České vysoké učení technické v Praze
Fakulta biomedicínského inženýrství
Katedra biomedicínské techniky

Extrakce, redukce dimenze a klasifikace příznakového prostoru EEG

Extraction, reduction dimension and classification of the EEG feature space

Bakalářská práce

Autor bakalářské práce: Eva Černá
Vedoucí bakalářské práce: Ing. Marek Piorecký

květen 2018

České vysoké učení technické v Praze, Fakulta biomedicínského inženýrství

Katedra biomedicínské techniky

Akademický rok: 2017/2018

Zadání bakalářské práce

Student: **Eva Černá**

Obor: Biomedicínský technik

Téma: **Extrakce, redukce dimenze a klasifikace příznakového prostoru EEG**

Téma anglicky: Extraction, reduction dimension and classification of the EEG feature space

Zásady pro vypracování:

Pomocí redukce dimenze extrahujte nové informace z příznakového prostoru EEG záznamů 10 pacientů s podezřením na epilepsii. V programovém prostředí MATLAB vytvořte metodiku pro simulaci EEG příznakového prostoru, na základě níž prostudujete vztahy mezi příznaky. Na reálném příznakovém prostoru proveďte redukci dimenze pomocí lineární a nelineární metody. Statisticky zhodnoťte rozdíly v klasifikaci na základě shlukovacího algoritmu na redukovaném reálném i simulovaném příznakovém prostoru.

Seznam odborné literatury:

[1] Krajča V., Mohylová J. , Číslicové zpracování neurofyziologických signálů, ed. 1st, ČVUT Praha, 2011, ISBN 9788001047217

[2] J. Birjandtalab, M. Baran Pouyan, M. Nourani, Nonlinear dimension reduction for EEG-based epileptic seizure detection, Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on, ročník 21, číslo 15937139, 2016, 10.1109/BHI.2016.7455968

[3] Vinay K. Ingle, John G. Proakis, Digital signal processing using MATLAB, ed. Third edition, CENGAGE Learning, 2012, ISBN 978-1-111-42737-5

Zadání platné do: 20.09.2019

Vedoucí: Ing. Marek Piorecký

Konzultant: Ing. Jan Štrobl

.....
vedoucí katedry / pracoviště

.....
děkan

V Kladně dne 19.02.2018

Obsah

Seznam použitých symbolů a zkratek

EEG	elektroencefalograf
EMG	Elektromyografické artefakty
EPI	Epileptická aktivita
PHYSIO	Fyziologická aktivita
WRONGEL	Artefakty ze špatné elektrody
23D	23-dimensionální
2D	2-dimensionální
FFT	Rychlá Fourierova transformace
WF	Wave-Finder
MATLAB	Matrix Laboration
PC	Hlavní komponenta
PCA	Analýza hlavních komponent
SVD	Singulární rozklad
t-SNE	t-Distributed Stochastic Neighbor Embedding
ROC	Receiver operating characteristics
TP	True positive
FP	False positive
TN	True negative
FN	False negative
PPV	Pozitivní prediktivní hodnota
PDF	Pravděpodobnostní hustotní funkce

Seznam tabulek

Seznam obrázků

Abstrakt

Pomocí elektroencefalografu zaznamenáváme elektrickou aktivitu mozku. Aby nám EEG signál poskytl užitečné informace o funkci mozku, je nutné jej předzpracovat, segmentovat a u segmentů vypočítat příznaky. K analýze EEG signálu jsou používány různé příznaky. Počet vypočítaných příznaků odpovídá počtu dimenzí výsledného příznakového prostoru. Protože vysokodimenzionální prostory jsou náročné na počítačové zpracování, je vhodné dimenze příznakového prostoru redukovat. Analyzovala jsem vlastnosti reálného příznakového EEG prostoru pomocí boxplotových grafů a pravděpodobnostních hustotních funkcí. V programovém prostředí MATLAB byla vytvořena funkce k simulaci příznakového prostoru EEG. Reálný i simulovaný příznakový prostor jsem redukovala lineární technikou analýzy hlavních komponent (PCA), metodou kovariance a metodou singulárního rozkladu, a nelineární technikou t-distributed Stochastic Neighbor Embedding (t-SNE). Redukované prostory jsem klasifikovala shlukovacím algoritmem k-means a následně provedla statistické hodnocení této klasifikace ROC analýzou. Byla vytvořena metodika na simulaci příznakového prostoru na základě vlastností reálného prostoru, která je využitelná k testování klasifikačních algoritmů. Epileptickou aktivitu je vhodnější klasifikovat do dvou tříd. Algoritmus t-SNE dokáže lépe separovat jednotlivé třídy EEG signálu, můžeme tedy předpokládat, že v příznakovém prostoru mají segmenty nelineární vztahy. K-means se nejeví jako vhodný klasifikační algoritmus pro EEG příznakové prostory.

Klíčová slova

EEG, k-means, PCA, příznaky, redukce dimenze, simulace příznakového prostoru EEG, t-SNE

Abstract

We record electrical activity of the brain using electroencephalograph. The preprocessing is necessary to provide useful information about brain function. We segment signal and calculate features. A variety of features are used to analyse the EEG signal. The number of computed features corresponds to the number of dimensions of the resulting feature space. It is advisable to reduce the size of the feature space because the large-dimensional spaces are difficult to computer processing. I analysed the properties of a real EEG feature space by using box-plot graphs and probability density functions. It has been created the function to simulation EEG feature space in the MATLAB. I reduced real and simulated feature space by linear techniques Principal Component Analysis (PCA), by method covariance and singular value decomposition, and nonlinear technique t-distributed Stochastic Neighbor Embedding (t-SNE). I have classified the reduced spaces with the clustering algorithm and then performed a statistical evaluation of this classification by ROC analysis. A methodology has been developed to simulate the feature space based on the real space properties. It can be used to test the classification algorithms. Epileptic activity is more appropriate to classify into two classes. The t-SNE algorithm is better to separate the clusters. We can assume that segments have non-linear relationships in the feature space. K-means does not appear to be a suitable classification algorithm for EEG feature spaces.

Key words

EEG, features, k-means, PCA, reduction dimension, simulation feature space of EEG, t-SNE

Poděkování

Děkuji mému vedoucímu, panu Ing. Marku Pioreckému, za velkou podporu a příkladné vedení mé bakalářské práce. Znovu děkuji panu Ing. Marku Pioreckému, dále paní Ing. Václavě Piorecké, panu doc. Ing. Vladimíru Krajčovi, CSc. a Ing. Vlastimilu Koudelkovi, Ph.D. za spolupráci při publikaci práce Simulation, modification and dimension reduction of EEG feature space na World Congress on Medical Physics and Biomedical Engineering 2018 v Praze. Také děkuji celému "BRAIN Teamu"za pomoc a věcné připomínky. Děkuji své partnerce, své rodině a svým pracovním kolegyním za podporu po celou dobu studia.

Prohlášení

Prohlašuji, že jsem bakalářskou práci s názvem **Extrakce, redukce dimenze a klasifikace příznakového prostoru EEG** vypracovala samostatně a použila k tomu úplný výčet citací použitých pramenů, které uvádím v seznamu přiloženém k bakalářské 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), ve znění pozdějších předpisů.

V . . . . . dne . . . .

Eva Černá

1 Úvod

Mozkové neurony při své aktivitě produkují a šíří elektrický potenciál. Při běžném neinvazivním vyšetření elektroencefalografie (EEG) z povrchu skalpu nejsme schopni zachytit jednotlivé akční potenciály, ale jsme schopni snímat sumační elektrický signál z celého mozku. Tento signál nám dává představu o funkci mozku a jeho aktivitě. EEG vyšetření je tedy funkčním vyšetřením mozku.

Aby nám EEG signál mohl poskytnout užitečné informace, musíme ho umět analyzovat. Analýza signálu probíhá v několika krocích. Jelikož je snímaný signál velice slabý a obsahuje pro nás i nedůležité frekvence, musí se signál zesílit, převést do digitálního tvaru a filtrovat. Takto předzpracovaný signál následně segmentujeme (dělíme na kratší úseky se stejnou nebo podobnou charakteristikou), extrahujeme příznaky a klasifikujeme.

Příznak je veličina, která popisuje konkrétní vlastnost jednotlivých segmentů EEG signálu. Příznaků si můžeme vymyslet teoreticky neomezené množství, ale prakticky by měly být schopny signál popsat jak z amplitudového, tak z frekvenčního hlediska. Čím více příznaků zvolíme, tím více-dimenzionální prostor vznikne. Segmentace a výběr příznaků ovlivňují klasifikaci signálu.

Pro svou práci jsem si vybrala systém 23 příznaků aplikovaných v programu WaveFinder (WF), jehož autorem je doc. Ing. Vladimír Krajča, CSc., a který je využíván v klinické praxi. Jak jsem uvedla výše, n příznaků vytvoří n-dimenzionální prostor. V mém případě se bude jednat o 23-dimenzionální prostor (23D). Není v lidských možnostech si v takto dimenzionálním prostoru představit rozložení jednotlivých segmentů, aby byla možná jejich optická separace na základě blízkosti (shluků v prostoru), a také s rostoucí dimenzionalitou prostoru rostou nároky na výpočetní techniku a na čas zpracování signálu. Proto existují snahy o redukování dimenze příznakového prostoru ještě před klasifikací signálu.

Analyzovala jsem metody redukce dimenzí používané při zpracování EEG signálů. Rozhodla jsem se porovnat lineární analýzu hlavních komponent (PCA), jako nejvíce používanou metodu redukce dimenze v EEG prostoru, s nelineární t-distributed Stochastic Neighbor Embedding (t-SNE). Toto porovnání lineární a nelineární techniky nám poskytne lepší představu o vztazích mezi segmenty užitečných signálů a artefaktů v EEG prostoru. Další možností, která nám může poskytnout informace o EEG prostoru, jsou boxplotové grafy a pravděpodobnostní hustotní funkce. Obojí jsem využila k analýze vlastností reálného příznakového prostoru.

Ve vědě a výzkumu se kvalita vyvíjených programů a algoritmů nejprve testuje na simulovaných signálech (datech). Simulovaný EEG signál se i přes vysokou snahu neshoduje s reálným. Simulací signálu je tedy zaneseno zkreslení dat. Proto jsme společně s Ing. Markem Pioreckým, Ing. Václavou Pioreckou, doc. Ing. Vladimírem Krajčou, CSc. a Ing. Vlastimilem Koudelkou, Ph.D. na základě znalostí o reálném příznakovém prostoru vyvinuli algoritmus na simulaci tohoto prostoru. Tento algoritmus jsme implementovali do programového prostředí MATLAB a vytvořili tak program, který může pomoci dalším studentům a vědcům.

Vytvořenou funkci jsem použila k simulaci 23D příznakového prostoru. Tento simulovaný i reálný 23D prostor jsem redukovala na 2D prostory a následně je klasifikovala shlukovacím algoritmem k-means. Klasifikaci jsem statisticky zhodnotila ROC analýzou.

1.1 Přehled současného stavu

Existují různé metody, jak redukovat dimenze prostoru EEG. Autoři v [1] se zabývali porovnáváním metod redukce dimenze a metod výběru příznaků pro detekci Event Related Potential (ERP). Redukovali dimenze analýzou hlavních komponent (PCA), Sparse PCA (SPCA), Empirical Mode Decomposition (EMD), a Local Mean Decomposition (LMD). Redukovaná EEG data klasifikovali lineární diskriminační analýzou (LDA) a seřadili použité kanály EEG na základě klasifikačního výkonu. Autoři v [2] porovnávali PCA,

analýzu nezávislých komponent (ICA) a lineární diskriminační analýzu (LDA) jako metody používané ke snížení rozměru dat. Pomocí podpůrných vektorových strojů (SVM) redukovaný prostor klasifikovali do dvou diskrétních skupin: epileptická aktivita a normální aktivita. Redukcí dimenze se také zabývali autoři v [3] kvůli analýze mozkových oscilací. Jejich myšlenkou je použití redukce dimenzionality s prostorově-spektrálním rozkladem (SSD) namísto běžně a téměř výlučně používané metody PCA. V práci [4] autoři použili t-distributed Stochastic Neighbor Embedding (t-SNE) pro nelineární redukci dimenze. Zvolený původní počet příznaků snížili pomocí analýzy hlavních komponent autoři v [5] na 2D prostor a ten klasifikovali pomocí algoritmu k-nejbližších sousedů (k-NN). V článku [6] autoři použili ke klasifikaci EEG dat metodu PCA a metodu k-means.

Nenalezla jsem žádnou dostupnou literaturu, která by se zabývala simulací EEG příznakového prostoru.

1.2 Cíle práce

Cílem práce je zhodnotit matematické vztahy příznaků v prostoru, ověřit použití metod redukce dimenze s ohledem na linearitu a nelinearitu vztahů v příznakovém prostoru EEG. Zkoumané příznaky byly vybrány s ohledem na použití u dat pacientů s podezřením na epilepsii. Chceme navrhnout metodiku pro simulaci příznakového EEG prostoru a implementovat ji jako funkci v programovém prostředí MATLAB. Vlastnosti a vztahy příznaků ověřím na reálném i simulovaném příznakovém prostoru. Klasifikačními algoritmy ověřím distribuci shluků v redukovaném reálném i simulovaném příznakovém prostoru. Výsledky vyhodnotím pomocí ROC analýzy.

2 Teoretická část

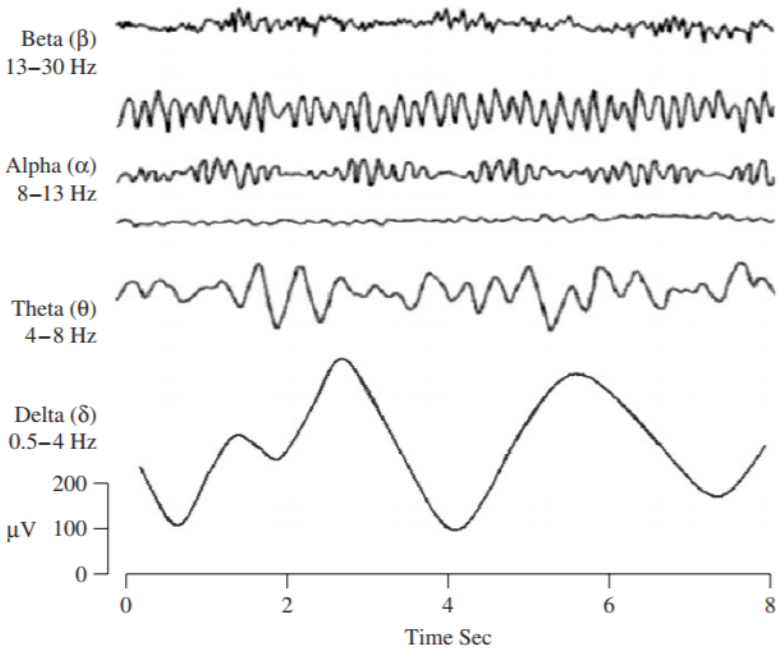
2.1 Elektroencefalogram

Elektroencefalogram (EEG) je záznam elektrické aktivity mozku. Má náhodný charakter a obsahuje užitečné informace o stavu mozku. Mnoho poruch mozku je diagnostikováno vizuální kontrolou EEG signálů. U zdravých dospělých se mění hodnoty amplitudy a frekvence takových signálů během změny jednotlivých stavů, jako je bdění a spánek. Charakteristiky jednotlivých vln se také mění s věkem. Signál jako celek je nestacionární, ale můžeme jej rozdělit do úseků stacionárních průběhů. Můžeme tedy tvrdit, že EEG signál je kvazistacionární (po částech stacionární). Tento předpoklad usnadňuje vyhodnocování záznamů. Amplitudový rozsah EEG je 2 - 100 µV (při epileptických záchvatech až 300 µV) a frekvenční rozsah standardně do 100 Hz. Největší část výkonu se nachází mezi 0,5 a 30,0 Hz. [7, 8, 9]

Jedním z kritérií pro popis EEG signálu jsou v něm obsažené frekvence. Tyto frekvence se rozdělují do čtyř základních frekvenčních pásem označovaných písmeny řecké abecedy, jak je vidět na obrázku 2.1. [10]

Delta rytmus

Delta vlny leží v rozmezí 0,5 až 4,0 Hz. Nachází se převážně u kojenců do 1 roku a u hlubokých spánkových stádií dospělých. Delta vlny jsou primárně spojeny s hlubokým spánkem. Nachází-li se v EEG záznamu dospělého ve stavu bdění, vždy se jedná o patologický projev. Čím je delta vlna spektrálně čistší a má vyšší amplitudu, tím je její patologický význam větší. Vlny delta se vyskytují i v transu a hypnóze. Ve spánku mají amplitudu i 100 µV. [8, 10]



Obrázek 2.1: Čtyři typické dominantní mozkové rytmy [8].

Theta rytmus

Theta vlny leží v rozmezí 4,0-7,5 Hz. Objevují se při usínání, hrají důležitou roli v dětství. Změny v rytmu theta vln jsou zkoumány v emočních studiích. Theta vlny se objevují v centrální, temporální (spánkové) a parietální (temenní) oblasti. Jestliže je amplituda theta vlny dvakrát vyšší než amplituda alfa vlny (případně 30 μV, jestliže alfa aktivita chybí), jedná se o patologický stav (proto jsou obvyklé u nižších stupňů kómatu). [8, 10]

Alfa rytmus

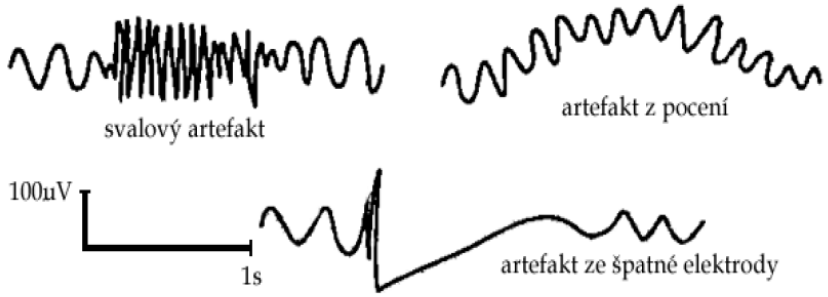
Alfa vlny mají frekvenční rozsah od 8 Hz do 13 Hz. Amplituda je většinou menší než 50 μV. V bdělém stavu bez duševní činnosti a při fyzické relaxaci se maximum alfa vln nachází nad zadními oblastmi mozkových hemisfér (nejlépe při zavřených očích). Alfa rytmus je především aktivitou optického analyzátoru. Alfa aktivita je snížena nebo eliminována otevřením očí, slyšením neznámých zvuků, úzkostí nebo duševní koncentrací a pozorností. Původ a fyziologický význam alfa vln je stále neznámý. [7, 10, 8]

Beta rytmus

Beta rytmus je elektrická aktivita mozku, která se pohybuje v rozmezí 14-26 Hz (ačkoli v některé literatuře není uvedena žádná horní hranice). Beta rytmus je obvyklý bdělý rytmus mozku spojený s aktivním myšlením, aktivní pozorností, zaměřením na vnější svět nebo řešením konkrétních problémů a nachází se u normálních dospělých. Vysoká amplituda beta vln se může vyskytovat, když je člověk v panickém stavu. Rytmická beta aktivita se vyskytuje hlavně v čelních a centrálních oblastech. Beta rytmus se obvykle netlumí zrakovým vjemem. Amplituda je nejčastěji 10-30 µV. [8, 10]

Artefakty

V záznamu EEG se kromě užitečného signálu (signálu vhodného k analýze) také objevují rušivé signály, tzv. artefakty. Artefakty mohou být způsobeny samotným pacientem, v tom případě hovoříme o biologických artefaktech. Technické artefakty jsou způsobeny okolím pacienta. Typickými artefakty jsou: rušení síťovým kmitočtem o frekvenci 50 Hz, artefakt ze špatné elektrody (kolísání isolinie, ztráta kontaktu), svalové artefakty, pocení pacienta apod. Typické signály některých artefaktů jsou uvedeny na obrázku 2.2. [10, 9]



Obrázek 2.2: Artefakty vyskytující se v EEG záznamech [10].

2.2 Zpracování EEG signálu

EEG data se zaznamenávají na digitálních (číslicových) EEG přístrojích [9]. Předzpracování EEG signálu probíhá ještě v přístroji. Signál se zesílí, převede do digitálního tvaru a filtruje. Takto předzpracovaný signál můžeme použít k analýze. Analýza EEG signálu probíhá v několika krocích. Nejprve signál segmentujeme, abychom získali interval, kde je signál stacionární. Poté provádíme extrakci příznaků a dle těchto příznaků signál klasifikujeme.

2.2.1 Segmentace

EEG signál nemá stacionární charakter. Segmentace je proces, při kterém se signál rozdělí do úseků o menší délce. V intervalech těchto úseků neboli segmentů předpokládáme, že je signál stacionární.

Obecně existují dva druhy segmentace - fixní a adaptivní. Fixní segmentace rozděluje záznam do úseků konstantní délky, ale hranice těchto úseků nerespektují charakter signálu. Proto je mnohem lepší rozdělit signál do po částech stacionárních úseků variabilní délky v závislosti na výskytu úseků stacionarit v signálu. To provádí adaptivní segmentace signálu. [9]

Adaptivní segmentace funguje lépe než fixní. To potvrzuje používání adaptivních segmentačních metod při zpracování nestacionárních signálů jako je EEG. [11]

Některé metody adaptivní segmentace dokáží pracovat jen se signály z jednoho kanálu, např. adaptivní segmentace na základě lineární predikce nebo adaptivní segmentace na základě autokorelační funkce. Nevýhodou těchto metod je jejich výpočetní složitost a zejména nemožnost nezávisle segmentovat více kanálů současně. [9]

Naproti tomu existuje metoda adaptivní segmentace na základě dvou spojených oken a jednoduché míry diference, která umožňuje segmentaci vícekanálových signálů v reálném čase. Segmentace každého kanálu je nezávislá na segmentaci ostatních kanálů. [12]

2.2.2 Výpočet příznaků

Tabulka 2.1: Příznaky používané pro klasifikaci [13]

Pořadí	Příznak	Popis příznaku
1	SIGM	variabilita signálu v daném segmentu
2	APOS	maximální pozitivní amplituda v daném segmentu
3	ANEG	maximální negativní amplituda v daném segmentu
4	DELT1	FFT hodnota v 1. části delta frekvenčního pásma (0,5 - 1,5 Hz)
5	DELT2	FFT hodnota v 2. části delta frekvenčního pásma (2 - 3,5 Hz)
6	THET1	FFT hodnota v 1. části theta frekvenčního pásma (4,0 - 5,5 Hz)
7	THET2	FFT hodnota v 2. části theta frekvenčního pásma (6,0 - 7,5 Hz)
8	ALPH1	FFT hodnota v 1. části alfa frekvenčního pásma (8 - 10 Hz)
9	ALPH2	FFT hodnota v 2. části alfa frekvenčního pásma (10,5 - 12,5 Hz)
10	SIGMA	FFT hodnota signálu v sigma frekvenčním pásmu (18 - 29 Hz)
11	BETA	FFT hodnota signálu v beta frekvenčním pásmu (13,5 - 17,5 Hz)
12	MAX1D	maximální hodnota první derivace v segmentu
13	MAX2D	maximální hodnota druhé derivace v segmentu
14	mf	hodnota střední frekvence v segmentu
15	MD1	střední hodnota první derivace v segmentu
16	MD2	střední hodnota druhé derivace v segmentu
17	mob	Hjorthův parametr mobility
18	comp	Hjorthův parametr komplexity
19	act	Hjorthův parametr aktivity
20	LOfC	délka křivky v segmentu
21	NlinE	nelineární energie segmentu
22	ZC	počet průchodů signálu nulou
23	Peaks	frekvence dominantní složky výkonu spektra

Ve své práci využívám příznaky implementované v programu WF. Tyto příznaky byly ověřeny v lékařské praxi pro klasifikaci záznamů pacientů s podezřením na epilepsii. Všechny příznaky získané z programu WF jsou následně normovány. Normalizace je provedena na základě maxima-minima. Příznaky tak nabývají hodnot od 0 do 1. [9]

Seznam těchto příznaků je uveden v tabulce 2.1. Na základě těchto 23 příznaků vytvořím simulovaný 23D příznakový prostor.

Pozitivní část signálu značí záporné hodnoty amplitudy a negativní část signálu znamená kladné hodnoty. Maximální pozitivní amplituda (APOS) je nejnižší hodnota napětí v daném segmentu, od které je odečtená stejnosměrná složka napětí, viz vzorec 1. [9]

$$APOS = A_{max} - \frac{\sum_{i=1}^{L} x_i}{L}, \tag{1}$$

kde L je délka segmentu a x_i je hodnota amplitudy i vzorku segmentu.

Maximální negativní amplituda (ANEG) je nejvyšší hodnota napětí v daném segmentu, od které je odečtená stejnosměrná složka napětí, viz vzorec 2. [9]

$$ANEG = A_{min} - \frac{\sum_{i=1}^{L} x_i}{L}, \tag{2}$$

kde L je délka segmentu a x_i je hodnota amplitudy i vzorku segmentu.

Maximální hodnota první derivace v segmentu (MAX1D) určuje sklon křivky a vypočítá se dle vztahu 3. [9]

$$MAX1D = max(x_{i+1} - x_i), \tag{3}$$

kde x_i je hodnota amplitudy i vzorku segmentu.

Maximální hodnota druhé derivace (MAX2D) určuje špičatost křivky v segmentu a vypočítá se dle vztahu 4. [9]

$$MAX2D = max(x_{i+4} - 2x_{i+2} - x_i), \tag{4}$$

kde x_i je hodnota amplitudy i vzorku segmentu.

Hjorthův parametr aktivita (act) je definována jako rozptyl signálu, který představuje jeho energii. Vzhledem k tomu, že směrodatná odchylka je definována jako [14]

$$\sigma(x) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}, \tag{5}$$

kde N je počet vzorků, x_i je hodnota amplitudy i vzorku segmentu a $\bar{x}$ je aritmetický průměr, může být aktivita zapsána jako [14]

$$act = \sigma(x)^2, \tag{6}$$

kde $\sigma(x)$ je směrodatná odchylka.

Hjorthův parametr mobility (mob) představuje střední hodnotu frekvence v segmentu a vypočítá se dle rovnice 7 [14]

$$mob = \frac{\sigma(x\frac{dx}{dt})}{\sigma(x)}, \tag{7}$$

kde $\sigma(x\frac{dx}{dt})$ je směrodatná odchylka první derivace signálu v segmentu a $\sigma(x)$ je směrodatná odchylka signálu v segmentu.

Hjorthův parametr komplexity (comp) představuje porovnání signálu v segmentu s harmonickým sinusovým signálem. Vypočítá se dle rovnice [14]

$$comp = \frac{mob(x\frac{dx}{dt})}{mob(x)}, \tag{8}$$

kde $mob(x\frac{dx}{dt})$ je Hjorthův parametr mobility derivace signálu v segmentu a $mob(x)$ je Hjorthův parametr mobility.

Délku křivky v segmentu si můžeme představit tak, že bychom změřili délku křivky nataženou jako niť. Matematicky se vypočítá dle rovnice [9]

$$LOfC = \sum_{i=1}^{N} abs[x(i) - x(i-1)], \tag{9}$$

kde N je počet vzorků, $x(i)$ je hodnota amplitudy i vzorku segmentu.

Nelineární energie segmentu se vypočítá dle vztahu [9]

$$NlinE = x^2(i) - x(i-1) \cdot x(i+1), \tag{10}$$

kde $x(i)$ je hodnota amplitudy i vzorku segmentu.

Počet průchodů signálu nulou ZC počítáme ze signálu, ze kterého je odečten průměr. Jako nula je definována velikost napětí 0,01 µV. Počet průchodů nulou se během epileptického záchvatu mění. [9]

Frekvence dominantní složky výkonu $Peaks$ spektra je frekvence dominantního vrcholu spektra signálu (frekvence spektra s největším výkonem). [9]

2.2.3 Klasifikace

Klasifikace je zařazování objektů do různých tříd. Slouží k rozdělování EEG segmentů podle podobnosti jejich charakteristik. Jejím cílem je pomoc lékařům při diagnostice. Klasifikace se v základu dělí na dvě hlavní metody, učení s učitelem a učení bez učitele.

Učení s učitelem

Učení s učitelem znamená poskytnout klasifikátoru informace o charakteristikách segmentů a o jejich zařazení do tříd. Takové segmenty považujeme za etalony. Klasifikátor poté rozděluje neznámé segmenty na základě podobnosti s etalonem. Mezi metody učení s učitelem patří k-nearest neighbors (k-NN), fuzzy k-nearest neighbor (fuzzy k-NN). Velkou výhodou učících se klasifikátorů je možnost on-line klasifikace. Naopak jejich nevýhoda spočívá v nutnosti předchozího trénování klasifikátorů. [9]

Učení bez učitele

Metody učení bez učitele nepotřebují trénovací množinu v podobě etalonů. Třídění segmentů EEG probíhá matematicky na základě podobnosti. Mezi metody učení bez učitele patří shluková analýza (cluster analysis) a fuzzy množiny.

Shlukové analýzy hledají přirozenou strukturu dat. S velkou výhodou je používáme ke klasifikaci neznámých objektů. Jejich nevýhodou je nemožnost on-line klasifikace. [9]

Jednou z nejznámějších metod shlukových analýz je klasifikace k-means. Tato metoda se v klinické praxi používá při klasifikaci EEG signálu, proto se jí budu více věnovat v podkapitole 3.4 K-means.

2.3 Redukce dimenze

Existují dva různé přístupy k řešení problému s vysokými rozměry v signálech EEG, výběr kanálů, redukce dimenze. Pro svou práci jsem si zvolila druhý z uvedených přístupů.

Výběr kanálů

Tyto techniky vybírají účinnou podmnožinu původních kanálů. Nemusí být vždy účinné, zvláště pokud existuje obrovský počet funkcí extrahovaných ze šumového prostředí, jakým je EEG na pokožce hlavy. [15]

Redukce dimenze

Vědci zkoumali, jak mapovat původní prostor funkce do menšího reprezentativního prostoru, aby se snížil ohromný počet funkcí. [15]

Techniky redukce dimenze umožňují vizualizaci dat ve dvou nebo třech dimenzích pro jejich lepší porozumění a zlepšení přesnosti klasifikačních modelů. Jednou metodou je promítání velkých rozměrových dat do nižšího euklidovského prostoru pomocí tradičních lineárních modelů jako je analýza hlavních komponent (PCA), singulární rozklad (SVD) a lineárních i nelineárních modelů jako je analýza nezávislých komponent (ICA). Na druhou stranu, nelineární algoritmy redukce dimenze jsou schopny rozvinout rozmanité části dat distribuovaných ve velkém rozměru získáním významné struktury dat založených na sousedních vzorcích. Zástupcem nelineárních algoritmu je t-Distributed Stochastic Neighbor Embedding (t-SNE). Je široce používán k vizualizaci vysoce dimenzionálních biologických dat, minimalizuje rozdíl mezi vysokým rozměrem a distribucí dat v malých rozměrech. Tímto způsobem jsou klíčové vztahy mezi datovými body zachovány. [4]

3 Metody

3.1 Analýza reálného příznakového prostoru

Data, použitá v této práci, jsou normované hodnoty 23 příznaků z tabulky 2.1 rozdělené do 4 tříd dle tabulky 3.1. Tato data pocházejí z měření pacientů ve Fakultní nemocnici Bulovka v Praze a byla zaznamenána pomocí přístroje Brain-Quick od firmy Micromed. Data jsem získala na základě návrhu projektu, který byl schválen etickou komisí Fakultní nemocnice Bulovka dne 28. června 2011. Jednalo se o klinické vyšetření trvající 15 až 40 minut. Měření bylo provedeno u 10 pacientů, 6 mužů a 4 žen ve věku od 26 do 60 let, s podezřením na epilepsii.

Pro zpracování získaných dat od těchto pacientů byl použit software C ++ a MATLAB R2015a. Vzorkovací frekvence v záznamu byla 128 Hz. Jako filtr byla použita pásmová propust 0,4 Hz a 70,0 Hz. Byla použita montáž se společnou elektrodou (average montage) a adaptivní odhad střední hodnoty. Po filtraci byla použita vícekanálová adaptivní segmentace na základě dvou spojených oken a jednoduché míry diference. Adaptivní segmentace je současně zpracována pro všechny kanály. Parametry adaptivní segmentace byly: délka okna - 128 vzorků, délka okna pro lokální identifikaci maxima - 30 vzorků, pohyblivý krok dvou připojených oken - 1 vzorek, minimální délka segmentu - 70 vzorků.

Tabulka 3.1: Seznam klasifikačních tříd

Číslo třídy	Název třídy	Zkratka v algoritmu
1	Fyziologická aktivita	PHYSIO
2	Epileptická aktivita	EPI
3	Elektromyografické artefakty	EMG
4	Artefakty ze špatné elektrody	WRONGEL

K analýze reálného příznakového prostoru jsem použila čisté (bez technického šumu či pohybových artefaktů) i zašumělé segmenty všech čtyř tříd. Čisté segmenty se hodí jako etalony k učení klasifikátorů. Počty analyzovaných čistých segmentů jsou uvedeny v tabulce 4.1. Počty analyzovaných zašumělých segmentů jsou uvedeny v tabulce 4.2.

K simulaci příznakového prostoru jsem použila pouze zašumělé segmenty, z důvodu větší reálnosti simulovaného prostoru (protože povrch hlavy, ze kterého je EEG snímáno, je šumové prostředí).

Ze všech příznaků všech tříd (čistých i zašumělých) jsem v programovém prostředí MATLAB vytvořila boxplotové grafy, uvedené v kapitole 4.1 Reálný příznakový prostor. Výhoda boxplotových grafů spočívá v tom, že dolní okraj každého boxu představuje dolní kvartil, horní okraj představuje horní kvartil a v každém boxu je označen medián. Pomocí '+' symbolů jsou jednotlivě vykreslovány odlehlé hodnoty. Tyto odlehlé hodnoty jsme ze simulace vyřadili z důvodu možné chybné klasifikace segmentu.

3.2 Metoda simulace příznakového prostoru

Na základě znalostí reálného příznakového prostoru jsme vytvořili a do programového prostředí MATLAB implementovali kód pro simulaci příznakového prostoru. Simulovaný příznakový prostor obsahuje normované hodnoty 23 příznaků z tabulky 2.1 rozdělené do 4 tříd dle tabulky 3.1. Na tvorbě této metody se podílel Ing. Marek Piorecký, já, Ing. Václava Piorecká, doc. Ing. Vladimír Krajča, CSc. a Ing. Vlastimil Koudelka, Ph.D.

Ke generování simulovaného příznakového prostoru jsme použili jádrový odhad pravděpodobnostní hustoty (kernel density estimation) každé třídy a příznaku očištěného od odlehlých hodnot. Odlehlé hodnoty jsme ze simulace vyřadili z důvodu možné chybné klasifikace segmentu.

Jádrový odhad je neparametrická reprezentace funkce hustoty pravděpodobnosti náhodné proměnné. Distribuci jádra můžeme použít, pokud parametrická distribuce nemůže správně popsat data, nebo pokud se chceme vyhnout předpokladům o distribuci dat. Jádrový odhad je klouzavý vážený průměr, u kterého nastavujeme šířku vyhlazovacího okna. [16]

Jádrový odhad hustoty $f(x)$ je definován vztahem

$$\widehat{f_h}(x) = \frac{1}{n \cdot h} \cdot \sum_{i=1}^{n} K \cdot (\frac{x - x_i}{h}), \tag{11}$$

kde $(x_1, x_2, \ldots, x_n)$ je náhodný výběr, n je počet vzorků, K je jádro a h je šířka vyhlazovacího okna. [16]

Jako hodnotu šířky vyhlazovacího okna jsme použili výchozí šířku vyhlazovacího okna funkce fitdist v programovém prostředí MATLAB, která je teoreticky optimální pro odhad hustoty pro normální distribuci. [17]

Poté jsme dle tohoto jádrového odhadu náhodně generovali data mezi lower adjacent value a upper adjacent value každého příznaku a třídy, viz algoritmus 1.

Data: reálný příznakový prostor
Result: simulace příznakového prostoru
for *1:počet tříd* **do**
  for *1:počet příznaků* **do**
    a = dolní přilehlá hodnota;
    b = horní přilehlá hodnota;
    $data = data \geq a$ a $data \leq b$;
    pd = jádrový odhad pravděpodobnostní hustoty $data$;
    $trida$ = simulace příznaků mezi a a b, rozložení dle pd;
  end
  $SimulovanyProstor = trida$;
end

Algoritmus 1: Simulace příznakového prostoru pro jednu třídu

V případě simulace "celého záznamu"(všechny třídy najednou) doporučujeme poměr tříd 1 : 14 : 242 : 6 - EMG : EPI : PHYSIO : WRONGEL. Segmentů třídy artefaktů ze špatné elektrody je v porovnání s EMG segmenty více, protože do třídy WRONGEL řadíme artefakty způsobené odpojenou elektrodou, špatným kontaktem elektrody s kůží díky potním artefaktům, nebo jakékoliv technické problémy s elektrodami.

3.3 Metody redukce dimenze

Na základě analýzy metod snižování dimenzí a současného stavu této problematiky jsem si pro svou práci zvolila dvě metody redukce dimenze. Analýzu hlavních komponent (PCA) jako zástupce lineárních technik a t-Distributed Stochastic Neighbor Embedding (t-SNE) jako zástupce nelineárních technik. Metodu PCA jsem zvolila, protože je jednou z nejvíce využívaných metod redukce dimenze, zvláště v EEG prostoru [1]. Výpočet PCA metodou kovariance jsem implementovala do prostředí MATLAB a na simulovaném příznakovém prostoru EEG jsem ověřila jeho funkčnost porovnáním s originální funkcí PCA v MATLAB, která jako výchozí metodu výpočtu využívá singulární rozklad (SVD). Algoritmus t-SNE nemá MATLAB implementován, nicméně jeho autoři, Laurens van der Maaten a Geoffrey Hinton, ho poskytují na svých webových stránkách `https://lvdmaaten.github.io/tsne/` jako bezplatný a otevřený zdrojový software, distribuovaný pod licencí FreeBSD.

3.3.1 Analýza hlavních komponent

Analýza hlavních komponent (PCA) je lineární technika redukce dimenze, kterou vyvinul a pojmenoval Harold Hotelling v roce 1933. V [18] je PCA detailně popsána, zde vysvětlím základní principy výpočtu hlavních komponent, které jsem ke své práci použila.

Hlavní myšlenkou analýzy hlavních komponent (PCA) je snížit dimenzi datové sady, která se skládá z velkého množství vzájemně propojených proměnných, a současně zacho-

vat co největší možnou odchylku v množině dat. Toho je dosaženo přeměnou na novou sadu proměnných, hlavních komponent (PC), které nejsou korelovány a které jsou uspořádány tak, že prvních pár hlavních komponent zachovává většinu odchylek přítomných ve všech původních proměnných. [19]

Analýza hlavních komponent je hojně využívanou metodou snižování dimenze. PCA hledá lineární kombinace multivariačních dat, které zachycují maximální množství rozptylu. Projekce, které PCA hledá, však nemusí nutně souviset s klasifikačními třídami, proto nemusí být pro problémy s klasifikací optimální. [1]

Existují tři různé způsoby, jakými lze vypočítat hlavní komponenty [20]:

1. Metoda kovariance (EIG)
2. Metoda singulárního rozkladu (SVD)
3. Algoritmus alternujících nejmenších čtverců (ALS)

3.3.2 Analýza hlavních komponent metodou kovariance

Výpočet PCA metodou kovariance sestává z několika kroků. Pro jednoduchost tuto metodu vysvětlím na dvoudimenzionálním prostoru. Předpokládejme, že máme matici $\mathbf{X}$ o rozměrech $(n \times p)$. Řádky matice $\mathbf{X}$ představují jednotlivá pozorování a sloupce matice $\mathbf{X}$ představují jednotlivé dimenze. V našem případě to budou dvě dimenze - x a y. V následujícím textu vycházím z [21], není-li uvedeno jinak.

Postup výpočtu hlavních komponent metodou kovariance:

1. Normalizace dat
2. Výpočet kovarianční matice
3. Výpočet vlastních vektorů a vlastních čísel kovarianční matice

4. Výběr komponent a vytvoření matice vlastních vektorů

5. Odvození nových dat

Normalizace dat

Aby PCA fungovala správně, musíme odečíst průměr každé dimenze z každého prvku příslušné dimenze. Takže od všech hodnot dimenze x odečteme $\bar{x}$ (průměr všech hodnot dimenze x) a od všech hodnot dimenze y odečteme $\bar{y}$ (průměr všech hodnot dimenze y). Výsledkem je normalizovaná datová sada, jejíž průměr je v každé dimenzi nula. [21]

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \tag{12}$$

Výpočet kovarianční matice

Kovariance je mírou toho, jak se dvě veličiny vzájemně mění. Pokud je hodnota kovariance kladná, znamená to, že obě dimenze společně rostou. Je-li hodnota kovariance záporná, pak jedna dimenze roste a druhá klesá. V posledním případě, je-li kovariance nulová, znamená to, že obě dimenze jsou na sobě nezávislé. Kovariance mezi 2 dimenzemi x a y je definována v rovnici 13. [21]

$$cov(x, y) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \tag{13}$$

Užitečný způsob, jak získat všechny možné hodnoty kovariance mezi všemi různými rozměry, je vypočítat všechny kovariance a dát je do matice. Vzhledem k tomu, že naše myšlená data jsou dvourozměrná, bude mít kovarianční matice $\mathbf{C}$ velikost 2×2. [21]

$$C = \begin{pmatrix} cov(x, x) & cov(x, y) \\ cov(y, x) & cov(y, y) \end{pmatrix} \tag{14}$$

Výpočet vlastních vektorů a vlastních čísel kovarianční matice

V matematice označuje vlastní vektor (anglicky eigenvector) dané transformace nenulový vektor, jehož směr se při transformaci nemění. Koeficient, o který se změní velikost vektoru, se nazývá vlastní číslo (anglicky eigenvalue). Vlastní vektory lze nalézt pouze pro singulární matice. Všechny vlastní vektory matice jsou ortogonální (kolmé), bez ohledu na to, kolik rozměrů matice má. Naše kovarianční matice je velikosti 2×2, má tedy 2 vlastní vektory. [22]

$$det(C - \lambda E) = 0 \tag{15}$$

V rovnici číslo 15 představuje λ vlastní číslo a **E** jednotkovou matici, což lze rozepsat rovnicí 16. [22]

$$\begin{vmatrix} cov(x,x) - \lambda & cov(x,y) \\ cov(y,x) & cov(y,y) - \lambda \end{vmatrix} = 0 \tag{16}$$

Vlastní vektor vypočítáme dle

$$Cu = \lambda u, \tag{17}$$

kde λ je vlastní číslo a u je vlastní vektor [22].

Výběr komponent a vytvoření matice vlastních vektorů

Vlastní vektor s nejvyšším vlastním číslem je hlavní složkou dat a ten, který prezentuje nejvýznamnější vztah mezi rozměry dat. Obecně platí, že jakmile jsou z kovarianční matice nalezeny vlastní vektory, je dalším krokem jejich uspořádání podle vlastního čísla, od nejvyššího až po nejnižší. To nám řadí komponenty podle jejich významu. Pokud vynecháme některé vlastní vektory, konečná data budou mít menší rozměry než původní.

Nyní je třeba vytvořit matici vlastních vektorů, jejíž sloupce budou vlastní vektory seřazené sestupně dle klesajícího vlastního čísla.

Odvození nových dat

To je poslední krok v PCA. Jakmile jsme si vybrali komponenty (vlastní vektory), které si přejeme uchovat v našich datech a vytvořili jsme matici těchto vlastních vektorů, jednoduše tuto matici transponujeme a násobíme jí vlevo od transponovaných původních normalizovaných dat. Tím získáme matici nových dat. [21]

$$NewData = MatrixEigenVectors^T \times NormalizedOriginalData^T, \tag{18}$$

kde **NewData** je matice nových dat, ve které nyní řádky matice představují jednotlivé dimenze a sloupce matice představují jednotlivá pozorování. **MatrixEigenVectors** je matice vlastních vektorů, jejíž sloupce jsou vlastní vektory seřazené sestupně dle klesajícího vlastního čísla a **NormalizedOriginalData** je matice normalizovaných původních dat.

Data: příznakový prostor
Result: redukce dimenze příznakového prostoru
for *1:počet příznaků* **do**
  $normX = X$ - $mean$;
end
$CovX$ = kovarianční matice $normX$;
výpočet vlastních čísel a vlastních vektorů $CovX$;
PC = výběr vlastních vektorů jako hlavních komponent ;
redukovaný prostor = PC * $normX'$;

Algoritmus 2: Redukce dimenze PCA metodou kovariance

3.3.3 Analýza hlavních komponent metodou singulárního rozkladu

PCA určuje množinu ortogonálních vektorů nazvaných hlavní komponenty, které jsou definovány lineární kombinací původních proměnných a seřazeny podle množství rozptylu vysvětleného ve směrech komponent. Koeficienty proměnných pro určení hlavních komponent jsou uloženy v loading matici. Vzhledem k množině I vzorků a J proměnných uspořádaných ve dvourozměrné matici $\mathbf{X}$ ($I \times J$) se význam hlavních komponent vypočítá singulární rozkladem (SVD) kovarianční matice $\mathbf{C}$ [23]:

$$C = \frac{X^T X}{I-1} = LS^2L^T = Z\Lambda Z^T, \tag{19}$$

kde $\mathbf{Z}$ ($J \times J$) je ortogonální matice, $\mathbf{S}$ ($J \times J$) je diagonální matice s nenulovými singulárními hodnotami na její diagonále, $\mathbf{L}$ ($J \times J$) je loading matice, jejíž každý j sloupec je vlastní vektor, koeficienty J proměnných pro definici j-té hlavní komponenty a Λ je diagonální matice, která obsahuje nezáporná vlastní čísla v sestupném pořadí ($\lambda_1 \geq \lambda_2 \geq ...\lambda_J \geq 0$). Každé vlastní číslo kóduje rozptyl týkající se příslušné hlavní komponenty. Z vlastních čísel lze vypočítat množství rozptylu (EV) a kumulativní množství rozptylu (CEV) přidružené ke každé m-té složce [23]:

$$EV_m = \frac{\lambda_m}{\sum_{j=1}^{J} \lambda_j} \tag{20}$$

$$CEV_m = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{J} \lambda_j} \tag{21}$$

Vzhledem k tomu, že v modelu PCA jsou zachovány pouze významné komponenty M, pak se rozměr loading matice L snižuje z ($J \times J$) na ($J \times M$) a vzorky se promítají

do nízkodimenzionálního prostoru definovaného významnými hlavními komponentami tímto způsobem [23]:

$$T = XL, \tag{22}$$

kde $\mathbf{T}$ $(I \times M)$ je matice skóre, která shromažďuje na každém m sloupci souřadnice I vzorků do m hlavní komponenty.

3.3.4 t-Distributed Stochastic Neighbor Embedding

Technika nazvaná t-Distributed Stochastic Neighbor Embedding (t-SNE) přeměňuje rozsáhlý soubor dat do matice dvojic podobností a následně tyto podobnosti vizualizuje. Metoda t-SNE je schopna velmi dobře zachytit velkou část místní struktury velkokapacitních dat a také odhalit globální strukturu, jako je přítomnost klastrů v několika měřítkách. [24]

Cílem t-SNE je zaujmout řadu bodů ve velkorozměrovém prostoru a najít věrnou reprezentaci těchto bodů v nižším rozměrovém prostoru, typicky 2D rovině. Tento algoritmus je nelineární, přizpůsobuje se podkladovým datům a provádí různé transformace v různých oblastech. [25]

Algoritmus t-SNE provádí následující obecné kroky při redukci dimenzí [26]:

1. Výpočet párových vzdálenosti mezi daty ve velkém rozměru.
2. Výpočet směrodatné odchylky σ_i pro každý vysoce rozměrný bod i tak, aby perplexita každého bodu byla na předem stanovené úrovni.
3. Výpočet matice podobností. To je společné rozdělení pravděpodobnosti P definované rovnicí 24.
4. Vytvoření počáteční sady nízkorozměrných bodů (typicky 2D [25]).

5. Iterativně aktualizovat body s nízkým rozměrem, aby se minimalizovala Kullback-Leiblerova divergence mezi Gaussovou distribucí ve vysokodimenzionálním prostoru a studentovou t-distribucí v nízkodimenzionálním prostoru. Tato optimalizace je nejvíce časově náročná část algoritmu.

V následujícím textu vycházím z [24], kde byla metoda t-SNE představena autory, a [15, 4], kde byla technika t-SNE použita při zpracování signálu EEG.

Technika t-SNE začíná převedením velkokapacitních euklidovských vzdáleností mezi datovými daty na podmíněné pravděpodobnosti, které představují podobnosti. Podobnost datového bodu x_j s datovým bodem x_i je podmíněná pravděpodobnost $p_{j|i}$, že x_i si vybere x_j jako svého souseda, pokud sousedé byli vybráni v poměru k jejich hustotě pravděpodobnosti pod gaussovským středem x_i. Pro blízké datové body je $p_{j|i}$ poměrně vysoká, zatímco u velice oddělených datových bodů bude $p_{j|i}$ téměř nekonečně malá (pro rozumné hodnoty rozptylu Gaussovy funkce, σ_i). Matematicky je podmíněná pravděpodobnost $p_{j|i}$ daná rovnicí 23. [24]

$$p_{j|i} = \frac{e^{-\|x_i - x_j\|^2/2\sigma_i^2}}{\sum_{k \neq i} e^{-\|x_i - x_k\|^2/2\sigma_i^2}}, \tag{23}$$

kde σ_i je odchylka Gaussovy funkce, která je soustředěna na datovém bodu x_i. Není pravděpodobné, že existuje jediná hodnota σ_i, která je optimální pro všechny datové body v datovém souboru, protože hustota dat se pravděpodobně bude lišit. V hustých oblastech je obvykle nižší hodnota σ_i vhodnější než v prořídlých oblastech. Konkrétní hodnotu σ_i nepřímo určuje sám uživatel prostřednictvím perplexity.

Perplexita říká (volně), jak vyvážit pozornost mezi lokálními a globálními aspekty našich dat. Parametr je v jistém smyslu odhadem počtu blízkých sousedů, které má každý bod. [25].

V tomto případě to znamená, že čím vyšší počet sousedů, tím vyšší hustota shluků. Výkonnost t-SNE je poměrně robustní vůči změnám v perplexitě a typické hodnoty jsou

mezi 5 a 50. Protože se zajímáme pouze o modelování párových podobností, nastavíme hodnotu $p_{i|i}$ na nulu.

Hodnoty $p_{i,j}$ jsou definovány jako symetrické podmíněné pravděpodobnosti, viz rovnice 24. [15, 4]

$$p_{i,j} = \frac{p_{j|i} + p_{i|j}}{2 \cdot n} \tag{24}$$

U nízkodimenzionálních protějšků y_i a y_j velkokapacitních datových bodů x_i a x_j je možné vypočítat pravděpodobnost, kterou označujeme $q_{i,j}$. Hodnoty $q_{i,j}$ jsou získány pomocí studentovi t-distribuce s jedním stupněm volnosti. [24]

$$q_{i,j} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \tag{25}$$

Používáme studentovu t-distribuci s jediným stupněm volnosti, protože jeho vlastností je, že $(1+\|y_i-y_j\|^2)^{-1}$ klesá se čtvercem vzdálenosti pro velké párové vzdálenosti $\|y_i-y_j\|$ v nízkodimenzionální mapě. Toto dělá mapové znázornění společných pravděpodobností (téměř) invariantní ke změnám v měřítku mapy pro body map, které jsou daleko od sebe. Jelikož se opět zajímáme pouze o modelování párových podobností, nastavíme hodnotu $q_{i,i}$ na nulu.

Pokud mapové body y_i a y_j správně modelují podobnost mezi vysokodimenzionálními datovými body x_i a x_j, pravděpodobnosti $p_{i,j}$ a $q_{i,j}$ budou stejné. Motivováno tímto zjištěním, t-SNE usiluje o nalezení nízkodimenzionální reprezentace dat, která minimalizuje nesoulad mezi $p_{i,j}$ a $q_{i,j}$. Přírodním měřítkem věrnosti, s níž $q_{i,j}$ modeluje $p_{i,j}$, je Kullback-Leiblerova divergence. Minimalizujeme jedinou Kullback-Leiblerovu divergenci mezi společným rozdělením pravděpodobnosti P, ve velkokapacitním prostoru a společným rozdělením pravděpodobnosti Q v nízkorozměrovém prostoru. [24]

$$KL(P\|Q) = \sum_i \sum_j p_{i,j} \cdot log\frac{p_{i,j}}{q_{i,j}} \tag{26}$$

Minimální hodnota Kullback-Leiblerovi divergence KL je počítána pomocí gradientu. [24]

$$\frac{\delta KL}{\delta y_i} = 4 \cdot \sum_j (p_{i,j} - q_{i,j}) \cdot (y_i - y_j) \cdot (1 + \|y_i - y_j\|^2)^{-1} \tag{27}$$

Algoritmus t-SNE přizpůsobuje svůj pojem "vzdálenost"regionálním změnám hustoty v datové sadě. Výsledkem je, že přirozeně rozšiřuje husté clustery a uzavírá spletité clustery. Vyrovnávání hustoty se děje podle předpokladu a je předvídatelnou vlastností t-SNE. Špatnou zprávou je, že vizualizace globální geometrie (vzdáleností mezi clustery) vyžaduje doladění. Neexistuje žádná hodnota perplexity, která by zachytila reálné vzdálenosti mezi všemi clustery. Oprava tohoto problému může být zajímavou oblastí pro budoucí výzkum. Základní zpráva je, že vzdálenosti mezi dobře oddělenými skupinami v grafu t-SNE nemusí mít pro náš výzkum význam. [25].

3.4 K-means

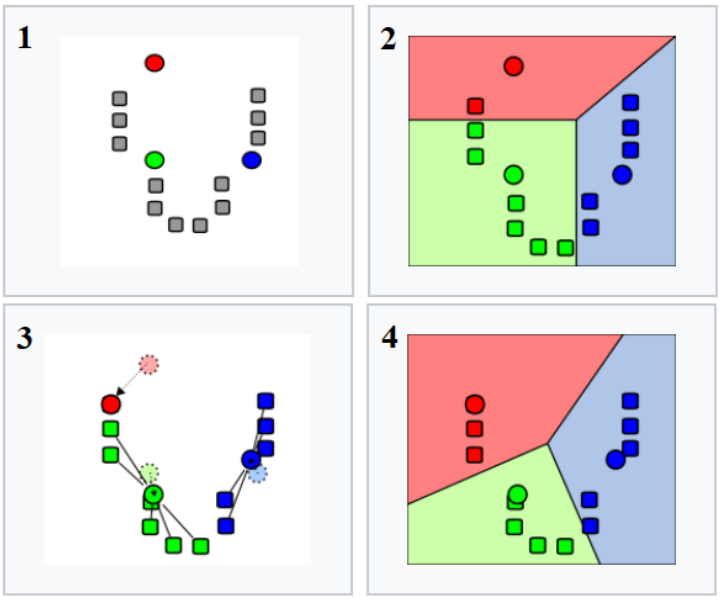
Po redukci dimenze reálného i simulovaného příznakového prostoru použiji shlukovací metodu k-means ke klasifikaci redukovaných 2D prostorů. Tuto metodu jsem si vybrala z důvodu jejího běžného klinického využití při analýze EEG signálů.

K-means patří mezi nehierarchické metody shlukové analýzy. Tyto metody iterativně hledají optimální rozdělení dat, kterým minimalizují určitou kriteriální funkci. [9]

Je to metoda kvantifikace vektoru a je velmi populární metodou clusterové analýzy. Hlavním cílem metody k-means je rozdělení n rozdílných pozorování do k skupin clusterů. Datový prostor tak může být rozdělen do mnoha užitečných buněk nazvaných jako Vo-

roného diagramy. K-means má vždy sklon nalézat shluky, které mají více nebo méně srovnatelný prostorový rozsah. [6]

Vstupním parametrem k-means je počet shluků k, do kterých mají být data rozdělena. Shluky jsou definovány svými centroidy (těžišti), což jsou středy shluků (každý shluk má své těžiště). [27]



Obrázek 3.1: Demonstrace algoritmu k-means. Čísla obrázků odpovídají číslům kroků v postupu výpočtu [27].

Postup výpočtu k-means se skládá z několika kroků [6] jak je vidět na obrázku 3.1 [27]:

1. Náhodná iniciace center shluků.
2. Přiřazení každého bodu x_i k nejbližšímu centru clusteru C_k nejčastěji pomocí výpočtu euklidovské vzdálenosti.
3. Každé centrum clusteru C_k je přepočítáno tak, aby se nacházelo v těžišti clusteru (v místě průměru všech bodů x_i, které ke clusteru patří).
4. Kroky 2-4 se opakují do konvergence algoritmu (dokud se centra clusterů C_k nestanou stabilní a všechny body x_i budou nejblíže k těžišti, které ke shluku patří).

3.5 Statistické zhodnocení

Statisticky jsem zhodnotila rozdíly v klasifikaci pomocí k-means na redukovaném reálném i simulovaném příznakovém prostoru. K tomuto zhodnocení rozdílů jsem využila ROC (Receiver operating characteristics) analýzy.

ROC analýza je technika pro vizualizaci, organizaci a výběr klasifikátorů na základě jejich výkonu. ROC křivky jsou běžně používány v lékařství i v oboru strojového učení. [28]

ROC analýza je využívána i při zpracování signálů EEG. Autoři v [29] testovali účinnost permutační entropie jako užitečného algoritmu pro detekci epileptických událostí v EEG. Využili ROC analýzu při hodnocení oddělitelnosti amplitudových rozdělení permutační entropie vyplývající z preiktální a interiktální fáze.

ROC analýza, stejně jako mnoho statistik používaných k vyhodnocení výkonnosti klasifikátorů, se nejlépe zobrazuje v konfuzní matici, viz 3.2. Tato matice porovnává výsledky testu s výsledky testu referenční klasifikace, v ideálním případě poskytuje skutečné zařazení objektu do třídy. [30]

		REFERENČNÍ KLASIFIKÁTOR	
		POZITIVNÍ	NEGATIVNÍ
TESTOVANÝ KLASIFIKÁTOR	POZITIVNÍ	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
	NEGATIVNÍ	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

Obrázek 3.2: Konfuzní matice porovnávající výsledky referenčního klasifikátoru a testovaného klasifikátoru.

True positive (TP)

Jako true positive je označen takový segment, který je klasifikátorem klasifikován jako pozitivní a reálně je také pozitivní. Například pokud k-means klasifikuje segment do třídy epileptické aktivity a tento segment skutečně do této třídy patří.

False positive (FP)

Jako false positive je označen takový segment, který je klasifikátorem klasifikován jako pozitivní, ale reálně je negativní. Například pokud k-means klasifikuje segment do třídy epileptické aktivity, ale tento segment ve skutečnosti patří do jiné třídy.

False negative (FN)

Jako false negative je označen takový segment, který je klasifikátorem klasifikován jako negativní, ale reálně je pozitivní. Například pokud k-means klasifikuje segment do jiné třídy než je třída fyziologické aktivity, ale tento segment ve skutečnosti do fyziologické aktivity patří.

True negative (TN)

Jako true negative je označen takový segment, který je klasifikátorem klasifikován jako negativní a reálně je také negativní. Například pokud k-means klasifikuje segment do jiné třídy než je třída fyziologické aktivity a tento segment ve skutečnosti také do jiné třídy patří.

Specificita

Specificita je schopnost klasifikátoru identifikovat true negative segmenty. Nízké množství FP segmentů indikuje vysokou specificitu klasifikátoru. Je to podíl všech TN segmentů, které klasifikátor odhalí, vůči součtu TN a FP segmentů. [28, 31]

Pokud má klasifikátor 100% specificitu znamená to, že dokáže odhalit všechny true negative segmenty a ani jeden segment nebude false positive.

$$Specificita = \frac{TN}{TN + FP} \tag{28}$$

Sensitivita

Sensitivita je schopnost klasifikátoru identifikovat true positive segmenty. Nízké množství FN segmentů naznačuje vysokou sensitivitu (citlivost) klasifikátoru. Je to podíl všech TP segmentů, které klasifikátor odhalí, vůči součtu TP a FN segmentů. [28, 31]

Pokud má klasifikátor 100% sensitivitu, znamená to, že dokáže odhalit všechny true positive segmenty a ani jeden segment nebude false negative.

$$Sensitivita = \frac{TP}{TP + FN} \tag{29}$$

Pozitivní prediktivní hodnota

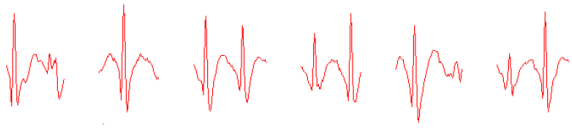
Pozitivní prediktivní hodnota (PPV) je podíl true positive segmentů a všech pozitivních segmentů. PPV indikuje preciznost klasifikátoru a má pro nás spolu se sensitivitou nejvíce vypovídající hodnotu. [28, 30]

$$PPV = \frac{TP}{TP + FP} \tag{30}$$

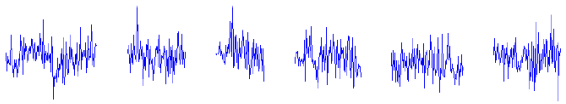
4 Výsledky

4.1 Reálný příznakový prostor

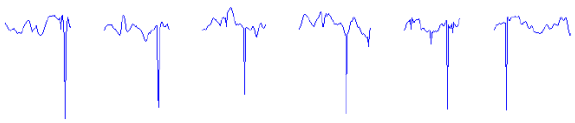
Ukázky signálů použitých klasifikačních tříd jsou uvedeny na obrázcích níže. Na obrázku 4.1 je pro názornost uvedena ukázka segmentů epileptické aktivity. Epileptická aktivita je z klinického hlediska nejdůležitější třída. Na obrázku 4.2 je uvedena ukázka segmentů svalových artefaktů. Svalová aktivita zcela zkresluje původní signál. Na obrázku 4.3 jsou segmenty artefaktů ze špatné elektrody. Tyto artefakty se většinou vyskytují jen v jednom kanálu, protože bývají způsobeny jen jednou elektrodou. Na obrázku 4.4 jsou segmenty fyziologické aktivity. Ze segmentů těchto tříd jsem analyzovala reálný příznakový prostor.



Obrázek 4.1: Ukázka segmentů epileptické aktivity, zobrazeno ve WF [32].

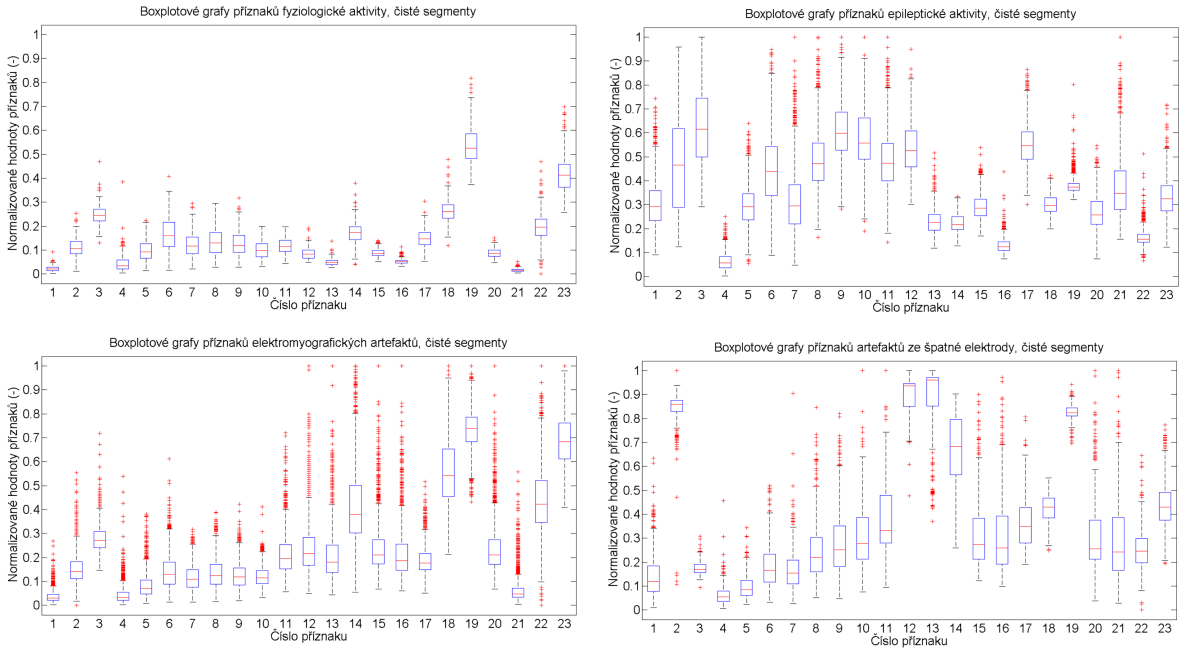


Obrázek 4.2: Ukázka segmentů elektromyografických artefaktů, zobrazeno ve WF [32].

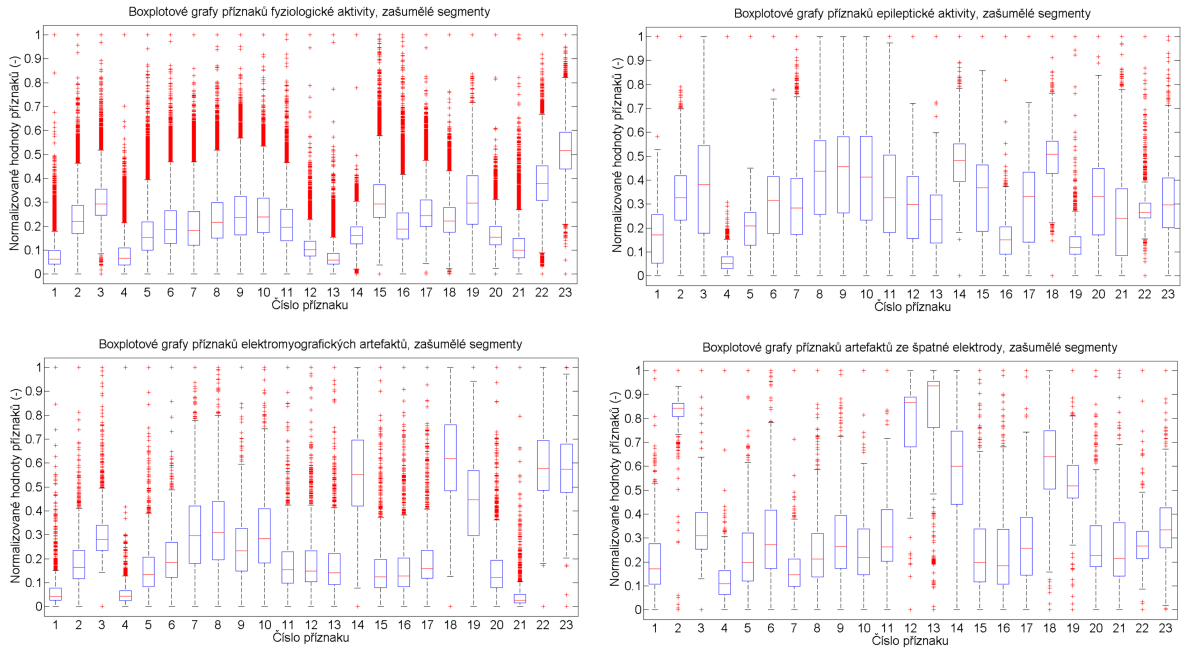


Obrázek 4.3: Ukázka segmentů artefaktů ze špatné elektrody, zobrazeno ve WF [32].

Obrázek 4.4: Ukázka segmentů fyziologické aktivity, zobrazeno ve WF [32].



Obrázek 4.5: Boxplotové grafy příznaků čistých segmentů tříd fyziologická aktivita, epileptická aktivita, elektromyografické artefakty a artefakty ze špatné elektrody.



Obrázek 4.6: Boxplotové grafy příznaků zašumělých segmentů tříd fyziologická aktivita, epileptická aktivita, elektromyografické artefakty a artefakty ze špatné elektrody.

Ze získaných dat jsem vytvořila boxplotové grafy zvlášť pro čisté segmenty a zvlášť pro zašumělé segmenty. Boxplotové grafy příznaků čistých segmentů tříd fyziologická aktivita, epileptická aktivita, elektromyografické artefakty a artefakty ze špatné elektrody jsou uvedeny na obrázku 4.5. Boxplotové grafy příznaků zašumělých segmentů tříd fyziologická aktivita, epileptická aktivita, elektromyografické artefakty a artefakty ze špatné elektrody jsou uvedeny na obrázku 4.6. Počty použitých čistých segmentů a jejich odlehlých hodnot jsou uvedeny v tabulce 4.1. Počty použitých zašumělých segmentů a jejich odlehlých hodnot jsou uvedeny v tabulce 4.2.

Tabulka 4.1: Velikost příznakového prostoru čistých segmentů

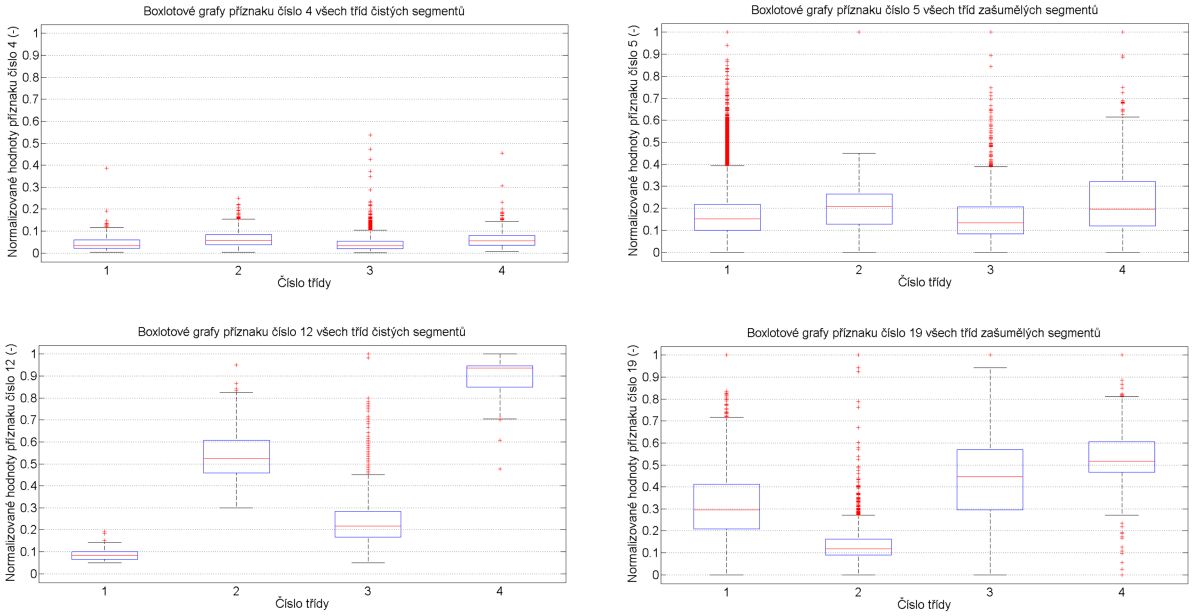
Třída	Počet segmentů	Počet příznaků	Počet odlehlých hodnot
Fyziologická aktivita	200	4 600	124
Epileptická aktivita	1 492	34 316	591
Elektromyografické artefakty	1453	33 419	1366
Artefakty ze špatné elektrody	483	11 109	411

Tabulka 4.2: Velikost příznakového prostoru zašumělých segmentů

Třída	Počet segmentů	Počet příznaků	Počet odlehlých hodnot
Fyziologická aktivita	19 763	454 549	13 755
Epileptická aktivita	2200	50 600	584
Elektromyografické artefakty	1 239	28 497	1 572
Artefakty ze špatné elektrody	495	11 385	450

Porovnala jsem boxplotové grafy každého příznaku čistých segmentů napříč všemi třídami. Nejvíce se překrývají boxy 4. příznaku u všech 4 tříd, viz obrázek 4.7 vlevo nahoře. Naopak u příznaku číslo 12 se nepřekrývaly boxy ani jedné z tříd, viz obrázek 4.7 vlevo dole.

Při porovnávání boxplotových grafů každého příznaku zašumělých segmentů napříč všemi třídami jsem zjistila, že nejvíce se překrývají boxy 5. příznaku u všech 4 tříd, viz obrázek 4.7 vpravo nahoře. Nenalezla jsem žádný příznak, u kterého by se nepřekrývaly žádné třídy. Ale u příznaku číslo 19 (Hjorthův parametr aktivity) se oddělil box třídy



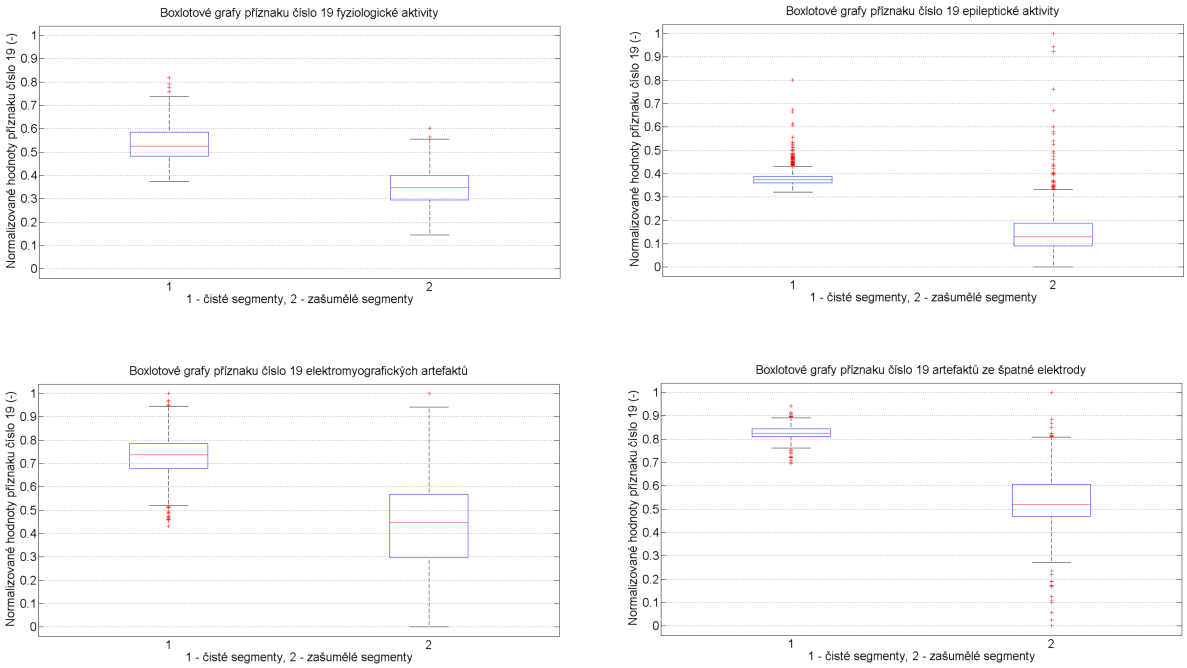
Obrázek 4.7: Vlevo jsou boxplotové grafy příznaku číslo 4 a příznaku číslo 12 všech tříd čistých segmentů. Vpravo jsou boxplotové grafy příznaku číslo 5 a příznaku číslo 19 všech tříd zašumělých segmentů.

epileptická aktivita, zatímco boxy ostatních tříd se překrývaly, viz obrázek 4.7 vpravo dole.

Tabulka 4.3: Porovnání kvartilů Hjorthova parametru aktivity mezi čistými a zašumělými segmenty všech tříd

Třída		Dolní kvartil	Medián	Horní kvartil
Fyziologická aktivita	čistá	0,4820	0,5262	0,5848
	zašumělá	0,2947	0,3463	0,4004
Epileptická aktivita	čistá	0,3605	0,3736	0,3886
	zašumělá	0,0902	0,1305	0,1874
Elektromyografické artefakty	čisté	0,6784	0,7366	0,7855
	zašumělé	0,2964	0,4474	0,5686
Artefakty ze špatné elektrody	čisté	0,8104	0,8226	0,8437
	zašumělá	0,4694	0,5191	0,6051

Porovnala jsem odpovídající příznaky mezi čistými a zašumělými segmenty. Kvůli nezávislosti změn hodnot kvartilů na počtu segmentů jsem v rámci třídy použila shodný počet segmentů. Příznak číslo 19 Hjorthův parametr aktivity je ovlivňován šumem nejvýrazněji ze všech příznaků. Na obrázku 4.8 jsou vidět rozdíly v boxech mezi čistými



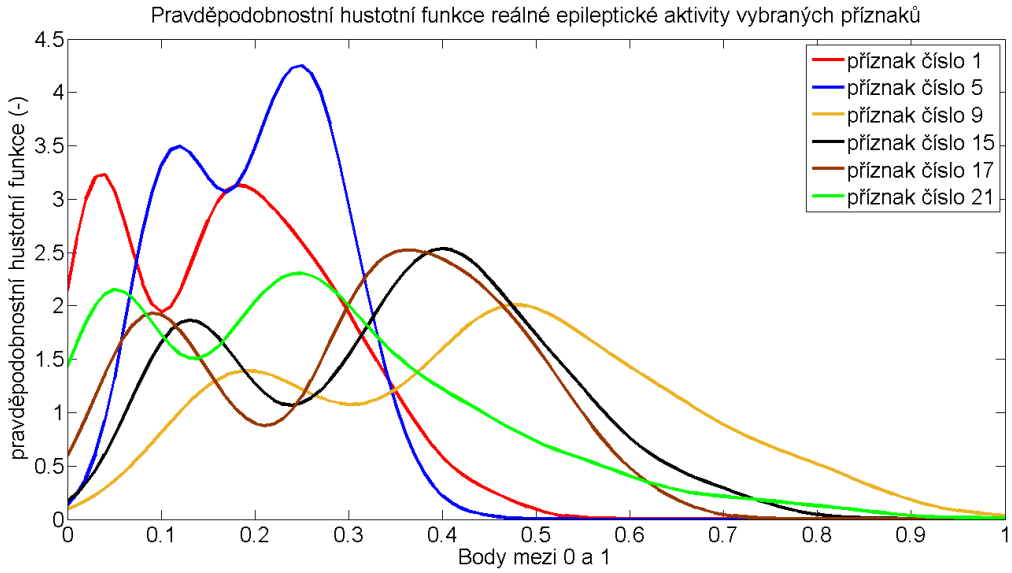
Obrázek 4.8: Boxplotové grafy příznaku číslo 19 Hjorthův parametr aktivity všech tříd

Tabulka 4.4: Procentuální rozdíl mezi kvartily příznaku Hjorthův parametr aktivity čistých a zašumělých segmentů

Třída	Dolní kvartil	Medián	Horní kvartil
Fyziologická aktivita	18,73 %	17,99 %	18,44 %
Epileptická aktivita	27,03 %	24,31 %	20,12 %
Elektromyografické artefakty	38,02 %	28,92 %	21,69 %
Artefakty ze špatné elektrody	34,10 %	30,35 %	23,86 %

a zašumělými segmenty. V tabulce 4.3 jsou vidět rozdíly v hodnotách všech kvartilů a v tabulce 4.4 jsou vypočítané procentuální rozdíly v kvartilech mezi čistými a zašumělými segmenty. U všech tříd došlo k poklesu hodnot kvartilů tohoto příznaku.

Na obrázku 4.9 jsou pravděpodobnostní hustotní funkce (PDF) reálné epileptické aktivity vybraných příznaků. Na křivkách jsou jasně patrné dvě špičky (peaky). V tabulce 4.5 jsou vypsány všechny příznaky, u kterých PDF tvoří dva peaky. Z 23 používaných příznaků v programu WF se jedná o 15 příznaků.



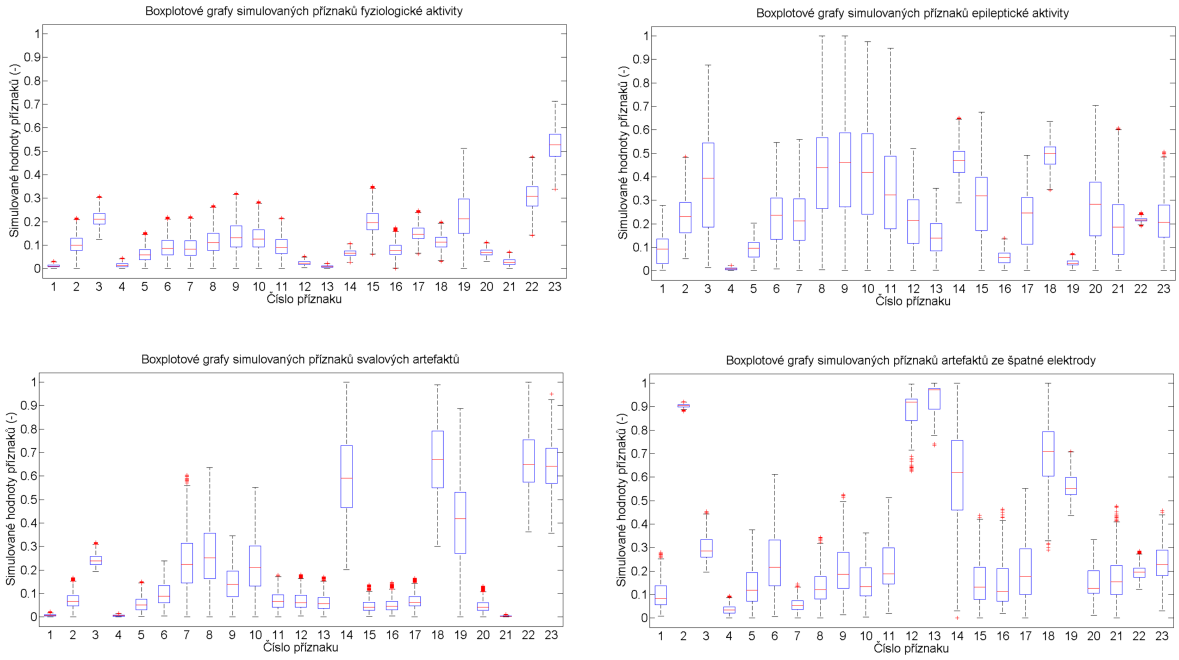
Obrázek 4.9: Pravděpodobnostní hustotní funkce reálné epileptické aktivity vybraných příznaků.

Tabulka 4.5: Příznaky reálné epileptické aktivity, u kterých křivka pravděpodobnostní hustotní funkce tvoří dvě špičky

Pořadí	Příznak
1	variabilita signálu v daném segmentu
3	maximální negativní amplituda v daném segmentu
5	FFT hodnota v 2. části delta frekvenčního pásma (2,0 - 3,5 Hz)
6	FFT hodnota v 1. části theta frekvenčního pásma (4,0 - 5,5 Hz)
8	FFT hodnota v 1. části alfa frekvenčního pásma (8 - 10 Hz)
9	FFT hodnota v 2. části alfa frekvenčního pásma (10,5 - 12,5 Hz)
10	FFT hodnota signálu v sigma frekvenčním pásmu (18 - 29 Hz)
11	FFT hodnota signálu v beta frekvenčním pásmu (13,5 - 17,5 Hz)
12	maximální hodnota první derivace v segmentu
13	maximální hodnota druhé derivace v segmentu
15	střední hodnota první derivace v segmentu
17	Hjorthův parametr mobility
18	Hjorthův parametr komplexity
20	délka křivky v segmentu
21	nelineární energie segmentu

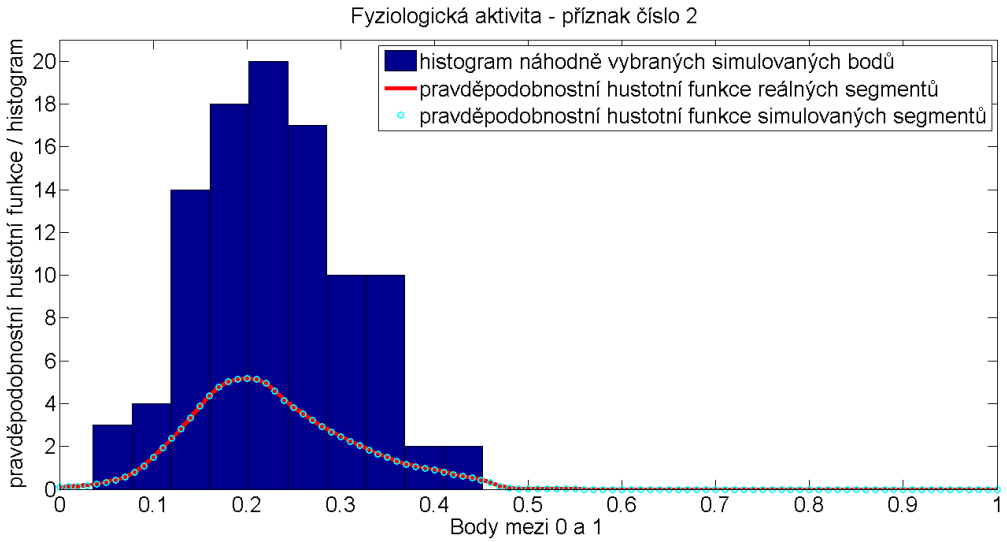
4.2 Simulovaný příznakový prostor

Na základě vědomostí o reálném příznakovém prostoru jsme vytvořili algoritmus v programovacím prostředí MATLAB pro simulaci příznakového prostoru. Tato metoda je popsána v kapitole 3.2 Metoda simulace příznakového prostoru. Následně jsem s jeho pomocí simulovala příznakový prostor pro všechny aktivity a artefakty z tabulky 3.1. K této simulaci jsem využila všechny příznaky z tabulky 2.1. Simulovala jsem u každého příznaku z každé třídy stejný počet segmentů, jako měl velikostně odpovídající reálný příznakový prostor. Boxplotové grafy simulovaného příznakového prostoru všech tříd jsou uvedeny na obrázku 4.10

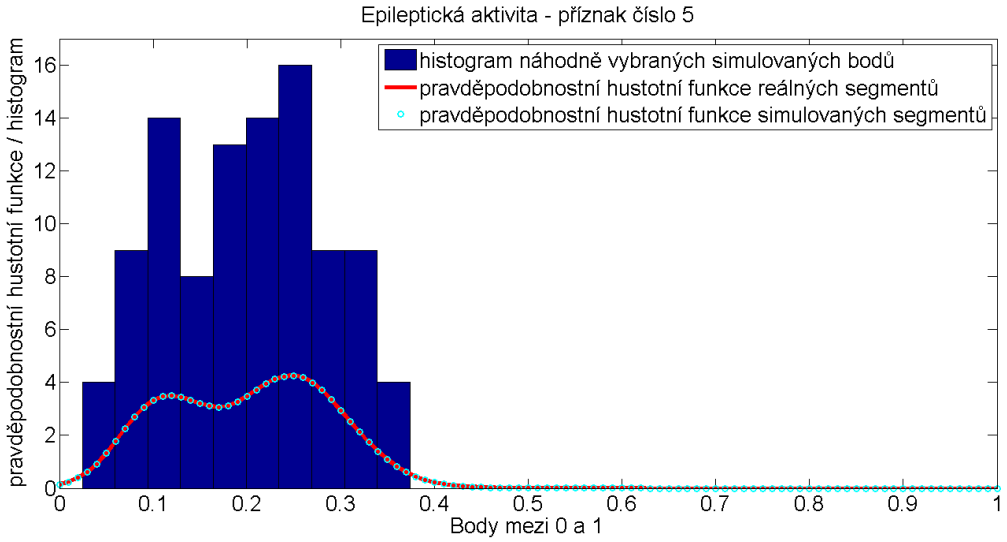


Obrázek 4.10: Boxplotové grafy příznaků simulovaných segmentů tříd fyziologická aktivita, epileptická aktivita, elektromyografické artefakty a artefakty ze špatné elektrody.

Jedním z mých úkolů bylo porovnat reálný příznakový prostor a příznakový prostor simulovaný pomocí naší metody. Z tohoto důvodu jsem vytvořila boxplotové grafy, viz obrázek 4.10, a také grafy pravděpodobnostních hustotních funkcí. Pro každou třídu jsem vytvořila graf s histogramem náhodně vybraných simulovaných segmentů a s pravděpodobnostními hustotními funkcemi pro reálné segmenty i simulované segmenty. Graf

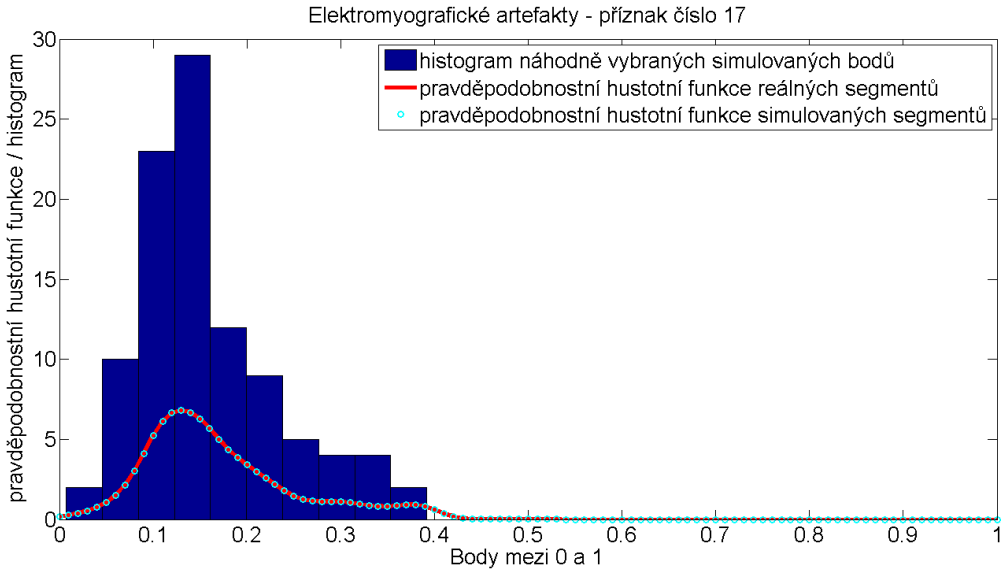


Obrázek 4.11: Histogram náhodně vybraných simulovaných segmentů a pravděpodobnostní hustotní funkce reálné i simulované fyziologické aktivity příznaku číslo 2.

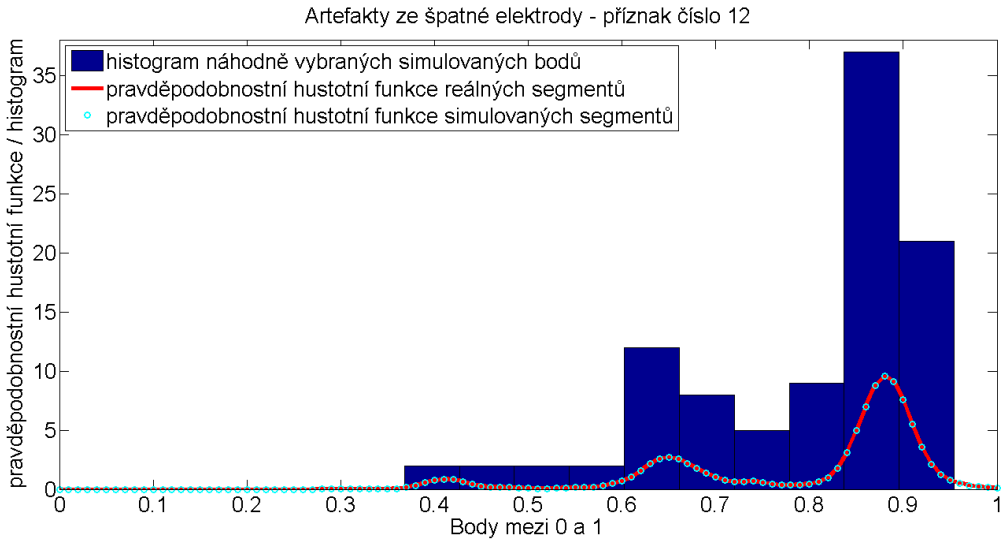


Obrázek 4.12: Histogram náhodně vybraných simulovaných segmentů a pravděpodobnostní hustotní funkce reálné i simulované epileptické aktivity příznaku číslo 5.

pro fyziologickou aktivitu příznaku číslo 2 je vidět na obrázku 4.11, pro epileptickou aktivitu příznaku číslo 5 na obrázku 4.12, pro elektromyografické artefakty příznaku číslo 17 na obrázku 4.13 a pro artefakty ze špatné elektrody příznaku číslo 12 na obrázku 4.14.



Obrázek 4.13: Histogram náhodně vybraných simulovaných segmentů a pravděpodobnostní hustotní funkce reálných i simulovaných elektromyografických artefaktů příznaku číslo 17.



Obrázek 4.14: Histogram náhodně vybraných simulovaných segmentů a pravděpodobnostní hustotní funkce reálných i simulovaných artefaktů ze špatné elektrody příznaku číslo 12.

4.3 Redukce dimenze reálného i simulovaného příznakového prostoru pomocí analýzy hlavních komponent

Nejprve jsem redukovala dimenze reálného zašumělého příznakového prostoru a simulovaného příznakového prostoru o stejném počtu a stejném poměru segmentů. Při redukci dimenze reálného i simulovaného příznakového prostoru pomocí mnou implementované metody PCA jsem se zaměřila na 2 aspekty. Tím prvním jsou procenta zachování informací při redukci na 2-dimenzionální (2D) prostor, viz tabulka 4.6. A druhým je velikost redukovaného příznakového prostoru při zachování minimálně 95% informací, viz tabulka 4.7. Stejné aspekty mě zajímaly při redukci dimenze reálného i simulovaného příznakového prostoru pomocí originální PCA, která je jednou z funkcí MATLAB, viz tabulka 4.6 a 4.7. Mnou implementovaná metoda PCA využívá k výpočtu hlavních komponent metodu kovariance. Originální PCA v MATLAB k výpočtu hlavních komponent využívá algoritmus singulárního rozkladu (SVD).

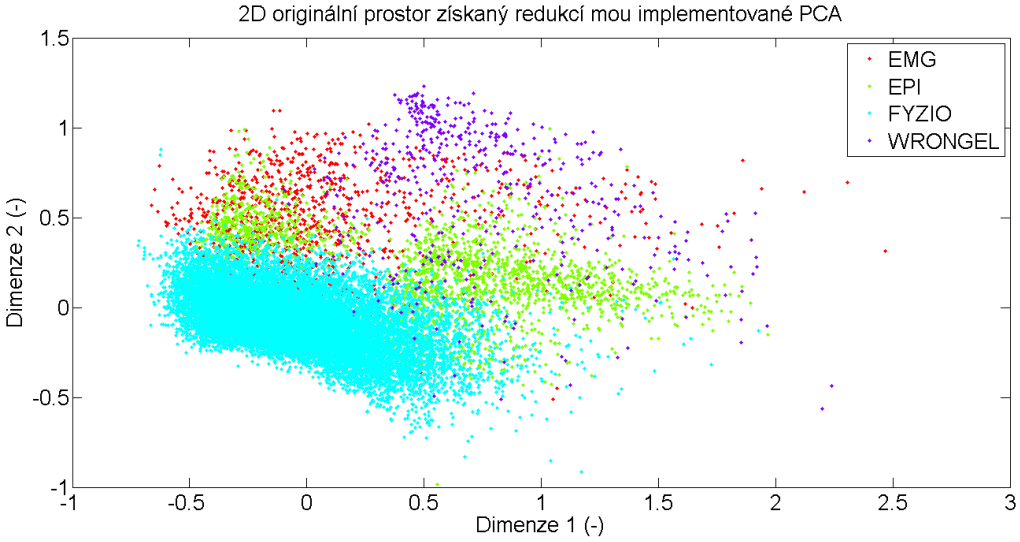
Tabulka 4.6: Procenta zachování informací při redukci příznakového prostoru na 2D prostor

Příznakový prostor	Mnou implementovaná PCA	Originální PCA v MATLABu
Reálný	56,96 %	56,96 %
Simulovaný	66,91 %	66,91 %

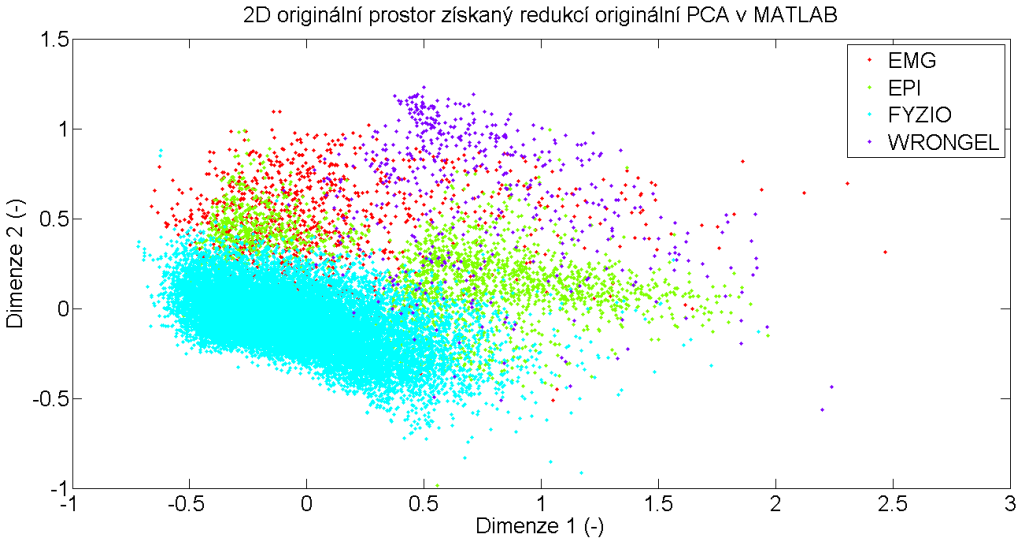
Tabulka 4.7: Velikost redukovaného příznakového prostoru při zachování minimálně 95% informací

Příznakový prostor	Mnou implementovaná PCA	Originální PCA v MATLABu
Reálný	12 D	12 D
Simulovaný	11 D	11 D

Jedním z důvodů redukce dimenze příznakového prostoru na 2D nebo 3D prostor je ten, že se můžeme podívat na rozložení jednotlivých segmentů různých klasifikačních tříd v příznakovém prostoru. Graf redukovaného reálného příznakového prostoru pomocí mnou implementované PCA na 2D prostor je uveden na obrázku 4.15. Graf redukovaného reálného příznakového prostoru pomocí originální PCA v MATLAB na 2D prostor je uveden na obrázku 4.16. Graf redukovaného simulovaného příznakového prostoru o stejném

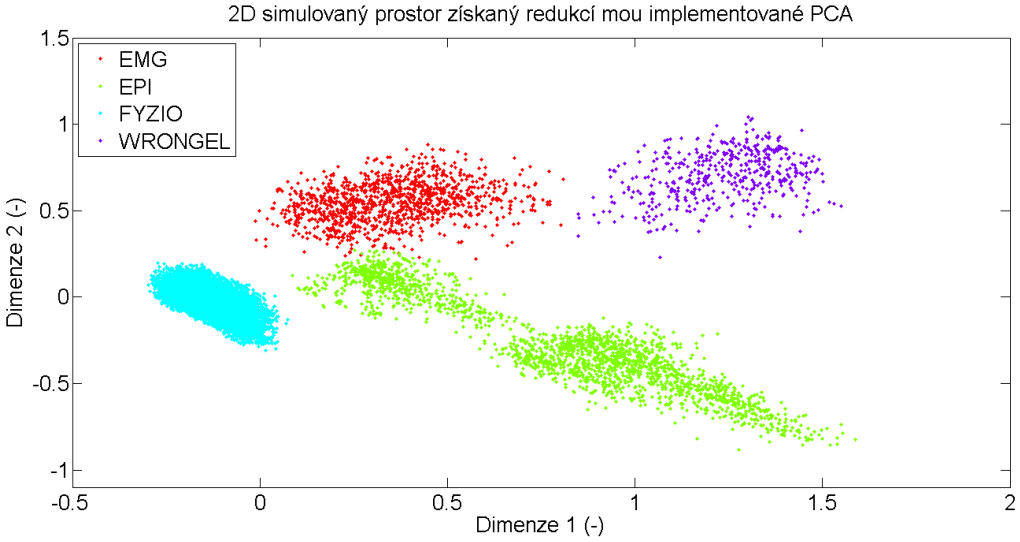


Obrázek 4.15: Rozložení klasifikačních tříd po redukci dimenze reálného příznakového prostoru na 2D prostor pomocí mnou implementované PCA

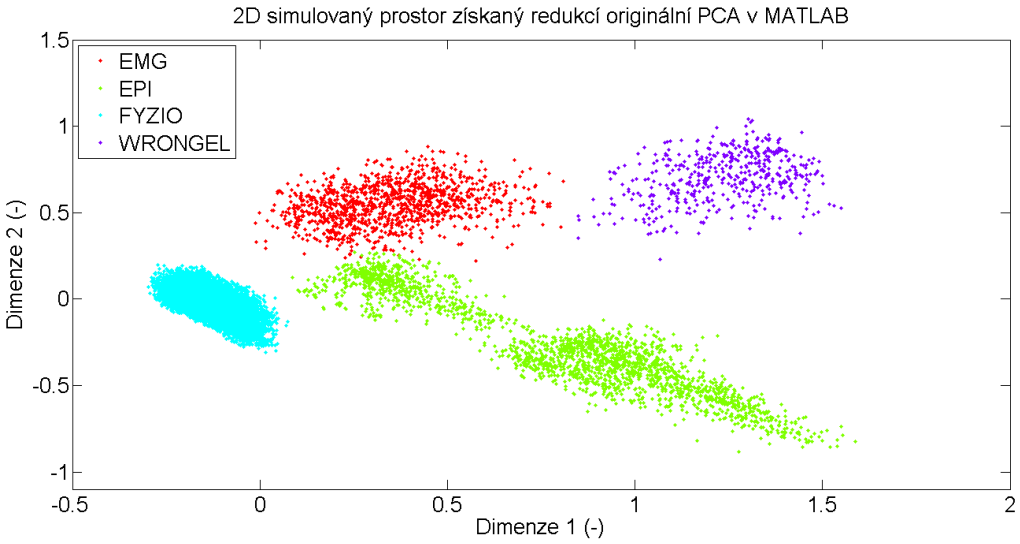


Obrázek 4.16: Rozložení klasifikačních tříd po redukci dimenze reálného příznakového prostoru na 2D prostor pomocí originální PCA v MATLAB

počtu segmentů jako reálný prostor pomocí mnou implementované PCA na 2D prostor je uveden na obrázku 4.17. Graf redukovaného simulovaného příznakového prostoru o stejném počtu segmentů jako reálný prostor pomocí originální PCA v MATLAB na 2D prostor je uveden na obrázku 4.18.



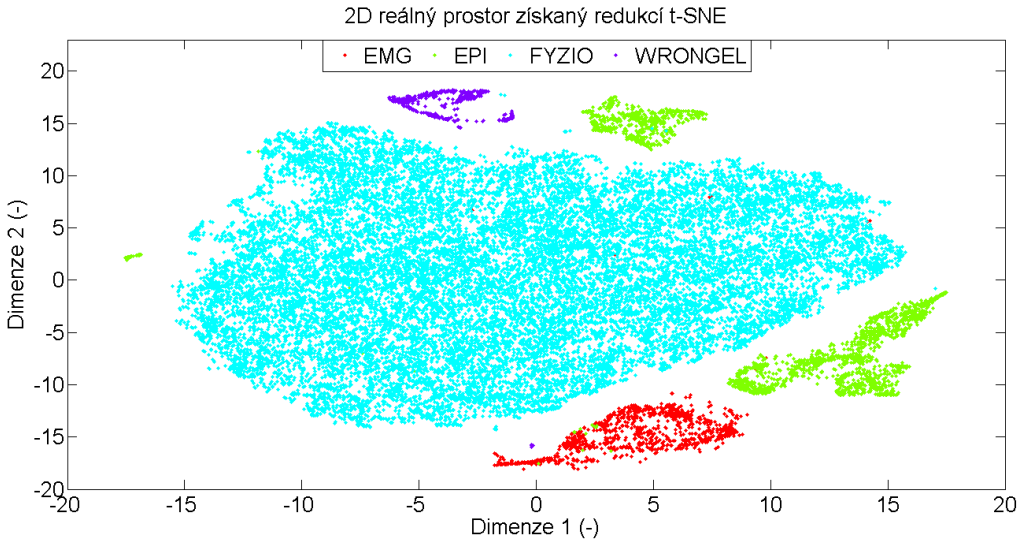
Obrázek 4.17: Rozložení klasifikačních tříd po redukci dimenze simulovaného příznakového prostoru na 2D prostor pomocí mnou implementované PCA



Obrázek 4.18: Rozložení klasifikačních tříd po redukci dimenze simulovaného příznakového prostoru na 2D prostor pomocí originální PCA v MATLAB

4.4 Redukce dimenze reálného i simulovaného příznakového prostoru pomocí t-Distributed Stochastic Neighbor Embedding

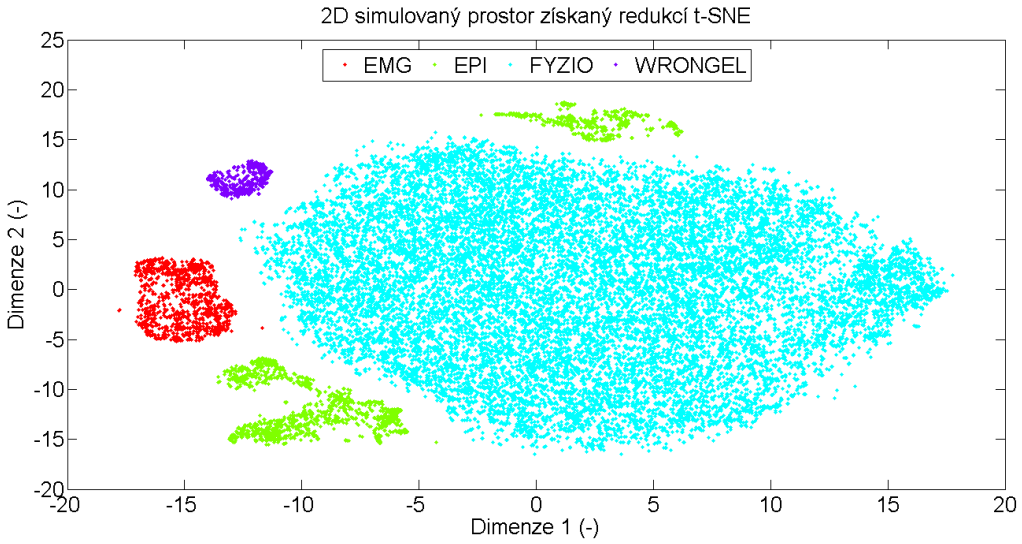
Při redukci dimenze pomocí nelineární metody t-SNE uživatel na vstupu funkce zadává hodnotu perplexity. Perplexita může být interpretována jako efektivní počet sousedů. Typické hodnoty perplexity jsou mezi 5 a 50. [24]. Kód t-SNE má v MATLAB nastavenou výchozí hodnotu perplexity na hodnotu 30, kterou jsem při redukci také použila. Na obrázku 4.19 je 2D redukovaný reálný příznakový prostor. Na obrázku 4.20 je 2D redukovaný simulovaný příznakový prostor. Oba grafy jsou zobrazením po 200 iteracích t-SNE.



Obrázek 4.19: Rozložení klasifikačních tříd po redukci dimenze reálného příznakového prostoru na 2D prostor pomocí t-SNE

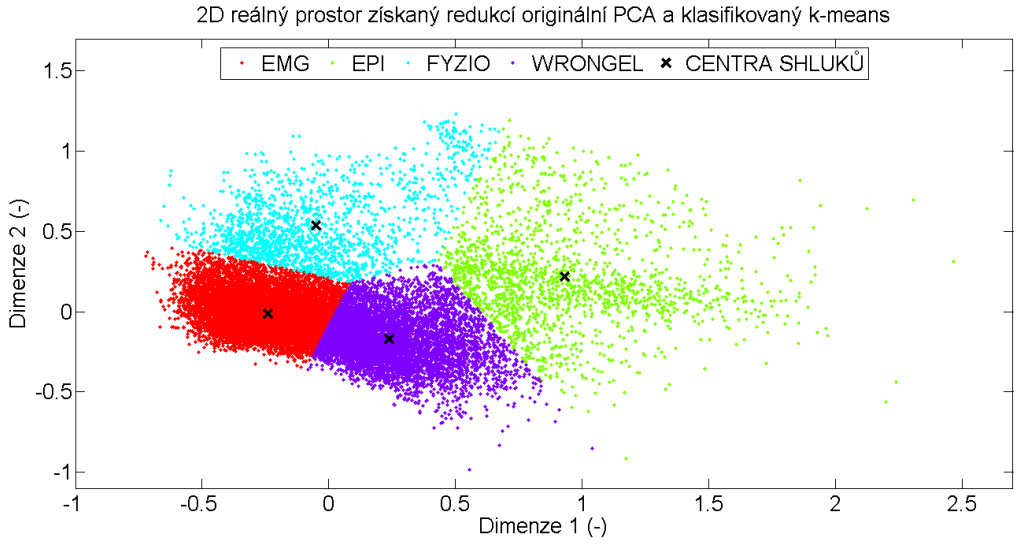
4.5 Klasifikace příznakového prostoru pomocí k-means

Redukovaný reálný příznakový prostor jsem klasifikovala pomocí shlukovacího algoritmu k-means. Výchozí počet shluků k jsem nastavila na hodnotu 4, protože v příznakovém prostoru máme 4 klasifikační třídy, viz tabulka 3.1. Klasifikovaný reálný příznakový prostor



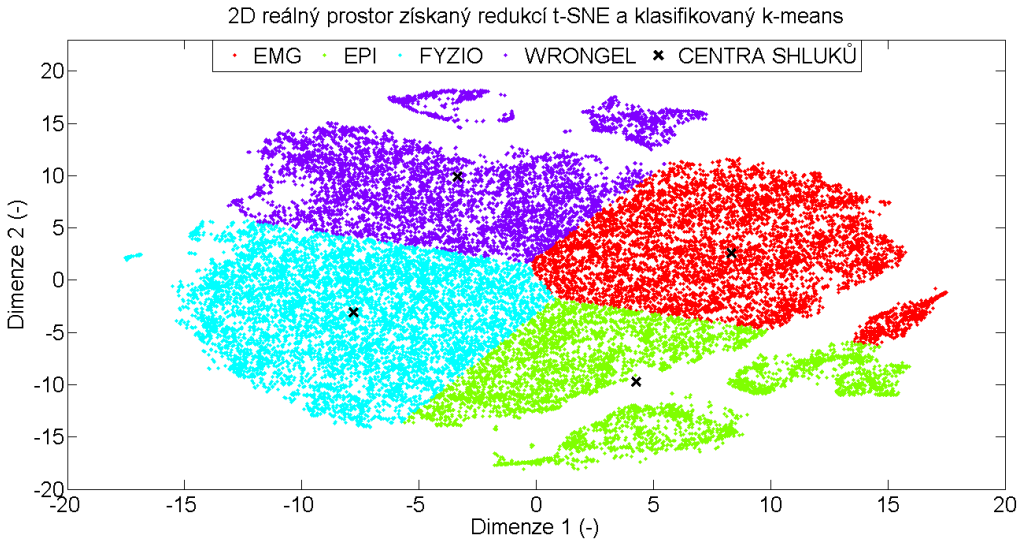
Obrázek 4.20: Rozložení klasifikačních tříd po redukci dimenze simulovaného příznakového prostoru na 2D prostor pomocí t-SNE

redukovaný pomocí originální PCA v MATLAB je uveden na obrázku 4.21. Klasifikovaný reálný příznakový prostor redukovaný pomocí t-SNE je uveden na obrázku 4.22.



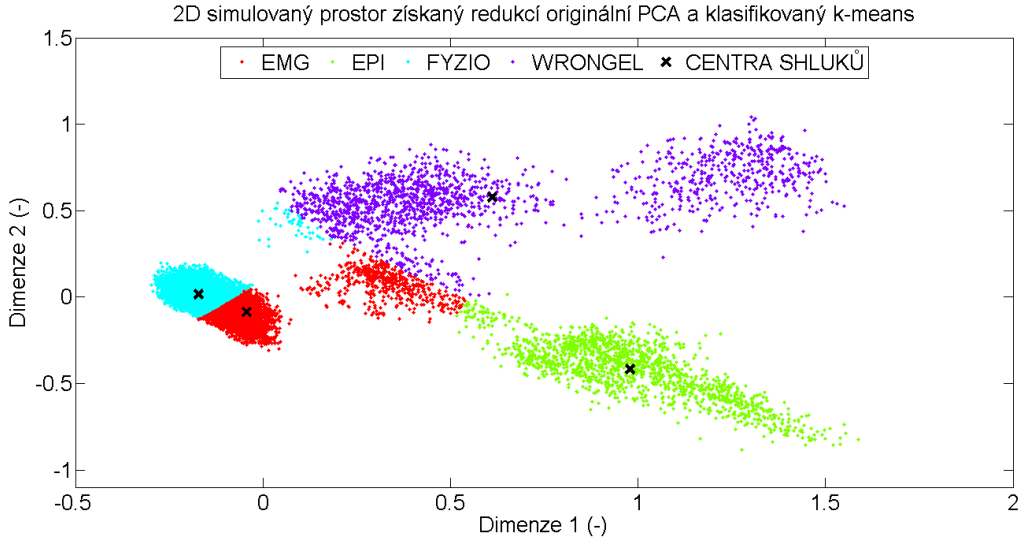
Obrázek 4.21: Klasifikovaný reálný příznakový prostor redukovaný pomocí originální PCA v MATLAB

Také jsem pomocí k-means klasifikovala simulovaný redukovaný příznakový prostor. Klasifikovaný simulovaný příznakový prostor redukovaný pomocí originální PCA v MAT-



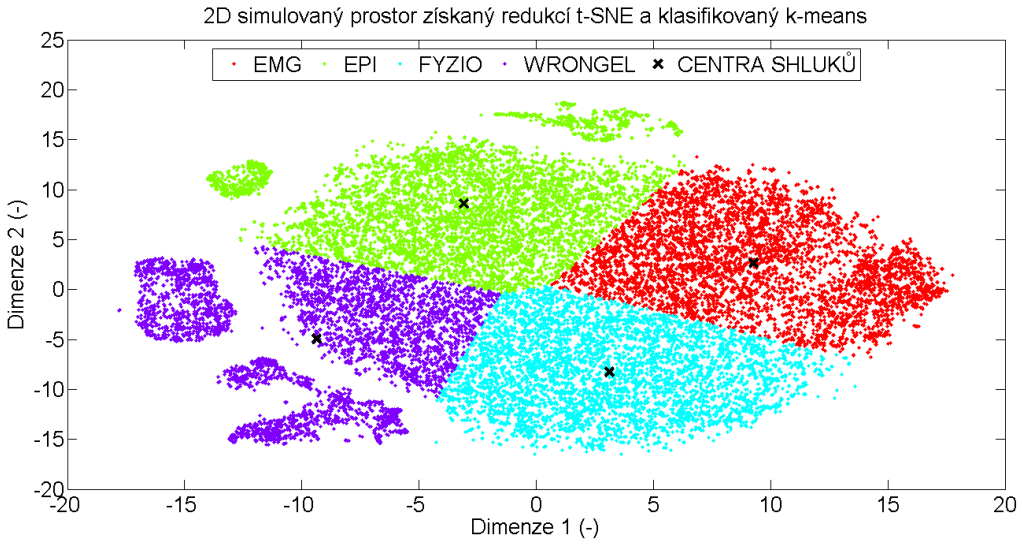
Obrázek 4.22: Klasifikovaný reálný příznakový prostor redukovaný pomocí t-SNE.

LAB je uveden na obrázku 4.23. Klasifikovaný simulovaný příznakový prostor redukovaný pomocí t-SNE je uveden na obrázku 4.24.



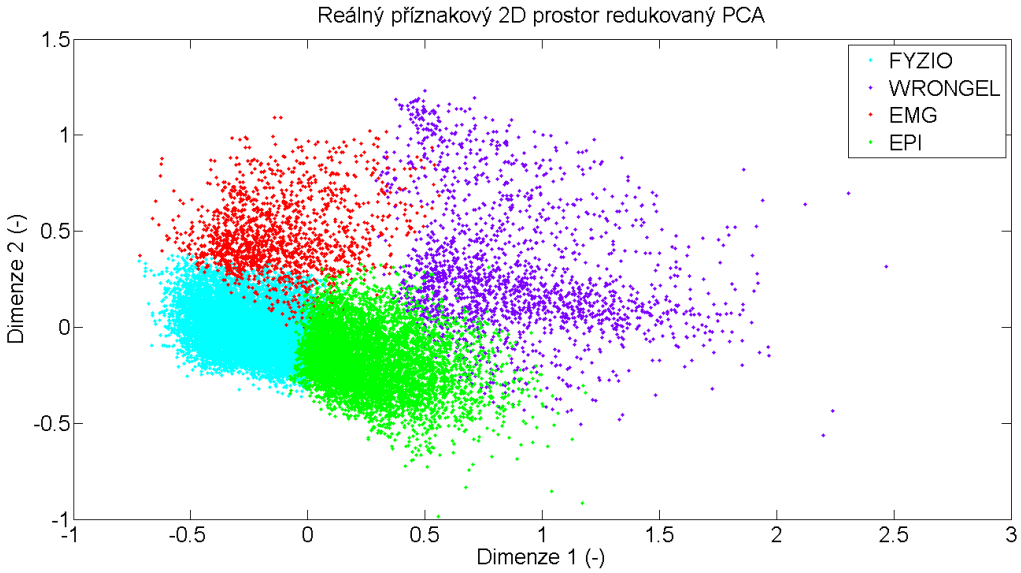
Obrázek 4.23: Klasifikovaný simulovaný příznakový prostor redukovaný pomocí originální PCA v MATLAB.

Pomocí k-means jsem klasifikovala reálný neredukovaný (23D) příznakový prostor. Indexy značící číslo třídy jsem poté použila na obarvení segmentů redukovaných 2D reálných



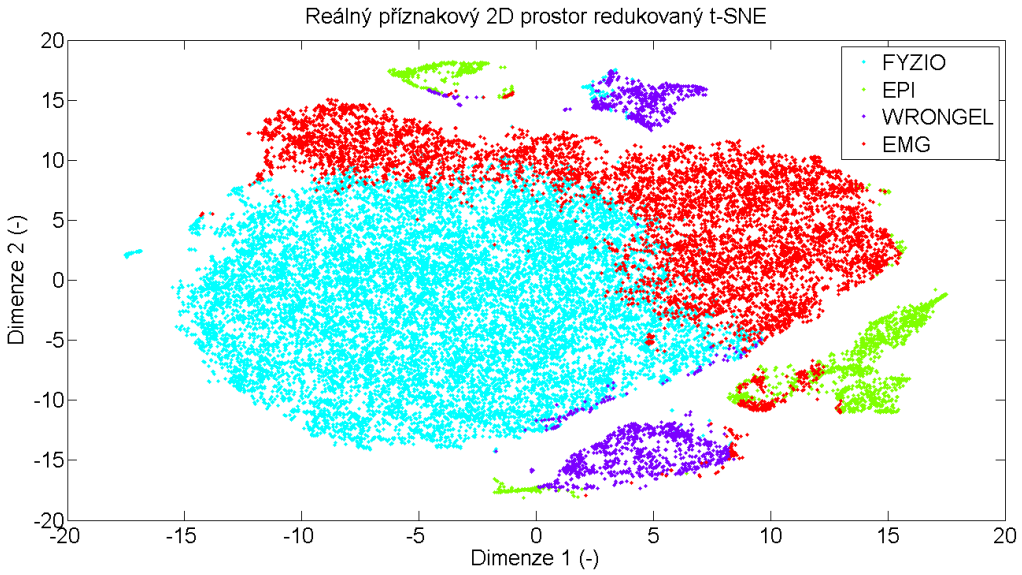
Obrázek 4.24: Klasifikovaný simulovaný příznakový prostor redukovaný pomocí t-SNE.

příznakových prostorů. Na obrázku 4.25 je vidět klasifikovaný reálný prostor redukovaný PCA a na obrázku 4.26 je zobrazený prostor po redukci t-SNE.



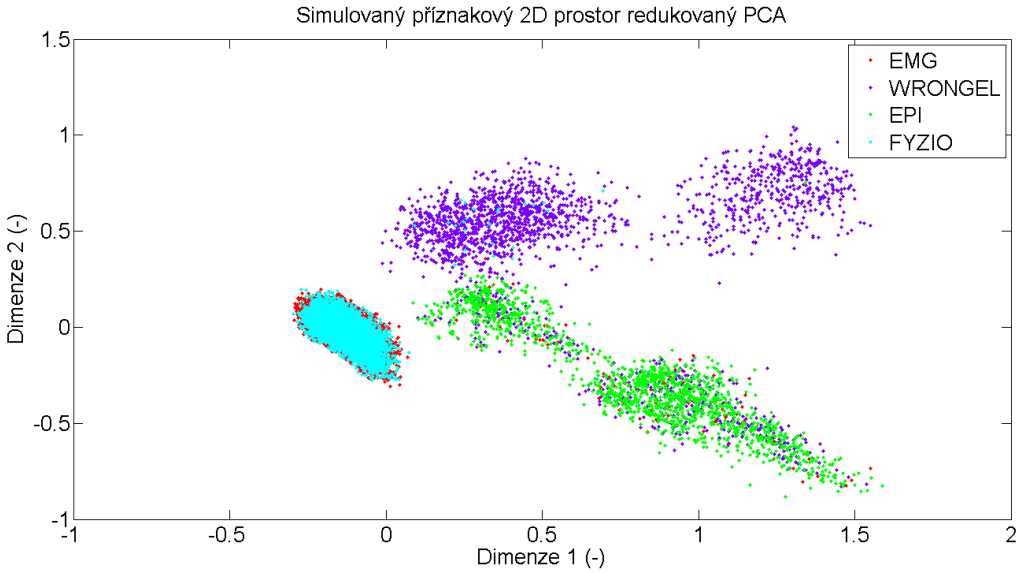
Obrázek 4.25: Klasifikovaný reálný 23D příznakový prostor zobrazený po redukci PCA na 2D prostor.

Pomocí k-means jsem klasifikovala i simulovaný 23D příznakový prostor. Indexy značící číslo třídy jsem opět použila na obarvení segmentů redukovaných 2D simulovaných přízna-

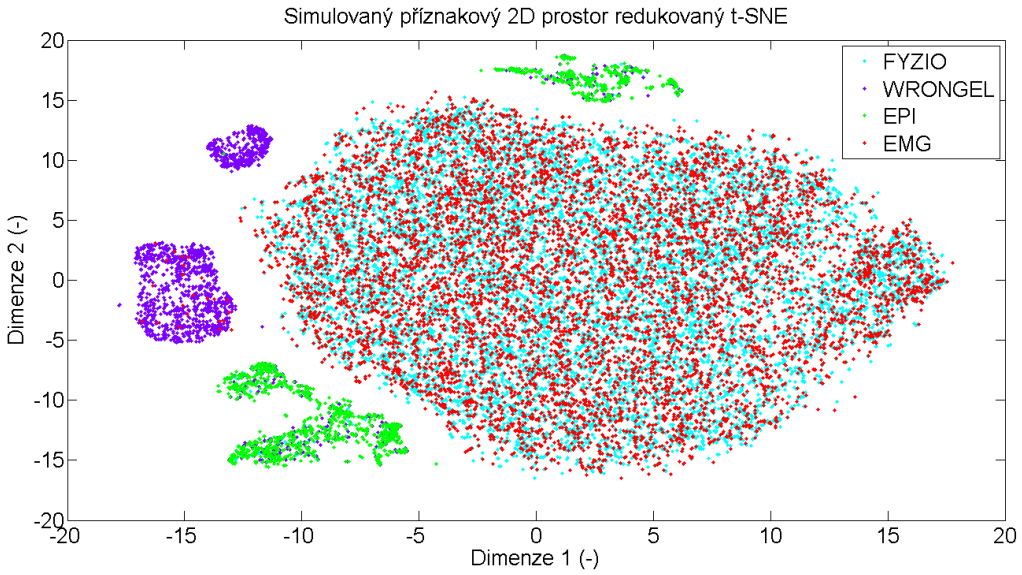


Obrázek 4.26: Klasifikovaný reálný 23D příznakový prostor zobrazený po redukci t-SNE na 2D prostor.

kových prostorů. Na obrázku 4.27 je klasifikovaný simulovaný prostor po redukci PCA a na obrázku 4.28 je prostor po redukci t-SNE.



Obrázek 4.27: Klasifikovaný simulovaný 23D příznakový prostor zobrazený po redukci PCA na 2D prostor.



Obrázek 4.28: Klasifikovaný simulovaný 23D příznakový prostor zobrazený po redukci t-SNE na 2D prostor.

4.6 Statistické zhodnocení klasifikace

Statisticky jsem zhodnotila klasifikaci redukovaného reálného i simulovaného příznakového prostoru. Porovnala jsem klasifikaci na prostoru redukovaném pomocí originální PCA v MATLAB a pomocí nelineární techniky t-SNE. Nejprve jsem vytvořila konfuzní matice získané klasifikací k-means třídy epileptické aktivity reálného příznakového prostoru redukovaného pomocí originální PCA v MATLAB, viz tabulka 4.8, a reálného příznakového prostoru redukovaného pomocí t-SNE, viz tabulka 4.9. Třídu epileptické aktivity jsem si vybrala z důvodu její nenahraditelnosti z hlediska diagnostiky epilepsie.

Tabulka 4.8: Konfuzní matice získaná klasifikací k-means třídy epileptické aktivity reálného příznakového prostoru redukovaného pomocí originální PCA v MATLAB

		Skutečný stav	
		Pozitivní	Negativní
k-means	Pozitivní	TP = 1129	FP = 651
	Negativní	FN = 1071	TN = 20846

Dále jsem vytvořila konfuzní matice získané klasifikací k-means třídy epileptické aktivity simulovaného příznakového prostoru redukovaného pomocí originální PCA v MATLAB,

Tabulka 4.9: Konfuzní matice získaná klasifikací k-means třídy epileptické aktivity reálného příznakového prostoru redukovaného pomocí t-SNE

		Skutečný stav	
		Pozitivní	Negativní
k-means	Pozitivní	TP = 993	FP = 4020
	Negativní	FN = 1207	TN = 17477

viz tabulka 4.10, a simulovaného příznakového prostoru redukovaného pomocí t-SNE, viz tabulka 4.11.

Tabulka 4.10: Konfuzní matice získaná klasifikací k-means třídy epileptická aktivita simulovaného příznakového prostoru redukovaného originální PCA v MATLAB

		Skutečný stav	
		Pozitivní	Negativní
k-means	Pozitivní	TP = 1598	FP = 0
	Negativní	FN = 602	TN = 21497

Tabulka 4.11: Konfuzní matice získaná klasifikací k-means třídy epileptické aktivity simulovaného příznakového prostoru redukovaného pomocí t-SNE

		Skutečný stav	
		Pozitivní	Negativní
k-means	Pozitivní	TP = 679	FP = 6065
	Negativní	FN = 1521	TN = 15432

Vypočítala jsem sensitivitu, specificitu a pozitivní prediktivní hodnotu (PPV) klasifikace k-means pro všechny čtyři klasifikační třídy reálného příznakového prostoru redukovaného pomocí originální PCA v MATLAB, viz tabulka 4.12, a reálného příznakového prostoru redukovaného pomocí t-SNE, viz tabulka 4.13.

Tabulka 4.12: Sensitivita, specificita a pozitivní prediktivní hodnota klasifikace k-means reálného příznakového prostoru redukovaného pomocí originální PCA v MATLAB

Název shluku	**Specificita**	**Sensitivita**	**PPV**
Elektromyografické artefakty	0,4152	0,1178	0,0110
Epileptická aktivita	0,9697	0,5132	0,6343
Fyziologická aktivita	0,6004	0,0108	0,1193
Artefakty ze špatné elektrody	0,7058	0,5250	0,0038

Tabulka 4.13: Sensitivita, specificita a pozitivní prediktivní hodnota klasifikace k-means reálného příznakového prostoru redukovaného pomocí t-SNE

Název shluku	**Specificita**	**Sensitivita**	**PPV**
Elektromyografické artefakty	0,7240	0,0056	0,0011
Epileptická aktivita	0,8130	0,4514	0,1981
Fyziologická aktivita	0,9792	0,3497	0,9883
Artefakty ze špatné elektrody	0,7847	0,9899	0,0893

Dále jsem vypočítala sensitivitu, specificitu a pozitivní prediktivní hodnotu klasifikace k-means pro všechny čtyři klasifikační třídy simulovaného příznakového prostoru redukovaného pomocí originální PCA v MATLAB, viz tabulka 4.14, a simulovaného příznakového prostoru redukovaného pomocí t-SNE, viz tabulka 4.15.

Tabulka 4.14: Sensitivita, specificita a pozitivní prediktivní hodnota klasifikace k-means simulovaného příznakového prostoru redukovaného pomocí originální PCA v MATLAB

Název shluku	**Specificita**	**Sensitivita**	**PPV**
Elektromyografické artefakty	0,7012	0,0065	0,0012
Epileptická aktivita	1,0000	0,7264	1,0000
Fyziologická aktivita	0,9878	0,6867	0,9965
Artefakty ze špatné elektrody	0,9454	1,0000	0,2811

Tabulka 4.15: Sensitivita, specificita a pozitivní prediktivní hodnota klasifikace k-means simulovaného příznakového prostoru redukovaného pomocí t-SNE

Název shluku	**Specificita**	**Sensitivita**	**PPV**
Elektromyografické artefakty	0,7396	0,0000	0,0000
Epileptická aktivita	0,7179	0,3086	0,1007
Fyziologická aktivita	0,9990	0,2859	0,9993
Artefakty ze špatné elektrody	0,7651	0,0000	0,0000

5 Diskuze

Ze získaných dat jsem vytvořila boxplotové grafy 23 příznaků všech tříd čistých segmentů i zašumělých segmentů. Mediány jednotlivých příznaků třídy fyziologická aktivita čistých segmentů jsou kromě čtyř příznaků všechny pod hodnotou 0,2. K analýze této třídy jsem měla k dispozici pouze 200 segmentů, proto je možné, že při vyšším počtu segmentů by vzrostl počet příznaků s hodnotou mediánu vyšší než 0,2 a také by se mohlo objevit více odlehlých hodnot.

Naopak mediány jednotlivých příznaků třídy epileptická aktivita čistých segmentů jsou kromě tří příznaků všechny nad hodnotou 0,2. Data této třídy obsahují větší množství odlehlých hodnot, což mohlo být způsobeno tím, že k analýze epileptické aktivity jsem měla k dispozici nejvíce segmentů ze všech 4 tříd čistých segmentů.

Pro analýzu příznakového prostoru fyziologické aktivity zašumělých segmentů jsem měla k dispozici nejvyšší množství segmentů. Z celkového množství příznaků jsem z pozdější simulace příznakového prostoru vyřadila 3,02 % odlehlých hodnot. Odlehlé hodnoty ze simulace vyřazuji z důvodu možné chybné klasifikace segmentu. Pouze 2 příznaky mají hodnotu mediánu nad 0,3: příznak číslo 22 a 23. Pouze u osmi příznaků je horní přilehlá hodnota vyšší než 0,5. Je to logický důsledek vlastností fyziologické aktivity mozku. Aby byl signál označen jako fyziologický, musí se jeho frekvenční i amplitudové vlastnosti pohybovat v mezích hodnot, které jsou odbornou literaturou a lékaři považovány za fyziologické. Proto jsou normalizované hodnoty příznaků nízké a rozpětí přilehlých hodnot příznaků malé.

Zašumělé segmenty epileptické aktivity mají největší možné rozpětí hodnot příznaků, tedy 1, u 4 příznaků: příznak číslo 3, 8, 9 a 10. U těchto příznaků jsem tudíž nezaznamenala žádné odlehlé hodnoty. Nejmenší rozpětí hodnot příznaků má příznak číslo 4. Tento příznak je také jediný, jehož nejvyšší hodnota příznaku se nachází pod 0,2. Z celkového počtu hodnot příznaků jsem pro simulaci příznakového prostoru odstranila pouze 1,15 % odlehlých hodnot.

S ohledem na množství analyzovaných segmentů čistých i zašumělých elektromyografických artefaktů jsem v této třídě zaznamenala největší množství odlehlých hodnot. Toto velké množství mohlo být způsobeno amplitudovými a frekvenčními vlastnostmi svalových artefaktů. Pro simulaci příznakového prostoru elektromyografických artefaktů jsem vyřadila celkem 5,52 % odlehlých hodnot. Celkem 4 příznaky mají hodnotu mediánu nad 0,5: příznak číslo 14, 18, 22 a 23. Jak je vidět na obrázku 4.2, tyto hodnoty jsou naprosto v souladu s tím, jak artefakt EMG v záznamu vypadá. Tento artefakt plně zkresluje původní signál, proto je nutné jej před analýzou EEG signálu detekovat a z analýzy vyřadit.

Artefakty ze špatné elektrody zašumělé segmenty mají největší možné rozpětí hodnot příznaků, tedy 1, u příznaku číslo 14. Z boxplotových grafů je patrné, že nejvyšších hodnot mediánů dosahují příznak číslo 2, 12 a 13. Vysoké hodnoty mediánů u těchto příznaků jsou logickým důsledkem vlastností tohoto artefaktu: velmi vysoké hodnoty amplitudy, velice strmý vzestup a pokles amplitudy s téměř bodovým hrotem. Z celkového počtu hodnot všech příznaků jsem odstranila 3,95 % odlehlých hodnot.

U příznaku číslo 19 Hjorthův parametr aktivity došlo k poklesu všech kvartilů u všech tříd zašumělých segmentů oproti čistých segmentům, nejvíce to bylo zřetelné ve třídě artefakty ze špatné elektrody. Z výsledků je patrné, že šum nejvíce ovlivňuje právě tento příznak. Dále šum výrazně ovlivňuje příznak číslo 9 a 10. Mediány těchto dvou příznaků tříd fyziologická aktivita a elektromyografické artefakty jsou u čistých segmentů nižší než u zašumělých segmentů. U třídy epileptická aktivita jsou naopak mediány u čistých segmentů těchto dvou příznaků vyšší než u zašumělých segmentů.

Porovnala jsem boxplotové grafy každého příznaku zašumělých segmentů napříč všemi třídami. Nejvíce se překrývají boxy příznaku číslo 5 u všech 4 tříd. Nenalezla jsem žádný příznak, u kterého by se nepřekrývaly žádné třídy. Ale u příznaku číslo 19 se oddělil box třídy epileptická aktivita, zatímco boxy ostatních tříd se překrývaly. Příznaky číslo 2 a 13 dokáží účinně separovat třídu artefakty ze špatné elektrody. Důsledkem překrývání boxů v různých třídách může být chybná klasifikace segmentu. Proto ke klasifikaci segmentů

používáme více příznaků. Příznaky musí popsat signál jak z amplitudového, tak z frekvenčního hlediska. Naopak zbytečně velký počet příznaků vytváří vysocedimenzionální příznakové prostory, které je technicky a časově náročnější zpracovat.

Při analýze reálného příznakového prostoru jsem se samostatně zaměřila na třídu epileptická aktivita. Pravděpodobností hustotní funkce 15 příznaků této třídy má zřetelné dvě špičky (peaky). Z 23 používaných příznaků v programu WF se tedy jedná o 65 % příznaků.

Vytvořila jsem boxploté grafy i simulovaného příznakového prostoru. Rozložení hodnot jednotlivých příznaků všech tříd simulovaného prostoru kopíruje rozložení reálného prostoru, ze kterého byly odstraněny odlehlé hodnoty. To je dáno tím, že jsme rozložení hodnot příznaků nesimulovali náhodně a rovnoměrně, ale použili jsme k tomu jádrový odhad pravděpodobnostní hustoty. Abych potvrdila správný postup naší simulace, porovnala jsem histogramy náhodně simulovaných bodů s pravděpodobnostními hustotními funkcemi (PDF) reálného i simulovaného příznakového prostoru. Z tohoto porovnání vyplývá, že PDF simulovaného prostoru kopíruje svým průběhem PDF reálného prostoru.

Při redukci dimenze reálného i simulovaného příznakového prostoru EEG jsem porovnávala analýzu hlavních komponent metodou kovariance, kterou jsem implementovala do programovacího prostředí MATLAB, s analýzou hlavních komponent metodou singulárního rozkladu, která je originální funkcí MATLAB. Použití obou těchto metod vedlo ke stejným výsledkům. Velkou výhodou metody PCA je možnost výpočtu procent zachovaných informací po redukci dimenzí. Toho jsem využila při porovnání redukcí oběma metodami. Při redukci dimenze reálného příznakového prostoru na 2D prostor zůstalo u obou metod zachováno stejné procento původních informací, u simulovaného prostoru taktéž. Pro lékaře je velmi důležité zachování co největšího procenta původních informací, v praxi je to většinou 95 %. S podmínkou zachování 95 % informací jsem redukovala reálný příznakový prostor u obou metod na 12 D a simulovaný na 11 D. U reálného příznakového prostoru to znamená redukce o 11 dimenzí, tedy o 52 %.

Analýza hlavních komponent metodou kovariance a metodou singulárního rozkladu redukuje reálné i simulované příznakové prostory stejně a zobrazené prostory se shodují. Tedy obě metody PCA jsou shodné. Na obrázcích redukovaných prostorů je třída epileptické aktivity rozdělena do dvou shluků, protože v PDF simulovaného prostoru epileptické aktivity jsou také vidět dvě špičky.

Zobrazila jsem 2D reálný příznakový prostor redukovaný pomocí t-SNE. Na obrázku 4.19 jsou jasně vidět oddělené shluky tříd. Třída fyziologické aktivity obsahovala 19 763 segmentů, jednalo se tedy o poměrně hustý shluk. Jak bylo uvedeno v metodách, přirozenou vlastností t-SNE je to, že rozšiřuje ("zoomuje") husté clustery. Epileptická aktivita je opět separována do dvou shluků.

Při zobrazení redukovaného simulovaného příznakového prostoru o velikosti 2D pomocí t-SNE je jasně vidět pět oddělených shluků odpovídající čtyřem použitým třídám, přičemž epileptická aktivita je opět separována do dvou clusterů.

Epileptická aktivita tvoří v reálném redukovaném 2D příznakovém prostoru dva shluky a pravděpodobnostní hustotní funkce 15 příznaků této třídy tvoří dva peaky. Z těchto důvodu se domnívám, že klasifikací epileptické aktivity do dvou tříd by se mohla zlepšit klasifikace a detekce epileptické aktivity.

Analýza hlavních komponent je rychlá a účinná metoda redukce dimenzí. Nicméně je to metoda lineární a její další nevýhodou je neschopnost účinně opticky separovat shluky jednotlivých tříd. T-SNE je velice účinná metoda redukce dimenzí v EEG prostoru. Dokáže opticky separovat všechny shluky, rozšiřuje velmi husté clustery, kde dále hledá vnitřní struktury. Jedná se o nelineární metodu. Nevýhodou t-SNE je časová náročnost algoritmu a vyšší nároky na výpočetní techniku. EEG záznam o velikosti desítek tisíc segmentů se může redukovat i několik hodin. Proto není vhodný k běžnému použití na ambulantní EEG záznamy. Při porovnání lineární a nelineární metody redukce dimenze jsem zjistila, že mnohem lepší na redukci dimenze v EEG prostoru je nelineární metoda t-SNE. Z toho vyplývá, že segmenty v EEG příznakovém prostoru zřejmě mají mezi sebou nelineární

vztahy. Domnívám se, že tento algoritmus by se dal využít jako klasifikátor vysokodimenzionálních dat.

Třída fyziologická aktivita reálného příznakového prostoru obsahovala nejvyšší počet segmentů a po redukci lineární metodou PCA byla klasifikována k-means minimálně do dvou shluků. Fyziologická aktivita simulovaného příznakového prostoru po redukci lineární metodou PCA vytvořila malý hustý cluster. K-means shluk klasifikoval jako dva.

Metoda redukce dimenze t-SNE rozšiřuje husté shluky, proto simulovaná i reálná fyziologická aktivita v 2D prostoru zaujímá největší procento plochy. Přestože t-SNE dokáže opticky separovat shluky tříd, k-means fyziologickou třídu klasifikuje do všech 4 tříd. Je to způsobeno tím, že metoda k-means rozděluje klasifikační prostor od středu do přibližně srovnatelných rozsahů.

Algoritmem k-means jsem klasifikovala neredukovaný (23D) reálný a simulovaný příznakový prostor. Indexy klasifikace segmentu do určité třídy jsem poté použila k obarvení segmentů v redukovaném 2D prostoru. Ve všech případech byla třída fyziologické aktivity klasifikována minimálně do dvou shluků.

Statistické zhodnocení klasifikace jsem provedla pomocí ROC analýzy. U reálného i simulovaného příznakového prostoru třídy epileptická aktivita redukovaného pomocí PCA dokázal algoritmus k-means označit více true positive i true negative segmentů než u prostoru redukovaném pomocí t-SNE.

U redukované simulované epileptické aktivity metodu PCA nebyl pomocí k-means označen ani jeden segment jako false positive. Následkem toho je specificita klasifikátoru a jeho pozitivní prediktivní hodnota rovna 1. Simulovaná třída artefaktů ze špatné elektrody redukovaná pomocí PCA měla sensitivitu rovnou 1. U třídy artefaktů ze špatné elektrody a elektromyografických artefaktů redukovaných pomocí t-SNE byla sensitivita a pozitivní prediktivní hodnota nulová.

Z výsledků klasifikace a jejího statistického zhodnocení vyplývá, že metoda k-means lépe klasifikuje prostor redukovaný PCA než t-SNE. Je to očekávaný výsledek, protože PCA i k-means jsou metody lineární. Nicméně obecně se k-means nehodí ke klasifikaci reálného i simulovaného 2D příznakového prostoru a nejeví se ani jako vhodný klasifikátor na neredukované prostory, protože je možné, že EEG segmenty příznakových prostorů mezi sebou mají nelineární vztahy. Řešením problému klasifikace by mohlo být využití jiné metody, např. hustotně založené metody DBSCAN či některé její modifikace.

6 Závěr

Cílem práce mé práce bylo redukovat dimenze příznakového prostoru EEG a extrahovat nové informace z příznakového prostoru EEG záznamu 10 pacientů s podezřením na epilepsii. Na jejím začátku jsem analyzovala metody redukcí dimenzí a současný stav této problematiky. Jako vhodné metody redukce dimenze v EEG prostoru jsem zvolila analýzu hlavních komponent (PCA), jako zástupce lineárních metod, a t-distributed Stochastic Neighbor Embedding (t-SNE), jako zástupce nelineárních metod. PCA metodou kovariance jsem implementovala do programovacího prostředí MATLAB.

Z porovnání boxplotových grafů všech tříd čistých a simulovaných segmentů je patrné, že šum nejvíce ovlivňuje příznak Hjortův parametr aktivity. Medián tohoto příznaku je všech tříd čistých segmentů zřetelně vyšší (minimálně o 23 %) než u zašumělých segmentů. Dále šum výrazně ovlivňuje příznak FFT hodnota v 2. části alfa frekvenčního pásma (10,5 - 12,5 Hz) a FFT hodnota v sigma frekvenčním pásmu (18 - 29 Hz).

Z výsledků mé práce dále vyplývá, že je vhodnější klasifikovat epileptickou aktivitu do dvou tříd, protože v reálném příznakovém prostoru tvoří dva velmi mírně se prolínající shluky a křivka pravděpodobnostní hustotní funkce 15 příznaků z 23 tvoří dvě špičky.

Na základě analýzy reálného příznakového prostoru jsme spolu s Ing. Markem Piorecký, Ing. Václavou Pioreckou, doc. Ing. Vladimírem Krajčou, CSc. a Ing. Vlastimilem Koudelkou, Ph.D. vyvinuli algoritmus na simulaci tohoto prostoru. Tento algoritmus jsme implementovali do programového prostředí MATLAB. Výsledky naší práce budou předneseny na konferenci World Congress on Medical Physics and Biomedical Engineering 2018 v Praze a bude publikován článek Simulation, modification and dimension reduction of EEG feature space.

PCA a t-SNE metoda redukce dimenze byla ověřena na simulovaných a reálných EEG příznacích. Při porovnání lineární a nelineární metody redukce dimenze jsem zjistila, že

mnohem lepší na redukci dimenze v EEG prostoru je nelineární metoda t-SNE. Z toho vyplývá, že segmenty v EEG příznakovém mají mezi sebou nelineární vztahy.

Z výsledků klasifikace redukovaných i neredukovaných reálných i simulovaných příznakových prostorů a jejich statistického zhodnocení vyplývá, že metoda k-means se nehodí ke klasifikaci 2D příznakového prostoru a nejeví se ani jako vhodný algoritmus na klasifikaci neredukovaných EEG prostorů. Řešením tohoto problému by mohlo být využití jiné metody klasifikace, např. hustotně založené metody DBSCAN či některé její modifikace.

Použitá literatura

[1] Lan, T.; Erdogmus, D.; Black, L.; aj.: A comparison of different dimensionality reduction and feature selection methods for single trial ERP detection. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, IEEE, 2010, ISBN 978-1-4244-4123-5, s. 6329–6332, doi:10.1109/IEMBS.2010.5627642. Dostupné z: `http://ieeexplore.ieee.org/document/5627642/`

[2] Subasi, A.; Gursoy, M. I.: EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, ročník 37, č. 12, 2010: s. 8659–8666, ISSN 09574174, doi:10.1016/j.eswa.2010.06.065. Dostupné z: `http://linkinghub.elsevier.com/retrieve/pii/S0957417410005695`

[3] Haufe, S.; Dähne, S.; Nikulin, V. V.: Dimensionality reduction for the analysis of brain oscillations. *NeuroImage*, ročník 101, 2014: s. 583–597, ISSN 10538119, doi:10.1016/j.neuroimage.2014.06.073. Dostupné z: `http://linkinghub.elsevier.com/retrieve/pii/S1053811914005503`

[4] Birjandtalab, J.; Pouyan, M. B.; Nourani, M.: Nonlinear dimension reduction for EEG-based epileptic seizure detection. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, 2016, ISBN 978-1-5090-2455-1, s. 595–598, doi:10.1109/BHI.2016.7455968. Dostupné z: `http://ieeexplore.ieee.org/document/7455968/`

[5] Smart, O.; Chen, M.: Semi-automated patient-specific scalp EEG seizure detection with unsupervised machine learning. In *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE, 2015, ISBN 978-1-4799-6926-5, s. 1–7, doi:10.1109/CIBCB.2015.7300286. Dostupné z: `http://ieeexplore.ieee.org/document/7300286/`

[6] Prabhakar, S. K.; Rajaguru, H.: PCA and K-means clustering for classification of epilepsy risk levels from EEG signals — A comparitive study between them. In *2015 International Conference on Intelligent Informatics and Biomedical Sciences*

(ICIIBMS), IEEE, 2015, ISBN 978-1-4799-8562-3, s. 83–86, doi:10.1109/ICIIBMS.2015.7439467. Dostupné z: `http://ieeexplore.ieee.org/document/7439467/`

[7] Subha, D. P.; Joseph, P. K.; U, R. A.; aj.: EEG Signal Analysis. *Journal of Medical Systems*, ročník 34, č. 2, 2010: s. 195–212, ISSN 0148-5598, doi:10.1007/s10916-008-9231-z. Dostupné z: `http://link.springer.com/10.1007/s10916-008-9231-z`

[8] Sanei, S.; Chambers, J.: *EEG signal processing*. Hoboken, NJ: John Wiley & Sons, 2007, ISBN 978-0-470-02581-9.

[9] Krajča, V.; Mohylová, J.: *Číslicové zpracování neurofyziologických signálů*. V Praze: České vysoké učení technické, 2011, ISBN 978-80-01-04721-7.

[10] Rieger, J.; Lhotská, L.; Krajča, V.: Zpracování dlouhodobých EEG záznamů. *Advances in electrical and electronic engineering*, ročník 4, č. 3, 2005: s. 151–156, ISSN 1336-1376. Dostupné z: `http://hdl.handle.net/10084/83686`

[11] Ahmadi, B.; Aimrfattahi, R.; Negahbani, E.; aj.: Comparison of Adaptive and Fixed Segmentation in Different Calculation Methods of Electroencephalogram Time-series Entropy for Estimating Depth of Anesthesia. In *2007 6th International Special Topic Conference on Information Technology Applications in Biomedicine*, IEEE, 2007, ISBN 978-1-4244-1867-1, s. 265–268, doi:10.1109/ITAB.2007.4407398. Dostupné z: `http://ieeexplore.ieee.org/document/4407398/`

[12] Krajca, V.; Petranek, S.; Varri, A.; aj.: On-line Multichannel Adaptive Segmentation As A Basement For Long-term Eeg Processing. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society Volume 13: 1991*, IEEE, 1991, ISBN 0-7803-0216-8, s. 445–446, doi:10.1109/IEMBS.1991.684018. Dostupné z: `http://ieeexplore.ieee.org/document/684018/`

[13] Schaabova, H.; Krajca, V.; Sedlmajerova, V.; aj.: Application of Artificial Neural Networks for Analyses of EEG Record with Semi-Automated Etalons Extraction.

In *Engineering Applications of Neural Networks*, Cham: Springer International Publishing, 2016, ISBN 978-3-319-44187-0, s. 94–107, doi:10.1007/978-3-319-44188-7_7. Dostupné z: `http://link.springer.com/10.1007/978-3-319-44188-7_7`

[14] Rizal, A.; Hidayat, R.; Nugroho, H. A.: Determining lung sound characterization using Hjorth descriptor. In *2015 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC)*, Aug 2015, s. 54–57, doi:10.1109/ICCEREC.2015.7337053.

[15] Birjandtalab, J.; Pouyan, M. B.; Cogan, D.; aj.: Automated seizure detection using limited-channel EEG and non-linear dimension reduction. *Computers in Biology and Medicine*, ročník 82, 2017: s. 49–58, ISSN 00104825, doi:10.1016/j.compbiomed.2017.01.011. Dostupné z: `http://linkinghub.elsevier.com/retrieve/pii/S0010482517300185`

[16] McSwiggan, G.; Baddeley, A.; Nair, G.: Kernel Density Estimation on a Linear Network. *Scandinavian Journal of Statistics*, ročník 44, č. 2: s. 324–345, doi:10.1111/sjos.12255, `https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjos.12255`. Dostupné z: `https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12255`

[17] Bowman, A. W.; Azzalini, A.: *Applied smoothing techniques for data analysis*. New York: Oxford University Press, 1997, ISBN 9780198523963.

[18] Hotelling, H.: Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, ročník 24, č. 6, 1933: str. 417.

[19] Jolliffe, I. T.: *Principal component analysis*. New York: Springer, druhé vydání, 2002, ISBN 0-387-95442-2.

[20] The MathWorks, Inc.: PCA. 1994-2018, online; cit. [2018-03-31]. Dostupné z: `https://www.mathworks.com/help/stats/pca.html`

[21] Smith, L. I.; aj.: A tutorial on principal components analysis. *Cornell University, USA*, ročník 51, č. 52, 2002: str. 65.

[22] Wikipedia: the free encyclopedia: Vlastní číslo. 2001-2018, online; cit. [2018-03-31]. Dostupné z: `https://goo.gl/e5Fj0u`

[23] Ballabio, D.: A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. *Chemometrics and Intelligent Laboratory Systems*, ročník 149, 2015: s. 1 – 9, ISSN 0169-7439, doi:https://doi.org/10.1016/j.chemolab.2015.10.003. Dostupné z: `http://www.sciencedirect.com/science/article/pii/S0169743915002476`

[24] van der MAATEN, L.; HINTON, G.: Visualizing data using t-SNE. *Journal of Machine Learning Research*, ročník 2008, č. 9.Nov: s. 2579–2605.

[25] Wattenberg, M.; Viégas, F.; Johnson, I.: How to Use t-SNE Effectively. *Distill*, 2016, doi:10.23915/distill.00002. Dostupné z: `http://distill.pub/2016/misread-tsne`

[26] The MathWorks, Inc.: T-SNE. 1994-2018, online; cit. [2017-12-30]. Dostupné z: `https://www.mathworks.com/help/stats/t-sne.html`

[27] Wikipedia: the free encyclopedia: K-means clustering. 2001-2018, online; cit. [2018-03-31]. Dostupné z: `https://en.wikipedia.org/wiki/K-means_clustering`

[28] Fawcett, T.: An introduction to ROC analysis. *Pattern Recognition Letters*, ročník 27, č. 8, 2006: s. 861–874, ISSN 01678655, doi:10.1016/j.patrec.2005.10.010. Dostupné z: `http://linkinghub.elsevier.com/retrieve/pii/S016786550500303X`

[29] Bruzzo, A. A.; Gesierich, B.; Santi, M.; aj.: Permutation entropy to detect vigilance changes and preictal states from scalp EEG in epileptic patients. A preliminary study. *Neurological Sciences*, ročník 29, č. 1, 2008: s. 3–9, ISSN 1590-1874, doi:10.1007/s10072-008-0851-3. Dostupné z: `http://link.springer.com/10.1007/s10072-008-0851-3`

[30] O'Reilly, C.; Nielsen, T.: 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)*, IEEE, 2013, ISBN 978-1-4673-5540-7,

s. 413–420, doi:10.1109/WoSSPA.2013.6602401. Dostupné z: `http://ieeexplore.ieee.org/document/6602401/`

[31] Sovierzoski, M. A.; de Azevedo, F. M.; Argoud, F. I. M.: Performance Evaluation of an ANN FF Classifier of Raw EEG Data using ROC Analysis. In *2008 International Conference on BioMedical Engineering and Informatics*, IEEE, 2008, ISBN 978-0-7695-3118-2, s. 332–336, doi:10.1109/BMEI.2008.220. Dostupné z: `http://ieeexplore.ieee.org/document/4548687/`

[32] Piorecký, M.: *Automatická klasifikace EEG segmentů metodou DBSCAN*. Diplomová práce, České vysoké učení technické v Praze, Fakulta biomedicínského inženýrství, Kladno, 2016.

A Obsah přiloženého CD

- Klíčová slova v ČJ
- Klíčová slova v AJ
- Abstrakt práce v ČJ
- Abstrakt práce v AJ
- Naskenované zadání bakalářské práce
- Kompletní bakalářská práce
- Publikovaný článek
- MATLAB soubory
  - Anonymizovaný reálný EEG příznakový prostor
  - Skript pro analýzu EEG příznakového prostoru
  - Skript pro simulaci EEG příznakového prostoru
  - Simulovaný EEG příznakový prostor
  - Skripty pro redukci dimenzí
  - Redukované EEG příznakové prostory
  - Skript pro klasifikaci k-means a statistické zhodnocení klasifikace ROC analýzou

B Publikovaný článek

Článek Simulation, modification and dimension reduction of EEG feature space publikovaný ve sborníku konference World Congress on Medical Physics and Biomedical Engineering 2018.

Simulation, modification and dimension reduction of EEG feature space

Marek Piorecký[1], Eva Černá[1], Václava Piorecká[1], Vladimír Krajča[1], and Vlastimil Koudelka[2]

[1] Faculty of Biomedical Engineering, CTU in Prague, Nám. Sítná 3105, Kladno 272 01, Czech Republic,
marek.piorecky@fbmi.cvut.cz
[2] National institute of Mental Health, Topolová 748, Klecany 250 65, Czech Republic

Abstract. An automate classification of EEG time segments is frequently used technique across many neuro-scientific fields. Generally, segment classification results in labeled EEG time segments (e.g. physiological brain activity, epileptic activity, muscle artifacts or electrode artifacts). However, currently used methods are usually tested on artificial surrogate data and more general validation approach is needed. Here, a generalized statistical model of commonly used discriminating features obtained from real EEG data is presented for the first time. Multivariate probability density functions (PDFs) of classes are fitted on more than twenty thousand of testing segments from human EEG. An unique testing set is designed using a recent non-linear dimension reduction technique. Parametric and non-parametric PDF estimators are applied and compared in sense of feature space model.

Keywords: EEG, dimension reduction, feature space

1 Introduction

Electroencephalography (EEG) is most commonly used method of examining of the electric brain activity. Records are examined in the time and frequency domain [4]. Signal preprocessing and automatic segmentation and classification help expert to recognize the pathological segments in EEG records [7].

Adaptive segmentation creates segments containing one kind of characteristic waveform and so it helps to make the classification easier [10, 6].

Features are counted for each segment of the EEG record. EEG segments are classified into the same classes based on similarity of their features.

There are many features, some of them are typically used to determine concretely waveform [4, 8].

Automatic methods were used for the classification of segments as well as learning methods, which required user's intervention [4]. Generally, learning methods require usage of etalons, which describe the typical content of the class.

Dimension reduction helps to speed up the data processing and brings a new perspective on them.

Principal component analysis (PCA) is one of most commonly used linear method, which can be applied on EEG. t-Distributed Stochastic Neighbor Embedding (tSNE) is a nonlinear method, which constructs a probability distribution over pairs of high-dimensional objects in such way that high probability is assigned to similar objects, while low probability is connected with dissimilar points [9].

2 Methods

2.1 Data

The data comes from the measurement of patients from Bulovka Hospital in Prague. The data was recorded by using the Brain-Quick (Micromed). It was obtained on the basis of the project proposal, which was approved by the ethics committee of Bulovka Hospital on June 28 2011. These are clinical examinations lasted from 15 to 40 minutes. Test data were measured on patients who had been diagnosed with suspected epilepsy disease. Patients were 6 men and 4 women aged between 26 and 60 years.

2.2 Segmentation

Our software in C++ and Matlab R2015a were used to process the data. The sampling frequency in the record was 128 Hz. The filter was set on a bandpass (0.4 Hz and 70 Hz). Computational operations were applied in the average montage. We used an adaptive estimation of mean value described by Grießbach [3]. Multichannel adaptive segmentation [10] was used after the filtration. The Varri's adaptive segmentation method is based on two joint windows sliding along the signal and detecting local maxima in the total difference if the measurement calculated from the amplitude and the frequency difference (eqs. in [10, 6]). The adaptive segmentation is simultaneously processed for all the channels.

Adaptive segmentation parameters were: the window length - 128 samples, the window length for local maxima identification - 30 samples, the moving step of the two connected windows - 1 sample, minimum segment length - 70 samples. The fluctuation eliminating threshold varied from patient to patient.

2.3 Features

23 features based on time and frequency domain were used in practice. The features were: (1) signal variability, (2) maximal positive and (3) negative value (amplitude) in segment, delta (4, 5) band as square root of power spectral density (PSD) in band 0.5 Hz − 2.0 Hz as delta 1 and 2.0 Hz − 3.5 Hz as delta 2. The same partition was in (6, 7) theta (4.0 Hz − 6.0 Hz, 6.0 Hz − 7.5 Hz) and (8, 9) alpha (8.0 Hz − 10.5 Hz, 10.5 Hz − 12.5 Hz) band. Sigma (10) was in 18.0 Hz − 29.0 Hz and beta (11) in 13.5 Hz − 29.0 Hz. To estimate the spectrum was used Welch method [11] of averaged modified periodograms instead of single

periodogram (by FFT) to get a better estimation of the spectra. Next features were maximum of the first (12) and second derivation (13), medium frequency (14), medium of the first (15) and second derivation (16), Hjorths parameter [8]: mobility (17), complexity (18) and activity (19), length of the curve (20), nonlinear energy (21), number of the passes by zero (22) and the maximum peak frequency in the spectrum (23). [4, 5].

2.4 Training set development

We obtained 23D space after feature computation. The goal is to separate four classes distributed within the feature space: physiological activity (PHYSIO), epileptic activity (EPI), EMG artifacts (EMG) and wrong electrode contact artifacts (WRONGEL), see Fig. 1.

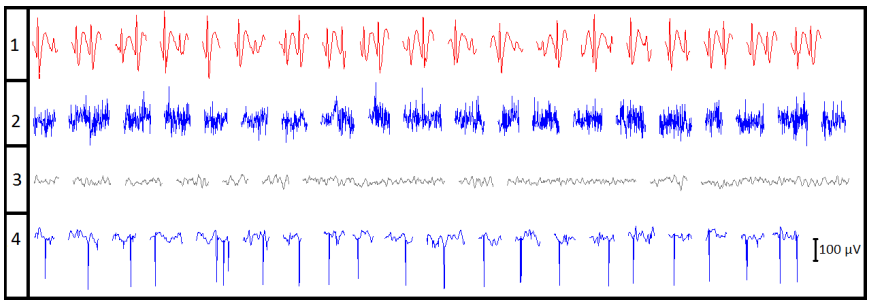


Fig. 1. A sample of segments of each created class, from the top: epilepsy activity, EMG artifacts, physiological activity and wrong electrode contact.

At first, a homogenous cluster was selected by using DBSCAN and k-means algorithm. Further processing can be summarized in following 4 points [12, 7]:

1. When DBSCAN ranked point (segment described by 23 features) as a center in the specific cluster and the k-means included this segment near the center of the same cluster, the segment was selected as a representative of that cluster (class).
2. Classes across patients and across measurements were merged and created one big class of specific segments of EEG signal.
3. The expert discarded segments whose character did not match the specific class.
4. Final classified dataset consisted of 335 segments of EMG, 1239 EMG and noise, 2200 epileptic activity, 17390 physiological activity, 19763 physiological activity with noise, 495 wrong electrode contact segments. The dataset with noise was created for realistic simulation feature space of EEG.

Pleas note that each feature was normalized by its minimum and maximum values within each class. We had to use normalization was necessary to use because of the different ranges of the original features.

2.5 Training set validation

At first, a linear PCA method based on singular-value decomposition was applied to visually inspect testing dataset in two dimensional space, see Fig. 2.

Since PCA method was sufficient for our purpose, we further investigated tSNE dimension reduction technique in order to depict the data structure. tSNE is a non-linear projection technique developed by van Maaten and Hinton [9]. 2D space was created by using parameters of tSNE: perplexity = 30 and initial dimension = 23.

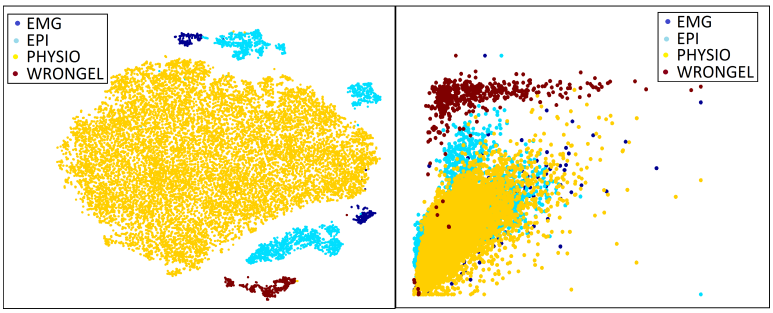


Fig. 2. EEG feature space reduced by the nonlinear and linear method. Left: Feature space after tSNE dimension reduction, 700 iteration steps. Right: Reduced space after PCA. Axes are modified features by dimension reduction.

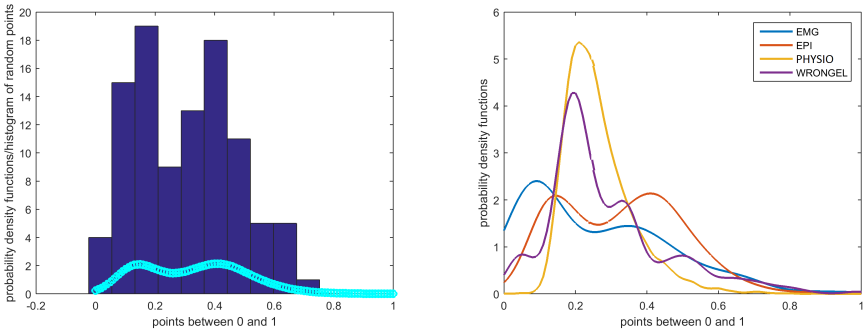


Fig. 3. Left: fitted kernel distribution on segments of epilepsy activity of feature 20 (white circles) and histogram shows the distribution of randomly generated dataset from the fitted function (dark bars); Right: Fitted kernel distribution on EMG artifacts, epilepsy and physiological activity and wrong electrode contacts segments.

2.6 Feature space modeling

In order to show all feature distributions, we calculated and depicted BOX plots. This is a standard approach to data analysis in EEG studies [1, 2]. Based on

tSNE map, data points lying out of the interquartile span were rejected from the training set. The cleaned dataset was than used in fitting the PDF function. Both normal distribution and kernel density estimator were fitted to the training set. Finally, obtained probability distributions (kernel and normal) were used to simulate general feature space.

3 Resuts and Discussion

The investigated EEG segments were randomly selected from various datasets, subjects, and events (e.g. EPI, PHYSIO, EMG, etc.). This way an appropriate generalization of the model could be obtained.

Physiological activity has the most outliers, see Fig. 4. More segments with a wider range of the features (e.g. freq. wave spectrum) are assigned to the physiological activity. Medium frequency and second derivation features separate well physiological activity and EMG artifacts. Hjorth parameter activity effectively isolates Epilepsy from all others classes, see Fig. 4.

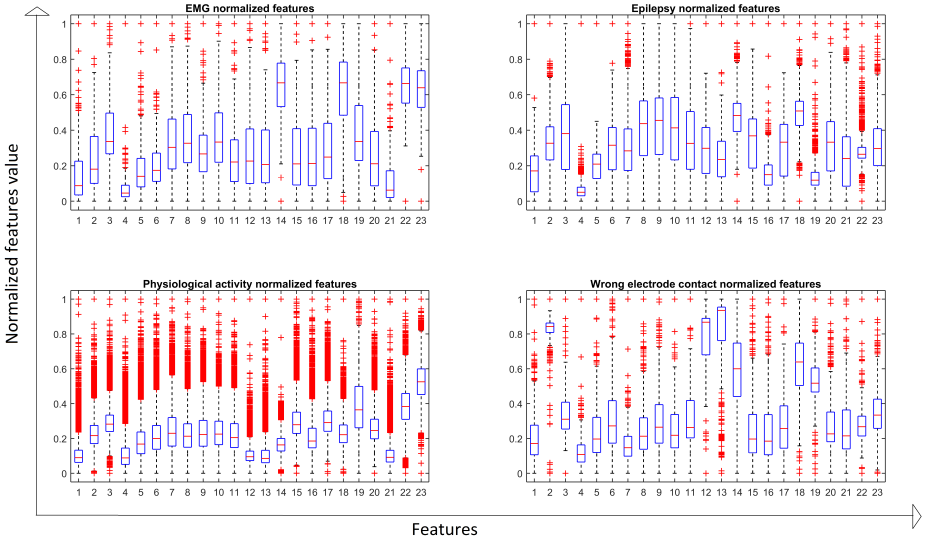


Fig. 4. Boxplot shows the distribution of features. The bottom and top of the box are the first and third quartiles, and the band inside the box is the second quartile - the median. The ends of the whiskers are the minimum and maximum of all of the data. Red markers represent outliers.

Nonlinear methods create 2D feature space separated from clusters created separable classes. From this point of view, DBSCAN seems to be a better option (opposite k-means) as it is able to distinguish interlacing clusters. Physiological segments match normal distribution, see Fig. 3. Kernel distribution has 2

local maximum in EMG class, Epilepsy class and wrong electrode class, see Fig. 3. There are two possibilities to use the simulation. Interquartile span enables its users to create clear etalons for learning classifiers. The use of the entire span offers the creation of an infinite number of segments that can be used as a simulated realistic feature space to test algorithms. The user can simulate any number of segments of the selected activity. In the case of simulation of the "whole record" (all classes at once), the recommended ratio of classes is 1:14:6:242 - EMG:EPI:WRONGEL:PHYSIO.

We assume that EEG records contain also noise from the recording device. In order to compare data across the device, it was appropriate to deduce general technical noise. General noise was subtracted from each feature. In our case segment white noise (power 75%) was of the same length as EEG segment.

4 Conclusion

A methodology for simulation feature space has been developed. The database of physiological, epileptic and artifact segments was created together with this work. Segments database offers the possibility to examine other features. We assume that nonlinear methods are better for examining EEG space dimension reduction as they maintain segments class separability. The simulation of feature space is needed to test the effectiveness of the selected features and to verify the classification methods. The Kernel distribution could be used for simulation EMG and Epileptic activity feature space because we do not want to suppress the second local maximum that occurs in this activity.

Acknowledgement:

This work was supported by the Grant Agency of Czech Republic with the topic: Temporal context in analysis of long-term non-stationary multidimensional signal, register number 17-20480S, by the Grant Agency of the CTU in Prague, registration number SGS18/159/OHK4/2T/17 with the topic: Feature space analysis using linear and nonlinear reduction of EEG space dimensions.

Conflict of interest declaration: The authors declare that there is no conflict of interest regarding the publication of this article.

Protection of human subjects and animals in research: The procedures followed were in compliance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects.

Statement of informed consent: The study protocol and patient informed consent have been approved by the Bulovka Hospital.

References

1. De Haan, W., Al Pijnenburg, Y., Lm Strijers, R., Van Der Made, Y., Van Der Flier, W. M., Scheltens, P., Stam, C. J.: Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory. In: BMC Neuroscience. 2009, vol. 10(1), 101. doi:10.1186/1471-2202-10-101., ISSN 1471-2202.
2. Easwaramoorthy, D., Uthayakumar. R.: Analysis of EEG signals using Advanced Generalized Fractal Dimensions. In: 2010 Second International conference on Computing, Communication and Networking Technologies. IEEE, 2010, pp. 1-6. doi: 10.1109/ICCCNT.2010.5591775. ISBN 978-1-4244-6591-0.
3. Grießbach, G., Witte, H.: Complex adaptive procedures for EEG monitoring. In: Witte, H., Zwiener, U., Schack, B., Doering, A. (eds.) Quantitative and Topological EEG and MEG Analysis, In: Druckhaus Mayer Verlag Gmbh Jena - Erlangen, 1997, pp. 295-297.
4. Krajca, V., Petranek, S., Patakova, I., Värri, A.: Automatic identification of significant graphoelements in multichannel EEG recordings by adaptive segmentation and fuzzy clustering. In: International Journal of Bio-Medical Computing, 1991, vol. 28(1-2), pp. 7189. doi:10.1016/0020-7101(91)90028-D.
5. Krajca, V., Petranek, S., Pietilä, T., Frey, H.: WaveFinder: A new system for an automatic processing of long-term EEG recordings. Quantitative EEG analysisclinical utility and new methods, 1993, pp. 103-106.
6. Paul, K., Krajca, V., Roth, Z., Melichar, J., Petranek, S.: Comparison of quantitative EEG characteristics of quiet and active sleep in newborns. In: Sleep Medicine, 2003, vol. 4(6), pp. 543-552. doi:10.1016/j.sleep.2003.08.008.
7. Prabhakar, S., K., Rajaguru, H.: Pca and k-means clustering for classiffcation of epilepsy risk levels from eeg signals - A comparitive study between them. In: 2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 2015, pp. 83-86. IEEE. doi:10.1109/ICIIBMS.2015.7439467.
8. Sana Tmar-Ben Hamida, Beena Ahmed, and Thomas Penzel. A novel insomnia identiffcation method based on hjorth parameters. In: 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Abu Dhabi, 2015, pp. 548-552. IEEE, ISBN 9781509004812. doi:10.1109/ISSPIT.2015.7394397.
9. Van der Maaten, L., Hinton, G.: Visualizing Data using t-SNE. In: Journal of Machine Learning, 2008, vol. 9, pp. 2579-2605.
10. Värri, A.: Algorithms and systems for the analysis of long-term physiological signals. In: Tampereen teknillinen korkeakoulu. Julkaisuja. Tampere University of Technology, 1992.
11. Welch, P.: The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modiffed periodograms. In: IEEE Transactions on Audio and Electroacoustics, 1967, vol. 15(2), pp. 70-73. doi:10.1109/TAU.1967.1161901.
12. Ware, S., V., Bharathi HN.: Study of density based algorithms. In: International Journal of Computer Applications, 2013, vol. 69(26). doi:10.5120/12132-8235.