementy. Segmenty s EMG grafoelementy byly klasifikovány s vysokou účinností (PPV = 1,000, senzitivita = 0,975). Do třídy EMG grafoelementů nebyly mylně zařazeny žádné segmenty s jinou aktivitou, pouze 6 segmentů EMG bylo mylně zařazeno do jiné třídy. Hodnota PPV, byla vysoká (0,938). Epileptická aktivita byla opět rozdělena do 2 tříd. Jedna třída obsahovala zároveň pomalé oční artefakty. V této třídě se vyskytoval vyšší počet epileptických grafoelementů než ve třídě druhé. Druhá epileptická třída by měla vyšší PPV, neboť obsahovala pouze epileptické grafoelementy, ale měla by nižší hodnoty senzitivity, protože by převažoval počet neklasifikovaných segmentů (FN).

GRIDBSCAN detekoval 4 třídy v signálu u pacienta 3. Fyziologická aktivita byla opět spolu s pomalými očními artefakty v jedné třídě. Část fyziologické aktivity bylo odděleno a chybně tvořilo další třídu spolu s epileptickými grafoelementy s nízkou amplitudou. Pomalé oční artefakty nebyly klasifikovány do samostatné třídy. EMG grafoelementy byly klasifikovány s vysokou přesností (PPV = 1,000, senzitivita = 0,840). EMG třídu tedy tvořily pouze grafoelementy EMG. 39 segmentů, které by do EMG třídy měly patřit, bylo chybně zařazeno do jiných tříd. Mylně řazené segmenty představovaly zašuměný signál. U takto řazených segmentů je mnohdy sporné, kdy se jedná o již velmi zašuměný signál a kdy by signál ještě mohl náležet do fyziologické aktivity či jiných tříd. Jak již bylo zmíněno, epileptické grafoelementy byly rozděleny podle amplitudy do 2 tříd. Díky tomu je nižší senzitivita pro tuto třídu (0,610), ale PPV zůstává vysoké (0,970).

Celkově na datech ze 3 pacientů k-means mylně dělí fyziologickou aktivitu do několika tříd. V jednom případě u k-means, ve všech u algoritmu GRIDBSCAN došlo ke sloučení



tříd fyziologické aktivity a pomalých očních artefaktů. EMG aktivita je oběma algoritmy úspěšně klasifikována. K chybnému zařazení dochází u zašuměných segmentů, které jsou na hranici mezi klasifikací k šumové a EMG aktivitě nebo k fyziologickým či jiným segmentům. Pulzní artefakty jsou také dobře separovatelné od ostatních tříd pro oba algoritmy. Epileptické grafoelementy tvoří u algoritmu GRIDBSCAN více homogenní třídy, než u algoritmu k-means (cca o 0,010 je PPV vyšší).

## Shrnutí

Algoritmus GRIDBSCAN je časově náročnější, neboť každý bod je v algoritmu navštíven několikrát a každá návštěva znamená další výpočetní operace. Hodnoty časů klasifikace jsou ovlivněny především typem procesoru a velikostmi pamětí počítače, na kterém je algoritmus spuštěn. Aktuální výkon počítače je dále ovlivněn počtem dalších spuštěných procesů, okolní vlhkostí i teplotou. Hodnoty časů klasifikace nemají tedy reálnou výpovědní hodnotu, je třeba se dívat pouze na trend křivek. Výpočetní náročnost lze snížit zvýšením počtu buněk, nebo snížením počtu příznaků. Pokud bychom parametry zadávali (nebyly by počítány automaticky), dojde ke zrychlení algoritmu (výpočet v řádu 2 minut). Data do jednotlivých tabulek byla počítána ihned po sobě, tudíž za prakticky shodných podmínek a nemělo by tak docházet ke zkreslení výsledků.

Třídu rovných čar je možné z klasifikace vyloučit, neboť se nejedná o projev mozkové aktivity a její výskyt je v řádu setin procenta z testovaných segmentů. Pomalé oční artefakty jsou lépe detekovány algoritmem k-means, neboť k-means vytváří dostatečný počet tříd i pro nadbytečně dělené shluky. Fyziologickou aktivitu klasifikuje lépe algoritmus GRIDBSCAN, neboť vytváří jeden velký shluk, který obsahuje pomalé oční artefakty a fyziologickou aktivitu. EMG aktivita a pulzní artefakty jsou dobře klasifikovány oběma algoritmy. K-means nalezne více segmentů s danou aktivitou, GRIDBSCAN vybírá skupiny pouze s daným gafoelementem za cenu, že některé segmenty zůstanou v jednom velkém shluku (fyziologické aktivity) zapomenuty. Proto má celkovou PPV vyšší než algoritmus k-means.

K zpřesnění klasifikace by molo dojít změnou parametrů, pokud by byly všechny parametry nastavovány pro konkrétní signál. Tento postup ale zvyšuje výpočetní náročnost (viz DMDBSCAN). Vhodný výběr příznaků pro metodu GRIDBSCAN by také zefektivnil klasifikaci. Různé příznaky mohou přispět ke zvýšení klasifikační úspěšnosti u jednotlivých tříd.

Algoritmus GRIDDBSCAN je možné použít pro zjištění počtu tříd. Případně by mohlo být využito jeho „ukázkových“ tříd pro epileptické, EMG a pulzní artefakty (všechny mají vysoké PPV). Tyto segmenty by mohly sloužit jako trénovací data pro naučení neuronových sítí. Pro jednoduchou klasifikaci ambulantních dat se zdá být algoritmus k-means lepším, neboť má podstatně nižší výpočetní náročnost. K chybné klasifikaci dochází především nadměrným dělením jedné třídy do více shluků, ale rozdělené shluky většinou zachovávají své PPV nad hranicí 0,75, kterou bereme jako mez úspěšné klasifikace.

## 6 Závěr

Pro automatickou klasifikaci segmentů EEG záznamu na základě příznaků jsem použil hustotní metody (DBSCAN a jeho modifikace). V programovém prostředí MATLAB jsem vytvořil uživatelsky přívětivé prostředí pro načtení a klasifikaci analyzovaných dat. Implementoval jsem modifikovanou verzi algoritmu DBSCAN (GRIDBSCAN) pro automatickou klasifikaci EEG signálu na základě příznaků. Tři verze algoritmů (DBSCAN, DMDBSCAN a GRIDBSCAN) jsem otestoval na mnou vytvořených simulovaných datech. Na těchto datech jsem provedl kvalitativní i kvantitativní analýzu implementovaných algoritmů. Modifikovaný algoritmus GRIDBSCAN, který dosahoval nejlepších klasifikačních výsledků, jsem aplikoval na reálné EEG záznamy pacientů. Celkem jsem klasifikoval 341913 segmentů. Následně jsem provedl kvalitativní zhodnocení a analyzoval jsem výpočetní (časové) náročnosti algoritmů. Statistické vyhodnocení účinnosti algoritmů jsem provedl pomocí ROC analýzy a jejích parametrů senzitivity, specificity a PPV. Ve své práci jsem vypracoval jsem všechny body zadání.

Vytvořil jsem plně funkční modul kompatibilní s aktuálně používaným softwarem v lékařských zařízeních, který nabízí alternativní klasifikační metodu k dosavadní metodě k-means. GRIDBSCAN má vyšší výpočetní náročnost, neboť veškeré vstupní parametry jsou počítány automaticky ze vstupních dat. Ve specifických případech je v klasifikaci úspěšnější algoritmus k-means. Menší úspěšnost v počtu nalezených segmentů patřících do třídy je u GRIDBSCANu kompenzována vysokou hodnotou PPV (nad 0,9), což považuji za hlavní výhodu tohoto klasifikačního algoritmu. Další z výhod algoritmu GRIDBSCAN je výstupní parametr počet tříd. Lze tedy algoritmy využít k odhadu počtu tříd v signálu.

Vysoká hodnota PPV (nad 0,9) nabízí alternativní možnost použití GRIDBSCANu pro klasifikační metody s učitelem. Algoritmus GRIDBSCAN by mohl být použit k rozdělení tříd grafoelementů s vysokou hodnotou PPV. Ty by následně fungovaly jako trénovací množiny pro neuronové sítě a další učící se klasifikátory. Extrakce etalonů proběhne jednorázově, časová náročnost není tedy omezujícím faktorem. Pro ověření všech těchto předpokladů je potřeba studie na více subjektech.

## Reference

- [1] Khushnandan Rai, Varun Bajaj, and Anil Kumar. Novel feature for identification of focal eeg signals with k-means and fuzzy c-means algorithms. In *Digital Signal Processing (DSP), 2015 IEEE International Conference on*, pages 412–416. IEEE, 2015.
- [2] Paschalis A. Bizopoulos, Dimitrios G. Tsalikakis, Alexandros T. Tzallas, Dimitrios D. Koutsouris, and Dimitrios I. Fotiadis. Eeg epileptic seizure detection using k-means clustering and marginal spectrum based on ensemble empirical mode decomposition. In *13th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–4. IEEE, 2013. ISBN 9781479931637. doi: 10.1109/BIBE.2013.6701528. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6701528>.
- [3] Tatsuya Teramae, Daisuke Kushida, Fumiaki Takemori, and Akira Kitamura. Estimation of feeling based on eeg by using nn and k-means algorithm for massage system. In *SICE Annual Conference 2010, Proceedings of*, pages 1542–1547. IEEE, 2010.
- [4] Vladimír Krajča and Svojmír Petránek. Wave-finder”: a new system for an automatic processing of long-term eeg recordings. In *Quantitative EEG Analysis - Clinical Utility and New Methods*, pages 103–106, Jena, 1993. Universitätsverlag GmbH.
- [5] Kdd 2014, 2014. URL <http://www.kdd.org/kdd2014/>.
- [6] Josef Faber. *Elektroencefalografie a psychofyziologie*. ISV, Praha, vyd. 1. edition, 2001. ISBN 80-858-6674-9.
- [7] Vladimír Krajča and Jitka Mohylová. *Číslíkové zpracování neurofyziologických signálů*. České vysoké učení technické, V Praze, 1. vyd. edition, 2011. ISBN 978-80-01-04721-7.
- [8] Zdeněk Wilhelm and a kolektiv. *Stručný přehled fyziologie člověka pro bakalářské studijní programy*. Masarykova universita - Lékařská fakulta, Brno, 4. edition, 2010. ISBN 978-80-210-5283-3.

- [9] Koji Sakai, Kenta Shimba, Kiyoshi Kotani, and Yasuhiko Jimbo. Microfabricated multi-electrode device for detecting oligodendrocyte-regulated changes in axonal conduction velocity. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7127–7130. IEEE, 2015. ISBN 9781424492718. doi: 10.1109/EMBC.2015.7320035. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7320035>.
- [10] J Rieger, Lenka Lhotská, and V Krajča. *Zpracování dlouhodobých eeg záznamů*, 2004.
- [11] Saeid Sanei and Jonathon A Chambers. *EEG signal processing*. John Wiley & Sons, 2013. ISBN 139780470025819.
- [12] M Moráň. Spánek a epilepsie. *Interní Med*, 1:26–31, 2006.
- [13] Y Aghakhani, AP Bagshaw, CG Benar, C Hawco, F Andermann, F Dubeau, and J Gotman. fmri activation during spike and wave discharges in idiopathic generalized epilepsy. *Brain*, 127(5):1127–1144, 2004.
- [14] Marek Penhaker. *Lékařské terapeutické přístroje*. VŠB - Technická univerzita Ostrava, Ostrava, 1. vyd. edition, 2007. ISBN 9788024815589.
- [15] Josef Faber. *EEG*. Triton, Praha, vyd. 1. edition, 1997. ISBN 8085875519.
- [16] Shyh-Yueh Cheng and Hong-Te Hsu. *Mental Fatigue Measurement Using EEG*. INTECH Open Access Publisher, 2011.
- [17] B Ahmadi, R Aimrfattahi, E Negahbani, M Mansouri, and M Taheri. Comparison of adaptive and fixed segmentation in different calculation methods of electroencephalogram time-series entropy for estimating depth of anesthesia. In *Information Technology Applications in Biomedicine, 2007. ITAB 2007. 6th International Special Topic Conference on*, pages 265–268. IEEE, 2007.
- [18] Forrest Sheng Bao, Ya-Liang Li, Jue-Ming Gao, and Jin Hu. Performance of dynamic features in classifying scalp epileptic interictal and normal eeg. In *Engineering in*

- Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 6308–6311. IEEE, 2010.
- [19] F S Bao, Ya-Liang Li, Jue-Ming Gao, and Jin Hu. Performance of dynamic features in classifying scalp epileptic interictal and normal eeg. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 6308–6311. IEEE, 2010. ISBN 9781424441235. doi: 10.1109/IEMBS.2010.5628091. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5628091>.
- [20] Francesca Finotello, Fabio Scarpa, and Mattia Zanon. Eeg signal features extraction based on fractal dimension. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4154–4157. IEEE, 2015. ISBN 9781424492718. doi: 10.1109/EMBC.2015.7319309. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7319309>.
- [21] John S Barlow. *The electroencephalogram*. MIT Press, Cambridge, Mass., c1993. ISBN 0262023547.
- [22] Sana Tmar-Ben Hamida, Beena Ahmed, and Thomas Penzel. A novel insomnia identification method based on hjorth parameters. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 548–552, Abu Dhabi, 2015. IEEE. ISBN 9781509004812. doi: 10.1109/ISSPIT.2015.7394397. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7394397>.
- [23] Hao Qu and J. Gotman. A patient-specific algorithm for the detection of seizure onset in long-term eeg monitoring. *IEEE Transactions on Biomedical Engineering*, vol. 44(issue 2):115–122. ISSN 00189294. doi: 10.1109/10.552241. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=552241>.
- [24] F Lotte, M Congedo, A Lécuyer, F Lamarche, and B Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 4(2):R1, 2007. URL <http://stacks.iop.org/1741-2552/4/i=2/a=R01>.

- [25] Salih Güneş, Kemal Polat, and Şebnem Yosunkaya. Efficient sleep stage recognition system based on eeg signal using k-means clustering based feature weighting. *Expert Systems with Applications*, vol. 37(issue 12):7922–7928, 2010. ISSN 09574174. doi: 10.1016/j.eswa.2010.04.043. URL <http://linkinghub.elsevier.com/retrieve/pii/S095741741000343X>.
- [26] Zhengmao Ye, Yongmao Ye, H. Mohamadian, P. Bhattacharya, and Kai Kang. Fuzzy filtering and fuzzy k-means clustering on biomedical sample characterization. In *Proceedings of 2005 IEEE Conference on Control Applications, 2005. CCA 2005*, pages 90–95. IEEE, 2005. ISBN 0780393546. doi: 10.1109/CCA.2005.1507106. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1507106>.
- [27] Vivek S Ware and HN Bharathi. Study of density based algorithms. *International Journal of Computer Applications*, 69(26), 2013.
- [28] George Karypis, Eui-Hong Han, and Vipin Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.
- [29] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1, 2007.
- [30] Sunil Kumar Prabhakar and Harikumar Rajaguru. Pca and k-means clustering for classification of epilepsy risk levels from eeg signals??? a comparative study between them. In *2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pages 83–86. IEEE, 2015.
- [31] Vasileios Kavvadias, George Epitropou, Niki Georgiou, Fani Grozou, Minas Paschopoulos, and Costas Balas. A novel endoscopic spectral imaging platform integrating k-means clustering for early and non-invasive diagnosis of endometrial pathology. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 4442–4445. IEEE, 2013.

- [32] Ahmed Mahmoud Hamad and Norimichi Tsumura. Silhouette extraction based on time-series statistical modeling and k-means clustering. In *Pattern Recognition (ACPR), 2011 First Asian Conference on*, pages 584–588. IEEE, 2011.
- [33] Juntao Wang and Xiaolong Su. An improved k-means clustering algorithm. In *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on*, pages 44–46. IEEE, 2011.
- [34] Kamran Khan, Saif Ur Rehman, Khurram Aziz, Simon Fong, and S Sarasvady. Dbscan: Past, present and future. In *Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the*, pages 232–238. IEEE, 2014.
- [35] Mohammed TH Elbatta and Wesam M Ashour. A dynamic method for discovering density varied clusters. *Int. Journal of Signal Processing, Image Processing and Pattern Recognition*, 6(1):123–134, 2013.
- [36] Samir Kumar Bandyopadhyay and Tuhin Utsab Paul. Segmentation of brain tumour from mri image-analysis of k-means and dbscan clustering. *International Journal of Research in Engineering and Science*, 1(1):48–57, 2013.
- [37] Wavefinder. doc. Ing. Vladimír Krajča, CSc., 1991. Release: 2.3.5.
- [38] AT Tzallas, CD Katsis, PS Karvelis, DI Fotiadis, S Konitsiotis, and S Giannopoulos. Classification of transient events in eeg recordings. In *Proceedings of the IEE Medical Signal and Information Processing Conference, MEDSIP*, volume 4, pages 5–8, 2004.
- [39] MATLAB®. The MathWorks, Inc., 2015. Release: 2015a.
- [40] Khushali Mistry, Swapnil Andhariya, and Sahista Machchhar. Ndcmd: A novel approach towards density based clustering using multidimensional spatial data. In *International Journal of Engineering Research and Technology*, volume 2. ESRSA Publications, 2013.



- [41] Amin Karami and Ronnie Johansson. Choosing dbscan parameters automatically using differential evolution. *International Journal of Computer Applications*, 91(7), 2014.
- [42] Ali Touka. Choosing parameters of dbscan algorithm, 2012. URL [https://github.com/alitouka/spark\\_dbscan/wiki/Choosing-parameters-of-DBSCAN-algorithm](https://github.com/alitouka/spark_dbscan/wiki/Choosing-parameters-of-DBSCAN-algorithm).
- [43] Nadia Rahmah and Imas Sukaesih Sitanggang. Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra. In *IOP Conference Series: Earth and Environmental Science*, volume 31, page 012012. IOP Publishing, 2016.
- [44] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
- [45] Chun-jiang Pang. Research of grid-similarity-based clustering algorithm. In *Information Engineering, 2009. ICIE'09. WASE International Conference on*, volume 2, pages 33–36. IEEE, 2009.
- [46] Cheng-Fa Tsai and Jun-Hao Zhang. Grid clustering algorithm with simple leaping search technique. In *Computer, Consumer and Control (IS3C), 2012 International Symposium on*, pages 938–941. IEEE, 2012.
- [47] Ozge Uncu, William A Gruver, Dilip B Kotak, Dorian Sabaz, Zafeer Alibhai, and Colin Ng. Gridbscan: grid density-based spatial clustering of applications with noise. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, volume 4, pages 2976–2981. IEEE, 2006.
- [48] Shaaban Mahran and Khaled Mahar. Using grid for accelerating density-based clustering. In *Computer and Information Technology, 2008. CIT 2008. 8th IEEE International Conference on*, pages 35–40. IEEE, 2008.

- [49] Miguel Antonio Sovierzoski, Fernando Mendes de Azevedo, and Fernanda Isabel Marques Argoud. Performance evaluation of an ann ff classifier of raw eeg data using roc analysis. In *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, volume 1, pages 332–336. IEEE, 2008.
- [50] Nadav David Marom, Lior Rokach, and Armin Shmilovici. Using the confusion matrix for improving ensemble classifiers. In *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, pages 000555–000559. IEEE, 2010. ISBN 9781424486816. doi: 10.1109/EEEI.2010.5662159. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5662159>.
- [51] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8): 861–874, 2006.
- [52] Matt Hancock. What is the roc curve?!, 2015. URL <http://notmatthancock.github.io/2015/08/18/what-is-the-roc-curve.html>.
- [53] Christian O’Reilly and Tore Nielsen. Revisiting the roc curve for diagnostic applications with an unbalanced class distribution. In *2013 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA)*, pages 413–420. IEEE, 2013. ISBN 9781467355407. doi: 10.1109/WoSSPA.2013.6602401. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6602401>.
- [54] Derya Birant and Alp Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.

## A Obsah CD

1. Elektronickou verzi této práce
2. Zadání práce
3. Abstrakt česky
4. Abstrakt anglicky
5. Klíčová slova
6. GUI pro načtení a export dat
7. Ukázkou testovacích 2D dat
8. Anonymizovaná reálná EEG data
9. Skripty pro jednotlivé algoritmy

## B Tabulky 12 pacientů vytvořené z náhodně vybraných 50 segmentů

Tabulka 27: Výsledky analýzy pacienta 1.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,425	1,000
	SPEC [-]	1,000	0,333
	PPV [-]	1,000	0,959
OČNÍ	SENZ [-]	1,000	1,000
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000
EMG	SENZ [-]	1,000	0,000
	SPEC [-]	1,000	-
	PPV [-]	1,000	-

Tabulka 28: Výsledky analýzy pacienta 2.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,356	0,500
	SPEC [-]	0,600	1,000
	PPV [-]	0,888	1,000
OČNÍ	SENZ [-]	0,6000	0,556
	SPEC [-]	0,956	0,600
	PPV [-]	0,600	0,926

Tabulka 29: Výsledky analýzy pacienta 3.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,343	0,857
	SPEC [-]	0,867	0,600
	PPV [-]	0,857	0,938
OČNÍ	SENZ [-]	0,667	0,583
	SPEC [-]	0,895	0,947
	PPV [-]	0,667	0,778
EMG	SENZ [-]	1,000	0,000
	SPEC [-]	1,000	-
	PPV [-]	1,000	-
EPIL	SENZ [-]	1,000	1,000
	SPEC [-]	1,000	0,625
	PPV [-]	1,000	0,400

Tabulka 30: Výsledky analýzy pacienta 4.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,302	1,000
	SPEC [-]	1,000	0,286
	PPV [-]	1,000	0,896
OČNÍ	SENZ [-]	1,000	1,000
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000
EMG	SENZ [-]	1,000	0,200
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000

Tabulka 31: Výsledky analýzy pacienta 5.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,420	0,953
	SPEC [-]	1,000	0,142
	PPV [-]	1,000	0,872
OČNÍ	SENZ [-]	0,833	0,167
	SPEC [-]	0,977	1,000
	PPV [-]	0,833	1,000
EMG	SENZ [-]	1,000	1,000
	SPEC [-]	0,959	0,898
	PPV [-]	0,333	0,167

Tabulka 32: Výsledky analýzy pacienta 6.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,553	1,000
	SPEC [-]	1,000	0,500
	PPV [-]	1,000	0,864
OČNÍ	SENZ [-]	0,500	1,000
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000
EMG	SENZ [-]	1,000	0,500
	SPEC [-]	0,898	1,000
	PPV [-]	0,167	1,000
EPIL	SENZ [-]	1,000	1,000
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000

Tabulka 33: Výsledky analýzy pacienta 7.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,400	0,917
	SPEC [-]	1,000	0,923
	PPV [-]	1,000	0,917
OČNÍ	SENZ [-]	1,000	0,333
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000
EMG	SENZ [-]	0,500	0,500
	SPEC [-]	1,000	0,978
	PPV [-]	1,000	0,750
EPIL	SENZ [-]	1,000	1,000
	SPEC [-]	0,878	0,878
	PPV [-]	0,143	0,143

Tabulka 34: Výsledky analýzy pacienta 8.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,304	0,870
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000
OČNÍ	SENZ [-]	1,000	1,000
	SPEC [-]	1,000	0,979
	PPV [-]	1,000	0,75
EMG	SENZ [-]	1,000	1,000
	SPEC [-]	0,959	0,980
	PPV [-]	0,333	0,500

Tabulka 35: Výsledky analýzy pacienta 9.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,667	1,000
	SPEC [-]	1,000	0,438
	PPV [-]	1,000	0,500
OČNÍ	SENZ [-]	0,357	0,250
	SPEC [-]	0,772	0,909
	PPV [-]	0,667	0,778
EMG	SENZ [-]	0,750	0,250
	SPEC [-]	0,978	1,000
	PPV [-]	0,750	1,000

Tabulka 36: Výsledky analýzy pacienta 10.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,400	0,500
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000
OČNÍ	SENZ [-]	1,000	0,036
	SPEC [-]	0,979	1,000
	PPV [-]	0,750	1,000
EMG	SENZ [-]	1,000	1,000
	SPEC [-]	1,000	0,978
	PPV [-]	1,000	0,833
EPIL	SENZ [-]	1,000	1,000
	SPEC [-]	0,896	0,896
	PPV [-]	0,286	0,286

Tabulka 37: Výsledky analýzy pacienta 11.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,459	1,000
	SPEC [-]	1,000	0,000
	PPV [-]	1,000	0,740
OČNÍ	SENZ [-]	0,636	0,000
	SPEC [-]	0,974	-
	PPV [-]	0,875	-
EMG	SENZ [-]	1,000	0,000
	SPEC [-]	1,000	-
	PPV [-]	1,000	-

Tabulka 38: Výsledky analýzy pacienta 12.

Parametry		K-means	GRIDBSCAN
FYZ	SENZ [-]	0,333	0,333
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000
OČNÍ	SENZ [-]	1,000	1,000
	SPEC [-]	1,000	1,000
	PPV [-]	1,000	1,000

## C Abstrakt publikovaný ve sborníku konference Seminář biomedicínského inženýrství 2016

### AUTOMATICKÁ KLASIFIKACE EEG SEGMENTŮ METODOU DBSCAN

Marek Piorecký

FBMI ČVUT V PRAZE, NÁM. SÍTNÁ 3105, KLADNO

Elektrickou aktivitu mozku zaznamenáváme pomocí EEG (elektroencefalografu). Nedílnou součástí vyšetření je detekce grafoelementů. Metody, které jsou založeny na matematickém principu klasifikace, je vždy nutné přizpůsobit stochastickému rázu EEG signálu. Hojně využívanou metodou je modifikovaný algoritmus k-means. Tento přístup ale skýtá omezení v prostorově prolnutých shlucích. U hustotně založeného algoritmu DBSCAN se tento problém neobjevuje. Zároveň nabízí algoritmus DBSCAN velké množství modifikací uzpůsobených pro více dimenzionální data.

Cílem diplomové práce je otestovat účinnost algoritmu DBSCAN na klasifikaci segmentů EEG signálu na základě 23 vypočtených příznaků, případně navrhnout vhodnou adaptaci algoritmu pro EEG signál.

Zpracovávaná data byla naměřena v Nemocnici Na Bulovce na 20 pacientech, kterým bylo indikováno vyšetření na základě podezření na nemoc epilepsii. Pacienti byli obou pohlaví ve věku mezi 26 – 60 roky. Délka záznamu se pohybuje od 15 minut do 24 hodin. Jednotlivé kanály EEG záznamu jsou klasifikovány zvláště pro lepší přehlednost při vyhodnocení klasifikace. Z vyhodnocených dlouhodobých záznamů jsou vybírány náhodné segmenty, u kterých 2 nezávislí odborníci validují příslušnost k dané třídě. Signál rozřazujeme do tříd odpovídající epileptické, svalové a sinusové aktivitě, artefaktům způsobeným špatnou elektrodou a na pomalé průběhy vln. Na základě klinického vyhodnocení bude provedena ROC analýza.



Z prvotních výsledků vyplývá, že DBSCAN není vhodný pro klasifikaci EEG záznamů, neboť není schopen odlišit zašuměné artefakty. Modifikovaný algoritmus, který 23D prostor dělí pomocí buněk adaptivních rozměrů, má lepší senzitivitu u segmentů se sinusovým charakterem. Tato metodika zároveň detekuje i epileptické grafoelementy s nízkou amplitudou. Nedostatky algoritmu se projevují v špatném zařazení pomalých očních artefaktů. K zefektivnění klasifikace by přispěl konkrétní výběr příznaků pro danou metodu. Modifikovaný DBSCAN lze využít k hodnocení záznamů EEG, ale stále je nutná intervence lékaře.

### **Klíčová slova**

EEG; DBSCAN; automatická klasifikace