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

Analýza příznaků automatické klasifikace epileptických EEG záznamů za pomoci algoritmu k-NN

Analysis of features for automated classification of epileptic EEG recordings using the k-NN algorithm

Bakalářská práce

Studijní program:	Biomedicínská a klinická technika
Studijní obor:	Biomedicínský technik

Autor bakalářské práce:	Barbora Balcarová
Vedoucí bakalářské práce:	Ing. Hana Schaabová
Konzultant práce:	doc. Ing. Vladimír Krajča, CSc.

květen 2018

České vysoké učení technické v Praze, Fakulta biomedicínského inženýrství

Katedra biomedicínské techniky

Akademický rok: 2017/2018

Zadání bakalářské práce

Student:	**Barbora Balcarová**
Obor:	Biomedicínský technik
Téma:	**Analýza příznaků automatické klasifikace epileptických EEG záznamů za pomoci algoritmu k-NN**
Téma anglicky:	Analysis of features for automated classification of epileptic EEG recordings using the k-NN algorithm

Zásady pro vypracování:

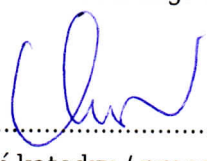
Navrhněte a implementujte algoritmus k-NN pro klasifikaci segmentů EEG signálu se zaměřením na rozlišení tříd epileptické aktivity. Analyzujte příznaky využívané pro tento druh klasifikace se zaměřením na epileptickou aktivitu a odlišení této aktivity od ostatních segmentů. Proveďte detailní posouzení výsledků klasifikace v závislosti na počtu vybraných příznaků a korelace mezi jednotlivými příznaky použitými ke klasifikaci. Diskutujte vhodnost jednotlivých příznaků pro klasifikaci epileptických EEG segmentů. Výsledky k-NN klasifikace v závislosti na výběru příznaků statisticky vyhodnoťte pomocí klasifikačních matic.

Seznam odborné literatury:

[1] Krajča V., Mohylová J., Číslicové zpracování neurofyziologických signálů, ed. 1., ČVUT Praha, 2011, 167 s., ISBN 978-80-01-04721-7

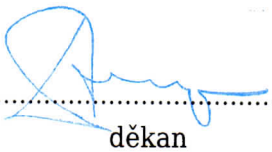
[2] Lotte, F., Congedo M., L'Ecuyer, A., Lamarche, F., Arnaldi, B., A review of classification algorithms for EEG-based brain-computer interfaces, Journal of Neural Engineering, ročník 4, číslo 1, 2007

Zadání platné do:	20.09.2019
Vedoucí:	Ing. Hana Schaabová
Konzultant:	doc. Ing. Vladimír Krajča, CSc.



.....

vedoucí katedry / pracoviště



.....

děkan

V Kladně dne 19.02.2018

Prohlášení

Prohlašuji, že jsem bakalářskou práci s názvem **Analýza příznaků automatické klasifikace epileptických EEG záznamů za pomoci algoritmu k-NN** vypracoval/a samostatně a použil/a k tomu úplný výčet citací použitých pramenů, které uvádím v seznamu přiloženém k bakalářské/diplomové práci.

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

V Kladně dne 17. 5. 2018 .....

Barbora Balcarová

Poděkování

Děkuji své vedoucí Ing. Haně Schaabové za podporu a vedení práce, ochotu pomoci téměř s čímkoli a neutuchající optimismus a doc. Vladimíru Krajčovi za konzultace. Dále děkuji prim. MUDr. Ing. Svojmilu Petránkovi a neurologickému oddělení nemocnice Na Bulovce z jejichž spolupráce pocházejí zpracovávaná dat. A také děkuji svým rodičům za všeobecnou podporu při studiu a Honzovi za trpělivost.

Abstrakt

Analýza příznaků automatické klasifikace epileptických EEG záznamů za pomoci algoritmu k-NN

Práce se zabývá automatickou klasifikací elektroenfalografického (EEG) signálu, jehož segmenty lze rozdělit do tříd podle typu mozkové aktivity. Za účelem lepší klasifikace do těchto tříd jsem analyzovala příznaky etalonů používané k detekci epileptické aktivity v EEG záznamech, jakých hodnot nabývají a jak se tyto hodnoty liší v závislosti na příslušnosti segmentu k třídě. V programovém prostředí MATLAB® jsem implementovala k-NN klasifikátor a s jeho použitím byly klasifikovány EEG záznamy pěti pacientů. Pro klasifikaci je použito nejprve 24 příznaků a následně po redukci na základě zjištěných korelací mezi příznaky je provedena klasifikace jen s 20, 15, 10 a dokonce 5 příznaky. Klasifikace je vyhodnocena pomocí klasifikačních (konfuzních) matic a pomocí statistických charakteristik: senzitivity, specificity a přesnosti. Výsledky klasifikace naznačují, že z důvodu vysoké korelace příznaků mezi sebou lze vybrat 5 příznaků tak, aby jejich klasifikace byla dostačující bez výrazného poklesu její přesnosti.

Klíčová slova

Klasifikace EEG, příznaky, korelace, algoritmus k-NN, trénovací množina.

Abstract

Analysis of features for automated classification of epileptic EEG recordings using the k-NN algorithm

The thesis deals with automated classification of electroencephalographic (EEG) signal, segments of the signal can be classified into classes depending on the type of brain activity. I analysed features used for detection of epileptic activity in EEG recordings, what their typical values are and what the difference is between the values depending on segment's class. I implemented the k-NN classifier myself in the programming environment MATLAB ® and it was used for classification of five EEG patient's recordings. Originally 24 features was used for classification, then it was reduced to 20, 15, 10 and 5 features, the features reduction is based on correlations found between the features. The classification is evaluated by confusion matrices and also by statistical characteristics like sensitivity, specificity and precision. The results show high correlation between some features and the classification indicates that using even only 5 features could be enough without significant decrease in precision.

Key words

EEG classification, features, correlation, k-NN algorithm, training set.

Obsah

Seznam použitých symbolů a zkratek

EEG	Elektroencefalografie
EMG	Elektromyografie
BCI	Brain-Computer Interface
LDA	Linear Discriminant Analysis
SVM	Support Vector Machine
MLP	Multi-Layer Perception
k-NN	k Nearest-Neighbours
FFT	Fast Fourier Transform
PCA	Principal Component Analysis
ICA	Independent Component Analysis

Seznam tabulek

Seznam obrázků

1 Úvod

Elektroencefalogram (EEG) je záznam elektrických potenciálů způsobených aktivitou mozku a obsahuje velké množství informací. Cílem digitální analýzy a zpracování EEG záznamů je zjednodušit a urychlit práci lékaře při vyhodnocování těchto záznamů [1]. EEG záznam jednotlivých kanálů lze rozdělit na segmenty, jednotlivé segmenty můžeme potom různými automatickými nebo poloautomatickými metodami klasifikovat do tříd, které odpovídají jak fyziologickým aktivitám (normální, epileptická, EMG a další), tak nefyziologickým aktivitám, jako jsou např. technické artefakty [1, 2]. Počet ani konkrétní obsah těchto tříd není ustálený a může se lišit dle zaměření studie nebo názoru lékaře.

Konkrétnější možnost postupu zpracování signálu lze shrnout do schématu (obrázek 1.1), přičemž důležitými články tohoto postupu, zkoumanými v této práci, jsou výběr vhodných příznaků, výběr etalonů pro klasifikaci a samotná klasifikace segmentů na základě příznaků (popsáno níže).

Obrázek 1.1: Řetězec zpracování EEG

1.1 Přehled současného stavu

1.1.1 Automatická klasifikace EEG segmentů

Jak bylo zmíněno v úvodu, EEG záznam lze rozdělit na segmenty a segmenty klasifikovat do tříd. Třídy segmentů jsou charakterizovány tzv. příznaky (features), tj. parametry popisujícími segment. Krokem předcházejícím klasifikaci segmentů je extrakce příznaků z jednotlivých segmentů. Příznaky mohou charakterizovat signál v časové, frekvenční, případně entropické oblasti [1].

Počet použitých příznaků pro klasifikaci se podle autorů a pro různé účely liší, například v [2] je použito 16 příznaků pro epileptické pacienty, v [3] pak 24 pro sledování novorozeneckého EEG a v [4] naopak pouze 6 pro klasifikaci transientů. V [4] např. používají dobu trvání, plochu pod EEG křivkou, průměrný sklon, špičatost, směrodatnou odchylku a dominantní frekvenci.

Výše uvedené zdroje pak používají tzv. normalizaci příznaků, díky které se příznaky s původně různými fyzikálními rozměry stanou bezrozměrné. Na základě příznaků má potom každý segment své místo v tzv. příznakovém prostoru.

Ve chvíli, kdy jsou ze segmentů EEG signálu extrahovány příznaky, může následovat jejich samotná klasifikace. Pro klasifikaci je možné použít široké spektrum klasifikačních algoritmů a to s učením jak s učitelem, tak bez učitele.

Algoritmy s učením bez učitele nepotřebují mít předem žádnou informaci o objektech klasifikace, hledají přirozenou strukturu dat na základě podobnosti resp. vzdálenosti objektů. Příkladem takové metody je shluková analýza, konkrétně např. nehierarchická metoda k-středů (k-means). [1]

Naproti tomu metody využívající učení s učitelem potřebují pro svou činnost trénovací množinu objektů (tzv. etalony, prototypy). Výhodou je, že umožňují online klasifikaci.

Trénovací množinu lze získat více způsoby, obecně ji můžeme vytvořit nebo vybrat z existujících dat ručně, nebo lze použít automatickou metodu např. shlukovou analýzu (algoritmus k-means) [1].

Pro využití při klasifikaci EEG segmentů je ruční výběr etalonů extrémně časově náročný a shluková analýza může dávat falešné klasifikace EEG (dle experta) [5], proto je vhodné použít kombinaci obou způsobů (semiautomatická metoda) [2]. Také by bylo velmi výhodné mít univerzální trénovací množinu pro často se vyskytující třídy segmentů, která by byla přenositelná a dala se použít pro jakýkoli EEG záznam.

1.1.2 Klasifikační algoritmy

Článek [6] uvádí rozsáhlý přehled klasifikačních algoritmů používaných pro tvorbu BCI systémů (Brain-Comuputer Interface) založených na snímání EEG aktivity. (BCI systémy slouží jako rozhraní mezi mozkem a elektronickým přístrojem a umožňují posílat přístroji pokyny pouze prostřednictvím mozkové aktivity [6].) Takové algoritmy lze rozdělit do několika kategorií. S většinou níže zmíněných klasifikátorů se často setkáme ve studiích zabývajících se automatickou klasifikací EEG rozhodně ne jen pro tvorbu BCI systémů, ale i pro nejrůznější další účely a některé z nich mohou dobře posloužit rovněž pro klasifikaci patologické (např. epileptické) mozkové aktivity.

První z kategorií jsou lineární klasifikátory, které pro rozlišení různých tříd používají lineární funkce. Pro oddělení dat vytvářejí tzv. nadroviny a tím v datech rozlišují třídy. Příkladem těchto klasifikátorů může být LDA (Linear Discriminat Analysis) nebo hojně využívaný SVM (Support Vector Machine). [6]

Druhou kategorií jsou umělé neuronové sítě, což je soubor několika umělých neuronů, který umožňuje vytvářet nelineární hranice pro rozhodování. Neuronové sítě mohou mít různé architektury, nejpoužívanějším typem je MLP (MultiLayer Perceptron), který se skládá z několika vrstev umělých neuronů (vstupní, výstupní, případně jedné či více skrytých mezi nimi) a výstupy neuronů v jedné vrstvě jsou vždy připojeny na vstupy neu-

ronů v následující vrstvě. Další kategorií jsou pak nelineární Bayesovské klasifikátory, které ovšem nejsou zdaleka tak rozšířené, jako lineární klasifikátory nebo neuronové sítě. [6]

Poslední uvedenou kategorií jsou klasifikátory hledající nejbližší sousedy, mezi které náleží algoritmus k-NN, ten patří svým principem (popsaný v následující kapitole 1.1.3) mezi relativně jednoduché klasifikátory [6]. Je pro svou jednoduchost vhodný pro srovnání s dalšími, většinou složitějšími, klasifikačními metodami, což se také často využívá, např. [2], [5] nebo [7].

1.1.3 Algoritmus k-NN

Algoritmus k-NN (k Nearest Neighbours) se řadí mezi příznakově orientované metody rozpoznávání obrazu a jedná se o klasifikátor využívající učení s učitelem [1]. Algoritmus při klasifikaci nového objektu nalezne v příznakovém prostoru jeho k nejbližších sousedů, zjistí, do kterých tříd patří, a zařadí ho do třídy, do které náleží nejvíce z jeho nejbližších sousedů. Algoritmus ze svého principu potřebuje pro svoji funkci trénovací množinu objektů (etalony). Tedy data, která jsou již předem roztříděna do tříd, a na jejich základě potom do těchto tříd klasifikuje objekty nové.

Tento algoritmus se díky své jednoduchosti a možnosti kontroly jeho činnosti hodí k analýze dalších článků zpracování EEG signálu. Proto byl algoritmus k-NN vybrán ke klasifikaci v této práci. Nejde zde totiž primárně o výběr nejlepšího klasifikačního algoritmu, ale o analýzu příznaků vhodných pro odlišení epileptické aktivity a výběr vhodných etalonových dat. Algoritmus k-NN se pro toto jeví jako vhodný nástroj.

Vzdálenost objektů klasifikace, která je využita klasifikátorem k-NN (ale případně i dalšími algoritmy), lze určit pomocí vícero různých metrik. Nejjednodušší je Euklidova vzdálenost [1], pro využití v klasifikaci EEG pro systémy BCI je dále zmiňována např. Mahalanobisova vzdálenost [6]. Při použití funkce k-NN v programu MATLAB® je dále možné použít např. Manningovu či blokovou vzdálenost [8], v [1] a [8] pak nalezneme mnoho dalších možností.

Ve studiích věnujících se klasifikaci EEG, které ke zpracování používají k-NN klasifikátor, je nejčastěji používaná Euklidova vzdálenost jako např. v [9], [10] nebo [11]. Ve studii [12] věnující se optimalizaci k-NN klasifikátoru pro rozlišení mezi normálními a dyslektickými dětmi na základě EEG záznamu byly testovány funkce pro vzdálenost: Euklidova, Kosinova a korelační. Nejlepších výsledků bylo dosaženo s Euklidovou vzdáleností.

Pro klasifikátor k-NN je základní parametr výše zmíněná hodnota k, tj. počet nejbližších sousedů, která ovlivňuje výsledky klasifikace. Pro zpracování EEG záznamů pomocí k-NN se používají různá k a jeho nejvhodnější hodnota byla i předmětem nebo součástí některých studií. Obecně vyšší hodnota k potlačuje vliv šumu a zároveň zesiluje hranice mezi třídami [13], což ovšem nemusí být žádoucí. Ve studii [11] byly vyzkoušeny hodnoty od 2 do 10 a nejlepších výsledků bylo dosaženo s hodnotou 5. Rovněž ve výše zmiňované studii [12] dosáhli nejlepších výsledků s hodnotou 5 (výběr z hodnot 1 – 13).

V článcích [13] a [7] byla použita hodnota $k = 3$, v článku [9] pro systém BCI $k = 6$ a v [14] (rovněž pro BCI) $k = 10$. Celkem v pěti nalezených článcích (včetně uvedených v předchozím odstavci) byla použita nebo vyhodnocena jako nejlepší hodnota $k = 5$ [2, 10, 11].

Při klasifikaci obecně se objevuje jeden problém tzv. prokletí dimenzionality. Týká se toho, že při klasifikaci používáme velké množství příznaků a příznakový prostor získává mnoho dimenzí. Přitom totiž platí, že pokud chceme řádně popsat rozdíly mezi jednotlivými třídami dat, tak se zvyšující se dimenzí exponenciálně roste počet potřebných trénovacích dat. Doporučuje se použít 5 až 10 krát více trénovacích dat (pro každou třídu), než je počet dimenzí. Podle [6] je algoritmus k-NN na prokletí dimenzionality citlivý. [6]

1.2 Cíle práce

Prvním cílem této práce je implementovat vlastní algoritmus k-NN navržený na základě literatury a implementaci provést v programovém prostředí MATLAB®, kde pro tento algoritmus sice existuje funkce, ale nelze všechny její parametry upravovat podle potřeb.

Dalším cílem je analyzovat příznaky využívané pro klasifikaci EEG segmentů a to konkrétně pro odlišení epileptické aktivity od ostatních segmentů a tuto analýzu provést nejen na základě literatury, ale také pomocí výpočtu statistických charakteristik příznaků a zkoumání vzájemných korelací.

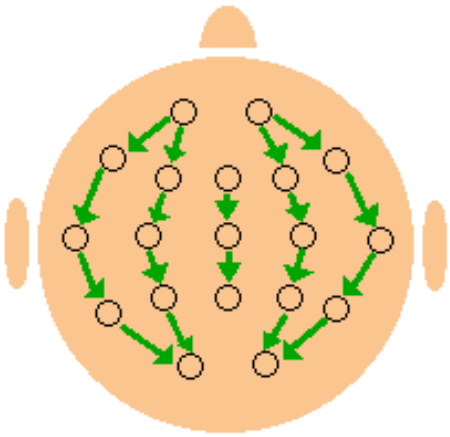
Následně chci provést klasifikaci segmentů v epileptických EEG záznamech pomocí algoritmu k-NN. Algoritmus k-NN díky svému jednoduchému principu umožňuje zaměřit svou pozornost na příznaky a minimalizovat vliv konkrétního klasifikátoru. Posledním krokem je statisticky vyhodnotit výsledky klasifikace EEG segmentů v závislosti na výběru příznaků pomocí klasifikačních matic.

2 Metody

2.1 Data

Do této práce byly použity reálné EEG záznamy naměřené ve spolupráci s Neurologickým oddělením v nemocnici Na Bulovce. Měření bylo provedeno digitálním systémem Brain-Quick se vzorkovací frekvencí 128 Hz a analogovým filtrem typu pásmová propust 0.5 – 30.0 Hz.

Byla použita data z 19 elektrod rozmístěných podle mezinárodního systému 10-20 v bipolární montáži, tzn. odečtení hodnot dvou sousedních elektrod a v konkrétně longitudinálním zapojení (obrázek 2.1), neboť dle odborníka je tato montáž vhodná pro detekci epileptické aktivity. Tím jsem dostala 18 kanálů záznamu. Z digitálních filtrů byl na záznam použit pouze tzv. mean removal.



Obrázek 2.1: Bipolární longitudinální montáž EEG elektrod

Všichni pacienti podepsali informovaný souhlas a měření bylo odsouhlaseno etickou komisí nemocnice Na Bulovce. Použitá data pochází od celkem pěti pacientů, informace o pacientech a podrobnosti o jejich záznamech jsou uvedeny v tabulce 2.1.

Tabulka 2.1: Podrobnosti o pacientech a jejich záznamech

Pacient č.	Věk	Délka záznamu (min)	Počet segmentů (celý záznam)	Počet segmentů (etalony)	Obsažené třídy (etalony)
1	66	30,9	38 063	350	0 – 6
2	18	18,1	13 694	328	0 – 6
3	24	8,4	3 311	256	0 – 5
4	19	13,7	13 244	300	0 – 5
5	33	9,4	7 042	266	0 – 5

Délka EEG záznamů se pohybovala od 8 do 32 minut, průměrná délka jednoho záznamu byla 16,1 minuta. V závislosti na délce záznamu se lišil počet segmentů celého záznamu. Jak je zřejmé z tabulky, pouze první dva pacienti měli třídu 6 (viz dále tabulka 2.2). Z každé třídy bylo vybráno 50 etalonových segmentů, pouze ve třídě 2 měli pacienti č. 2, 3 a 5 méně než 50 segmentů, jelikož EMG artefaktů nebylo v záznamu dostatek.

2.2 Předzpracování dat

V programu Wavefinder [15] doc. Krajči byla dále provedena adaptivní segmentace signálu (tj. rozdělení EEG záznamu na segmenty, které nemají stejnou délku, na základě Värriho metody dvou spojených oken [1]) a pro každý segment byly v programu spočteny hodnoty 24 příznaků (viz dále tabulka 2.3). Ze záznamu byly následně vybrány grafoelementy odpovídající etalonům jednotlivých tříd (se svými příznaky).

Tento výběr byl proveden ručně odborníkem, aby bylo možné výsledky klasifikace porovnat. Známe tak u všech vybraných segmentů jejich třídu a lze jejich část použít jako etalon, zbytek klasifikovat a určit, zda byly klasifikovány správně.

Etalony byly rozděleny do sedmi tříd (viz tabulka 2.2, včetně čísel tříd použitých v celé práci) a jak bylo uvedeno v popisu dat, pro každého pacienta bylo do každé třídy, kde to bylo možné, vybráno 50 prototypových segmentů.

Tabulka 2.2: Použité třídy segmentů EEG záznamu

0	nepatologická normální aktivita na pozadí
1	alfa aktivita
2	EMG a podobné šumy
3	oční artefakty a jim podobné transienty
4	epileptická aktivita typu hrot-vlna s nízkou amplitudou
5	epileptická aktivita typu hrot-vlna s vysokou amplitudou
6	impulsní elektrodové artefakty

Pro klasifikaci jsem se rozhodla nepoužít segmenty ze třídy 2 a 6 (tj. EMG a podobné šumy a impulsní elektrodové artefakty), jelikož se jedná o artefakty, které lze odfiltrovat. Pro výběr dat ke klasifikaci jsem tedy použila segmenty ze tříd 0, 1, 3, 4 a 5. Jednu třetinu segmentů z každé použité třídy jsem vždy považovala za neznámé (testovací množinu) a ty jsem klasifikovala vytvořenou funkcí. Pro vyhodnocení klasifikace jsem ovšem už nerozlišovala všechny třídy, ale rozdělila jsem data pouze na epileptickou aktivitu (třídy 4 a 5) a ostatní neepileptickou aktivitu (třídy 0, 1 a 3).

Jelikož příznaky použité pro klasifikaci mají různé jednotky, je nutné je před samotnou klasifikací znormovat. V některých dalších částech práce se pracuje pouze s vybranými segmenty odpovídajícími etalonům, pro normalizaci je však třeba použít celý pacientův EEG záznam. Normalizaci příznaků jsem provedla pomocí maximální a minimální hodnoty každého příznaku. Normovaná hodnota x_n každého příznaku pak byla vypočtena podle následujícího vztahu:

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{1}$$

kde vždy x je původní hodnota příznaku, x_{max} a x_{min} jsou maximální a minimální hodnota tohoto příznaku v rámci celého souboru dat jednoho pacienta.

2.3 Program MATLAB®

Většinu zpracování dat jsem provedla v programu MATLAB® společnosti MathWorks. Používala jsem jeho verzi R2014a (akademickou licenci pro ČVUT).

2.4 Použité příznaky

Z programu Wavefinder jsem měla k dispozici hodnoty celkem 24 příznaků [16, 17] pro každý segment, tyto příznaky včetně zkratek a jednotek jsou uvedeny v tabulce 2.3.

Tabulka 2.3: Příznaky použité v programu Wavefinder

Číslo	Zkratka	Jednotka	Popis
1	SIGM	μV	směrodatná odchylka hodnot amplitudy
2	APOS	μV	maximální pozitivní hodnota v segmentu
3	ANEG	μV	minimální negativní hodnota v segmentu
4	Delt1	—	hodnota FFT ve frekvenčním pásmu 0,5 – 1,5 Hz
5	Delt2	—	hodnota FFT ve frekvenčním pásmu 2,0 – 3,5 Hz
6	Thet1	—	hodnota FFT ve frekvenčním pásmu 4,0 – 5,5 Hz
7	Thet2	—	hodnota FFT ve frekvenčním pásmu 6,0 – 7,5 Hz
8	Alph1	—	hodnota FFT ve frekvenčním pásmu 8,0 – 10,0 Hz
9	Alph2	—	hodnota FFT ve frekvenčním pásmu 10,5 – 12,5 Hz
10	Sigma	—	hodnota FFT ve frekvenčním pásmu 18,0 – 29,0 Hz
11	Beta	—	hodnota FFT ve frekvenčním pásmu 13,0 – 17,5 Hz
12	MAX1D	μV/s	maximální absolutní hodnota první derivace
13	MAX2D	$\mu V/s^2$	maximální absolutní hodnota druhé derivace
14	mf	Hz	střední frekvence
15	MD1	μV/s	průměrná hodnota první derivace
16	MD2	$\mu V/s^2$	průměrná hodnota druhé derivace
17	mob	—	mobilita (2. Hjorthův parametr)
18	comp	—	komplexita (3. Hjorthův parametr)
19	act	μV^2	aktivita (1. Hjorthův parametr)
20	LOfC	—	délka křivky
21	NLinE	μV^2	nelineární energie
22	ZC	—	počet průchodů nulou
23	Peaks	Hz	dominantní spektrální vrchol
24	Infle	—	inflexní bod

2.5 Testování normality

Před dalším zpracováním dat jsem nejprve otestovala normální rozdělení hodnot každého z příznaků uvedených v předchozí kapitole a to jak pro celý zázanam, tak pro každou třídu prototypových segmentů zvlášť. Použila jsem jednovýběrový Smirnov-Kolmogorův test, konkrétně funkci v programu MATLAB® `kstest`. Nulová hypotéza předpokládá, že data jsou z normálního rozložení, proti alternativě, že data z normálního rozložení nejsou. Hladina významnosti pro zamítnutí nulové hypotézy je 5 %. [18]

Pokud data nemají normální rozdělení, nemá to vliv na samotnou klasifikaci, jelikož pracuji s klasifikátorem k-NN, který žádné konkrétní (tedy ani normální) rozdělení dat nepředpokládá. Má to ovšem vliv na volbu správného výpočtu odhadu korelačního koeficientu a také se jedná o zajímavou charakteristiku příznaku.

2.6 Výpočet korelace a metoda redukce příznaků

Jak bylo zmíněno výše, měla jsem k dispozici 24 příznaků, to je velké množství, pokud uvážíme, že 24 příznaků počítáme pro každý segment, kterých v EEG záznamu pacienta může být i desítky tisíc nebo více. Když se následně provede klasifikace, je to opět v prostoru o 24 dimenzích. I z těchto důvodu je vhodné počet příznaků zredukovat.

Často používanými metodami k redukci příznakového prostoru pro klasifikaci EEG jsou analýza hlavních komponent (PCA) a analýza nezávislých komponent (ICA)[19]. Jinou možností je provést redukci na základě korelací jednotlivých příznaků. Po jednou provedené redukci by tak na rozdíl od výše zmíněných metod už nebylo třeba počítat u nových pacientů všech 24 příznaků a provádět následnou redukci, ale počítaly by se pouze vybrané příznaky.

Pro zkoumání korelací mezi jednotlivými příznaky jsem použila funkci `corr` v programu MATLAB® a jelikož hodnoty příznaků nepocházely podle předcházejícího tes-

tování z normálního rozdělení (viz kapitola 3.3), zvolila jsem jako typ výpočtu korelace Spearmanův koeficient pořadové korelace ρ [20]. Spearmanův koeficient je neparametrický koeficient, který je robustní vůči odchylkám od normality a popisuje, jak vztah dvou náhodných vektorů odpovídá monotónní funkci [21]. Pracuje s pořadími pozorovaných hodnot a vypočte se podle následujícího vztahu:

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}, \tag{2}$$

kde X_i a Y_i jsou pořadí vzestupně uspořádaných hodnot ve vektoru X a Y, n je délka vektoru X resp. Y (oba vektory musí mít stejnou délku) a čísla $\overline{X}$ a $\overline{Y}$ jsou průměry hodnot X_i a Y_i (tj. průměrná pořadí) [21]. Program MATLAB® počítá Spearmanův koeficient právě podle vztahu 2 [20].

Korelační koeficienty jsem počítala pro každou dvojici příznaků a s použitím segmentů celého pacientského záznamu. Nejprve jsem je vypočetla pro každého pacienta zvlášť a potom jsem z hodnot pro jednotlivé pacienty určila aritmetický průměr.

Na základě volby limitu pro korelační koeficient (např. 0,9) jsem potom mohla postupně odebírat příznaky, jejichž korelace byla vyšší něž tento limit. Z dvojice příznaků, které spolu korelovaly, jsem vždy pro klasifikaci ponechala pouze jeden. Tímto způsobem jsem pomocí změny limitu pro koeficient odebírala i z dvojic příznaků, které měly postupně menší míru korelace, a mohla hodnotit výsledky klasifikace v závislosti na počtu použitých příznaků.

2.7 Implementace k-NN

2.7.1 Parametry algoritmu k-NN

Pro implementaci algoritmu k-NN jsem musela zvolit jeho parametry. Základním parametrem algoritmu je hodnota k, která určuje, kolik nejbližších sousedů algoritmus hledá.

Ve vytvořené funkci jsem pro její univerzálnost nechala hodnotu k jako nastavitelný vstupní argument (viz kapitola 2.7.2), na základě provedené rešerše jsem ale pro klasifikaci segmentů v této práci vždy používala hodnotu $k = 5$.

Druhým základním parametrem je použitá metrika vzdálenosti pro zjištění nejbližších sousedů. Opět na základě rešerše jsem se rozhodla použít Euklidovu vzdálenost.

2.7.2 Vstupní a výstupní parametry vytvořeného klasifikátoru k-NN

Pro zpracování dat v této práci jsem vytvořila v prostředí MATLAB® funkci `kNN_vCyklu2`. Pro funkci jsem zvolila následující vstupní a výstupní argumenty (jsou popsány v tabulce 2.4 a 2.5).

Tabulka 2.4: Vstupní argumenty funkce

Označení	Popis
x	Trénovací množina, matice, kde každý řádek obsahuje hodnoty příznaků pro segment. Počet sloupců odpovídá počtu použitých příznaků, počet řádků odpovídá počtu segmentů.
class	Sloupcový vektor obsahující čísla tříd segmentů v trénovací množině (matice x). Počet řádků musí být stejný jako u matice x.
Pvec	Klasifikované segmenty, matice, kde každý řádek obsahuje hodnoty příznaků pro segment určený ke klasifikaci. Počet sloupců (tj. příznaků) musí být stejný jako u matice x.
k	Hodnota k, tj. počet nejbližších sousedů.

Tabulka 2.5: Výstupní argumenty funkce

Označení	Popis
classNew	Sloupcový vektor obsahující čísla tříd segmentů testovací množiny (tedy segmentů z matice *Pvec*. Počet řádků odpovídá rozměru matice *Pvec*.

2.7.3 Popis algoritmu k-NN a pseudokód

Celý kód vytvořené funkce je uveden v příloze A. Zde uvádím pseudokód vytvořené funkce (algoritmus 1) a jeho stručný popis.

Input: příznaky trénovací množiny (`Pvec`), třídy segmentů trénovací množiny (`class`), příznaky testovací množiny (`x`), počet nejbližších sousedů (`k`)

```
i = 1;
for každý segment testovací množiny do
    P = příznaky segmentu i testovací množiny;
    spočti vzdálenosti testovacího segmentu od segmentů v trénovací množině;
    NN = výběr k segmentů s nejnižší vzdáleností od P (nejbližší sousedé);
    classNN = třídy NN segmentů;
    spočti kolik NN, je z které třídy;
    if existuje právě jedno maximum počtu NN z jedné třídy then
        classP = třída, z které je maximální počet segmentů ;
        else
            sečti vzdálenosti NN z každé maximální třídy;
            classP = třída NN, které mají nejnižší součet vzdáleností;
        end
    end
    P přidej do trénovací množiny;
    classP přidej do vektoru tříd segmentů trénovací množiny;
    i = i+1;
end
```

Output: třídy segmentů testovací množiny(`classNew`)

Algoritmus 1: Funkce k-NN

Funkce nejprve vybere z matice *Pvec* parametry (příznaky) prvního segmentu. Pomocí funkce `pdist2` vypočte vzdálenosti tohoto segmentu od všech ostatních v trénovací množině *x*. Vzdálenosti seřadí od nejmenší k nejvyšší a vybere *k* segmentů s nejmenší

vzdáleností (tj. k nejbližších sousedů). Dále u těchto nejbližších sousedů zjistí jejich třídy a spočte, kolik segmentů je ze které třídy.

Pokud počet nejbližších sousedů z jedné třídy je maximální, přiřadí funkce klasifikovaný segment do této třídy. V případě, že počet nejbližších sousedů ze dvou nebo více tříd je stejný a zároveň maximální, postupuje funkce následovně. Sečte vzdálenosti nejbližších sousedů z těchto tříd a segment přiřadí do třídy, pro kterou bude tento součet nejnižší.

Nakonec parametry segmentu přiřadí do posledního řádku matice x a přiřazenou třídu jako poslední prvek vektoru *class*. Algoritmus postupně takto klasifikuje jednotlivé segmenty z matice *Pvec*. Po ukončení cyklu vrátí ve vektoru *classNew* čísla tříd klasifikovaných segmentů.

2.7.4 Testování funkce

U výše popsaného vytvořeného algoritmu jsem musela nejprve otestovat jeho funkčnost. Toto testování bylo provedeno na výchozí sadě dat z programového prostředí MATLAB®: `dataset fisheriris` (https://www.mathworks.com/help/stats/sample-data-sets.html), která obsahuje 150 vzorků dat, každý z nich se čtyřmi příznaky a názvem třídy, do které vzorek patří. Pro ověření klasifikace ručně přidaného vzorku jsem provedla grafické zobrazení nalezených nejbližších sousedů a výsledné třídy klasifikace.

2.8 Klasifikace

Dále jsem provedla klasifikaci záznamu každého pacienta pomocí vlastního implementovaného algoritmu k-NN. Klasifikaci jsem prováděla s různým počtem vybraných příznaků na základě dříve zjištěných korelací.

Klasifikaci jsem prováděla pouze s vybranými (etalonovými) segmenty, kde jsem u každého segmentu věděla, do jaké třídy náleží a mohla tak výsledky klasifikace kontrolovat. Z těchto dat jsem pro každou z tříd (0, 1, 3, 4 a 5) použila 2/3 dat jako trénovací množinu a zbylou třetinu jako množinu testovací. Na trénovací a testovací jsem data rozdělovala náhodně pomocí funkce `crossvalind` v programu MATLAB®. Využila jsem metodu křížové validace a klasifikaci takto náhodně (a pokaždé různě) rozdělených dat jsem opakovala při každém testování vždy několikrát. Pro stanovení vhodného počtu opakování jsem zvolila následující metodu.

Po každé jednotlivé provedené klasifikaci jsem ji rovnou vyhodnotila pomocí klasifikační matice (viz kapitola 2.9). Při dalších klasifikacích jsem matici vypočítala jako aritmetický průměr všech dosud provedených vyhodnocení a sledovala jsem, jak se hodnoty v matici změnily. Pokud se hodnoty v matici během posledních tří opakování nezměnily o více než 0.2, ukončila jsem proces validace a další opakování klasifikace jsem už neprováděla. Minimální počet opakování jsem stanovila na pět.

Na základě zvolené podmínky byla klasifikace opakována obvykle např. patnáctkrát. Podmínku ukončení opakování jsem považovala za dostačující z toho důvodu, že klasifikační matice byla v tomto kroku počítána v absolutních počtech segmentů, počet klasifikovaných segmentů mírně překračoval 100, takže odchylka v momentě ukončení byla méně než 0,2 %.

2.9 Vyhodnocení klasifikace

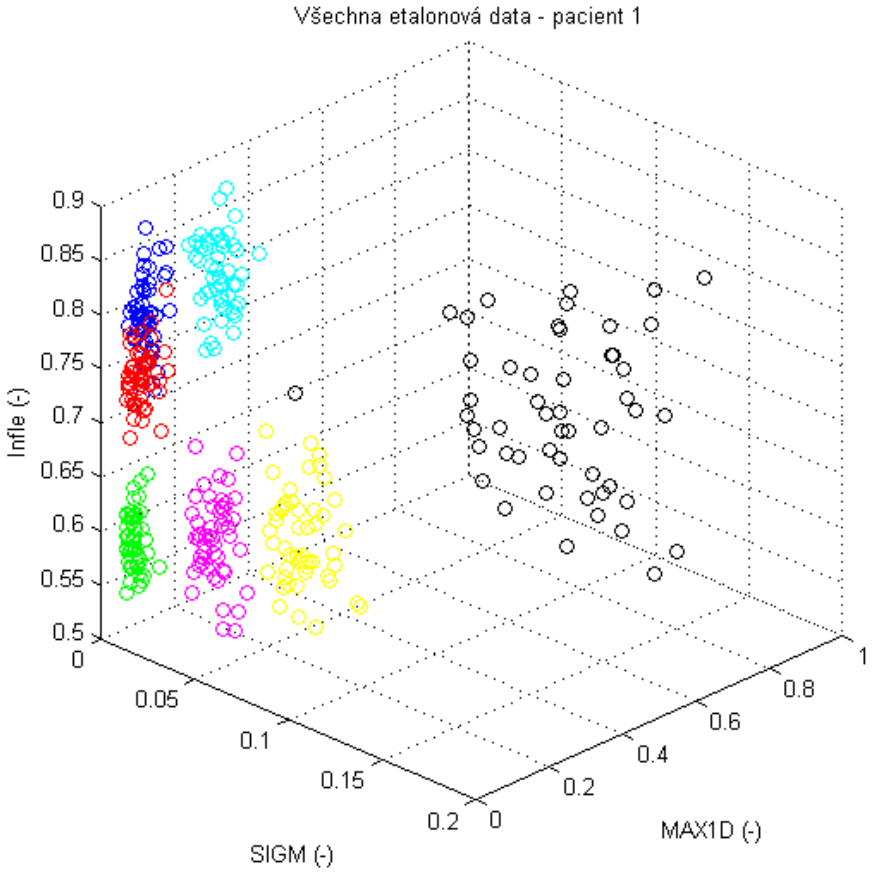
Vyhodnocení klasifikace EEG segmentů v závislosti na počtu vybraných příznaků jsem provedla pomocí tzv. klasifikační (konfuzní) matice. V klasifikační matici je znázorněno, kolik segmentů bylo klasifikováno správně (jako epileptická nebo neepileptická aktivita) a kolik nesprávně. Jako epileptickou aktivitu jsem brala segmenty ve třídách 4 a 5 (epileptická aktivita typu hrot-vlna s nízkou a vysokou amplitudou), jako neepileptickou pak segmenty ve třídách 0, 1 a 3 (viz tabulka 2.2).

Dále jsem kvalitu klasifikace ohodnotila pomocí vypočtení obvyklých statistických charakteristik: senzitivity a specificity. Také jsem určila přesnost klasifikace, kterou jsem definovala podle [4] jako podíl správně klasifikovaných segmentů ku všem klasifikovaným segmentům.

3 Výsledky

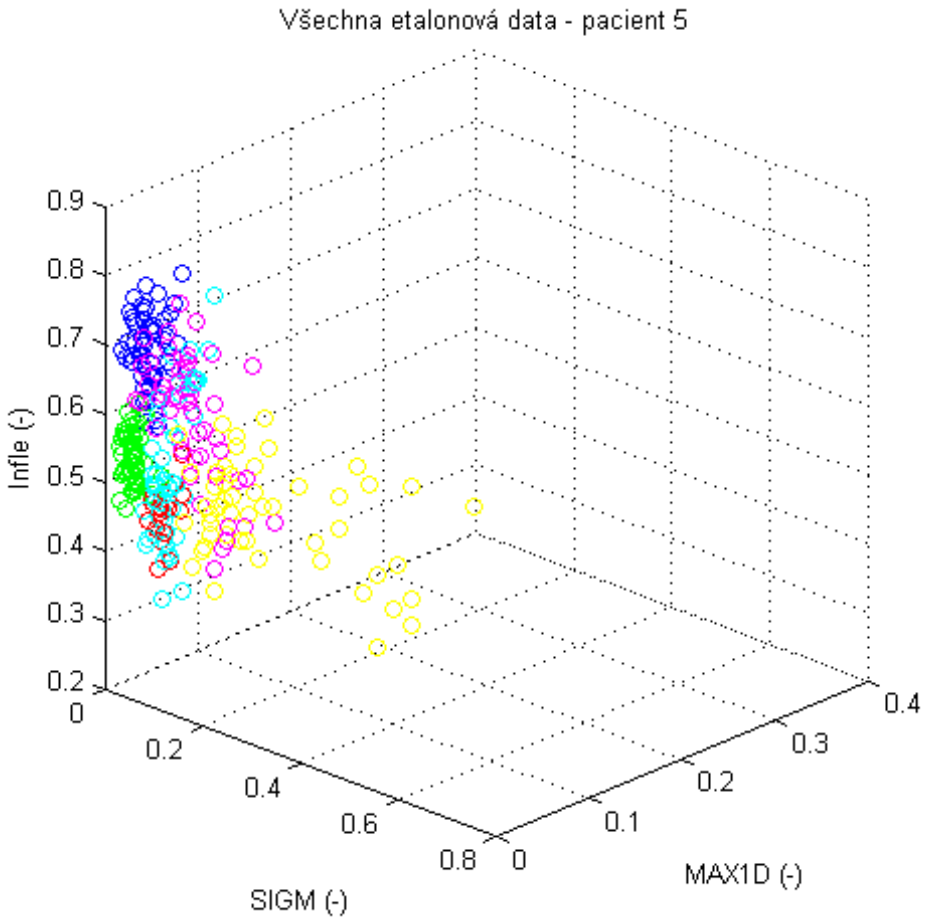
3.1 Příznakový prostor

Příznaky extrahované ze segmentů EEG záznamu utvoří tzv. příznakový prostor. Se všemi 24 námi použitými příznaky má tento prostor 24 dimenzí. Pro vhodné zobrazení (24 dimenzí nelze vhodně zobrazit) ho lze promítnout do 3D či 2D prostoru. Právě projekce do prostoru tří příznaků (SIGM, MAX1D a Infle) je ukázána na obrázcích 3.1 a 3.2 pro dva vybrané pacienty a jejich etalonové segmenty.



Obrázek 3.1: Etalonová data (pacient č. 1). Barvy znázorňují jednotlivé třídy segmentů: modrá – normální nepatoligická aktivita, zelená – alfa aktivita, červená – EMG a podobné šumy, tyrkysová – oční artefakty, fialová – nízká epileptická aktivita, žlutá – vysoká epileptická aktivita, černá – impulsní elektrodové artefakty.

Mezi pacienty můžeme vidět jasné rozdíly v hodnotách tří uvedených příznaků. Pacient 1 má většinu tříd i v této projekci do 3D prostoru dobře oddělených. Ovšem často pacienti nemají takto dobře oddělitelné třídy a data v příznakovém prostoru mohou vypadat tak jako u pacienta 5 (obrázek 3.2). Jsou zde vidět sice poměrně kompaktní shluky jednotlivých tříd, ale ty se různě prolínají, navíc se v záznamu pacienta 5 nevyskytují impulsní elektrodové artefakty (třída 6).



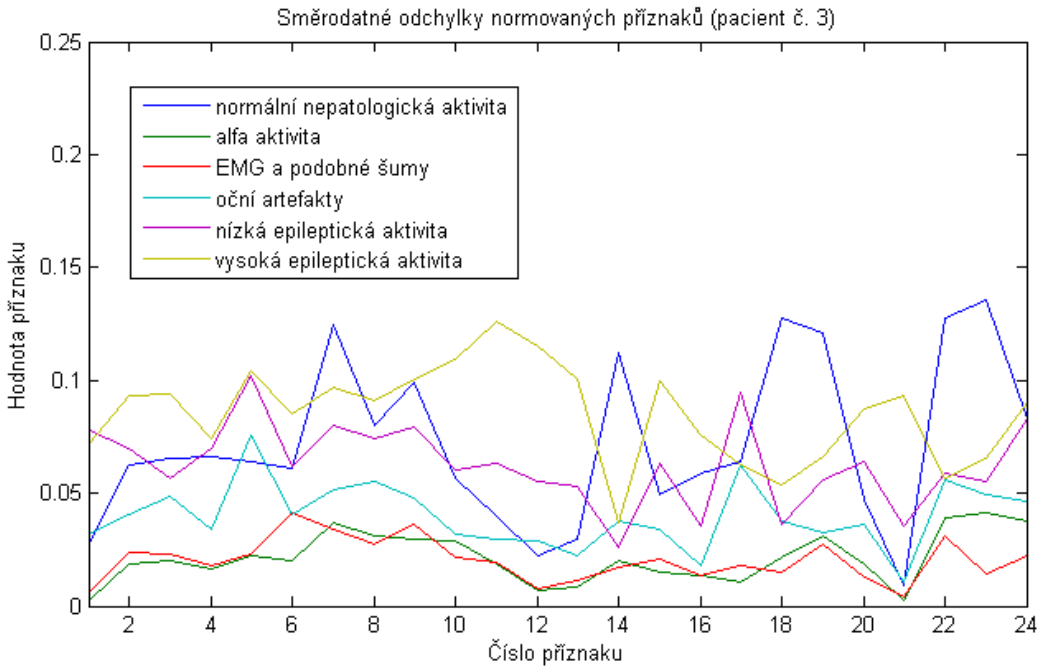
Obrázek 3.2: Etalonová data (pacient č. 5). Barvy znázorňují jednotlivé třídy segmentů: modrá – normální nepatoligická aktivita, zelená – alfa aktivita, červená – EMG a podobné šumy, tyrkysová – oční artefakty, fialová – nízká epileptická aktivita, žlutá – vysoká epileptická aktivita.

3.2 Analýza hodnot příznaků

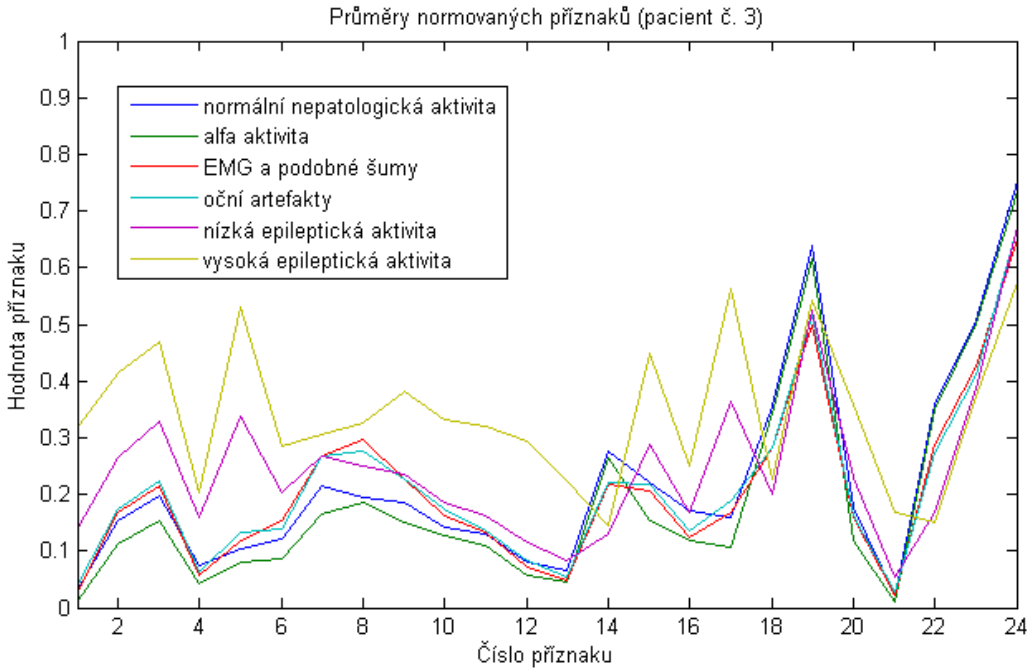
Analyzovala jsem příznaky etalonových segmentů tak, že jsem nejprve zjistila průměrné hodnoty pro každý příznak, dále jejich směrodatné odchylky a mediány s následným porovnáním těchto hodnot.

Toto jsem provedla u každého pacienta jak pro nenormované, tak normované hodnoty. Nenormované hodnoty lze kvůli různosti jednotek velmi obtížně zobrazit do přehledného grafu, proto jsou následující grafy vytvořené na základě normovaných hodnot.

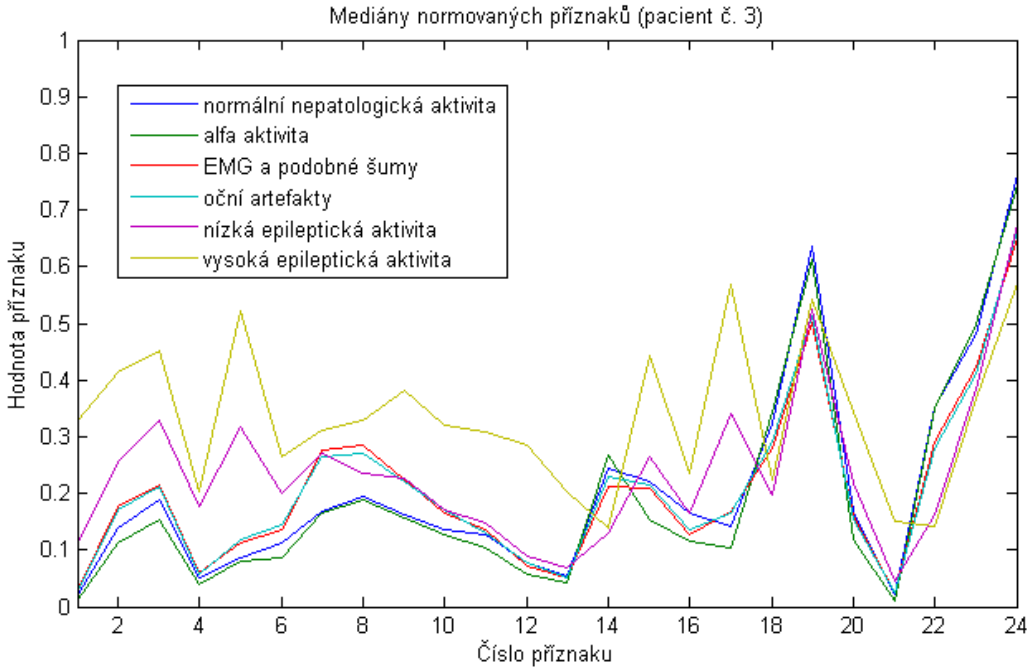
Na příkladu pacienta č. 3 (obrázky 3.3, 3.4 a 3.5) je vidět, že průměry příznaků se od mediánů lišily jen velmi málo, to platí i pro ostatní pacienty, proto dále už uvádím pouze mediány, což je vhodnější parametr pro data, která nemají normální rozdělení (viz níže, kapitola 3.3).



Obrázek 3.3: Směrodatné odchylky normovaných příznaků pro jednotlivé třídy (pacient č. 3)

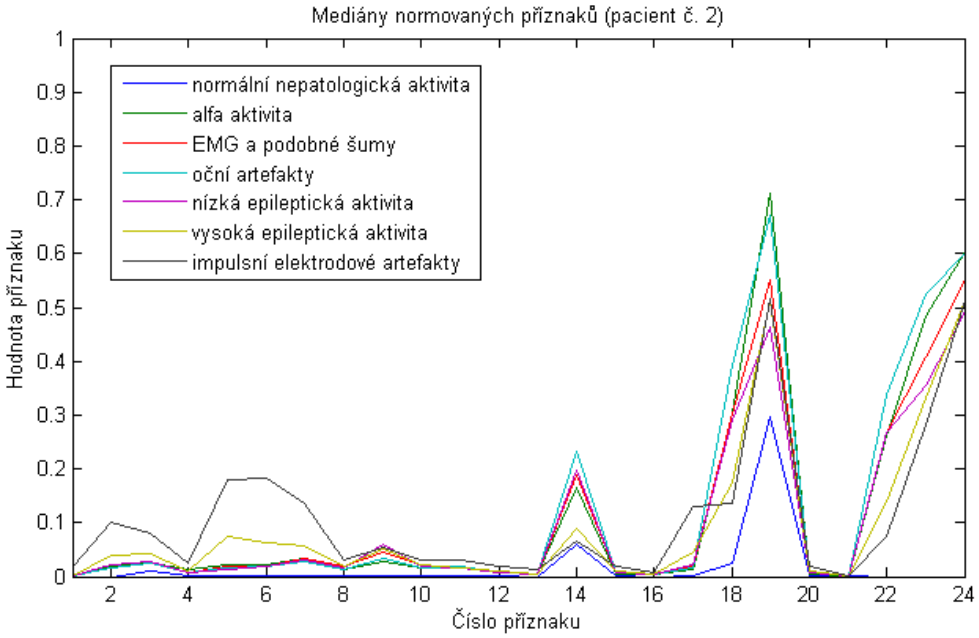


Obrázek 3.4: Průměry normovaných příznaků pro jednotlivé třídy (pacient č. 3)

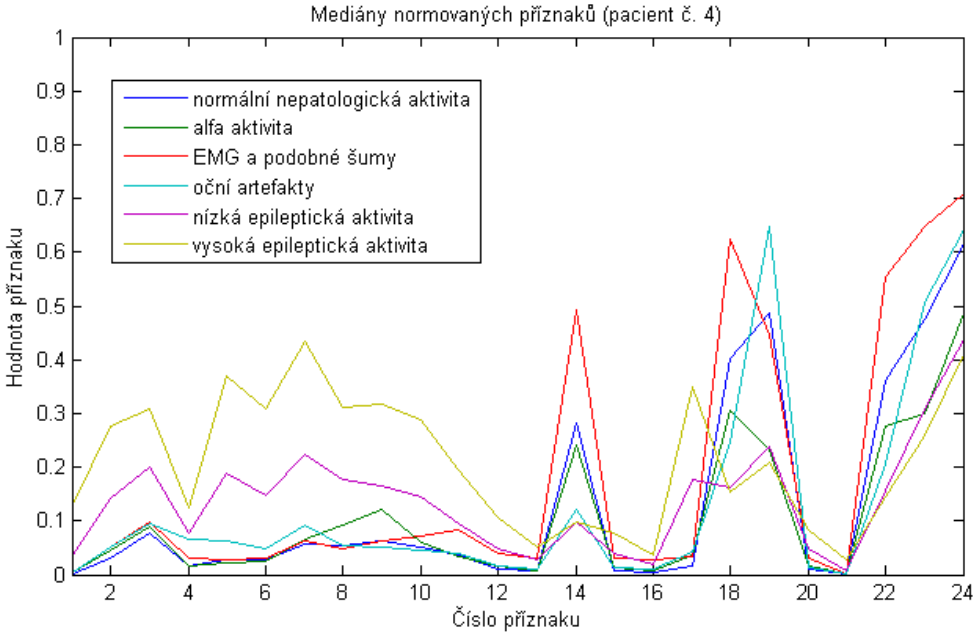


Obrázek 3.5: Mediány normovaných příznaků pro jednotlivé třídy (pacient č. 3)

Hodnoty příznaků se poměrně lišily pacient od pacienta. Pro srovnání uvádím ještě pacienty č. 2 a 4. Hodnoty příznaků ostatních pacientů jsou uvedeny v příloze B.

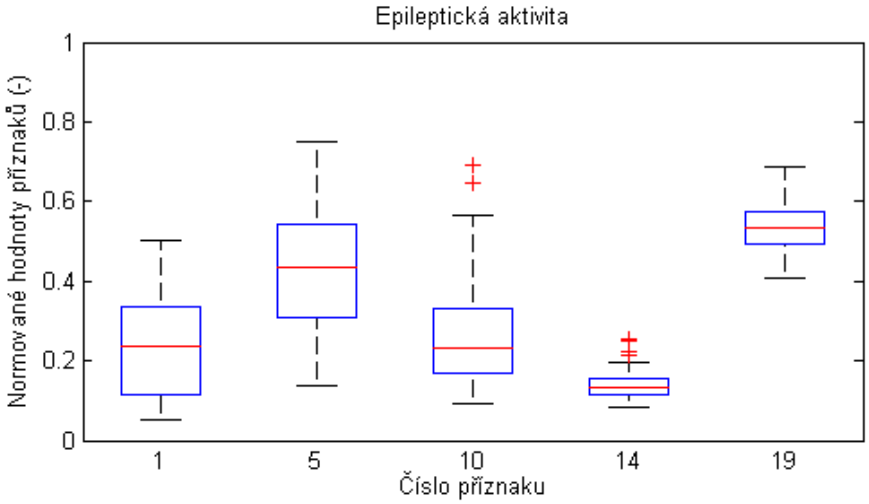


Obrázek 3.6: Mediány normovaných příznaků pro jednotlivé třídy (pacient č. 2)

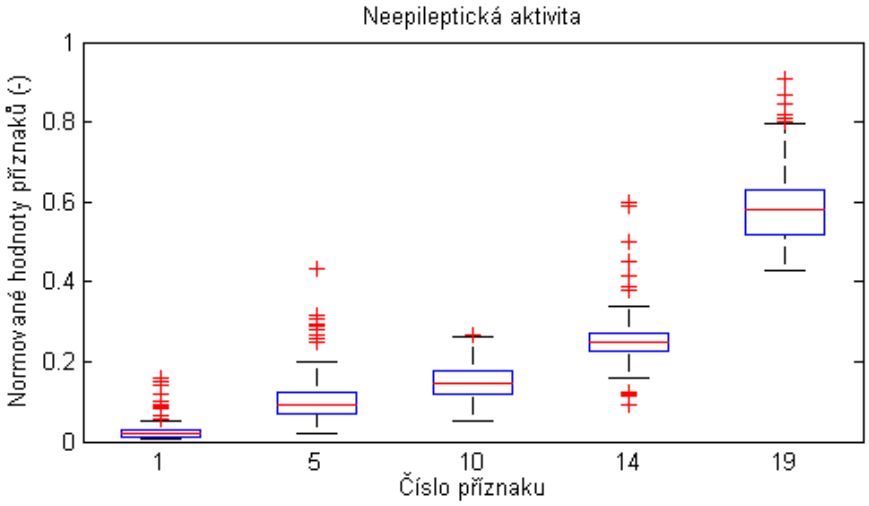


Obrázek 3.7: Mediány normovaných příznaků pro jednotlivé třídy (pacient č. 4)

Hodnoty pěti vybraných příznaků (vybrané dále na základě redukce v kapitole 3.5) také do tzv. boxplotů. Lze z nich na obrázcích 3.8 a 3.9 porovnat, jak se liší hodnoty těchto příznaků epileptické aktivity (třídy 4 a 5) a ostatní neepileptické aktivity (třídy 0, 1 a 3).



Obrázek 3.8: Hodnoty vybraných příznaků pro třídy s epileptickou aktivitou – třídy 4 a 5 (pacient č. 3). Příznaky: 1 – směrodatná odchylka, 5 – Delta 2, 10 – Sigma, 14 – střední frekvence EEG aktivity, 19 – Hjortův parametr aktivita.

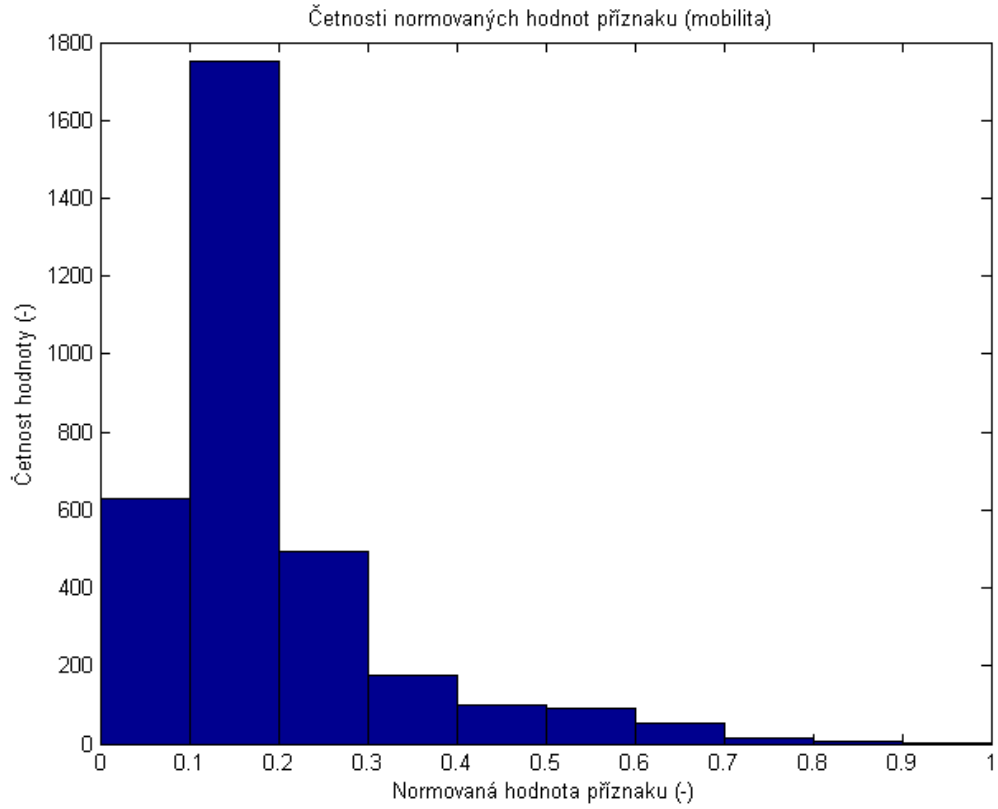


Obrázek 3.9: Hodnoty vybraných příznaků pro třídy s neepileptickou aktivitou – třídy 0, 1 a 3 (pacient č. 3). Příznaky: 1 – směrodatná odchylka, 5 – Delta 2, 10 – Sigma, 14 – střední frekvence EEG aktivity, 19 – Hjortův parametr aktivita.

3.3 Normalita dat

Pro každého pacienta jsem testovala, zda mají hodnoty všech příznaků v záznamu normální rozdělení. To jsem provedla pomocí Smirnov-Kolmogorova testu. Pro všechny příznaky a pacienty byla nulová hypotéza normality na hladině významnosti 5 % zamítnuta.

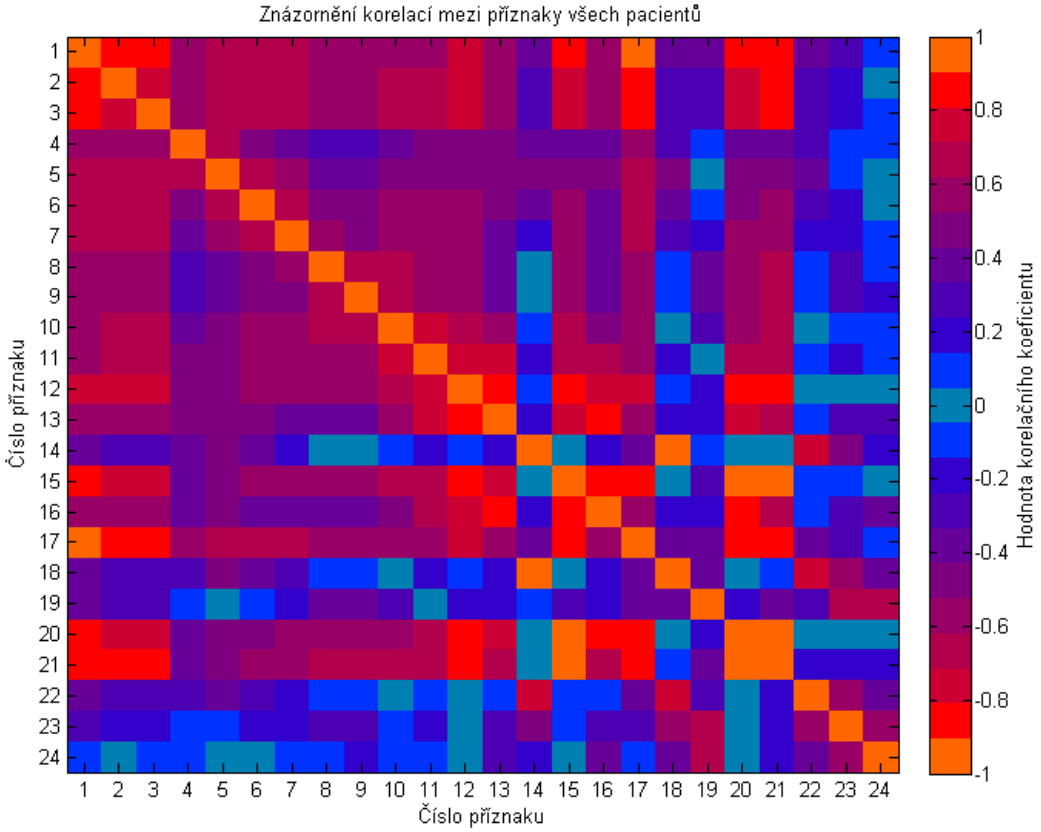
Hodnoty příznaků tedy nemají normální rozložení, což jsem si ověřila i graficky, když jsem si pro každý příznak vykreslila histogram, který zobrazuje četnost hodnot (tedy odhad rozdělení). Z těchto histogramů bylo zřejmé, že rozložení hodnot příznaků se normálnímu rozdělení vůbec nepodobá. Příkladem může být obrázek 3.10, který znázorňuje rozložení hodnot příznaku mobility u pacienta č. 3.



Obrázek 3.10: Histogram normovaných hodnot mobility (pacient č. 3)

3.4 Korelace mezi příznaky

Jak bylo zmíněno dříve (kapitola 2.6), pro každou dvojici příznaků jsem spočítala Spearmanův korelační koeficient a koeficienty všech pacientů jsem zprůměrovala. Podle čísel příznaků (viz tabulka 2.3) jsem je potom mohla uspořádat do matice a v té následně barevně znázornit (obrázek 3.11) míru korelace podle hodnoty korelačního koeficientu.



Obrázek 3.11: Znázornění korelací mezi jednotlivými příznaky.

Na první pohled můžeme z obrázku vidět, že na diagonále je vždy znázorněna korelace příznaku sama se sebou, proto je zde korelační koeficient roven jedné. Díky barevnému zobrazení můžeme také hned pozorovat, jak některé příznaky korelují s ostatními jen velmi málo (takové je potřeba zachovat) a jiné naopak mají velkou míru korelace s mnoha ostatními (z nich lze ponechat pouze některé, které budou kombinací ostatních).

Z obrázku sice není zřejmé znaménko koeficientu, já jsem ovšem ke koeficientům přistupovala jako k absolutní hodnotě, jelikož i výrazná nepřímo úměrná závislost je důvodem k redukci příznaku. Všechny korelace, se kterými jsem při redukci nakonec pracovala (tj. korelační koeficient o absolutní hodnotě vyšší než 0,6 viz. následující kapitola 3.5), byly ovšem kladné.

3.5 Redukce příznaků

Způsob redukce příznaků je popsán v kapitole 2.6. Limit pro korelační koeficient (z dvojic příznaků s vyšší korelací jsem vždy odebírala) jsem volila tak, aby zůstalo 20, 15, 10 a nakonec 5 příznaků. V tabulce 3.1 jsou uvedeny jednotlivé použité limity korelačního koeficientu, příznaky, které jsem oproti předchozímu kroku odebrala, počet příznaků, které po odebrání zůstaly pro klasifikaci a nakonec výčet těchto zbylých příznaků.

Tabulka 3.1: Redukce příznaků

Limit	Odebrané příznaky	Počet zbylých příznaků	Příznaky pro klasifikaci
1,00	—	24	1 – 24
0,90	17, 18, 20, 21	20	1 – 16, 19, 22 – 24
0,85	2, 3, 12, 13, 15	15	1, 4 – 11, 14, 16, 19, 22 – 24
0,67	4, 6, 9, 11, 22	10	1, 5, 7, 8, 10, 14, 16, 19, 23, 24
0,60	7, 8, 16, 23, 24	5	1, 5, 10, 14, 19

Pokud dále v práci uvádím, že byla klasifikace provedena podle např. 10 příznaků, konkrétně použité příznaky jsou vždy ty, uvedené v této tabulce.

3.6 Výsledky klasifikace

Podle vypočítaných korelací a provedené redukce jsem následně realizovala klasifikaci etalonových segmentů záznamu algoritmem k-NN. Tu jsem prováděla u každého pacienta celkem pětkrát, pokaždé s jiným počtem příznaků — 24, 20, 15, 10 a 5.

Každou klasifikaci jsem po provedení křížové validace vyhodnotila průměrnou klasifikační maticí. Jako příklad uvádím klasifikační matice pro pacienta č. 3 (tabulky 3.2 až 3.6). Výsledky klasifikace podle 15 a 10 příznaků jsou rovněž znázorněny na obrázcích 3.12 a 3.13 promítnuté do 2D prostoru podle příznaků .

Tabulka 3.2: Výsledky klasifikace podle 24 příznaků (pacient č. 3)

Klasifikace podle 24 příznaků		Správné třídy	
		epi	neepi
Výsledek klasifikace	epi	31,2	2,1
	neepi	0,8	45,9

Tabulka 3.3: Výsledky klasifikace podle 20 příznaků (pacient č. 3)

Klasifikace podle 20 příznaků		Správné třídy	
		epi	neepi
Výsledek klasifikace	epi	31,5	2,6
	neepi	0,5	45,4

Tabulka 3.4: Výsledky klasifikace podle 15 příznaků (pacient č. 3)

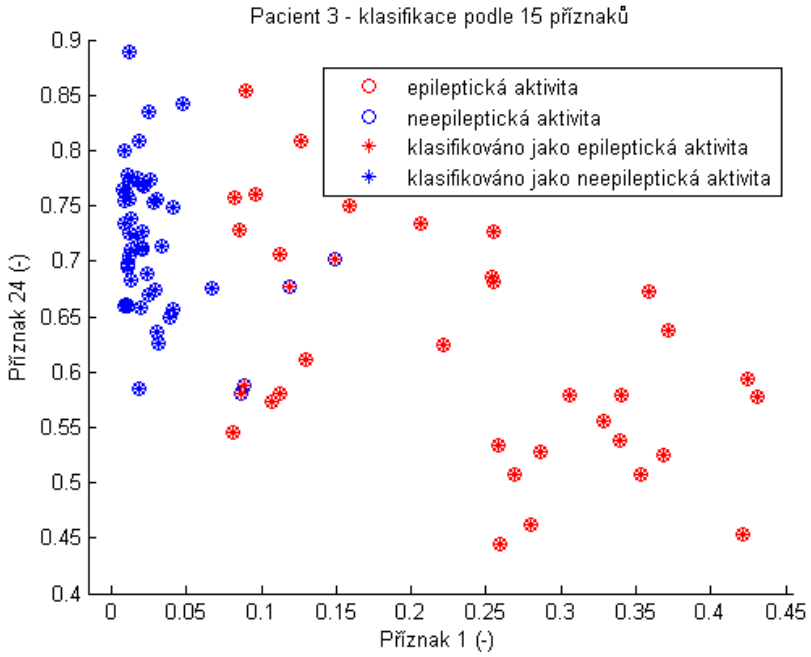
Klasifikace podle 15 příznaků		Správné třídy	
		epi	neepi
Výsledek klasifikace	epi	31,2	2,6
	neepi	0,9	45,4

Tabulka 3.5: Výsledky klasifikace podle 10 příznaků (pacient č. 3)

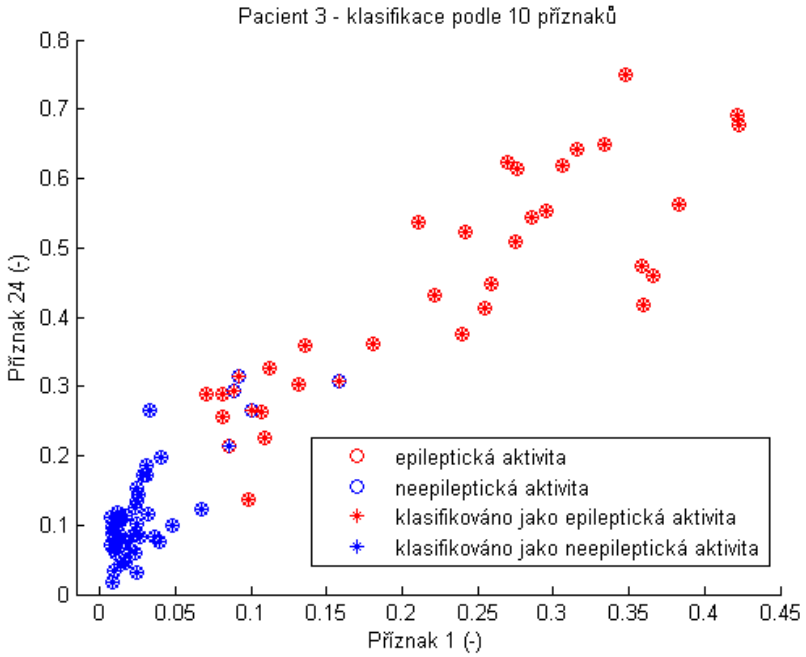
Klasifikace podle 10 příznaků		Správné třídy	
		epi	neepi
Výsledek klasifikace	epi	31,4	2,6
	neepi	0,6	45,4

Tabulka 3.6: Výsledky klasifikace podle 5 příznaků (pacient č. 3)

Klasifikace podle 5 příznaků		Správné třídy	
		epi	neepi
Výsledek klasifikace	epi	31,9	3,9
	neepi	0,1	44,1



Obrázek 3.12: Výsledky klasifikace podle 15 příznaků (pacient č. 3). Barva hvězdičky znázorňuje výsledek klasifikace segmentu a kruh okolo udává správné zařazení segmentu dle lékaře. Příznak 1 je směrodatná odchylka, příznak 24 je inflexní bod.



Obrázek 3.13: Výsledky klasifikace podle 10 příznaků (pacient č. 3). Barva hvězdičky znázorňuje výsledek klasifikace segmentu a kruh okolo udává správné zařazení segmentu.

Celkové výsledky klasifikace u všech pacientů výstižně popisují vypočítané parametry: senzitivita, specificita a přesnost, které jsou uvedeny v tabulkách 3.7, 3.8 a 3.9.

Tabulka 3.7: Senzitivita klasifikace

Počet příznaků	24	20	15	10	5
pacient 1	100,0 %	100,0 %	100,0 %	100,0 %	99,1 %
pacient 2	80,1 %	79,9 %	78,9 %	82,6 %	82,2 %
pacient 3	97,6 %	98,4 %	97,4 %	98,1 %	99,6 %
pacient 4	100,0 %	100,0 %	100,0 %	100,0 %	100,0 %
pacient 5	88,5 %	91,9 %	91,5 %	92,3 %	92,3 %
průměr	93,2 %	94,0 %	93,6 %	94,6 %	94,6 %

Tabulka 3.8: Specificita klasifikace

Počet příznaků	24	20	15	10	5
pacient 1	100,0 %	100,0 %	100,0 %	100,0 %	100,0 %
pacient 2	97,4 %	97,3 %	97,3 %	96,6 %	95,2 %
pacient 3	95,6 %	94,5 %	94,5 %	94,5 %	92,0 %
pacient 4	100,0 %	100,0 %	100,0 %	100,0 %	100,0 %
pacient 5	92,5 %	91,0 %	92,4 %	91,0 %	91,8 %
průměr	97,1 %	96,6 %	96,8 %	96,4 %	95,8 %

Tabulka 3.9: Přesnost klasifikace

Počet příznaků	24	20	15	10	5
pacient 1	100,0 %	100,0 %	100,0 %	100,0 %	99,6 %
pacient 2	90,5 %	90,3 %	89,9 %	91,0 %	90,0 %
pacient 3	96,4 %	96,1 %	95,7 %	95,9 %	95,0 %
pacient 4	100,0 %	100,0 %	100,0 %	100,0 %	100,0 %
pacient 5	90,9 %	91,4 %	91,4 %	91,5 %	92,0 %
průměr	95,6 %	95,6 %	95,4 %	95,7 %	95,3 %

4 Diskuze

Data pěti pacientů, se kterými jsem při analýze a klasifikaci pracovala, byla různorodá, lišila se jak délkou záznamu, tak obsaženými třídami nebo separabilitou tříd, jak je ukázáno na dvou 3D obrázcích v kapitole 3.1. Některé záznamy obsahovaly impulsní elektrodové artefakty, jiné ne.

Pacient č. 2 měl např. v záznamu velmi vysoké tyto impulsní elektrodové artefakty, což má zákonitě vliv na normalizaci příznaků, protože vysoké hodnoty příznaků těchto částí záznamu způsobí při mnou použitém způsobu normalizace snížení amplitudy všech ostatních příznaků. Snížení amplitudy se potvrdilo porovnáním mediánů či průměrů normovaných příznaků tohoto pacienta s ostatními. Přesnost klasifikace v tomto případě byla sice nižší než u ostatních pacientů, ale překvapivě ne příliš výrazně (maximálně 2 %).

Příznaky etalonových segmentů měly u každého pacienta jiné průměrné hodnoty a bylo obtížné vysledovat v nich nějakou zákonitost. Z hlediska odlišení epileptické aktivity od ostatní aktivity lze říci, že nejvýrazněji odlišné hodnoty měly příznaky 1 – 6, 17 a 24. Toto ovšem platí pouze omezeně pro pacienta č. 2, kvůli výše uvedeným důvodům. Obecně se od ostatní aktivity více liší příznaky vysoké epileptické aktivity než nízké epileptické aktivity.

Příznaky jsem analyzovala i z hlediska normálního rozdělení. U všech příznaků všech pacientů jsem otestovala, zda pocházejí z normálního rozdělení a u všech bylo toto zamítnuto na hladině významnosti 5 %. Je tedy nutné s tímto počítat při dalším zpracování a nepoužívat metody, které normální rozdělení dat vyžadují jako podmínku.

Analýza příznaků pomocí korelací ukázala zajímavé vztahy. Bylo z ní na první pohled zřejmé, které příznaky jsou jedinečné a které naopak nesou obdobnou informaci o segmentu. Až zarážející byly např. příznaky 1 (směrodatná odchylka) a 17 (Hjortův parametr mobilita) s průměrnou korelací u pacientů přes 0,99. Poměrně malou míru korelace s většinou ostatních i mezi sebou měly všechny příznaky obsahující hodnoty rychlé Fou-

rierovy transformace v různých frekvenčních pásmech (příznaky 4 – 11). A nízkou korelaci s ostatními měly i příznaky 19 (Hjortův parametr aktivita), 23 (dominantní spektrální vrchol) a 24 (inflexní bod).

Po korelační analýze bylo možné vhodně snížit počet 24 příznaků až na pět vybraných příznaků tak, aby nebyla významně ovlivněna klasifikace. Těchto pět posledních příznaků byly: směrodatná odchylka, FFT v pásmu 2 – 3,5 Hz (Delta2) a v pásmu 18 – 29 Hz (Sigma), střední frekvence EEG aktivity a Hjortův parametr aktivita. Tři z příznaků tedy charakterizují segment ve frekvenční oblasti a další dva popisují amplitudu v časové oblasti.

Vzhledem ke způsobu výběru příznaků měl počet příznaků na klasifikaci segmentů téměř zanedbatelný vliv. Lze očekávat, že pokud odebereme příznak, který není korelován s ostataními, dojde ke snížení přesnosti, senzitivity a specificity. Tedy dokud hodnoty těchto ukazatelů výrazně neklesají, pak jsem odebrala vhodně příznaky, které silně korelovaly s jinými a zůstala tak zachována většina podstatných informací o segmentu potřebných pro odlišení epileptické aktivity od ostatních segmentů.

U některých pacientů se může zdát, že hodnoty ukazatelů klesají, u jiných ovšem není výjimkou, že klasifikace s menším počtem příznaků dopadla mírně příznivěji, než ta s větším počtem. Rozdíly v řádu desetin procent mohou být ale způsobeny náhodným výběrem segmentů pro klasifikaci a následnou křížovou korelací.

Jelikož nemají všechny příznaky rovnocennou vypovídací hodnotu, tj. nenesou stejně významnou informaci a některé nesou opakující se informace, nebyl by výsledek klasifikace rozhodně stejný, pokud by záviselo pouze na počtu příznaků, ale konkrétní příznaky by byly vybírány nějakým náhodným způsobem.

Senzitivita klasifikace se v průměru pohybovala okolo 94 %. Dle mého soudu je důležitá co nejvyšší hodnota senzitivity, jelikož to lékaři umožní bez náročného hledání zkontrolovat epileptické segmenty a nižší hodnota specificity není tak velký problém, jelikož právě segmenty vyskytující se na hranici by měl lékař také zkontrolovat. Příznivá byla právě ale

i specificita průměrně okolo 96 % a průměrná přesnost přes 95 %. Pro porovnání uveďme v úvodu zmíněný článek [9], kde je provedena klasifikace rovněž i pomocí algoritmu k-NN, byla senzitivita klasifikace 87.4 %, specificita 97.6 % a přesnost 96.1 %.

Musí se ovšem připomenout, že klasifikace byla prováděna na etalonových (prototypových) segmentech, tedy segmentech, které typicky náleží do určité třídy a klasifikace celého záznamu se segmenty s nejasnou příslušností ke třídě by takovouto přesnost neměla.

V blízké budoucnosti bychom chtěly společně s mojí vedoucí celou analýzu příznaků a jejich redukci pomocí korelací rozšířit na alespoň 10 pacientů a výsledky publikovat. Také bychom chtěly tento způsob důkladně porovnat s běžnými metodami redukce příznakového prostoru, jakými jsou PCA a ICA.

5 Závěr

V programovém prostředí MATLAB® jsem implementovala algoritmus k-NN pro klasifikaci segmentů EEG záznamu. Implementaci jsem provedla na základě parametrů zjištěných v literatuře. V programu MATLAB® jsem vytvořila funkci `kNN_vCyklu2`, jejíž detaily jsou popsány v kapitole 2.7 a celý kód je v příloze A. Funkčnost jsem otestovala nejprve vizuálně s použitím výchozí sady dat v programu MATLAB® a po té přímo na reálných datech při odlišování tříd epileptické aktivity.

Analyzovala jsem příznaky využívané pro klasifikaci EEG segmentů a k detekci epileptické aktivity v EEG záznamu. Zabývala jsem se především 24 příznaky počítanými programem Wavefinder, jelikož počítá většinu parametrů, které jsem nalezla v literatuře. U každého příznaku jsem analyzovala, jakých hodnot nabývá v závislosti na tom, do jaké patří třídy a také, jakému pacientovi patří. Pro tento účel jsem vypočetla pro každou třídu a pacienta průměrnou hodnotu, medián a směrodatnou odchylku každého příznaku. Dále jsem testem normality prokázala, že příznaky nemají normální rozdělení.

Na datech z reálného EEG záznamu (příznaky segmentů 24 kanálového EEG záznamu od celkem pěti pacientů) jsem provedla klasifikaci segmentů. Pro klasifikaci jsem použila nejprve 24 příznaků a následně po redukci na základě zjištěných korelací mezi příznaky jsem provedla klasifikaci jen s 20, 15, 10 a dokonce 5 příznaky. Výsledky klasifikace jsem vyhodnotila pomocí klasifikačních (konfuzních) matic a z nich jsem dále určila senzitivitu, specificitu a přesnost klasifikace. Přesnost klasifikace neklesla pod 90 %, senzitivita se v průměru pohybovala okolo 94 % a specificita okolo 96 %.

Použitá literatura

[1] Krajča, V.; Mohylová, J.: *Číslicové zpracování neurofyziologických signálů*. Praha: České vysoké učení technické v Praze, první vydání, 2011, ISBN 9788001047217.

[2] Schaabova, H.; Krajca, V.; Sedlmajerova, V.; aj.: Supervised learning used in automatic EEG graphoelements classification. In *2015 E-Health and Bioengineering Conference (EHB)*, IEEE, 2015, ISBN 978-1-4673-7544-3, s. 1–4, doi:10.1109/EHB.2015.7391470, [cit. 2017-11-14]. Dostupné z: `http://ieeexplore.ieee.org/document/7391470/`

[3] Djordjevic, V.; Reljin, N.; Gerla, V.; aj.: Feature extraction and classification of EEG sleep recordings in newborns. In *2009 9th International Conference on Information Technology and Applications in Biomedicine*, IEEE, 2009, ISBN 978-1-4244-5379-5, s. 1–4, doi:10.1109/ITAB.2009.5394439, [cit. 2017-11-14]. Dostupné z: `http://ieeexplore.ieee.org/document/5394439/`

[4] Exarchos, T.; Tzallas, A.; Fotiadis, D.; aj.: A Data Mining Based Approach for the EEG Transient Event Detection and Classification. In *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, IEEE, 2005, ISBN 0-7695-2355-2, s. 35–40, doi:10.1109/CBMS.2005.7, [cit. 2017-11-15]. Dostupné z: `http://ieeexplore.ieee.org/document/1467664/`

[5] Rus, I. D.; Marc, P.; Dinsoreanu, M.; aj.: Classification of EEG signals in an object recognition task. In *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, IEEE, 2017, ISBN 978-1-5386-3368-7, s. 391–395, doi:10.1109/ICCP.2017.8117036, [cit. 2017-12-26]. Dostupné z: `http://ieeexplore.ieee.org/document/8117036/`

[6] Lotte, F.; Congedo, M.; L'ecuyer, A.; aj.: A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, ročník 4, 2007.

[7] Mehmood, R. M.; Lee, H. J.: Emotion classification of EEG brain signal using SVM and KNN. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, ISBN 978-1-4799-7079-7, s. 1–5, doi:10.1109/ICMEW.2015.7169786, [cit. 2017-12-26]. Dostupné z: `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7169786`

[8] THE MATHWORKS, I.: MATLAB Documentation: knnsearch. [cit. 2017-11-15]. Dostupné z: `https://www.mathworks.com/help/stats/knnsearch.html?requestedDomain=www.mathworks.com`

[9] Poorna, S. S.; Baba, P. M. V. D. S.; Ramya, G. L.; aj.: Classification of EEG based control using ANN and KNN — A comparison. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, IEEE, 2016, ISBN 978-1-5090-0612-0, s. 1–6, doi:10.1109/ICCIC.2016.7919524, [cit. 2017-12-26]. Dostupné z: `http://ieeexplore.ieee.org/document/7919524/`

[10] Isa, R. M.; Pasya, I.; Taib, M. N.; aj.: EEG brainwave behaviour due to RF Exposure using kNN classification. In *2013 IEEE 3rd International Conference on System Engineering and Technology*, IEEE, 2013, ISBN 978-1-4799-1030-4, s. 385–388, doi:10.1109/ICSEngT.2013.6650205, [cit. 2017-12-26]. Dostupné z: `http://ieeexplore.ieee.org/document/6650205/`

[11] Chatterjee, S.; Choudhury, N. R.; Bose, R.: Detection of epileptic seizure and seizure-free EEG signals employing generalised S -transform. *IET Science, Measurement & Technology*, 2017: s. 847–855, ISSN 1751-8822, doi:10.1049/iet-smt.2016.0443, [cit. 2017-12-26. Dostupné z: `http://digital-library.theiet.org/content/journals/10.1049/iet-smt.2016.0443`

[12] Zainuddin, A. Z. A.; Lee, K. Y.; Mansor, W.; aj.: Optimized KNN classify rule for EEG based differentiation between capable dyslexic and normal children. In *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, IEEE, 2016, ISBN 978-1-4673-7791-1, s. 685–688, doi:10.1109/IECBES.2016.7843537, [cit. 2017-12-26]. Dostupné z: `http://ieeexplore.ieee.org/document/7843537/`

[13] Zhang, T.; Chen, W.: LMD Based Features for the Automatic Seizure Detection of EEG Signals Using SVM. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, ročník 25, č. 8, 2017: s. 1100–1108, ISSN 1534-4320, doi: 10.1109/TNSRE.2016.2611601, [cit. 2017-12-26]. Dostupné z: `http://ieeexplore.ieee.org/document/7572197/`

[14] Seraj, E.; Karimzadeh, F.: Improved detection rate in motor imagery based BCI systems using combination of robust analytic phase and envelope features. In *2017 Iranian Conference on Electrical Engineering (ICEE)*, IEEE, 2017, ISBN 978-1-5090-5963-8, s. 24–28, doi:10.1109/IranianCEE.2017.7985458, [cit. 2017-12-26]. Dostupné z: `http://ieeexplore.ieee.org/document/7985458/`

[15] Krajča, V.; Petránek, S.; Pietielä, T.; aj.: Wave-finder: a new system for an automatic processing of long-term EEG recordings. *Quantitative EEG Analysis-Clinical Utility and New Methods*, 1993: str. 103–106, [cit. 2018-01-03].

[16] 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, [cit. 2017-11-14]. Dostupné z: `http://link.springer.com/10.1007/978-3-319-44188-7_7`

[17] Krajca, V.; Piorecka, V.; Schaabova, H.; aj.: Detection of sleep stages in neonatal EEG records. In *EMBEC & NBC 2017*, Singapore: Springer Singapore, 2017, ISBN 978-981-10-5121-0, s. 250–253, doi:10.1007/978-981-10-5122-7_63, [cit. 2017-12-27]. Dostupné z: `http://link.springer.com/10.1007/978-981-10-5122-7_63`

[18] THE MATHWORKS, I.: MATLAB Documentation: kstest. [cit. 2018-01-08]. Dostupné z: `https://www.mathworks.com/help/stats/kstest.html?requestedDomain=www.mathworks.com`

[19] Dhongade, D. V.; Rao, T.: Classification of sleep disorders based on EEG signals by using feature extraction techniques with KNN classifier. In *2017 International*

Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT), IEEE, 2017, ISBN 978-1-5090-5778-8, s. 1–5, doi:10.1109/IGEHT.2017.8093976, [cit. 2018-05-11]. Dostupné z: `http://ieeexplore.ieee.org/document/8093976/`

[20] THE MATHWORKS, I.: MATLAB Documentation: corr. [cit. 2018-03-28]. Dostupné z: `https://www.mathworks.com/help/stats/corr.html?requestedDomain=true#mw_e263787a-88e2-4603-8e9b-a3d5600c2a19`

[21] Holčík, J.; Komenda, M.: *Matematická biologie*. Brno: Masarykova univerzita, první vydání, 2015, ISBN 978-80-210-8095-9, [cit. 2018-04-17]. Dostupné z: `http://portal.matematickabiologie.cz/`

Příloha A: Kód funkce kNN_vCyklu2

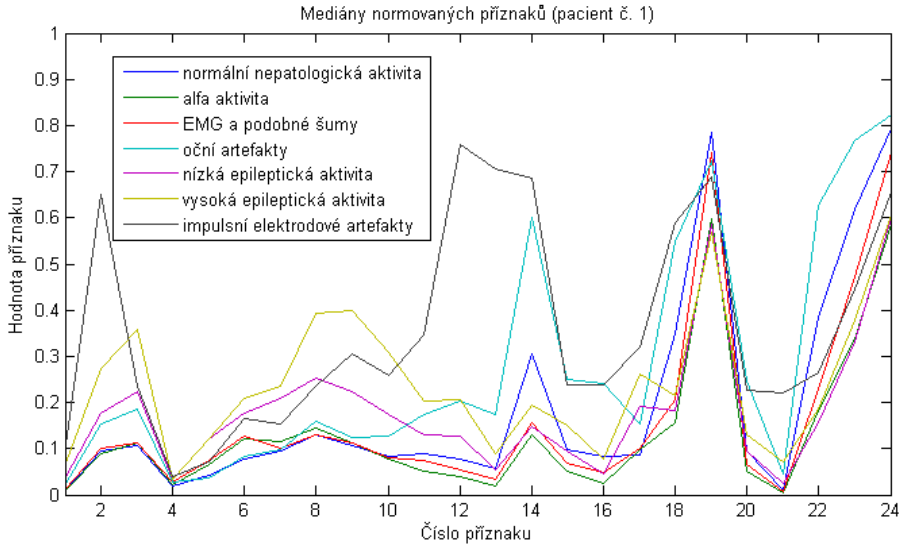
```
function [ classNew ] = kNN_vCyklu2(x, class, Pvec, k)

% KLASIFIKACE K-NN ALGORITMEM V CYKLU %
% INPUTS:
% x ... sada příznaků pro segmenty v trénovací množině (každý řádek = 1 segment)
% class ... vektor obsahující třídy segmentů z proměnné 'x'
% Pvec ... matice příznaků pro segmenty v testovací množině
% k ... argument funkce k-NN (kolik nejbližších sousedů uvažuji pro klasifikaci)
% OUTPUTS:
% classNew ... vektor obsahující třídu každého segmentu z testovací množiny
%
% Author: Barbora Balcarova
% Date: 2018-04-03, version: 2

[a1, ~] = size(Pvec); % zjistím počet nových segmentů
classtype = unique(class); % vypíše třídy, které se vyskytují v trénovací množině
Nclass = numel(classtype); % počet tříd v 'classtype'
for i1 = 1:a1 % cyklus se zopakuje pro každý nový segment
 P = Pvec(i1,:);% výběr segmentu, se kterým se pracuje v aktuálním cyklu
 % POČÍTÁNÍ VZDÁLENOSTI
dist = pdist2(x,P,'euclidean'); % 'dist' = vektor vzdáleností segmentu P od všech
ostatních v trénovací množině
% HLEDÁNÍ 'k' NEJBLIŽŠÍCH SOUSEDŮ
 [~, p] = sort(dist); % seřadíme vzdálenosti od nejmenší po největší, v proměnné
'p' je pořadí indexů
 pN = p(1:k); % do 'pN' vybereme indexy pouze 'k' nejbližších sousedů
% TŘÍDY NEJBLIŽŠÍCH SOUSEDŮ
```

```
classNN = class(pN,:); % vybereme třídy 'k' nejbližších sousedů
 classcount = zeros(Nclass ,1);
for i = 1:numel(classtype)
classcount(i) = sum(classNN==classtype(i)); % spočteme počet nejbližších sousedů
z každé třídy (pozice ve vektoru odpovídá číslu třídy)
 end
% URČENÍ TŘÍDY NOVÉHO SEGMENTU
% zjištění jestli je mezi nejbližšími sousedy nejvíce pouze z jedné třídy nebo
z více -> pak je třeba rozhodnout, kam tedy segment zařadit
 maxim = max(classcount); % kolik je maximum z jedné třídy
 maxclass = sum(classcount==maxim); % z kolika tříd je 'maxim'
 if maxclass==1
[~, classI] = max(classcount); % 'classI' je index třídy (ve vektoru 'classtype'),
ve které je největší počet nejbližších sousedů
 classP = classtype(classI);
else % tato varianta nastane, pokud 'maxclass' neni 1 tj. pokud nebude pouze
jedna maximální třída mezi nejbližšími sousedy
 maxposition = zeros(Nclass,1);
for i3 = 1:Nclass
maxposition(i3) = classcount(i3)== maxim; % pokud třída patří mezi ty s maximem,
tak na její pozici bude 1
 end
a = dist(pN); % 'a' je vektor vzdáleností 'k' nejbližších sousedů
 for i2=1:Nclass % cyklus projde všechny třídy
 if maxposition(i2)==1 % zjistíme, jestli třída patří mezi ty s maximem
 distNN(i2) = sum(a(classNN==classtype(i2))); % spočteme součet vzdáleností
segmentů z teto třídy, v 'distNN' jsou tyto součty vzdáleností pro každou max
třídu
 end
end
```

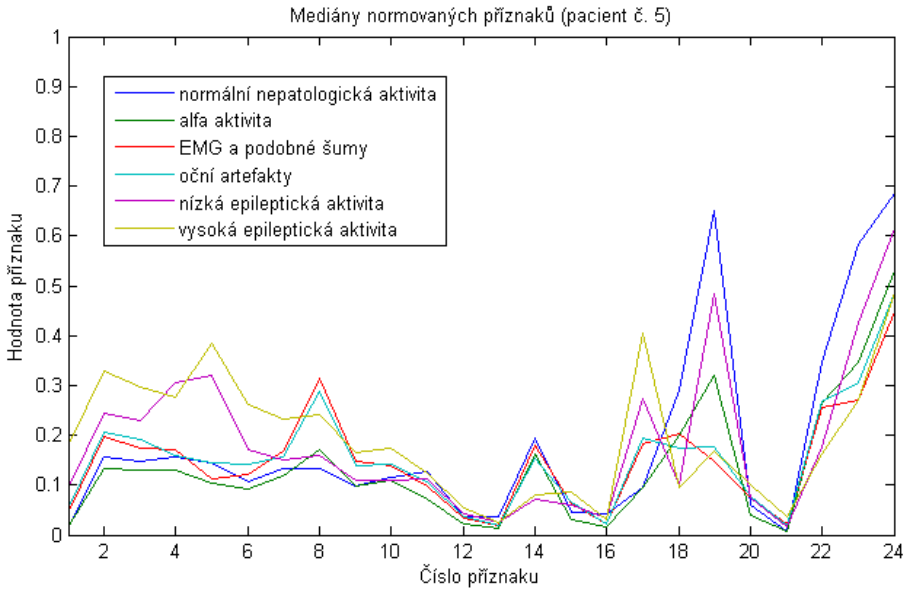
```
[∼, classI] = min(distNN(distNN>0)); % 'classI' je index třídy (ve vektoru
help), ve které je největší počet nejbližších sousedů
 help = classtype(distNN>0); % vybereme třídy, kde 'distNN' není nula
 classP = help(classI) ;
end
% PŘIŘAZENÍ DO URČENÉ TŘÍDY
 x = [x; P];
class = [class; classP];
end
classNew = class(end-a1+1:end,:); % výběr tříd odpovídajících testovací množině
 end
```

Příloha B: Hodnoty příznaků

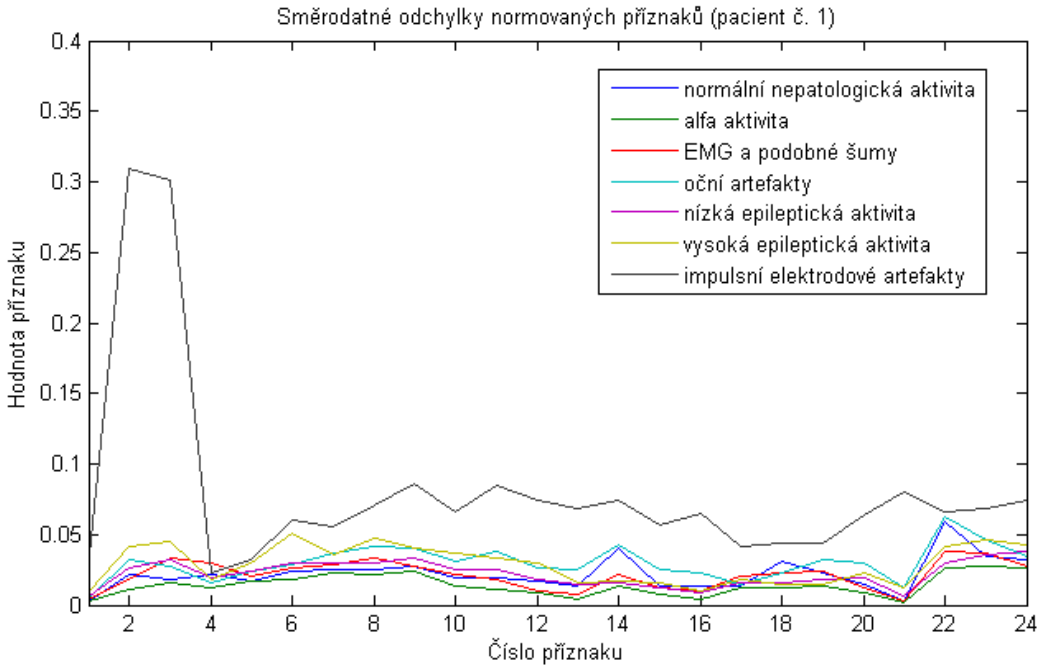
V této příloze jsou uvedeny grafy mediánů a směrodatných odchylek příznaků pro pacienty, kteří nebyli uvedeni ve výsledcích práce.



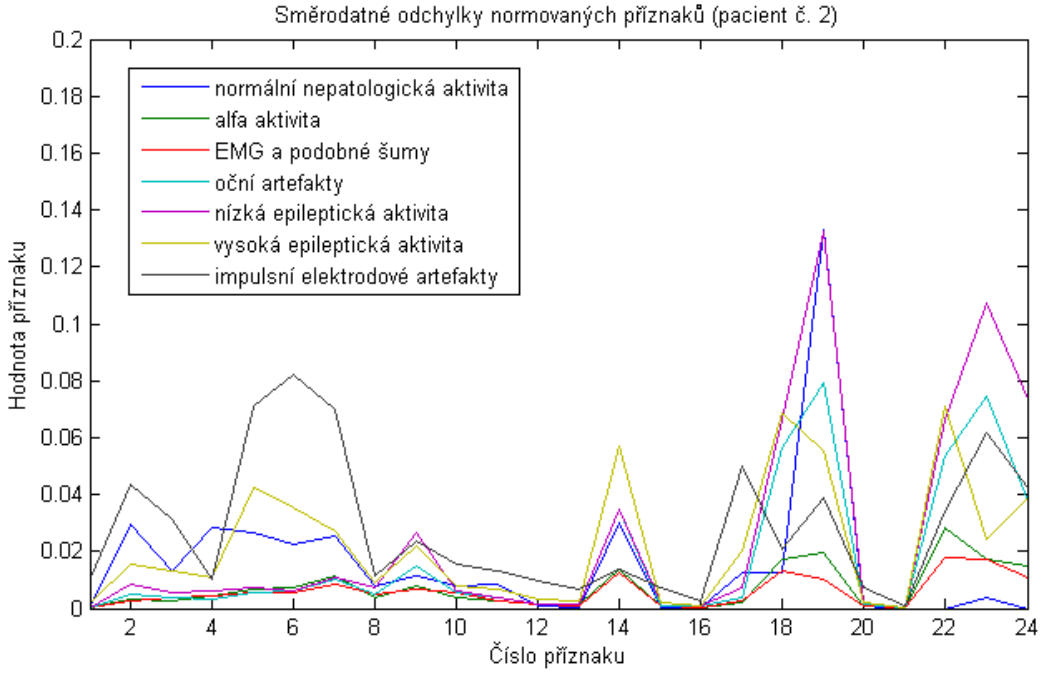
Obrázek B.1: Mediány normovaných příznaků pro jednotlivé třídy (pacient č. 1)



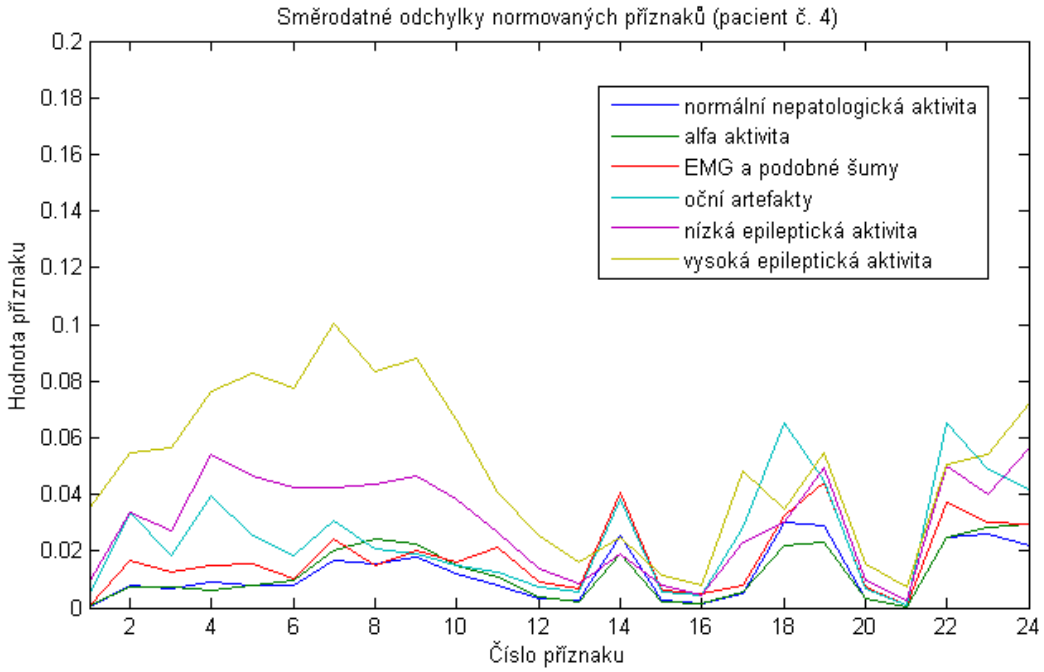
Obrázek B.2: Mediány normovaných příznaků pro jednotlivé třídy (pacient č.5)



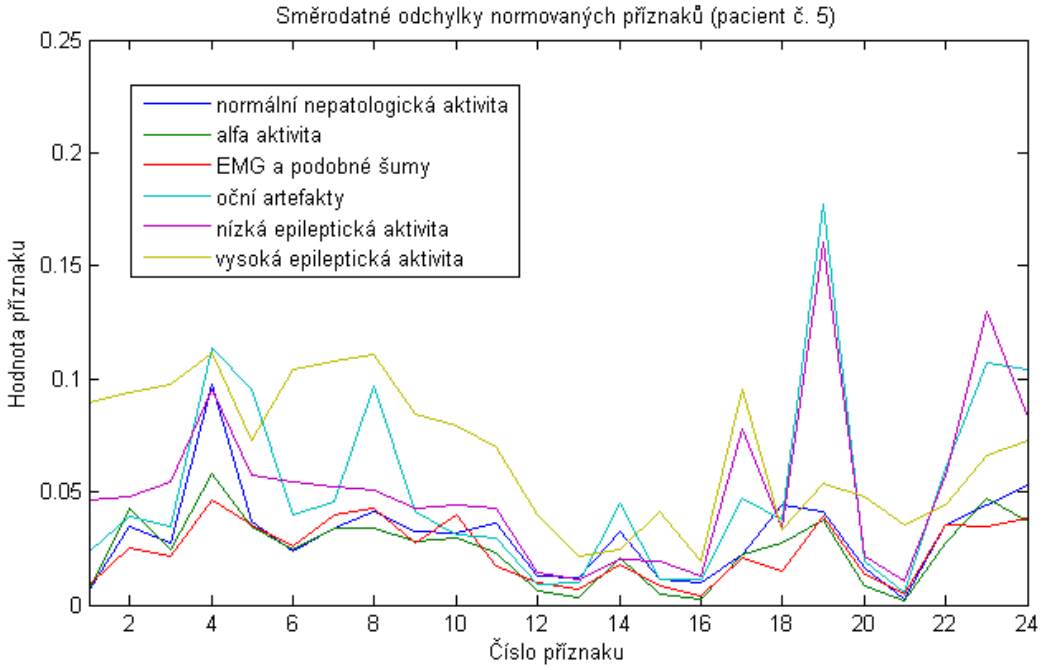
Obrázek B.3: Směrodatné odchylky normovaných příznaků (pacient č.1)



Obrázek B.4: Směrodatné odchylky normovaných příznaků (pacient č.2)



Obrázek B.5: Směrodatné odchylky normovaných příznaků (pacient č.4)



Obrázek B.6: Směrodatné odchylky normovaných příznaků (pacient č.5)

Příloha C: 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
- skripty vytvořených funkcí v programu MATLAB® ve složce
- přehled statistických charakteristik příznaků a výsledků klasifikace pro každého z pěti pacientů (ve formátu xlsx)