

Slovenská technická univerzita v Bratislave  
Fakulta informatiky a informačných technológií

FIIT-XXXX-86077

**Bc. Timotej Zaťko**

**Uplatnenie interpretovateľnosti a  
vysvetliteľnosti neurónových sietí pri  
vyhodnocovaní medicínskych obrazových  
dát**

Priebežná správa o riešení DP1

Študijný program: Inteligentné softvérové systémy

Študijný odbor: 18. Informatika

Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového  
inžinierstva (FIIT)

Vedúci práce: Ing. Martin Tamajka

máj 2020





Čestne vyhlasujem, že som túto prácu vypracoval samostatne, na základe konzultácií a s použitím uvedenej literatúry.

V Bratislave, 6. máj 2020

Timotej Zaťko



# **Anotácia**

Slovenská technická univerzita v Bratislave

Fakulta informatiky a informačných technológií

Študijný program: Inteligentné softvérové systémy

Autor: Bc. Timotej Zaťko

Diplomová práca: Uplatnenie interpretatívnosti a vysvetliteľnosti neurónových sietí pri vyhodnocovaní medicínskych obrazových dát

Vedúci diplomového projektu: Ing. Martin Tamajka

máj 2019

TODO:



# **Annotation**

Slovak University of Technology Bratislava  
Faculty of Informatics and Information Technologies  
Degree Course: Intelligent Software Systems

Author: Bc. Timotej Zatko

Diploma's Thesis: Application of interpretability and explainability of neural networks in the evaluation of medical images

Supervisor: Ing. Martin Tamajka

2019, May

TODO:



## **Pod'akovanie**

Ďakujem môjmu školiteľovi Ing. Martinovi Tamajkovi za odbornú pomoc a vedenie pri tvorbe tejto práce.



# Obsah

<b>1</b>	<b>Úvod</b>	<b>3</b>
<b>2</b>	<b>Analýza</b>	<b>5</b>
2.1	Alzheimerova choroba . . . . .	5
2.1.1	Diagnostika Alzheimerovej choroby . . . . .	6
2.1.2	Biologické ukazovatele . . . . .	6
2.1.3	Obrazové a rádiologické ukazovatele . . . . .	7
2.2	Neurónové siete . . . . .	8
2.2.1	Interpretovanie neurónovej siete . . . . .	9
2.2.2	Vysvetľovanie predikcie neurónovej siete . . . . .	12
2.2.2.1	Analýza senzitivity . . . . .	12
2.2.2.2	LRP (angl. layer-wiser relevance propagation) . . .	13
2.2.2.3	RISE - Randomized Input Sampling for Explanation	15
2.3	Využitie neurónových sietí pri odhalovaní Alzheimerovej choroby . .	17
2.3.1	Vysvetľovanie rozhodnutí neurónových sietí detegujúcich Alzheimerovu chorobu . . . . .	18
2.4	Zhrnutie . . . . .	19
<b>3</b>	<b>Návrh riešenia</b>	<b>21</b>
3.1	Zhrnutie . . . . .	21
<b>Literatúra</b>		<b>23</b>

## Dodatok A Plán práce

### A.1 Zimný semester

## Obsah

---

### A.2 Letný semester

# 1. Úvod

Umelá inteligencia sa už dávno stala súčasťou nášho každodenného života. Prichádzame s ňou do kontaktu neustále, keď odomykáme telefón vlastnou tvárou alebo keď pomocou prekladača prekladáme text to iného jazyka. Jej využitie je tiež rozšírené v oblasti medicíny, kde má potenciál zachraňovať životy. Využíva sa pri výrobe liekov, monitorovaní zdravia, analýze zdravotných plánov, chirurgických zákrokov a aj pri odhaľovaní chorôb [1]. Práve pri odhaľovaní chorôb sa častokrát využívajú hlboké neurónové siete, a to napríklad pri detekcii rakoviny kože, rakoviny pľúc alebo Alzheimerovej choroby z obrazových dát.

Neurónovým sieťam sa už podarilo dosiahnuť také dobré výsledky, že sú porovnatelné s expertmi v medicínskej oblasti. Ich problémom však je, že sa správajú ako "čierna skrinka", čo v oblasti medicíny nie je žiadúce. Preto je nevyhnutné, aby boli rozhodnutia neurónovej siete interpretovateľné a pacient s lekárom vedeli, na základe čoho sa neúronová sieť rozhodla. Lekári by si mali svoje rozhodnutia vedieť obhájiť. Aby sa teda neurónové siete mohli stať bežným pomocníkom lekárov, je vysvetliteľnosť ich rozhodnutí dôležitá.



## **2. Analýza**

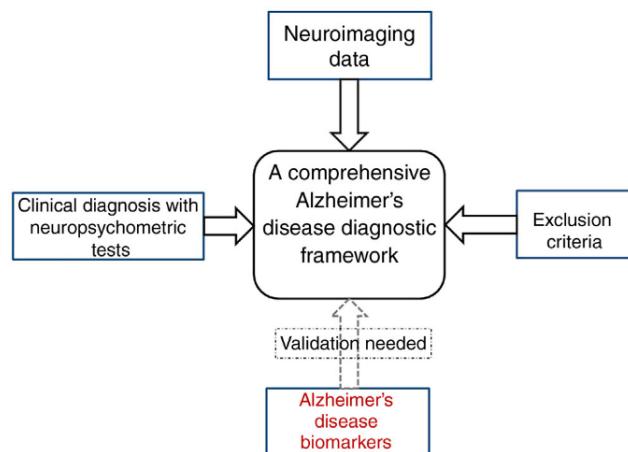
### **2.1 Alzheimerova choroba**

Alzheimerova choroba je najčastejšou príčinou demencie. Prvotné príznaky tejto choroby sú zhoršenie pamäti, zabúdanie nedávnych udalostí, mien, neschopnosť rozoznávať známe miesta či orientovať sa v čase [2]. Jej priebeh sa vyznačuje postupným poklesom kognitívnych funkcií, postupným zhoršením pamäte, myslenia, rozprávania a schopnosti učenia sa [3]. Najčastejšie sa vyskytuje u ľudí starších ako 65 rokov, s pravdepodobnosťou výskytu až 50% po dovršení 85 rokov života [3]. S narastajúcim vekom človeka sa zvyšuje pravdepodobnosť ochorenia. Pravdepodobnosť ochorenia zvyšujú taktiež úrazy hlavy, poruchy prekrvenia mozgu, pozitívna rodinná anamnéza či vzdelanie (protože ľudia s nižším vzdelaním majú väčšie riziko rozvoja tohto ochorenia) [2]. Toto ochorenie sa vyskytuje častejšie u žien ako u mužov, v pomere 2:1 [4].

Alzheimerova choroba nie je “iba“ o strate pamäti, ale aj šiestou najčastejšou príčinou smrti v USA [5]. Medzi rokmi 2000 až 2017 sa počet úmrtí v USA viac ako zdvojnásobil [5]. Ľudia starší ako 65 rokov ktorým bola diagnostikovaná táto choroba sa v priemere dožívajú 4 až 8 rokov po jej diagnóze [5].

### 2.1.1 Diagnostika Alzheimerovej choroby

Alzheimerova býva diagnostikovaná kombináciou viacerých ukazovateľov. Pri určovaní diagnózy sa používajú neuropsychometrické (kognitívne) testy, rádiologické snímky (angl. neuroimaging data), biologické ukazovatele a špecifické kritériá, na základe ktorých je možné vylúčenie iných chorôb u pacienta z jeho história vývoja ochorenia [4]. T. Khan zadefinoval tieto ukazovatele do tzv. komplexného rámca pre diagnózu Alzheimerovej choroby (Obr. 2.1). V súčasnosti sa v tejto oblasti skúmajú biologické ukazovateľe (ich identifikácia a použitie), keďže používanie (a teda aj vytvorenie) rádiologických ukazovateľov je drahé [4] (vyžaduje si to zaškolený personál a vybavenie). Biologické ukazovateľe zatiaľ niesú dostatočne spoľahlivé [4].



Obr. 2.1: **Komplexný rámec pre diagnózu alzheimerovej choroby.** Pozostáva z neuropsychometrických testov, rádiologických snímok (z PET, MRI...), biologických ukazovateľov (napr. úrovne hladín určitých proteínov v krvnej plazme) Alzheimerovej choroby a kritérií vylúčenia iných neurologických chorôb.[4]

### 2.1.2 Biologické ukazovatele

Biologické ukazovatele (angl. biomarkers) sú merateľné biologické ukazovatele slúžiace na detekciu prítomnosti choroby. National Institute of Health definguje bio-

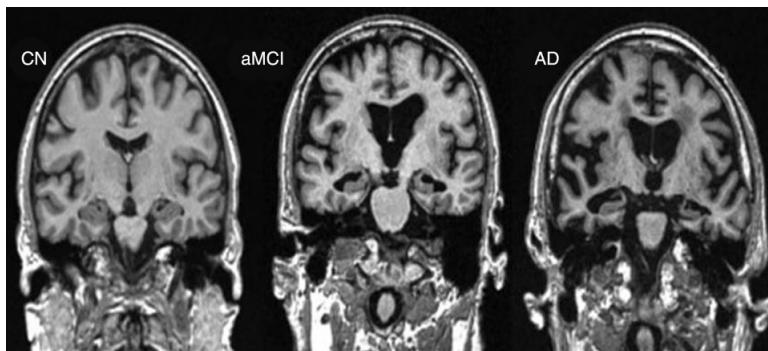
logický ukazovateľ ako indikátor určitého objektívneho merania a hodnotenia biologického procesu, patogénneho procesu alebo farmakologického hodnotenia terapeutickej účinnosti [6]. Alzheimerova choroba môže byť identifikovaná sledovaním týchto biologických ukazovateľov napríklad v krvnej plazme [4] alebo v mozgovo-miechovej tekutine (angl. cerebrospinal fluid) (ako úrovne hladín proteínov P-tau and A $\beta$ 42) [4] (angl. cerebrospinal fluid).

### 2.1.3 Obrazové a rádiologické ukazovatele

Identifikovanie Alzheimerovej choroby je v súčasnosti možné aj z rádiologických snímkov. Tvorba rádiologických snímkov je v súčasnosti možná pomocou techník akými sú počítačová tomografia s jednou fotónovou emisiou (angl. single-photon emission computed tomography - SPECT), pozitrónová emisná tomografia (angl. positron emission tomography PET), počítačová tomografia (angl. computed tomography - CT), magnetická rezonancia (magnetic resonance imaging - MRI) a magnetická rezonančná spektroskopia (angl. magnetic resonance spectroscopy - MRS) [4].

Snímky z magnetickej rezonancie (MRI) dokážu zachytiť odumieranie tkaniva (na základe biologických procesov), ktoré sa odohráva v rôznych častiach mozgu [4]. Príklad takéhoto snímku sa nachádza na obrázku 2.2.

Snímky z pozitrólovej emisnej tomografie (PET) dokážu zachytiť pokles mozgovej aktivity, ktorá je u pacientov s Alzheimerovou chorobou nižšia. Mozgová aktivita odráža úroveň metabolizmu glukózy v mozgu. Na miestach v mozgu, ktoré s sú touto chorobou postihnuté, je úroveň metabolizmu glukózy nižšia. Tento jav je znázornený na obrázku 2.3.

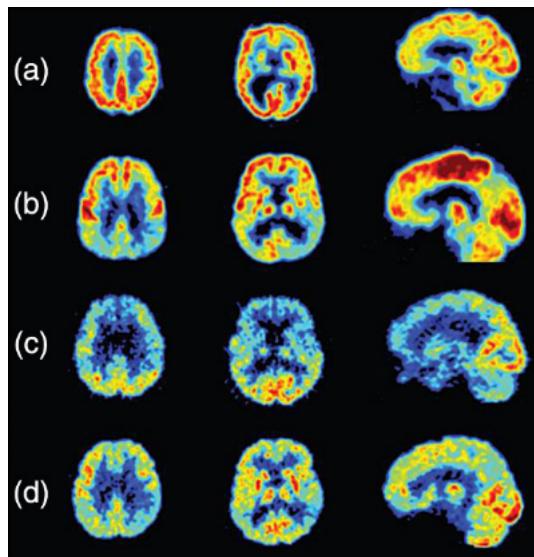


Obr. 2.2: **Typické odumieranie mozgového tkaniva zachytené magnetickou rezonanciou.** Obrázok zľava, označený ako CN (angl. cognitive normal), reprezentuje kognitívne normálneho jedinca. Obrázok v strede, označený ako aMCI (angl. amnestic mild cognitive impairment) reprezentuje jedinca s miernym kognitívnym poškodením - na obrázku je zreteľný úbytok mozgového tkaniva (šedá farba) najmä v strede mozgu (ale aj na jeho okrajoch) oproti kognitívne normálnemu jedincovi. Posledný obrázok označený ako AD (angl. Alzheimer's disease) reprezentuje jedinca s Alzheimerovou chorobou - na obrázku je zreteľný značný úbudok mozgového tkaniva. [4]

## 2.2 Neurónové siete

Neurónové siete patria medzi obľúbené techniky strojového učenia. Špeciálnou kategóriou sú hlboké neurónové siete (často označované skratkou DNN od angl. deep neural network), ktoré sa oproti obyčajným neurónovým sieťam odlišujú počtom vrstiev. Hlbokým neurónovým sieťam sa doteraz podarilo dosiahnuť v mnohých úlohách výnimočné výsledky, v ktorých častokrát už dokázali poraziť človeka.

V doméne obrazových dát sa využívajú predovšetkým konvolučné neurónové siete. Tie majú schopnosť naučiť sa rozpoznávať špecifické štruktúry/tvary z obrázka. Toto dokážu pomocou takzvaných konvolučných filtrov, ktoré sa v nižších vrstvách naučia rozoznávať jednoduchšie tvary, akými sú napríklad obrys alebo hrany. V tých vyšších vrstvách sú to zložitejšie štruktúry akými môžu byť celé objekty v závislosti od typu úlohy na ktorú boli trénované. Ak bola neurónová sieť trénovaná napríklad na klasifikáciu zvierat, môže tým objektom byť pes alebo morča, v prípade ak je úlohou neurónovej siete detekcia Alzheimerovej choroby môžu týmito

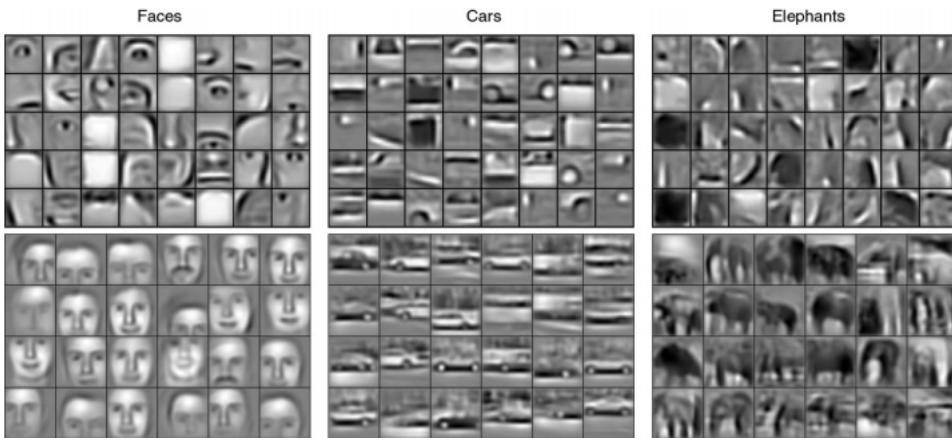


Obr. 2.3: **Snímky normálneho mozgu a mozgu postihnutého Alzheimerovou chorobou z pozitrónovej emisnej tomografie (PET).** [4] Na obrázkoch je viditeľná úroveň metabolizmu glukózy, u pacientov s Alzheimerovou chorobou je táto úroveň nižšia (žltá a modrá farba na obrázkoch). (a) Mozog kognitívne zdravého jedinca - vyznačuje sa vyššou mozgovou aktivitou. (b) Mozog vyznačujúci symptómy Alzheimerovej choroby - je vidieť nižšiu aktivitu v niektorých častiach mozgu oproti kognitívne zdravému jedincovi. (c) Mozog postihnutý frontotemporálnou demenciou (angl. frontotemporal dementia), tiež sa vyznačuje nižšou mozgovou aktivitou. (d) Mozog postihnutý Alzheimerovou chorobou.

objektami byť niektoré väčšie časti mozgu (napr. hippocampus).

### 2.2.1 Interpretovanie neurónovej siete

Montavon; Samek; Müller (2018) definujú interpretovanie ako mapovanie abstraktného konceptu (napríklad predikovanej triedy) do domény, ktorej človek dokáže porozumieť. Ako príklad domény, ktorá je interpretovateľná uvádzajú obrázky (pole pixelov) alebo text (sekvencia slov) [8]. Medzi domény, ktoré nie sú interpretovateľné zaraďujú napríklad latentné vektorové reprezentácie slov (angl. word embeddings) alebo iné abstraktné vektorové reprezentácie [8]. Na rozdiel od vstupných

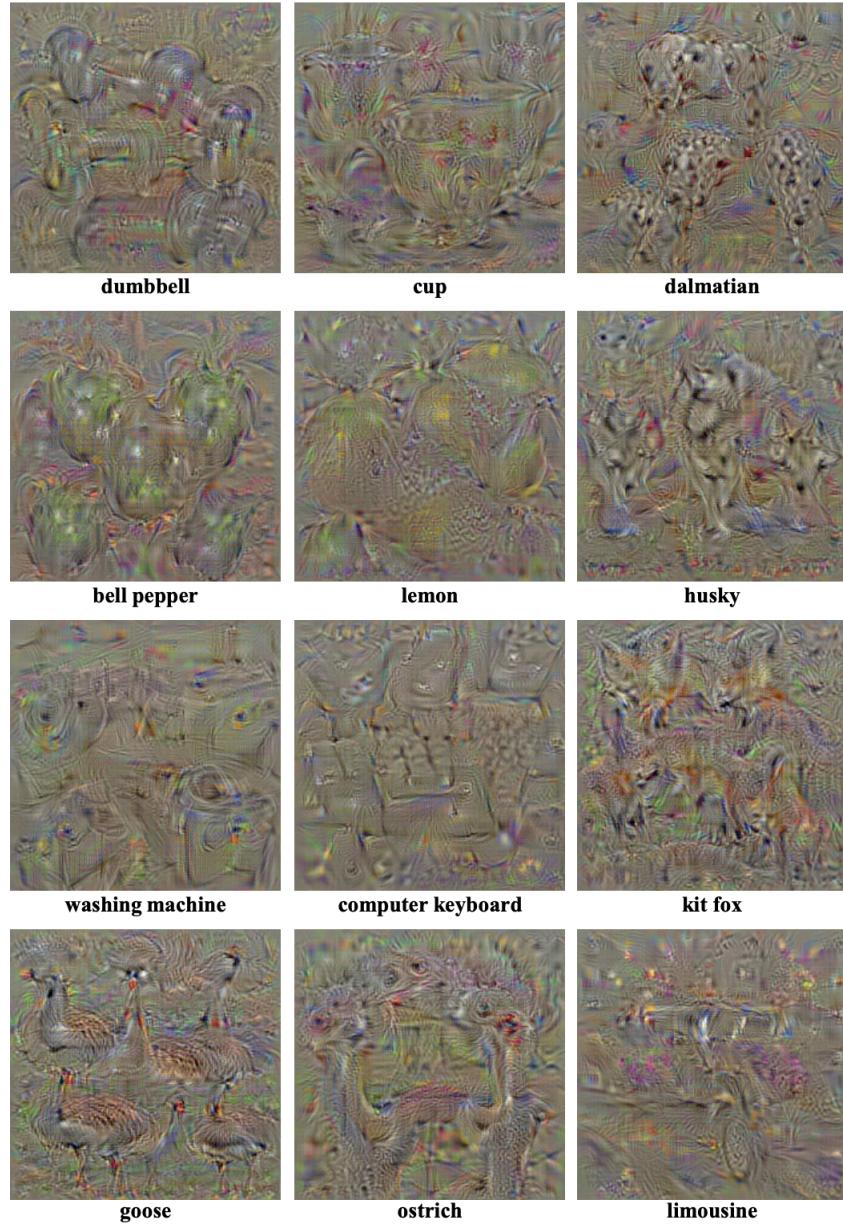


Obr. 2.4: **Vizualizácia druhej (hore) a tretej vrstvy (dole) konvolučných neurónových sietí naučených na špecifické kategórie objektov (tváre, autá a slony).** [7]

dát do neurónovej siete, ktoré sú zvyčajne interpretovatelné, neuróny na výstupnej vrstve a v skrytých vrstvách sú abstraktné a vyžadujú dodatočné úsilie na ich interpretovanie. Jedným zo spôsobov interpretovania týchto neurónov je maximalizácia aktivácie (angl. activation maximization).

**Maximalizácia aktivácie (angl. Activation maximization)** Maximalizácia aktivácie je metóda na nájdenie takého vstupného prototypu, ktorý vyprodukuje najväčšiu mieru aktivácie pre zvolený neurón (zvyčajne je to neurón hľadanej triedy na najvyššej vrstve). Takýto vstupný prototyp je nájdený tak, že neurónovej sieti je daný na vstup neutrálny obrázok, ktorý v danej doméne nereprezentuje žiadnu triedu (zvyčajne sa jedná o šedý obrázok) a je optimalizovaná funkcia maximalizácie aktivácie pomocou poklesu gradientu [8] (angl. gradient descent). Pri aplikovaní tejto metódy na obrazové dátá výsledné prototypy vyzerajú tak ako na obrázku 2.5.

**Maximalizácia aktivácie s expertom** Na získanie realistickejších prototypov (prototypov, ktoré sa viac podobajú vstupným dátam)  $l_2$ -regularizácia (používaná v maximalizácii aktivácie) je nahradená takzvaným “expertom“, ktorý sa



Obr. 2.5: Maximalizácia aktivácie aplikovaná na obrazové  
dáta. [9] Výsledné vzorové prototypy pre jednotlivé triedy  
nevyzerajú prirodzene, sú prevažne šedé s farebnými črtami objektov.  
Tieto vzorové prototypy nereprezentujú príklady vstupov "z reálneho  
sveta" ale ideálne vstupy pre jednotlivé triedy. Takéto vstupy  
nerónová sieť bežne nedostane.

snaží naučiť distribúciu hľadanej triedy [8]. Oproti  $l_2$ -regularizácii, ktorá hľadá vstup maximalizujúci pravdepodobnosť triedy, expert hľadá taký vstup, ktorý je najpravdepodobnejší pre zvolenú triedu. Ako “expert” môže byť použitý napríklad Gaussian RBM (angl. Restricted Boltzmann machine) [8].

### 2.2.2 Vysvetľovanie predikcie neurónovej siete

Montavon; Samek; Müller (2018) definujú vysvetľovanie ako kolekciu vlastností dát, ktoré sú z interpretovateľnej domény, ktoré prispeli k výslednému rozhodnutiu (napr. zaradenie do určitej triedy - klasifikácia) pre určité pozorovanie [8]. Rozdiel oproti interpretovaniu teda je, že pri interpretovaní hľadáme vzorový prototyp (vzorové pozorovanie) pre zvolenú triedu, zatiaľ čo pri vysvetľovaní sa snažíme zistiť prečo, a teda ktoré z vlastností vstupu najviac prispeli (t.j. sú najviac relevantné) k výslednej predikcii neurónovej siete (napr. zaradenie pozorovania do určitej triedy).

Niekteré metódy vysvetľovania fungujú na základe zakrývania častí obrázka a sledovaním zmeny predikcie predikovanej triedy – perturbačné metódy, iné zasa na základe spätného šírenia (angl. backpropagation) – napr. LRP, analýza senzitivity.

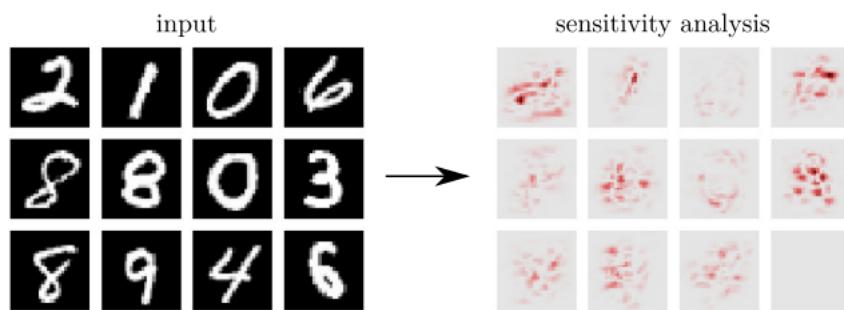
Každá z metód má svoje výhody a nevýhody, napríklad výhodou perturbačných metód je, že môžu byť použité na akýkoľvek model, keďže jediné čo potrebujú je výstup (predikciu) z modelu. Ich nevýhodou však je, že sú pomalé. Niektoré z metód vysvetľovania bližšie opíšeme v tejto kapitole.

#### 2.2.2.1 Analýza senzitivity

Analýza senzitivity slúži na vysvetľovanie predikcie neurónovej siete. Táto metóda identifikuje, ktoré z vlastností vstupného pozorovania najviac prispievajú, či už pre alebo proti, výslednej predikcii. Najviac dôležité sú také vlastnosti, ktorých zmenou sa najvýraznejšie zmení výsledná predikcia. Na takéto vlastnosti je výsledná

predikcia najviac senzitívna (resp. citlivá) [8].

Výsledok analýzy senzitivity znázornený v tepelnej mape (angl. heatmap) je zobrazený na obrázku 2.6. Analýza senzitivity zachytáva teda vlastnosti vstupného pozorovania, ktoré k výslednej predikcii prispievajú pozitívne aj negatívne (napr. zmenením určitej vlastnosti vstupu sa výrazne zníži zaradenie do danej triedy). Na výslednej tepelnej mape vlastnosti, ktoré k výslednej predikcii prispievajú pozitívne, a vlastnosti, ktoré k výslednej predikcii prispievajú negatívne (proti), nevieme rozlíšiť. Vieme len, že zmenením danej vlastnosti výrazne ovplyvníme predikciu.



Obr. 2.6: **Analýza senzitivity** aplikovaná na konvolučnú neurónovú sieť trénovanú na dátovej sade MNIST. [8]

Červenou farbou sú zobrazené miesta ktoré najviac prispievajú, či už pre alebo proti, výslednej predikcii. Čím je červená farba výraznejšia, tým viac je výsledok senzitívny na zmenu daného pixela.

### 2.2.2.2 LRP (angl. layer-wiser relevance propagation)

Metóda vrstvami propagovanej relevancie, ďalej len LRP (angl. layer-wise relevance propagation), sa od analýzy senzitivity odlišuje tým, že vo výslednej tepelnej mape dokáže odlišiť vlastnosti, ktoré prispeli pozitívne alebo negatívne k výslednej predikcii (v závislosti od použitých parametrov  $\alpha$  a  $\beta$ ).

Táto technika funguje tak, že vstupný obrázok dopredným šírením ”prejde” neurónovou sieťou, pričom sú zozbierané aktivácie neurónov v jednotlivých vrstvách.

Následne je neurónovou sietou spätným šírením propagované skóre z výstupu neurónovej siete v podobe relevancie až k vstupnému obrázku.

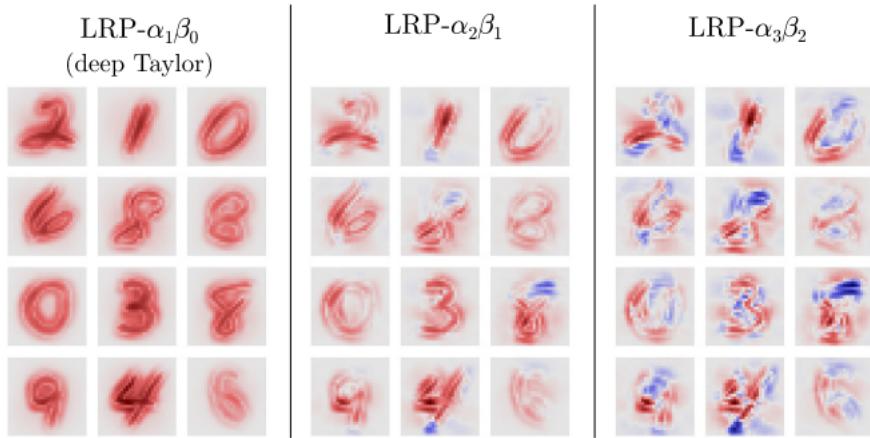
Nasledovné vzorce 2.1, 2.2, 2.3 [8] vyjadrujú spôsob výpočtu propagovanej relevancie medzi vrstvami.  $j$  a  $k$  sú jednotlivé vrstvy, pričom  $k$  je vrstva, z ktorej je relevancia  $R$  propagovaná. Parametre  $\alpha$  a  $\beta$  upravujú, koľko pozitívnej ( $\alpha$ ) alebo negatívnej ( $\beta$ ) relevancie je vytvorennej počas fázy spätného šírenia relevancie. Pri ich nastavovaní musí platiť, že  $\alpha - \beta = 1$  a zároveň  $\beta \geq 0$ . Súčet pozitívnej a negatívnej relevancie je však medzi vrstvami vždy rovnaký [8], výsledok použitia rôznych hodnôt  $\alpha$  a  $\beta$  je znázornený na obrázku 2.7.  $R_{j \leftarrow k}^+$  (Obr. 2.1) a  $R_{j \leftarrow k}^-$  (Obr. 2.3) vyjadrujú množstvo pozitívnej (+), resp. negatívnej (-) relevancie propagovanej z vrstvy  $k$  do vrstvy  $j$ .  $a_j$  je aktivácia neurónu, na ktorý je propagovaná relevancia.

$$R_{j \leftarrow k}^+ = \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} \quad (2.1)$$

$$R_{j \leftarrow k}^- = \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-} \quad (2.2)$$

$$R_j = \sum_k (\alpha R_{j \leftarrow k}^+ - \beta R_{j \leftarrow k}^-) R_k \quad (2.3)$$

Výhodou LRP oproti analýze senzitivity je, že vysvetlenie (výsledná tepelná mapa) vytvorené technikou LRP je pre rôzne obrázky vždy rôzne [10]. Naopak, pri analýze senzitivity je vysvetlenie vždy rovnaké pokial v architektúre neurónovej siete neboli použité združovacie vrstvy (angl. pooling layers) [10]. Ďaľším rozdielom je, že vo výslednom vysvetlení LRP rozlišuje, ktoré vlastnosti pozitívne alebo negatívne prispeli k negatívnej predikcii.



Obr. 2.7: Výsledné vysvetlenie (v podobe tepelnej mapy) vytvorené použitím LRP s rôznymi hodnotami  $\alpha$  a  $\beta$  na dátovej sade MNIST. [8] Pozitívna relevancia je zobrazená červenou farbou. [8] Negatívna relevancia je zobrazená modrou farbou. [8] V prípade, že použijeme  $\alpha = 1$  a  $\beta = 0$  stráčame informáciu o tom, ktoré pixely negatívne (tj. sú proti výslednej predikcii) prispeli k výslednej predikcii (a opačne).

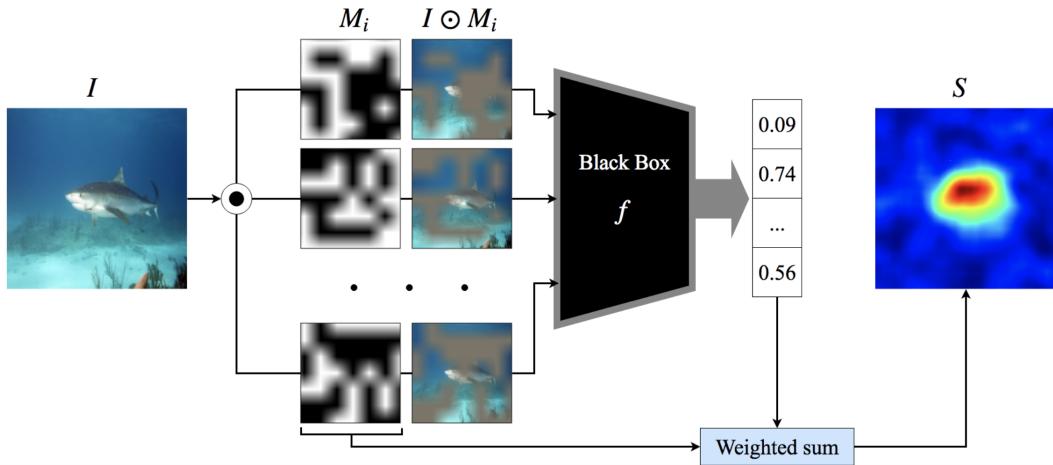
### 2.2.2.3 RISE - Randomized Input Sampling for Explanation

Túto metódu môžeme zaradiť medzi perturbačné metódy, keďže je tiež založená na zakrývaní jednotlivých častí obrazu a sledovaním zmeny výslednej predikcie modelu. Už z názvu modelu (*Randomized Input Sampling for Explanation*) je zrejmé, že táto metóda využíva náhodu na zakrývanie jednotlivých častí vstupného obrazu. Vstupný obraz je prekrytý náhodou maskou, ktorá je vytvorená nasledovne [11]:

- Je vytvorená náhodná binárna (tj. iba z bielej a čiernej farby) maska o malej veľkosti (napríklad 8px x 8px).
- Táto maska je zväčšená (angl. upscaled) pomocou bilineárnej interpolácie [11] (angl. bilinear interpolation) na veľkosť ktorá je mierne väčšia ako veľkosť obrázka s ktorým bude prekrytá (kvôli oreznávaniu). Tým sa zníži jej kvalita a ostré hrany medzi bielymi a čiernymi časťami sa zjemnia. Masky už teda nie sú binárne.

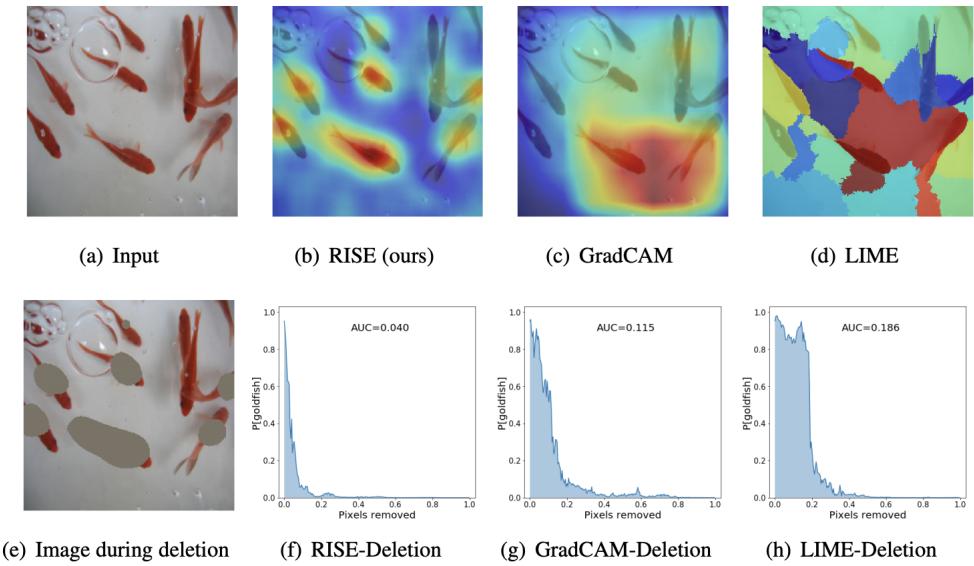
- Z masky je náhodne vyrezaná náhodná časť o veľkosť prekrývaného obrázka.

Toto sa opakuje  $N$  krát. Výsledná tepelná mapa je vypočítaná ako vážený priemer všetkých vygenerovaných masiek, kde váhy sú skóre (pravdepodobnosť predikovanej triedy) z modelu. Tento proces je zobrazený na obrázku 2.8.



Obr. 2.8: Metóda *Rise*. [11] Vygenerované masky nahradzajú vstupný obrázok na, ktorý sú aplikované. Z výstupných predikcií jednotlivých masiek je nakoniec vypočítaná tepelná mapa.

Autori porovnali túto metódu s metódami *GradCAM* (Selvaraju et al. 2017) a *LIME* (Ribeiro et al. 2016). Metóda *Rise* si oproti týmto dvom metódam počínala lepšie (Obr. 2.9). Vykonali niekoľko experimentov, v ktorých porovnali architektúry neurónových sietí *ResNet50* (He et al. 2016) a *VGG16* (Simonyan; Zisserman 2014) natrénované na dátových sadách PASCAL VOC07 (Everingham et al. 2016) a MSCOCO2014 (Lin et al. 2014). Sledovali metriky *insertion* a *deletion* (Obr. 2.9). Metrika *insertion* je vyjadrená ako plocha pod krivkou (angl. Area Under Curve - AUC) funkcie  $y = f(x)$ , kde  $y$  je istota predikcie a  $x$  je počet pridaných najdôležitejších pixelov, dôležitosť pixelov je určená metódou vysvetľovania predikcie neurónovej siete a môže byť zobrazené pomocou tepelnej mapy. Metrika *deletion* naopak odoberá najdôležitejšie pixely z obrázka.



Obr. 2.9: Porovnanie metódy *RISE* s *GradCAM* alebo *LIME*. [11] V prvom riadku sú tepelné mapy jednotlivých metód pre vstup. V druhom riadku je znázornená porovnávaná metrika *deletion*. Táto metrika sleduje vzťah medzi odobratím najdôležitejších pixelov a výslednou predikciou modelu. Je vyčíslená pomocou výpočtu plochy pod krivkou (angl. AUC - Area Under Curve). Na grafoch si môžeme všimnúť, že metóda *RISE* potrebuje odobrať menej pixelov na to aby klesla pravdepodobnosť predikovanej triedy. To znamená, že tepelná mapa (metódy *RISE* oproti ostatným metódam) lepšie zaznamenáva dôležité pixely pre predikovanú triedu.

## 2.3 Využitie neurónových sietí pri odhalovaní Alzheimerovej choroby

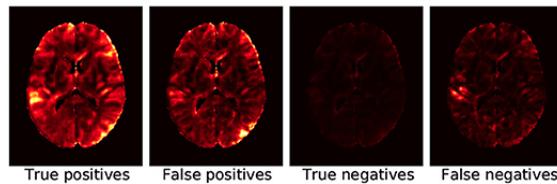
Neurónovým sietiam sa doposiaľ podarilo dosiahnuť veľmi dobré výsledky pri odhalovaní Alzheimerovej choroby. Medzi state-of-the-art riešenia patrí konvolučná neurónová sieť od Esmaeilzadeh et al. s presnosťou **94.1%** (a s  $F_2$  skóre 0.93%) na populárnej dátovej množine s názvom *ADNI-1*. Tento výsledok dosiahli v úlohe klasifikácie iba do CN a AD, pridaním MCI sa presnosť výrazne znižuje. Vsturom do tejto neurónovej siete boli snímky z magnetickej rezonancie (MRI) ale aj demografické informácie akými sú napríklad vek alebo pohlavie. Autor avšak

nereportuje úspešnosť modelu, ktorý bol natrénovaný iba z obrazových dát, táto úspešnosť by bola pravdepodobne o niečo nižšia.

S takto vysokou úspešnosťou môžu byť neurónové siete veľmi dobrým pomocníkom doktorov. Problémom však je, že sa správajú ako čierne skrinky, preto je potrebné ich rozhodnutia interpretovať, aby bolo pre doktora zrejmé na základe čoho sa neurónová sieť urobila svoju predikciu.

### **2.3.1 Vysvetľovanie rozhodnutí neurónových sietí detegujúcich Alzheimerovu chorobu**

Existujúce práce sa už zaobrali metódami vysvetľovania rozhodnutí neurónových sietí detekujúcich Alzheimerovu chorobu. Böhle; Eitel; Weygandt; Ritter uviedli možnosti analýzy rozhodnutí za účelom ich vysvetľovania. Konkrétnie sa zaobrali metódami vrstvami propagovanej relevancie (LRP) a vedenou spätnou propagáciou (angl. guided backpropagation). Uvádzajú LRP ako metódu na vysvetľovanie invidividuálnych rozhodnutí neurónovej siete kde naopak vedenú spätnú propagáciu ako metódu na zistenie oblastí, na ktoré je neurónová sieť senzitívna. Tieto metódy skúmali porovnávaním priemerov tepelných máp (angl. heatmaps) všetkých pozorovaní v predikovaných triedach (2 - AD, HC). Taktiež porovnávali priemerné tepelné mapy pozorovaní podľa spôsobu zaradenia výslednej predikcie (4 - true positive, true negative, false positive, false negative) (Obrázok 2.10). Okrem iného porovnávali mieru relevancie pri metóde LRP v jednotlivých častiach mozgu u pozorovaní s Alzheimerovou chorobou a u pozorovaní bez nej. Možným vylepšením tejto práce je vyskúšanie metódy LRP aj na pacientoch s miernym kognitívnym poškodením (angl. mild-cognitive impairment), nie len na pacientoch s Alzheimerovou chorobou a zdravých jedincoch.



Obr. 2.10: **Priemerná relevancia (z metódy LRP) pozorovaní podľa spôsobu zaradenia výslednej predikcie** Najviac relevancie je na miestach s červenou farbou.

## 2.4 Zhrnutie

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.



## **3. Návrh riešenia**

### **3.1 Zhrnutie**



# Literatúra

1. AMISHA, Paras Malik; PATHANIA, Monika; RATHAUR, Vyas Kumar. Overview of artificial intelligence in medicine. *Journal of family medicine and primary care*. 2019, roč. 8, č. 7, s. 2328.
2. 2019. Dostupné tiež z: <http://www.alzheimer.sk/informacie/alzheimerovachoroba.aspx>.
3. DUTHEY, Béatrice. Background paper 6.11: Alzheimer disease and other dementias. *A Public Health Approach to Innovation*. 2013, s. 1–74.
4. KHAN, Tapan. *Biomarkers in Alzheimer's Disease*. Academic Press, 2016.
5. 2017. Dostupné tiež z: <https://www.alz.org/alzheimers-dementia/facts-figures>.
6. WORKING, G Biomarkers Definitions. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*. 2001, roč. 69, č. 3, s. 89–95.
7. LEE, Honglak; GROSSE, Roger; RANGANATH, Rajesh; NG, Andrew. Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks. Dostupné z DOI: [10.1145/2001269](https://doi.org/10.1145/2001269).
8. MONTAVON, Grégoire; SAMEK, Wojciech; MÜLLER, Klaus-Robert. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*. 2018, roč. 73, s. 1–15.
9. SIMONYAN, Karen; VEDALDI, Andrea; ZISSERMAN, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*. 2013.

10. MÜLLER, Klaus-Robert; SAMEK, Wojciech; MONTAVON, Gregoire; LAPUSCHKIN, Sebastian; ARRAS, Leila. *Explaining and Interpreting Deep Neural Networks*. Dostupné tiež z: [http://iphome.hhi.de/samek/pdf/DTUSummerSchool2017\\_1.pdf](http://iphome.hhi.de/samek/pdf/DTUSummerSchool2017_1.pdf).
11. PETSIUK, Vitali; DAS, Abir; SAENKO, Kate. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*. 2018.
12. SELVARAJU, Ramprasaath R; COGSWELL, Michael; DAS, Abhishek; VEDANTAM, Ramakrishna; PARIKH, Devi; BATRA, Dhruv. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. 2017, s. 618–626.
13. RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why should i trust you? Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, s. 1135–1144.
14. HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, s. 770–778.
15. SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2014.
16. EVERINGHAM, Mark; VAN GOOL, Luc; WILLIAMS, Christopher KI; WINN, John; ZISSERMAN, Andrew. The pascal visual object classes (voc) challenge. *International journal of computer vision*. 2010, roč. 88, č. 2, s. 303–338.
17. LIN, Tsung-Yi; MAIRE, Michael; BELONGIE, Serge; HAYS, James; PERONA, Pietro; RAMANAN, Deva; DOLLÁR, Piotr; ZITNICK, C Lawrence. Microsoft coco: Common objects in context. In: *European conference on computer vision*. 2014, s. 740–755.
18. ESMAEILZADEH, Soheil; BELIVANIS, Dimitrios Ioannis; POHL, Kilian M; ADELI, Ehsan. End-to-end Alzheimer's disease diagnosis and biomarker iden-

## Literatúra

---

- tification. In: *International Workshop on Machine Learning in Medical Imaging*. 2018, s. 337–345.
19. BÖHLE, Moritz; EITEL, Fabian; WEYGANDT, Martin; RITTER, Kerstin. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer’s disease classification. *Frontiers in aging neuroscience*. 2019, roč. 11, s. 194.



# A. Plán práce

## A.1 Zimný semester

V tomto semestri plánujem pracovať na implementácii navrhnutej metódy, ktorú budem overovať v experimentoch a postupne vylepšovať. V tomto semestri plánujem:

- Natrénovať model na detekciu Alzheimerovej choroby z MRI snímkov
- Implementovať navrhnutú metódu
- Experimentovať s hyper-parametrami navrhnutej metódy
- Skúmať dosiahnuté výsledky, hľadať príčiny a možné vylepšenia
- Priebežne písat' prácu – implementáciu a dosiahnuté výsledky

## A.2 Letný semester

V tomto semestri budem pracovať na finalizácii tejto práce, navrhnutú metódu plánujem už iba vylepšovať a pracovať na záverečnom dokumente. V tomto semestri plánujem:

- Písat' prácu a jej jednotlivé časti - implementácia, technická dokumentácia, dosiahnuté výsledky, záver

## Dodatok A. Plán práce

---

- Vykonáť úpravy v navrhnutej metóde na základe doterajších výsledkov experimentov
- Vyhodnotiť a porovnať vykonalé experimenty
- Porovnať navrhnutú metódy s existujúcimi metódami
- Odovzdať prácu

Dodatok A. Plán práce

Dodatok A. Plán práce

Dodatok A. Plán práce