

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-182905-86077

Bc. Timotej Zatko

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

Diplomová práca

Máj 2021

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-182905-86077

Bc. Timotej Zaťko

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

Diplomová práca

Študijný program: Inteligentné softvérové systémy

Študijný odbor: Informatika

Miesto vypracovania: Ústav počítačového inžinierstva a aplikovanej informatiky
(FIIT)

Vedúci práce: Ing. Martin Tamajka

Bratislava Máj 2021

Návrh zadania diplomovej práce

Finálna verzia do diplomovej práce¹

Študent:

Meno, priezvisko, tituly: Timotej Zaťko, Bc.
Študijný program: Inteligentné softvérové systémy
Kontakt: timi.zatko@gmail.com

Výskumník:

Meno, priezvisko, tituly: Martin Tamajka, Ing.

Projekt:

Názov: Uplatnenie interpretateľnosti a vysvetliteľnosti neurónových sietí pri vyhodnocovaní medicínskych obrazových dát
Názov v angličtine: Application of interpretability and explainability of neural networks in the evaluation of medical images
Miesto vypracovania: Ústav počítačového inžinierstva a aplikovanej informatiky, FIIT STU
Oblast problematiky: počítačové videnie, hlboké neurónové siete, analýza medicínskych obrazových dát, vysvetliteľnosť a interpretateľnosť

Text návrhu zadania²

Umelá inteligencia a špeciálne hlboké neurónové siete sa za posledných desať rokov stali jedným z dominantných výskumných problémov, pričom v mnohých úlohách významne prekonávajú doterajšie prístupy. Zatial' čo vo výskume je prípustná istá miera neistoty alebo nepresnosti, v oblastiach ako je medicína je žiaduce, aby algoritmy umelej inteligencie poskytovali účinné mechanizmy kontroly správnosti predikcie. V medicínskej oblasti sa už umelá inteligencia uplatnila pri výrobe liekov, monitorovaní zdravotného stavu, chirurgických zákrokov a aj pri odhaľovaní chorôb. Práve pri odhaľovaní chorôb, akými sú napríklad rakovina plúc, rakovina kože alebo Alzheimerova choroba, sa využívajú hlboké neurónové siete za účelom získania klinicky relevantných informácií z medicínskych obrazových dát.

Analyzujte doménu medicínskych obrazových dát a súčasný stav problematiky interpretateľnosti a vysvetliteľnosti predikcie neurónovej siete. Navrhnite metódu na detekciu nesprávnych rozhodnutí alebo odhadovanie miery správnosti modelu neurónovej siete pri vyhodnocovaní medicínskych obrazových dát. Navrhnutú metódu implementujte a dosiahnuté výsledky vyhodnoťte na dostatočne veľkej dátovej množine. Dosiahnuté výsledky porovnajte s inými súčasnými riešeniami.

¹ Vytlačiť obojstranne na jeden list papiera

² 150-200 slov (1200-1700 znakov), ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

Literatúra³

- MONTAVON, Grégoire, Wojciech SAMEK and Klaus-Robert MÜLLER, 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* [online]. 2018, vol. 73, pp. 1-15.
- STURM, Irene, Sebastian LAPUSCHKIN, Wojciech SAMEK and Klaus-Robert MÜLLER, 2016. Interpretable deep neural networks for single-trial EEG classification. *Journal of Neuroscience Methods* [online]. 2016, vol. 274, pp. 141-145.

Vyššie je uvedený návrh diplomového projektu, ktorý vypracoval(a) Bc. Timotej Zaťko, konzultoval(a) a osvojil(a) si ho Ing. Martin Tamajka a súhlasi, že bude takýto projekt viesť v prípade, že bude pridelený tomuto študentovi.

V Bratislave dňa 29.1.2020



Pôpis študenta



Pôpis výskumníka

Vyjadrenie garanta predmetov Diplomový projekt I, II, III

Návrh zadania schválený: áno / nie⁴

Dňa: 17. 2. 2020



Pôpis garanta predmetov

³ 2 vedecké zdroje, každý v samostatnej rubrike a s údajmi zodpovedajúcimi bibliografickým odkazom podľa normy STN ISO 690, ktoré sa viažu k téme zadania a preukazujú výskumnú povahu problému a jeho aktuálnosť (uvedte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

⁴ Nehodiace sa prečiarknite

Zadanie diplomovej práce

Meno študenta: **Bc. Timotej Zaťko**

Študijný program: Inteligentné softvérové systémy

Študijný odbor: Softvérové inžinierstvo – hlavný študijný odbor
Umelá inteligencia – vedľajší študijný odbor

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

Samostatnou výskumnou a vývojovou činnosťou v rámci predmetov Diplomový projekt I, II, III vypracujte diplomovú prácu na tému, vyjadrenú vyššie uvedeným názvom tak, aby ste dosiahli tieto ciele:

Všeobecný cieľ:

Vypracovaním diplomovej práce preukážte, ako ste si osvojili metódy a postupy riešenia relativne rozsiahlych projektov, schopnosť samostatne a tvorivo riešiť zložité úlohy aj výskumného charakteru v súlade so súčasnými metódami a postupmi študovaného odboru využívanými v príslušnej oblasti a schopnosť samostatne, tvorivo a kriticky pristupovať k analýze možných riešení a k tvorbe modelov.

Špecifický cieľ:

Vytvorte riešenie zodpovedajúce návrhu textu zadania, ktorý je prílohou tohto zadania. Návrh bližšie opisuje tému vyjadrenú názvom. Tento opis je záväzný, má však rámcový charakter, aby vznikol dostatočný priestor pre Vašu tvorivosť.

Riadťe sa pokynmi Vášho vedúceho.

Pokiaľ v priebehu riešenia, opierajúc sa o hlbšie poznanie súčasného stavu v príslušnej oblasti, alebo o priebežné výsledky Vášho riešenia, alebo o iné závažné skutočnosti, dospejete spoločne s Vaším vedúcim k presvedčeniu, že niečo v texte zadania a/alebo v názve by sa malo zmeniť, navrhnite zmenu. Zmena je spravidla možná len pri dosiahnutí kontrolného bodu.

Miesto vypracovania: Ústav počítačového inžinierstva a aplikovanej informatiky, FIIT STU Bratislava

Vedúci práce: **Ing. Martin Tamajka**

Termíny odovzdania:

Podľa harmonogramu štúdia platného pre semester, v ktorom máte príslušný predmet (Diplomový projekt I, II, III) absolvovať podľa Vášho študijného plánu

Predmety odovzdania:

V každom predmete dokument podľa pokynov na www.fiit.stuba.sk v časti:
home > Informácie o > štúdiu > harmonogram štúdia > diplomový projekt.

V Bratislave dňa 17. 2. 2020

**SLOVENSKÁ TECHNICKÁ UNIVERZITA
V BRATISLAVE**

Fakulta Informatiky a informačných technológií
Ilkovičova 2, 842 15 Bratislava 4



1

doc. Ing. Peter Lacko, PhD.

riaditeľ Ústavu informatiky, informačných systémov
a softvérového inžinierstva

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

V Bratislave, 17. máj 2021

Timotej Zatko

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 2021

Súčasný vplyv umelej inteligencie na spoločnosť je nespochybniteľný. Využitie si už našla v rôznych oblastiach našich životov či už je to v smartfónoch pri odomykaní tvárou alebo najnovšie pri kontrole používania ochranného rúška pri vstupe do obchodov. Umelá inteligencia sa postupne dostáva do oblasti medicíny, kde má potenciál zachraňovať životy. Aby sa stala spoľahlivým pomocníkom doktorov pri diagnóze ochorení, je nevyhnutné, aby jej rozhodnutia bolo možné vysvetliť.

V oblasti medicíny je možné použitie neurónových sietí, pretože dokážu veľmi dobre pracovať s obrazovými dátami, a tak sa dajú využiť napríklad pri diagnóze Alzheimerovej choroby z rádiologických snímkov. Ich problémom však je, že sa správajú ako "čierne skrinky" čo bráni ich použitiu v bežnej praxi.

V tejto práci sme navrhli a implementovali novú metódu vysvetľovania rozhodnutí neurónových sietí, ktorá vychádza z už existujúcej metódy - RISE. Vytvorená metóda je vhodná pre obrazové dáta, ku ktorým vytvára tepelné mapy vysvetľujúce rozhodnutia modelu.

Metódu sme overili na neurónovej sieti deketujúcej Alzheimerovu chorobu z MRI snímkov a porovnali ju voči pôvodnej metóde a iným existujúcim metódam. Nová metóda síce dosiahla oproti pôvodnej lepšie výsledky, avšak sa neukázala ako vhodná pre rádiologické dátá.

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

2021, May

The current impact of artificial intelligence on society is undeniable. It has already been used in various areas of our lives, whether it is in smartphones for unlocking via face recognition or, most recently, for controlling the use of protective masks when entering shops or groceries. Artificial intelligence is entering the field of medicine, where it has the potential to save lives. Thus, in order to be a reliable assistant to doctors for example in the diagnosis of the disease, it is necessary that its decisions can be explained.

Since neural networks perform well on image data, they can be used in the field of medicine, for example, in the diagnosis of Alzheimer's disease from radiological images. However, their problem is that they behave like a "black box" which prevents their adoption into common practice.

In this work, we designed and implemented a novel method for explaining neural network decisions, which is based on other existing method - RISE. The created method is suitable for image data, for which it creates heatmaps explaining the decisions of the model.

We have validated the method on a neural network detecting Alzheimer's disease from MRI images and compared it to the original, and other existing methods. Although the new method achieved better results than the original, it did not prove to be suitable for radiological data.

Pod'akovanie

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

Obsah

1	Úvod	1
2	Analýza	3
2.1	Alzheimerova choroba	3
2.1.1	Diagnostika Alzheimerovej choroby	4
2.1.2	Biologické ukazovatele	4
2.1.3	Obrazové a rádiologické ukazovatele	5
2.2	Neurónové siete	7
2.2.1	Neurón	7
2.2.2	Dopredné neurónové siete	9
2.2.3	Konvolučné neurónové siete	9
2.2.4	Architektúry konvolučných neurónových sietí	12
2.2.5	Interpretovanie neurónovej siete	13
2.2.6	Vysvetľovanie predikcie neurónovej siete	15
2.2.6.1	Analýza senzitivity	15
2.2.6.2	LRP	16
2.2.6.3	Riadená spätná propagácia	17
2.2.6.4	GradCAM	18
2.2.6.5	Riadený GradCAM	19
2.2.6.6	RISE	19
2.3	Využitie neurónových sietí pri odhalovaní Alzheimerovej choroby .	21
2.3.1	Vysvetľovanie rozhodnutí neurónových sietí detegujúcich Alzheimerovu chorobu	25
2.4	Spracovanie obrazu	25

2.4.1	Rekonštrukcia obrazu	27
2.5	Zhrnutie	27
3	Ciele práce	31
3.1	Vytvorenie novej, alebo vylepšenie existujúcej metódy pre vysvetľovanie rozhodnutí neurónových sietí	31
3.2	Využitie vytvorenej metódy na určenie miery správnosti modelu neurónovej siete detegujúcej Alzheimerovu chorobu	32
4	Návrh riešenia	33
4.1	RISEI	34
4.2	Overenie riešenia	37
4.2.1	Dátová sada	37
4.3	Model	37
4.3.1	Experimenty	40
4.3.1.1	Určenie kvality metódy vysvetľovania rozhodnutí modelu	40
4.3.1.2	Určenie správnosti modelu	40
4.4	Zhrnutie	41
5	Implementácia	43
5.1	Metóda RISEI	43
5.1.1	Generovanie masiek	43
5.1.2	Vytvorenie tepelných máp	52
5.1.3	Vyhodnotenie tepelných máp	54
5.1.3.1	Metriky insertion & deletion	54
5.2	Model na detekciu Alzheimerovej choroby na základe MRI snímok .	55
5.2.1	Dátová sada	56
5.2.1.1	Predspracovanie	57
5.2.2	Model	58
5.2.3	Trénovanie	58
5.3	Zhrnutie	62

Obsah

6 Overenie riešenia	65
6.1 Experimenty	66
6.1.1 Výber architektúry neurónovej siete pre ďalšie experimenty	66
6.1.2 Overenie metódy RISEI	67
6.1.2.1 Stabilita tepelných máp	67
6.1.2.2 Experiment 1 (jedna snímka)	68
6.1.2.3 Experiment 2 (viacero snímok)	69
6.1.3 RISE vs RISEI (s rôznymi parametrami)	71
6.1.3.1 Nastavenie parametrov RISEI	76
6.1.4 Porovnanie s existujúcimi metódami	78
6.1.4.1 Metriky	79
6.1.4.2 Výsledky	80
6.2 Zhrnutie	84
7 Zhodnotenie	85
7.0.1 Zhodnotenie cieľov práce	86
7.0.2 Limitácie	87
7.0.3 Možné rozšírenia	87
Literatúra	89
Dodatok A Plán práce	
A.1 Letný semester - DP1	
A.2 Zimný semester - DP2	
A.2.1 Vyjadrenie k plneniu plánu	
A.3 Letný semester - DP3	
A.3.1 Vyjadrenie k plneniu plánu	
Dodatok B Technická dokumentácia	
B.1 Príprava vývojového prostredia	
B.2 Závislosti (použité knižnice)	
B.3 Technické riešenie	
B.4 Moduly	

Obsah

- B.4.1 Modul: src.risei
 - B.4.1.1 Trieda: RISEI
- B.4.2 Modul: src.heatmaps.evaluation
 - B.4.2.1 Trieda: HeatmapEvaluationV3
 - B.4.2.2 Trieda: HeatmapEvaluationHistory

Dodatok C Obsah priloženého digitálneho média

Zoznam použitých skratiek

AD angl. Alzheimier disease (Alzheimerova choroba) – používa sa na označenie pacientov trpiacich Alzheimerovou chorobou

AUC angl. area under curve (plocha pod krivkou)

CN angl. cognitive normal (kognitívne zdravý) – používa sa na označenie pacientov bez kognitívneho poškodenia (tj. zdravých jedincov)

MCI angl. mild cognitive impairment (mierne kognitívne poškodenie) – používa sa na označenie pacientov s miernym kognitívnym poškodením

MRI angl. magnetic resonance imaging (magnetická rezonancia) – je spôsob tvorby rádiologických snímkov ľudského tela, používa sa pri diagnostike Alzheimerovej choroby

PET angl. positron emission tomography (pozitrónová emisná tomografia) – je spôsob tvorby rádiologických snímkov ľudského tela, používa sa pri diagnostike Alzheimerovej choroby

TP angl. true positive (správne pozitívny) – výraz pre pozorovanie, ktoré je pozitívne a bolo označené správne

Obsah

TN angl. true negative (správne negatívny) – výraz pre pozorovanie, ktoré je negatívne a bolo označené správne

FP angl. false positive (nesprávne pozitívny) – výraz pre pozorovanie, ktoré je negatívne, ale bolo neprávne označené ako pozitívne

FN angl. false negative (nesprávne negatívny) – výraz pre pozorovanie, ktoré je pozitívne, ale bolo neprávne označené ako negatívne

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 sietiam 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. Avšak toto nie je jedinou motiváciou pre vysvetliteľnosť rozhodnutí neurónových sietí. Novovznikajúce regulácie, ako napríklad pripravovaná regulácia s názvom "Right to Explanation" od Európskej Únie [2] vyžadujú vysvetliteľnosť systémov umelej inteligencie. Motivácia je teda aj legislatívna.

V tejto práci sa zaoberáme uplatnením interpretovateľnosti a vysvetliteľnosti neurónových sietí pri vyhodnocovaní medicínskych obrazových dát. Navrhujeme metódu na vysvetľovanie rozhodnutí neurónových sietí, ktorá vychádza z existujúcej

Kapitola 1. Úvod

metódy RISE. Rozširujeme ju o podporu pre 3D volumetrické dáta a o parametri-zovateľnú možnosť prekryvu (aj pomocou dokreslenia), ktorý táto metóda využíva na tvorbu vysvetlení - tepelných máp.

Navrhnutú metódu vyhodnocujeme na MRI snímkoch pacientov s Alzheimerovou chorobou. Výsledky porovnávame s pôvodnou metódou a inými existujúcimi me-tódami na vysvetľovanie rozhodnutí neurónových sietí.

2. Analýza

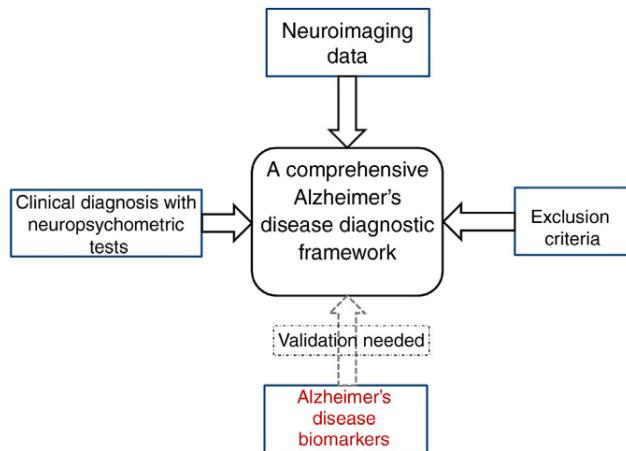
2.1 Alzheimerova choroba

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 [3]. Jej priebeh sa vyznačuje postupným poklesom kognitívnych funkcií, postupným zhoršením pamäte, myslenia, rozprávania a schopnosti učenia sa [4]. Najčastejšie sa vyskytuje u ľudí starších ako 65 rokov, s pravdepodobnosťou výskytu až 50% po dovršení 85 rokov života [4]. 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) [3]. Toto ochorenie sa vyskytuje častejšie u žien ako u mužov, v pomere 2:1 [5].

Alzheimerova choroba nie je “iba” o strate pamäti, ale aj šiestou najčastejšou príčinou smrti v USA [6]. Medzi rokmi 2000 až 2017 sa počet úmrtí v USA viac ako zdvojnásobil [6]. Ľudia starší ako 65 rokov ktorým bola diagnostikovaná táto choroba sa v priemere dožívajú 4 až 8 rokov po jej diagnóze [6].

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 [5]. 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 [5] (vyžaduje si to zaškolený personál a vybavenie). Biologické ukazovateľe zatiaľ nie sú dostatočne spoľahlivé [5].



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.[5]

2.1.2 Biologické ukazovatele

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

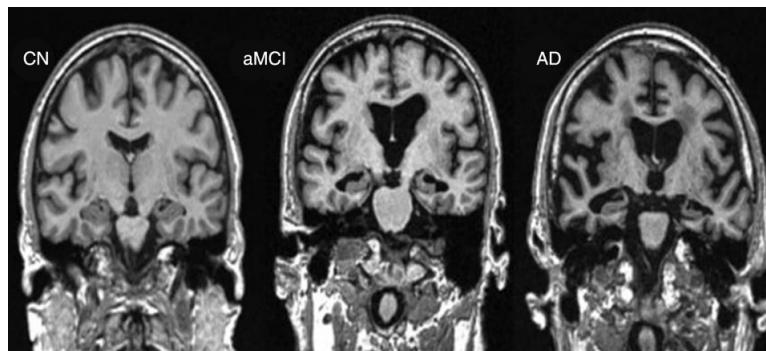
logický ukazovateľ ako indikátor určitého objektívneho merania a hodnotenia biologického procesu, patogénneho procesu alebo farmakologického hodnotenia terapeutickej účinnosti [7]. Alzheimerova choroba môže byť identifikovaná sledovaním týchto biologických ukazovateľov napríklad v krvnej plazme [5] alebo v mozgovo-miechovej tekutine (angl. cerebrospinal fluid) (ako úrovne hladín proteínov P-tau and A β 42) [5] (angl. cerebrospinal fluid).

2.1.3 Obrazové a rádiologické ukazovatele

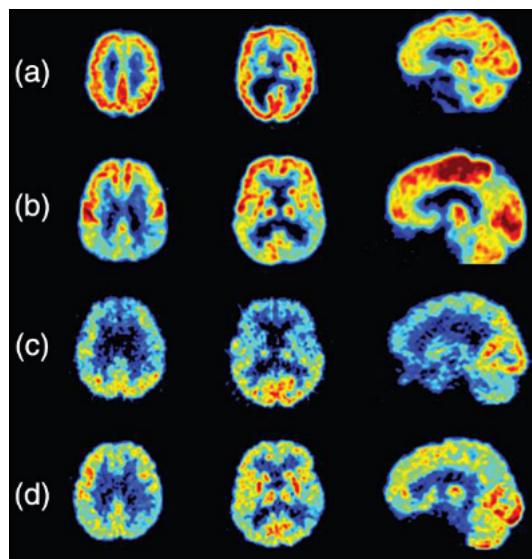
Identifikovanie Alzheimerovej choroby je v súčasnosti možné aj z rádiologických snímok. Tvorba rádiologických snímok je v súčasnosti možná pomocou techník akými sú počítačová tomografia s jednou fotónovou emisiou (angl. single-photon emission computed tomography - SPECT), pozitrónová emisná tomografia (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) [5].

Snímky z magnetickej rezonancie (MRI) dokážu zachytiť odumieranie tkaniva (na základe biologických procesov), ktoré sa odohráva v rôznych častiach mozgu [5]. Príklad takého 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 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. [5]



Obr. 2.3: **Snímky normálneho mozgu a mozgu postihnutého Alzheimerovou chorobou z pozitronovej emisnej tomografie (PET).** [5] 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.

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 prekonať človeka. V našej oblasti obrazových rádiologických dát sa používajú najmä konvolučné neurónové siete.

Haykin et al. [8] definujú neurónovú sieť nasledovne:

Neurónová sieť je veľký paralelný distribuovaný procesor tvorený jednoduchými procesorovými jednotkami, ktorý má prirodzený sklon ukladať poznatky a sprístupňovať ich na použitie. Ľudskému mozgu sa podobá v dvoch aspektoch:

1. Neurónová sieť získava vedomosti zo svojho prostredia prostredníctvom procesu učenia.
2. Na uchovanie získaných poznatkov sa používajú prepojenia medzi jednotlivými neurónami.

Neurónové siete sú teda inšpirované fungovaním mozgu človeka, keďže napodobňujú jeho fungovanie.

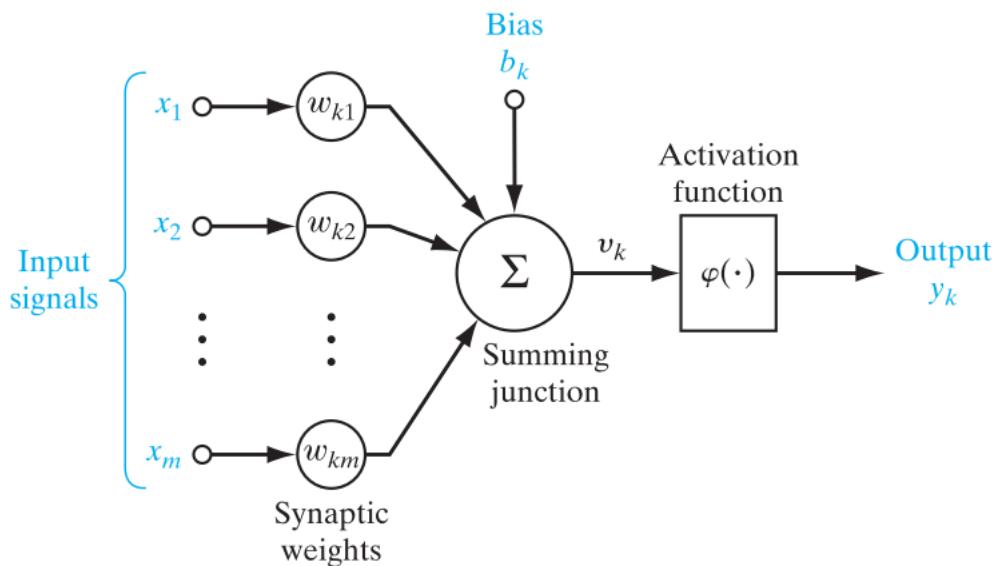
2.2.1 Neurón

Neurón (Obr. 2.4) je základnou stavebnou jednotkou neurónových sietí. Matematicky sa dá zapísť ako [8]:

$$y_k = \varphi(b_k + \sum_{j=1}^m w_{kj} \cdot x_j) \quad (2.1)$$

Kde:

- x_1, x_2, \dots, x_m sú vstupné signály
- $w_{k1}, w_{k2}, \dots, w_{km}$ sú váhy neurónu k
- b_k je sklon neurónu k
- $\varphi(\dots)$ je aktivačná funkcia
- y_k je výsupný signál neurónu k



Obr. 2.4: **Model neurónu.** [8] Neurón sa skladá zo vstupných signálov a váh, ktoré sú na tieto signály aplikované, sklon (b_k - angl. bias) a aktivačnej funkcie, ktorá zabezpečuje nelinearitu. Vzorec 2.1 matematicky popisuje správanie neurónu.

Parametrami, ktoré sa počas trénovania neurónovej siete menia sú váhy w_{kj} a sklon b_k , tieto parametre sú takzvané trénovateľné parametre. Tieto parametre sa upravujú pri spätnej propagácii (angl. backpropagation), kedy sa minimalizuje chybová funkcia (angl. loss function).

V neurónových sietiach s viac vrstvami sa stávajú výstupné signály y neurónov jednej vrstvy vstupom x do ďalšej.

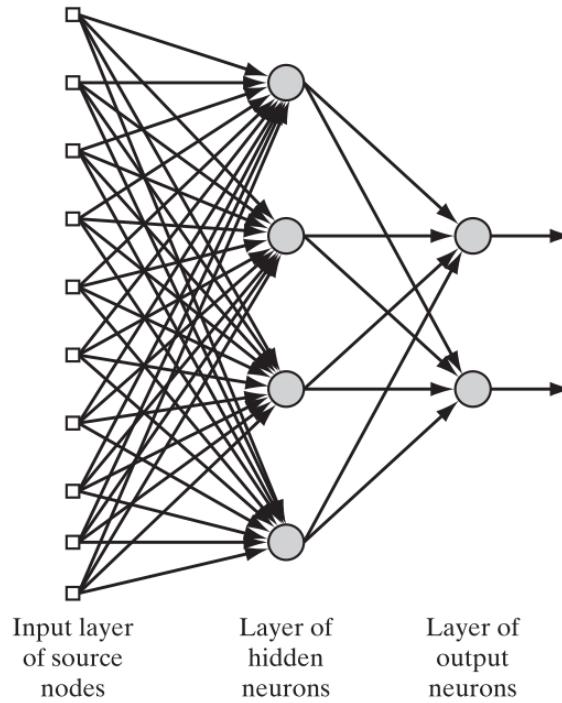
Aktivačná funkcia zabezpečuje nelinearitu neurónu, medzi najpoužívanejšie aktivačné funkcie patria Sigmoid ($S(x) = \frac{1}{1+e^{-x}}$), Tanh alebo ReLU ($ReLU(x) = \max(0, x)$). Jednotlivé neuróny si môžeme predstaviť ako nelineárne funkcie, ktorých spojením do viac vrstiev dokážu skladať ešte zložitejšie a komplexnejšie funkcie.

2.2.2 Dopredné neurónové siete

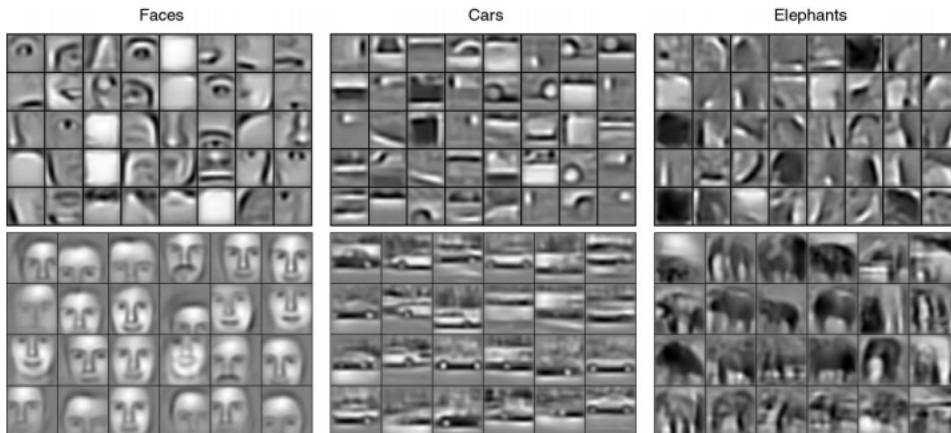
Dopredné neurónové siete (Obr. 2.5) sú jednou z mnoha architektúr neurónových sietí. V dopredných neurónových sieťach výstupný signál z jednej vrstvy nemôže byť vstupným signálom do jej predošej vrstvy. Signál je prenášaný iba v jednom smere – dopredu. Dopredné neurónové siete sa môžu skladať z viacerých vrstiev. Základom je vstupná a výstupná vrstva a ľubovoľný počet skrytých vrstiev. Ich počet nie je limitovaný, avšak v hlbokých neurónových sieťach (tj. sieťach s veľkým početom skrytých vrstiev) môže nastať problém miznúceho gradientu.

2.2.3 Konvolučné neurónové siete

Konvolučné neurónové siete sa používajú prevažne v doméne obrazových dát. Tieto siete 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 (Obr. 2.6). 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 možu týmito objektami byť niektoré väčšie časti mozgu (napr. hippocampus).



Obr. 2.5: **Model doprednej neurónovej siete.** [8] Dopredné neurónové siete sa skladajú zo vstupnej vrstvy, skrytých vrstiev a výstupnej vrstvy. Keď hovoríme o počte vrstiev vstupnú vrstvu nepočítame. Neurónová sieť na obrázku má teda dve vrstvy.

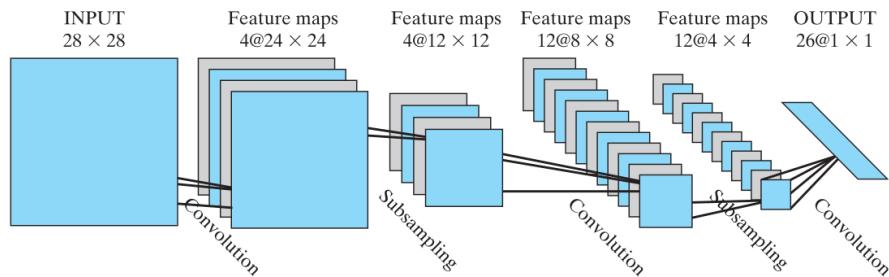


Obr. 2.6: Vizualizácia druhej (hore) a tretej vrstvy (dole) konvolučných neurónových sietí naučených na špecifické kategórie objektov (tváre, autá a slony). [9] Nižšie vrstvy rozoznávajú jednoduchšie štruktúry zatiaľ čo vyššie už dokážu rozoznať aj tie zložitejšie.
10

Základnými stavebnými blokmi konvolučných neurónových sietí sú konvolučné vrstvy (angl. convolutional layers) a združovacie vrstvy (angl. pooling layers).

Konvolučné vrstvy Pomocou konvolučných vrstiev sa neurónová sieť učí extrahovať črty z obrázka [8]. Konvolúcia prebieha tak, že tzv. jadro (angl. kernel) sa posúva po tzv. mape vlastností (angl. feature map) a matematickými operáciami z pôvodnej mapy vlastností a svojich parametrov vytvára novú mapu vlastností. Tieto parametre sú trénovateľné, čo umožňuje sa každému jadru naučiť určitú črtu - napr. hranu. Konvolučná vrstva tiež dokáže znižovať komplexitu modelu (a teda aj celkový počet jeho parametrov) jej hyper parametrami (angl: stride, padding, depth).

Združovacie vrstvy Cieľom združovacích vrstiev je postupne znižovať dimenzionalitu dát, tým znižovať počet počet parametrov modelu, a teda aj jeho komplexitu [10]. Najčastejšie sa používajú vrstvy združujúce maximom (angl. max-pooling), ale existujú aj vrstvy združujúce priemerom či súčtom.



Obr. 2.7: Príklad architektúry konvolučnej neurónovej siete.

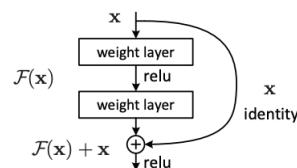
[8] V tejto architektúre neurónovej siete sa používajú tri konvolučné vrstvy (označené ako *convolution*) a dve združovacie vrstvy (označené ako *subsampling*). Môžeme si všimnúť, že konvolučné vrstvy postupne pridávajú mapy vlastností (tiež označované ako: angl. "volumes") a tiež mierne znižujú ich veľkosť. Združovacie vrstvy zasa výrazne znižujú ich veľkosť (až o polovicu) a tým aj počet parametrov v neurónovej sieti.

2.2.4 Architektúry konvolučných neurónových sietí

Architektúra neurónovej siete hovorí o tom, ako neurónová sieť vyzerá - koľko má vrstiev, z akých vrstiev sa skladá (konvolučné, združovacie, husté), koľko filtrov je v jednotlivých vrstvách a pod. Nie každá architektúra je vhodná na každý problém. Ak je problém jednoduchý, môže byť použitie veľmi hlbokej neurónovej siete zbytočné. Taktiež, jednoduchšia architektúra potrebuje menej výpočtových zdrojov na natrénovanie a je odolnejšia voči pretrénovaniu. Spomenieme niekoľko najznámejších architektúr, ktoré sú používané najmä pri klasifikácii obrazových dát.

- VGG [11] - hlboká neurónová sieť, so 16 alebo s 19 vrstvami. Skladá sa s konvolučných a združovacích vrstiev.
- ResNET [12] - hlboká neurónová sieť skladajúca sa z reziduálnych blokov. Reziduálne bloky obsahujú skracovacie spojenia, "skratky" (angl. shortcut connections) ako nástroj na zabránenie miznúcemu a explodujúcemu gradientu. Táto architektúra bola navrhnutá s 20, 32, 44, 56, 110 a 1202 vrstvami. Na klasifikáčnych úlohách na dátovej sade ImageNet táto architektúra prekonala architektúru VGG.
- Inception (GoogLeNet) [13] - hlboká neurónová sieť skladajúca sa s inception blokov. Každý blok robí niekoľko rôznych konvolúcií zo vstupe daného bloku, ktoré sú následne spojené v združovacom bloku.

Taktiež existuje niekoľko ďalších vylepšení Inception architektúry (Inception v1 až v4), dokonca aj kombinácia s architektúrou ResNet.



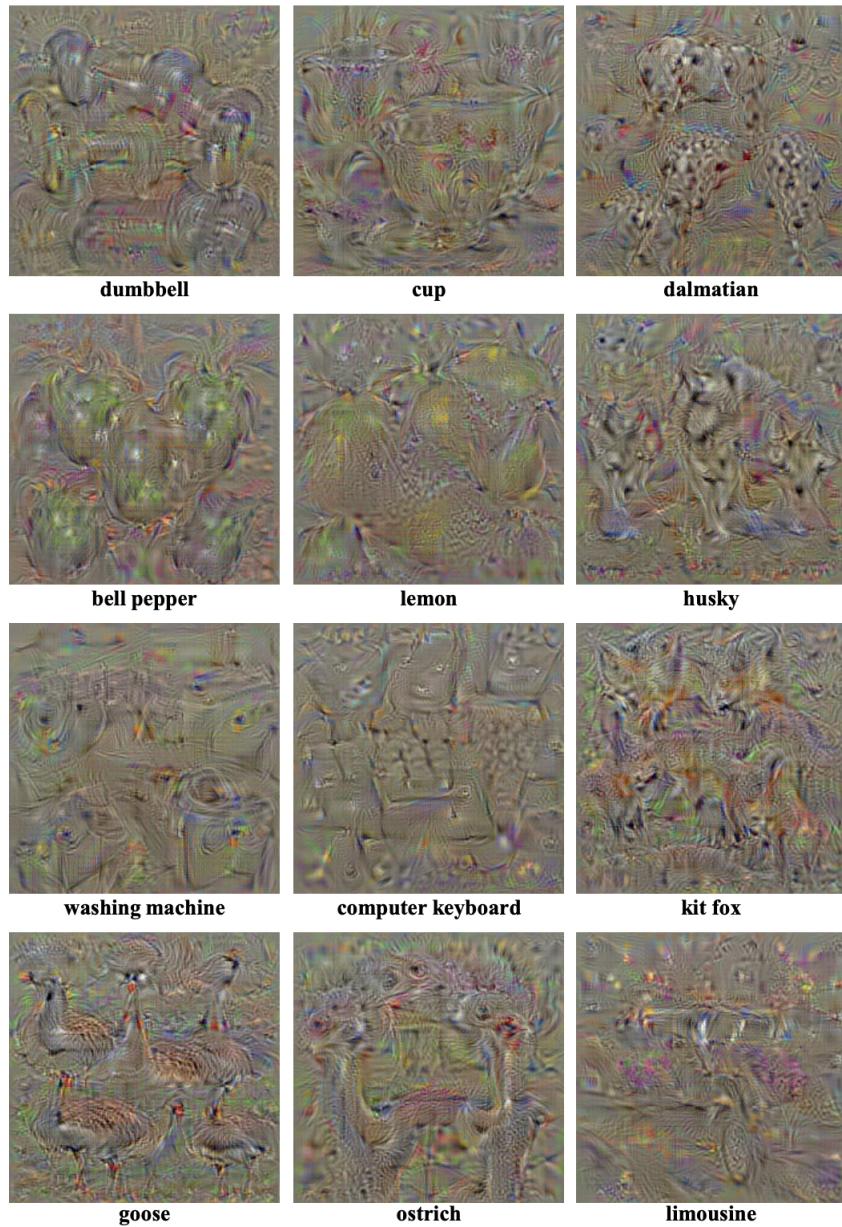
Obr. 2.8: Reziduálny blok v architektúre ResNET. Informácia z predhádzajúceho bloku je súčasťou výstupu aktuálneho bloku pomocou skracovacieho spojenia. [12]

2.2.5 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) [14]. Medzi domény, ktoré nie sú interpretovateľné zaradujú napríklad latentné vektorové reprezentácie slov (angl. word embeddings) alebo iné abstraktné vektorové reprezentácie [14]. Na rozdiel od vstupných 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 [14] (angl. gradient descent). Pri aplikovaní tejto metódy na obrazové dátá výsledné prototypy vyzerajú tak ako na obrázku 2.9.

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 snaží naučiť distribúciu hľadanej triedy [14]. 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) [14].



Obr. 2.9: Maximalizácia aktivácie aplikovaná na obrazové dátu. [15] 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.

2.2.6 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 [14]. 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 (tj. 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.

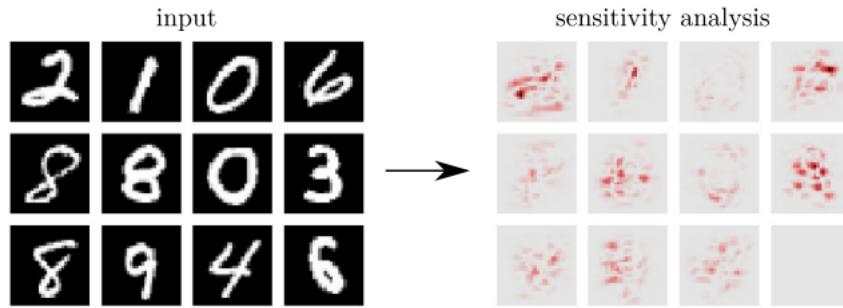
Každá z metód má svoje výhody a nevýhody, napríklad výhodou perturbačných metód je, že môžu byť použité na akýkoľvek model, keďže jediné čo potrebujú je výstup (predikciu) z modelu. Ich nevýhodou však je, že sú pomalé. Niektoré z metód vysvetľovania bližšie opíšeme v tejto sekcií.

2.2.6.1 Analýza senzitivity

Analýza senzitivity slúži na vysvetľovanie predikcie neurónovej siete. Táto metóda identifikuje, ktoré z vlastností vstupného pozorovania najviac prispievajú výslednej predikcii. Najviac dôležité sú také vlastnosti, ktorých zmenou sa najvýraznejšie zmení výsledná predikcia. Na takéto vlastnosti je výsledná predikcia najviac citlivá [14].

Výsledok analýzy senzitivity znázornený v tepelnej mape (angl. heatmap) je zobrazený na obrázku 2.10. 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 rozlísiť. Vieme len, že zmenením danej vlastnosti výrazne ovplyvníme predikciu.



Obr. 2.10: **Analýza senzitivitu aplikovaná na konvolučnú neurónovú sieť trénovanú na dátovej sade MNIST.** [14]

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.

2.2.6.2 LRP

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

Táto technika funguje tak, že vstupný obrázok (metóda sa dá použiť aj na iné ako obrazové dátu, keďže pracujeme práva s obrazovými dátami metódy budeme vysvetľovať práve na nich) dopredným šírením "prejde" neurónovou sieťou, pričom sú zozbierané aktivácie neurónov v jednotlivých vrstvách. Následne je neurónovou sieťou spätným šírením propagované skóre z výstupu neurónovej siete v podobe relevancie až k vstupnému obrázku.

Nasledovné vzorce 2.2, 2.3, 2.4 [14] vyjadrujú spôsob výpočtu propagovanej relevancie medzi vrstvami. j a k sú jednotlivé vrstvy, pričom k je vrstva, z ktorej je

relevancia R propagovaná. Parametre α a β upravujú, koľko pozitívnej (α) alebo negatívnej (β) 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 [14], výsledok použitia rôznych hodnôt α a β je znázornený na obrázku 2.11. $R_{j \leftarrow k}^+$ (Obr. 2.2) a $R_{j \leftarrow k}^-$ (Obr. 2.4) vyjadrujú množstvo pozitívnej (+), resp. negatívnej (-) relevancie propagovanej z vrstvy k do vrstvy j . a_j je aktivácia neurónu, na ktorý je propagovaná relevancia.

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

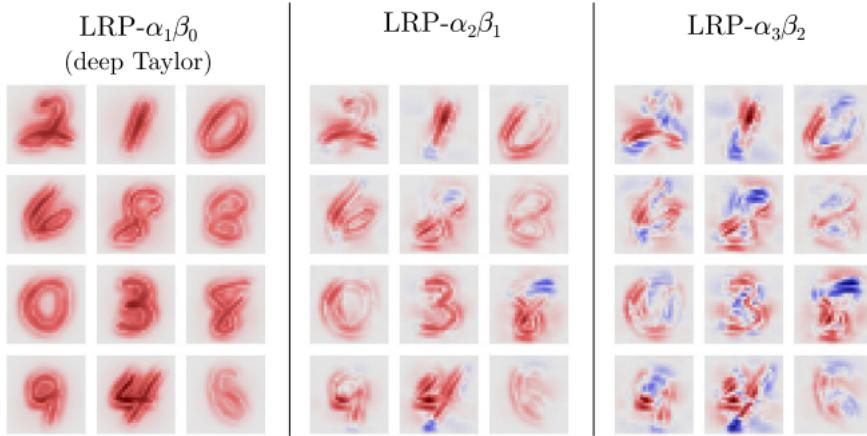
$$R_{j \leftarrow k}^- = \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-} \quad (2.3)$$

$$R_j = \sum_k (\alpha R_{j \leftarrow k}^+ - \beta R_{j \leftarrow k}^-) R_k \quad (2.4)$$

Výhodou LRP oproti iným metódam, ako napríklad dekonvolúciu je, že vysvetlenie (výsledná tepelná mapa) vytvorené technikou LRP je pre rôzne obrázky vždy rôzne [16]. Naopak, pri dekonvolúcii je vysvetlenie vždy rovnaké pokial v architektúre neurónovej siete neboli použité združovacie vrstvy (angl. pooling layers) [16]. Ďaľším rozdielom je (aj oproti analýze senzitivity), že vo výslednom vysvetlení LRP rozlišuje, ktoré vlastnosti pozitívne alebo negatívne prispeli k negatívnej predikcii.

2.2.6.3 Riadená spätná propagácia

Metóda riadená spätná propagácia (angl. Guided Backprop) je rozšírením metódy dekonvolúcie (angl. deconvolution) [14]. Metóda využíva aktivácie $ReLU$ pre smerovanie signálu (tj. výsledného "tepla") na príslušné miesta vstupného obrazu [14]. Pri spätnom šírení sú negatívne gradienty nahradené hodnotou nula. Rovnako aj gradienty z neurónov, ktoré mali pri doprednom šírení po aplikácii aktivačnej



Obr. 2.11: Výsledné vysvetlenie (v podobe tepelnej mapy) vytvorené použitím LRP s rôznymi hodnotami α a β na dátovej sade MNIST. [14] Pozitívna relevancia je zobrazená červenou farbou [14]. Negatívna relevancia je zobrazená modrou farbou [14]. V prípade, že použijeme $\alpha = 1$ a $\beta = 0$ strácamo informáciu o tom, ktoré pixely negatívne (tj. sú proti výslednej predikcii) prispeli k výslednej predikcii (a opačne).

funkcie ($ReLU$) hodnotu 0. Obrázok 2.12 zobrazuje výslednú tepelnú mapu po použití riadenej spätej propagácie, a porovnáva ju s inými metódami. Na rozdiel od metódy LRP, táto metóda nezobrazí na tepelnej mape oblasti, ktoré k výslednej predikcii prispievajú negatívne.

2.2.6.4 GradCAM

GradCAM [17], alebo Gradientom-vážené mapovanie aktivácií triedam (angl. Gradient-weighted Class Activation Mapping) je spôsob vysvetľovania rozhodnutí aplikovateľný na neurónové siete, ktoré používajú konvolučné vrstvy. Cieľom tejto metódy je vysvetliť, ktoré časti vstupu sú dôležité pre vybranú triedu (tj. jednu z tried, ktoré neurónová sieť predikuje). Metóda funguje nasledovne:

- Vstupný obraz "prejde" neurónovou sieťou (inferencia).
- Vypočítame gradient poslednej konvolučnej vrstvy voči výstupu na posled-

nej vrstve neurónovej siete pre vybranú triedu. Môžeme použiť aj inú ako poslednú konvolučnú vrstvu, tá sa však používa najčastejšie.

- Výsledok predchádzajúceho kroku má rovnaký rozmer ako veľkosť poslednej konvolučnej vrstvy, hodnoty v tejto matici sčítame cez dimenziu kanálov. Z konvolučnej vrstvy o rozmere (x, y, c) , kde c je dimenzia kanálov (angl. channels) vznikne matica o rozmere (x, y) .
- Na tento výsledok použijeme funkciu *ReLU* čím odstránime z tepelenej mapy časti obrázku, ktoré prispievajú proti predikovanej triede negatívne.
- Túto matice pomocou bilineárnej interpolácie zväčšíme na veľkosť vstupu. V prípade trojrozmerných dát sa používa trilineárna interpolácia. Taktiež sa môže použiť lineárna interpolácia.

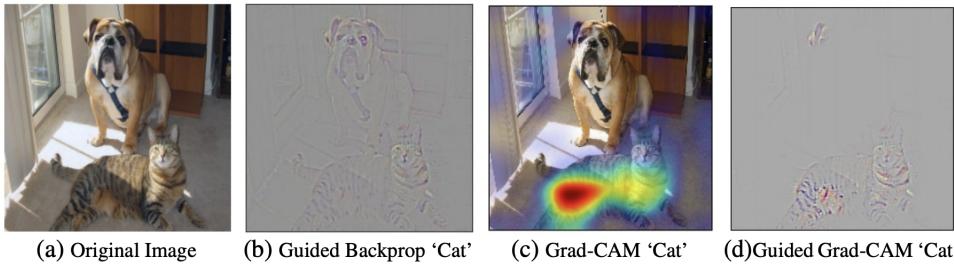
Výsledkom je tepelná mapa pre vstupný obrázok voči vybranej predikovanej triede. Vyššie hodnoty v tepelnej mape vyjadrujú doležitejšie časti obrazu pre vybranú triedu, nižšie hodnoty vyjadrujú tie menej doležité.

2.2.6.5 Riadený GradCAM

Riadený GradCAM (angl. Guided GradCAM), ďalej len Guided GradCAM, kombinuje metódu riadenej spätej propagácie s metódou GradCAM. Výsledné tepelné mapy z oboch metód, pre vstupný obraz a vybranú triedu sú navzájom vynásobené (nejedná sa o maticové násobenie ale násobenie po prvkoch) čím vznikne nová tepelná mapa [17] (Obr. 2.12).

2.2.6.6 RISE

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 metódy *Rise* - (*Randomized Input Sampling for Explanation*) je zrejmé, že táto metóda využíva náhodu na zakrývanie jednotlivých častí



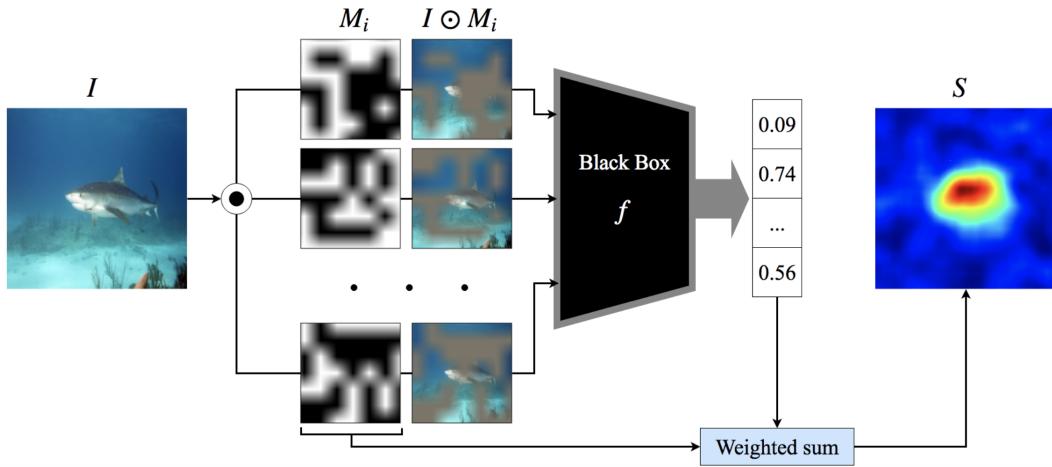
Obr. 2.12: Porovnanie metód Guided Backprop (b), GradCAM (c) a Guided GradCAM (d) [17]. Tepelná mapa (d) je výsledkom násobenia medzi tepelnými mapami (b) a (c), o tom svedčí väčšina "tepla" v spodnej časti obrázku. Teplo v nej má polohu z obrázku (c) a tvar z obrázku (b).

vstupného obrazu. Vstupný obraz je prekrytý náhodou maskou, ktorá je vytvorená nasledovne [18]:

- Je vytvorená náhodná binárna (tj. iba z bielej a čiernej farby) maska o malej veľkosti (napríklad 8px x 8px).
- Táto maska je zväčšená (angl. upsampled) pomocou bilineárnej interpolácie [18] (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 2.13.

Autori porovnali túto metódu s metódami *GradCAM* (Selvaraju et al. 2017) [17] a *LIME* (Ribeiro et al. 2016) [19]. Metóda *RISE* si oproti týmto dvom metódam počínala lepsie (Obr. 2.14). Vykonali niekoľko experimentov, v ktorých porovnali architektúry neurónových sietí *ResNet50* (He et al. 2016) [12] a *VGG16* (Simonyan; Zisserman 2014) [11] natrénované na dátových sadách PASCAL VOC07 (Everin-



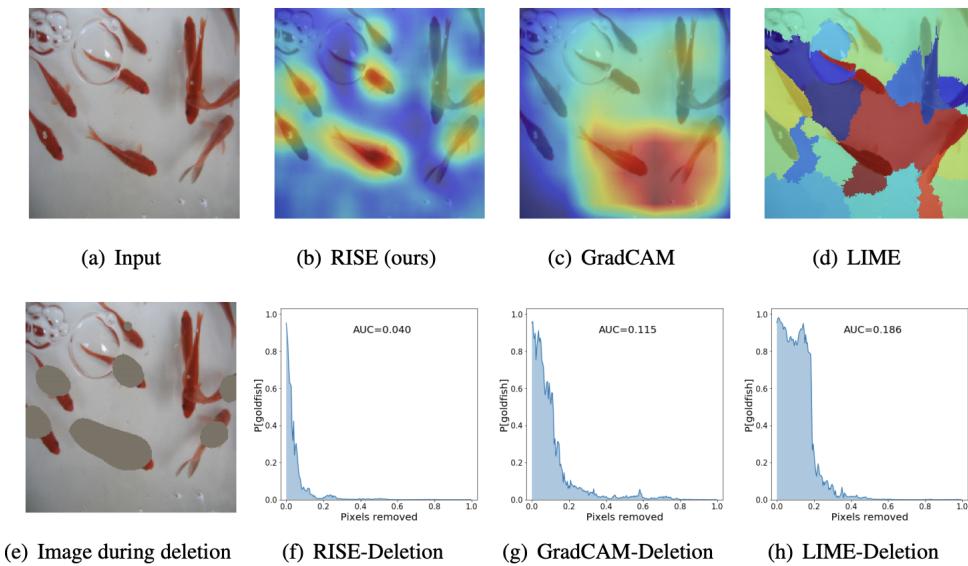
Obr. 2.13: Metóda *Rise*. [18] Vygenerované masky nahradzajú vstupný obrázok na, ktorý sú aplikované. Z výstupných predikcií jednotlivých masiek je nakoniec vypočítaná tepelná mapa.

gham et al. 2010) [20] a MSCOCO2014 (Lin et al. 2014) [21]. Sledovali metriky *insertion* a *deletion* (Obr. 2.14). Metrika *insertion* je vyjadrená ako plocha pod krivkou (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.

Výhodou tejto metódy je, že oproti bežným perturbačným metódam je výrazne rýchlejšia.

2.3 Využitie neurónových sietí pri odhalovaní Alzheimerovej choroby

Neurónovým sieťam sa doposiaľ podarilo dosiahnuť veľmi dobré výsledky pri odhalovaní Alzhemiemerovej choroby. Ako vstup používajú rádiologické snímky ako sú z MRI či PET. Tieto rádiologické ukazovateľe sme bližšie popísali v sekcií 2.1.3. Okrem rádiolgických snímok môžu byť vstupom do neurónovej siete demografické



Obr. 2.14: Porovnanie metódy *RISE* s *GradCAM* alebo *LIME*. [18] V prvom riadku sú tepelné mapy jednotlivých metód pre vstup. V druhom riadku je znázornená porovávaná metrika *deletion*. Táto metrika sleduje vzťah medzi odobratím najdôležitejších pixelov a výslednou predikciou modelu. Je vyčíslená pomocou výpočtu plochy pod krivkou (AUC). 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.

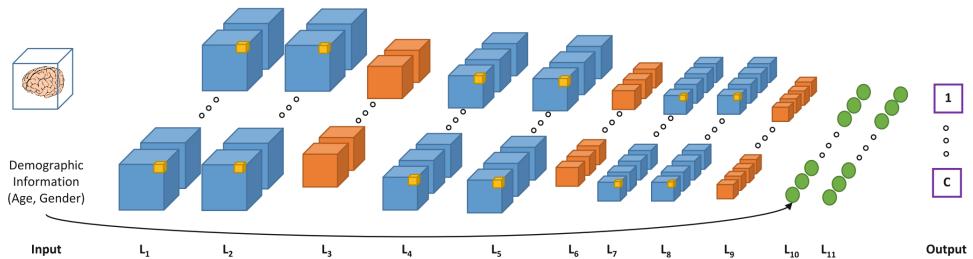
údaje o pacientovi, či výstupy z rôznych klinických alebo kongitívnych testov. Ta- kéto údaje o pacientoch obsahuje populárna dátová množina *ADNI-1* [22].

Neurónové siete natrénované na predikciu Alzheimerovej choroby sa líšia najmä v:

- **predspracovaní** - vstupné dátá sú zmenšené/zväčšené rôznymi algoritmi mi na rôzne veľkosti, častokrát sa z rádiologických snímok odstraňuje lebka
- **type vstupných dát** - môžu to byť rádiologické snímky (MRI, PET), vlasté črty extrahované z rádiologických snímok (MRI, PET), alebo kombinácia takého snímok/črt, s inými, napríklad demografickými údajmi

- **architektúre** - môžu to byť konvolučné siete s 2D konvolúciami (v prípade, že sa používa iba časť rádiologickej snímky, alebo vlastné črty) alebo 3D konvolúciami (ak je vstup celý rádiologický snímok, angl. "full volume"), alebo iné architektúry ako ResNET (reziduálne neurónové suete) alebo VGG
- **ako boli natrénované** - pri niektorých neurónových sieťach autori využili učenie prenosom (angl. transfer learning) a rôzne spôsoby augmentácie vstu-
pus

Ako príklad 3D konvolučnej neurónovej siete uvediem neurónovú sieť od Esmaeilzadeh et al. s presnosťou **94.1%** (a s F_2 skóre 0.93) na populárnej dátovej množine s názvom *ADNI-1* (Obr. 2.15). Tento výsledok dosiahli v úlohe klasifikácie iba do tried CN a AD (bez MCI). Vstupom do tejto neurónovej siete boli snímky z mag-
netickej rezonancie (MRI) ale aj demografické informácie, akými sú napríklad vek alebo pohlavie. Autori článku neuvádzajú úspešnosť modelu, ktorý bol natréno-
vaný iba z obrazových dát, táto úspešnosť by bola pravdepodobne nižšia, nakoľko
vek aj pohlavie sú významnými faktormi ovplyvňujúcimi rozvoj Alzheimerovej
choroby.



Obr. 2.15: Architektúra konvolučnej neurónovej siete použitej pri detekcii Alzheimerovej choroby. [23] Modré kocky sú konvolučné vrstvy, oranžové kocky sú *max-pooling* vrstvy, posledné dve (zelené) vrstvy sú plne prepojené vrstvy. Môžeme si všimnúť, že do posledných dvoch plne prepojených vrstiev okrem obrazových dát vstupujú aj informácie o veku a pohlaví.

V prípade klasifikácie do všetkých troch tried - CN, MCI a AD autorí tejto práce dosiahli horšie výsledky oproti binárnej klasifikácii. Ich model dokázal správne zaradiť pacienta s presnosťou **61.1%** (a s F_2 skóre 0.62) [23]. Pri dosiahnutí tohto

Kapitola 2. Analýza

výsledku použili tzv. učenie s prenosom (angl. transfer learning), ktoré im zlepšilo úspešnosť modelu až o 7.1% z pôvodných 54%. Model, z ktorého učili prenosom je už skôršie spomínaný model na binárnu klasifikáciu pacientov s Alzheimerovou chorobou.

Autori experimentovali trénovaním dvoch rôznych modelov, jedného jednoduchšieho a druhého zložitejšieho. Lepší bol jednoduchší model, pretože neboli tak náchylní na pretrénovanie. V týchto modeloch použili dropout, l_2 regularizáciu a augmentované dátá (obrázky otočili po osi x). Tieto "vylepšenia" pridávali postupne a sledovali rozdiel v úspešnosti modelu, každé jedno z týchto vylepšení výrazne zlepšilo úspešnosť modelu. V kroku predspracovania dát odstránili z obrázkov také časti, ktoré nepredstavovali tkáni mozgu (napr. lebka) technikou s názvom BET (Smith 2002) [24], pretože z nich sa Alzheimerova choroba nedá diagnostikovať.

Niekteré práce (Suk et al. 2016) sa zaoberali dokonca klasifikáciou do štyroch tried: AD, CN, pMCI (angl. progressive MCI - pacienti ktorí pokročili k AD do 18 mesiacov), sMCI (angl. stable MC - pacienti ktorí nepokročili k AD do 18 mesiacov). Táto úloha je samozrejme náročnejšia, najlepší model v tomto prípade dosahoval presnosť 53.72% [25]. V prípade binárnej klasifikácie (AD vs CN) sa autorom podarilo dosiahnuť presnosť až **95.09%**, oproti Esmaeilzadeh et al. však použili aj rádiologické snímky z PET. Táto práca sa ďalej vyznačuje adaptívou selekciou črt, vďaka ktorej sa autorom podarilo dosiahnuť tak dobré výsledky. V tejto práci autori taktiež vykonali odstránenie lebky zo vstupných snímok počas fázy predspracovania.

Učenia prenosom (angl. transfer learning) je veľmi dobrým spôsobom na zrýchlenie trénovania a zlepšenie úspešnosti modelu. Hosseini-Asl et al. využili učenie prenosom a to tak, že najskôr netrénovali 3D konvolučný autoenkovodér, ktorý mal za úlohy rekonštruovať vstup - tj. vstupný radiologický snímok. Z tohto autoenkodéra zobraťali jednu jeho časť - enkovodér za ktorý dali konvolučné vrstvy, ktoré dotrénovali na detekciu Alzheimerovej choroby. Enkovodér teda slúžil na ektrakciu črt.

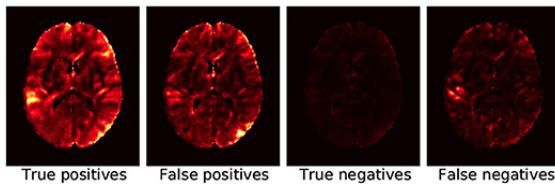
Neurónové siete sa v niektorých prácach používajú v kombinácii s inými algoritmami strojového učenia. Suk et al. použili kombináciu riedkych regresných modelov (angl. sparse regression models) a 2D konvolučnej neurónovej siete, kde výstupy z týchto regresiných modelov slúžili ako vstup do neurónovej siete.

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 2019 uviedli možnosti analýzy rozhodnutí za účelom ich vysvetľovania. Konkrétnie sa zaobrali metódami vrstvami propagovanej relevancie (LRP) a vedenou spätnou propagáciou (angl. Guided Backprop). 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) (Obr. 2.16). Okrem iného porovnávali mieru relevancie pri metóde LRP v jednotlivých častiach mozgu u pozorovaní s Alzheimerovou chorobou a u pozorovaní bez nej. Autori práce zistili, že metóda LRP môže byť v klinickom prostredí veľmi užitočná na hodnotenie inividuálnych prípadov. Zároveň v sledovaných matrikách dosiahla oproti Guided Backprop lepšie výsledky. Možným vylepšením tejto práce je vyskúšanie metódy LRP aj na pacientoch s miernym kognitívnym poškodením (angl. mild-cognitive impairment), nie len na pacientoch s Alzheimerovoch chorobou a zdravých jedincoch.

2.4 Spracovanie obrazu

Kedže pri diagnostike Alzheimerovej choroby sa pracuje s rádiologickými snímkami, čo sú trojrozmerné obrazové dátá, pri jej detekcii neurónovými sieťami je potrebné tieto dátá spracovať technikami spracovania obrazu.



Obr. 2.16: Priemerná relevancia (z metódy LRP - $\beta = 0$)
pozorovaní podľa spôsobu zaradenia výslednej predikcie
Najviac relevancie je na miestach so žltou farbou. [28]

Metódy spracovania obrazu podľa Chen [29] rozdeľujeme do nasledovných kategórií:

- vylepšovanie obrazu (angl. image enhancement)
- rekonštrukcia obrazu (angl. image restoration)
- analýza obrazu (angl. image analysis)
- kompresia obrazu (angl. image compression)

Pri **vylepšovaní obrazu** je obraz upravovaný predovšetkým heuristickými technikami [29], môže sa napríkald jednať o upravenie jasu, kontrastu alebo farieb. Cieľom **rekonštrukcie obrazu** je zrekonštruovať poškodené časti obrazu, napr. pri fotografiách to môžu byť ich vyblednuté časti. Metódy **analýzy obrazu** umožňujú obraz spracovať tak, že je možné z neho automaticky získať (extrahovať) informácie [29]. Príkladmi analýzy obrazu je segmentácia obrazu, extrakcia hrán alebo analýza textúry. **Kompresia obrazu** umožňuje zmenšenie veľkosti obrazu znižovaním počtom potrebných bitov na jeho reprezentáciu [29]. Môže sa jednať o zmenšenie rozmerov obrazu, alebo počtu farieb potrebných na jeho reprezentáciu.

V našej doméne budeme pracovať so všetkými týmito technikami. Ako príklad môžem uviesť odstránenie takých častí obrazu, ktoré nepredstavujú mozgové tkivo (BET - Smith 2002). Táto technika je kombináciou analýzy obrazu - identifikácia častí na odstránenie a vylepšenia obrazu - samotné odstránenie tých častí. Kompresia obrazu sa používa, v časti predspracovania pred tým ako je samotný snímok použitý ako vstup do neurónovej siete. Metódy rekonštrukcie obrazu sa bežne v

tejto oblasti nepoužívajú, avšak my by sme ich chceli v našej práci použiť pri vytváraní novej metódy, preto sa im budeme bližšie venovať.

2.4.1 Rekonštrukcia obrazu

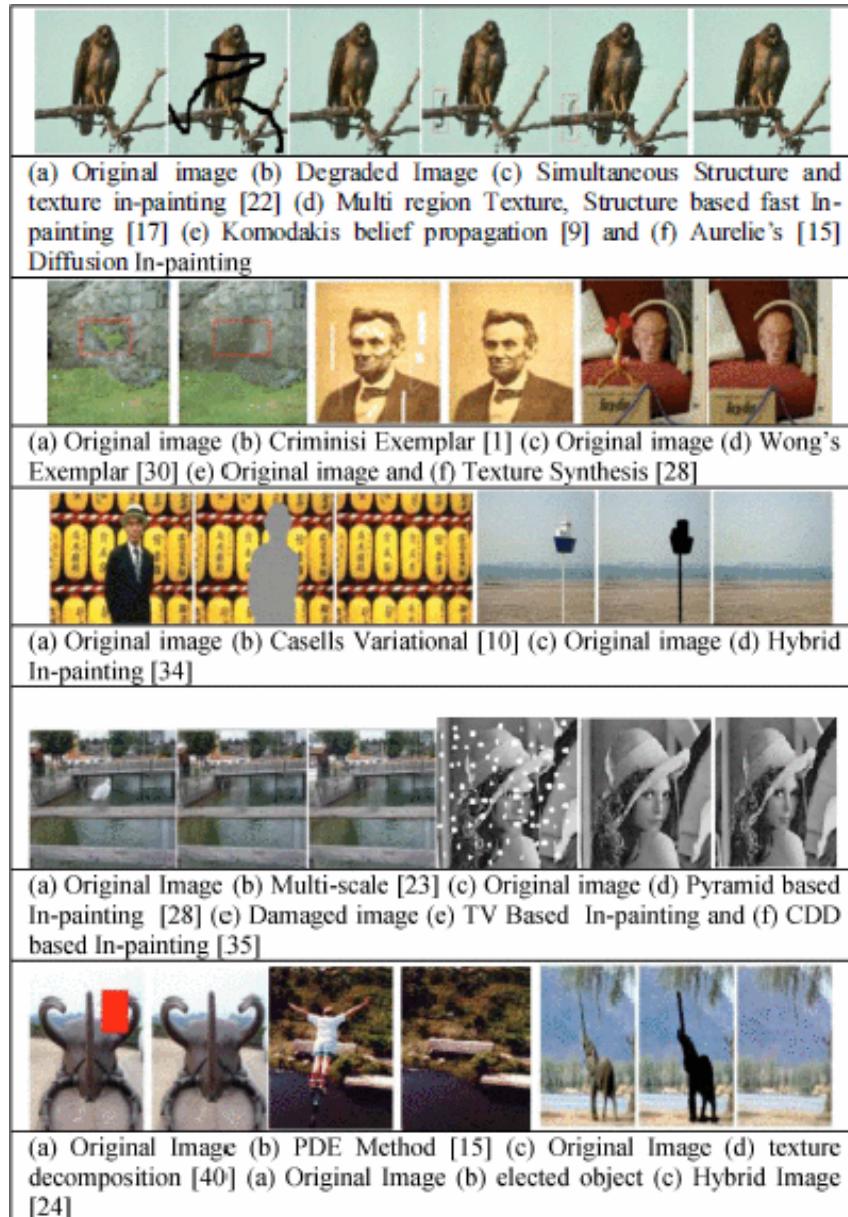
Metódy rekonštrukcie obrazu, alebo inak nazývané aj dokreslenia obrazu (angl. inpainting), podľa Ravi; Pasupathi; Muthukumar.; Krishnan [30] môžeme zaraďiť do nasledovných kategórií:

- dokresľovanie založené na syntéze textúr
- poloautomatické a rýchle digitálne dokresľovanie
- dokresľovanie založené na parciálnej diferenciálnej rovnici
- dokresľovanie na základe predlohy a vyhľadávania
- hybridné dokresľovanie

Tieto metódy sa líšia rýchlosťou dokresľovania, schopnosti dokreslovať veľké/malé plochy a predovšetkým kvalitou dokreslenia. Metódy dokresľovania založené na syntéze textúr fungujú dobre pre väčšie chýbajúce oblasti, avšak v ich výsledku môžu vzniknúť nežiadúce hrany [30]. Dokresľovanie na základe predlohy má zas problémy so zakrivenými štruktúrami [30]. Obr. 2.17 zobrazuje príklady použitia niektorých techník dokreslenia obrazu.

2.5 Zhrnutie

Alzhemierova choroba je bez pochyby veľmi nebezpečnou chorobou, keďže nie je "iba" o strate pamäti ale patrí k častým príčinám smrti (Sek. 2.1). Diagnostika tejto choroby pozostáva najmä z neuropsychometrických testov a analýzy rádiologických snímok (napr. z PET, MRI). V súčasnosti tieto rádiologické snímky posudzujú doktori samotný. Práve tu je priestor pre umelú inteligenciu, aby im pri posudzovaní týchto snímok pomohla.



Obr. 2.17: Príklady dokreslenia obrázkov rôznymi metódami [30].

V doméne obrazových dát sa používajú najmä konvolučné neurónové siete, pretože majú veľmi dobrú schopnosť naučiť sa rozoznávať špecifické objekty z obrázka. Konvolučné neurónové siete sa v nižších vrstvách naučia rozoznávať jednoduchšie tvary/hrany a vo vyšších zložitejšie šruktúry až celé objekty. Keďže jednou z

Kapitola 2. Analýza

možností diagnostiky Alzheimerovej choroby je diagnostika pomocou rádiologických snímok, je možné použiť neurónové siete práve pri detekcii tohto ochorenia.

Neurónovým sieťam sa doteraz podarilo dosiahnuť veľmi dobré výsledky pri detekcii Alzheimerovej choroby, niektoré state-of-the-art riešenia dosahujú presnosť až **95.09%** (Suk et al. 2016). S takto vysokou úspešnosťou môžu byť veľmi dobrým pomocníkom doktorov. Do úvahy však musíme zobrať, že tieto výsledky boli dosiahnuté bez klasifikácie MCI pacientov. V reálnom svete doktora navštívia všetky typy pacientov - CN, MCI a AD. V tomto prípade neurónové siete dosahujú rádovo nižšiu presnosť (**61.1%**, Böhle et al. 2019). Niektoré práce dosiahli tieto výsledky použitím informácií o veku a pohlaví pacienta. Keďže pravdepodobnosť výskytu Alzheimerovej choroby po dovršení 85 rokov života je až 50% (Sek. 2.1), je možné, že sa pri vyššom veku pacienta model začne rozhodovať najmä na základe tejto informácie a nie na základe obrazových dát. Zároveň to však môže neurónovej sieti pomôcť, ak nebude brať tento atribút ako hlavný indikátor Alzheimerovej choroby, ale skôr ako pomocný atribút, ktorý bude meniť jej správanie u rôznych typov pacientov. Tu je však dôležité, takúto neurónovú sieť podrobiť dôkladnej analýze jej rozhodnutí. Osobne si ale myslím, že v produkčnom modeli by sa tento atribút mal vynechať.

Ďalším problémom neurónových sietí je, že sa správajú ako čierne skrinky. Preto je potrebné ich rozhodnutia interpretovať, aby bolo pre doktora zrejmé na základe čoho neurónová sieť urobila svoju predikciu. V tomto práve môžu pomôcť metódy na vysvetľovanie rozhodnutí neurónovej siete, alebo iné metódy, ktoré sú nezávislé od použitého modelu (napr. RISE, LIME).

Bežnému používaniu neurónových sietí ako pomocníka pre doktorov, nebráni len ich vysvetliteľnosť, ale aj ich schopnosť detektie ochorenia, keďže aj tu je priestor na zlepšenie - napr. úspešnosti klasifikácie do tried CN, MCI a AD.

Pre pochopenie správania sa neurónových sietí poznáme metódy jej interpretovania a vysvetľovania jej rozhodnutí. Interpretovaním neurónovej siete zisťujeme, ako vyzerá vzorové pozorovanie pre jednu z tried, ktorú klasifikuje. Vysvetľovaním jej rozhodnutí zas zisťujeme na základe čoho neurónová sieť spravila svoje rozhod-

Kapitola 2. Analýza

nutie, a teda ktoré zo vstupných vlastností pozorovania ju navideli k zaradeniu do určitej triedy. Niektoré z týchto metód (LRP a vedená spätná propagácia) už boli použité pri vysvetľovaní rozhodnutí neurónových sietí detekujúcich Alzheimerovu chorobu, avšak zatiaľ len pri binárnej klasifikácii pacientov.

3. Ciele práce

Vychádzajúc zo zadania projektu a na základe poznatkov nadobudnutých z analýzy domény a problému, sme si stanovili nasledovné ciele.

3.1 Vytvorenie novej, alebo vylepšenie existujúcej metódy pre vysvetľovanie rozhodnutí neurónových sietí

Existujú rôzne metódy pre vysvetľovanie rozhodnutí neurónových sietí. Niektoré z nich potrebujú poznať model, ako napríklad LRP (ktorá pracuje iba s neurónovými sieťami), iné nepotrebuju, a je ich teda možné použiť na ľubovoľný typ modelu. Každá z metód má iné výhody/nevýhody preto je tu priestore na vytvorenie alebo vylepšenie existujúcej metódy. V prípade vylepšenia existujúcej metódy je nutné túto metódu porovnať najmä s vylepšovanou metódou a následne s inými metodami. Cieľom je teda vytvoriť novú metódu, ktorá vytvára presnejšie vysvetlenia ako iné metódy, alebo vylepšíť existujúcu metódu, ktorá vytvára presnejšie vysvetlenia ako metóda, z ktorej vychádza. Zároveň táto metóda ma byť použiteľná na medicínske obrazové dátá.

3.2 Využitie vytvorenej metódy na určenie miery správnosti modelu neurónovej siete detegujúcej Alzheimerovu chorobu

Pri neurónových sieťach detekujúcich Alzheimerovu chorobu je dôležité, aby sa naučili klasifikovať pacientov na základe relevantných črt z rádiologických snímkov. Práve preto je potrebné určiť mieru správnosti modelu podľa toho či sa model rozhoduje práve na základe týchto črt a nie iných. Na to sa využívajú metódy na vysvetľovanie rozhodnutí neurónových sietí, v tomto prípade sa použije novovytvorená metóda. Cieľom je teda určiť správnosť modelu detegujúceho Alzheimerovu chorobu pomocou vytvorenej metódy pre vysvetľovanie rozhodnutí neurónovej siete.

4. Návrh riešenia

Pre použitie neurónových sietí v bežnej praxi doktorov pri diagnostike Alzheimerovej choroby je nevyhnutné, aby sa rozhodnutia neurónových sietí dali vysvetliť. Preto navrhujeme metódu na vyvsetľovanie rozhodnutí neurónových sietí, ktorú overíme na MRI snímkach pri klasifikácii týchto snímok do dvoch skupín podľa diagnóz pacienta (CN, AD).

Vychádzajúc cieľa práce *3.1 Vytvorenie novej alebo vylepšenie existujúcej metódy pre vysvetľovanie rozhodnutí neurónových sietí* navrhujeme metódu, ktorá vychádza z už existujúcej metódy *RISE* (Sek. 2.2.6.6). Táto metóda dosiahla veľmi dobré výsledky oproti metódam GradCAM a LIME a považujem ju teda vhodný základ pre ďalšie vylepšenia. Metóda RISE funguje na princípe zakrývania častí obrázka (tak ako iné perturbačné/oklúzne metódy) jednou hodnotou (tj. farbou). Po takomto prekrytí nevznikajú žiadné ostré hrany, ktoré by mohli neurónovú sieť myliť ako u iných metódach, ktoré fungujú na princípe zakrývania častí obrazu.

Autori metódy RISE používali obrázky vo farebnom priestore RGB a prekryvali ich čiernou farbou ($r = 0, g = 0, b = 0$). MRI snímky nepoužívajú žiadnu farebnú schému, ale zachytávajú intenzitu (hodnoty sú zväčša reálne čísla). V tomto prípade môžeme zakrývať maximálnou alebo minimálnou hodnotou (minimálna hodnota je ekvivalentná RGB v prípade šedej). Toto zakrytie môže byť práve ďalším zdrojom zmätenia pre neurónovú sieť, keďže úbytky tkaniva sú vyjadrené nízkymi hodnotami na snímkoch.

Preto navrhujeme zakrývané miesta dokresliť určitou metódou spracovania ob-

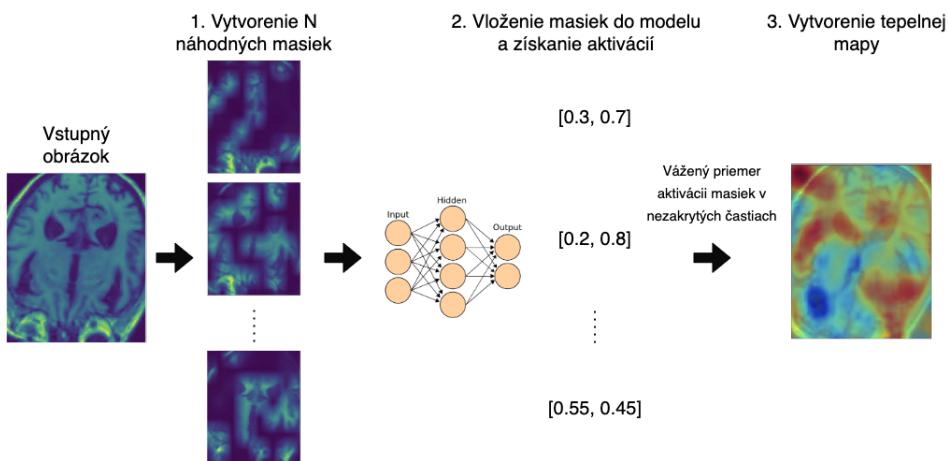
razu (Sek. 2.4) alebo na zakrytie použiť inú hodnotu. Pôvodná metóda bola ale narvhnutá pre obrázky (tj. 2D) a nie 3D volumetrické dátá, preto metódu RI-SEI upravujeme tak, aby vedela pracovať s 3D dátami - tj. budeme generovať 3D masky.

4.1 RISEI

Metódu sme pomenovali *Randomized Input Sampling for Explanation with In-painting* (tj. náhodné vzorkovanie vstupu pre vysvetlovanie s dokreslovaním) so skratkou RISEI.

Keďže metóda vychádza už z existujúcej metódy, časť našej metódy je samozrejme rovnaká. Proces vytvorenia vysvetlenia klasifikácie (Obr. 4.1) do triedy T pre obrázok O modelom je teda nasledovný:

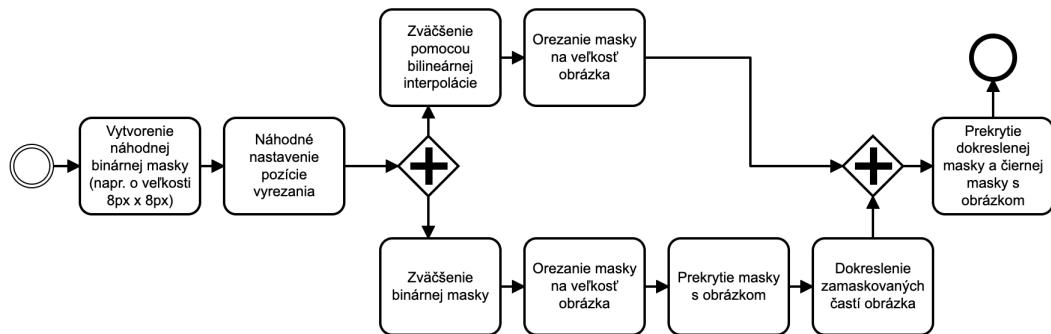
1. Vytvorenie N náhodne zamaskovaných obrázkov z obrázka O .
2. Vloženie zamaskovaných obrázkov do modelu a následné získanie aktivácie pre triedu T .
3. Vytvorenie a vizualizácia vysvetlenia pomocou tepelnej mapy.



Obr. 4.1: Proces vysvetlenia klasifikácie - vytvorenia tepelnej mapy.

Toto sú 3 hlavné kroky z ktorých pozostáva táto metóda, ďalej bližšie popíšeme jednotlivé z nich.

1. Vytvorenie N náhodne zamaskovaných obrázkov z obrázka O Vytvorenie náhodne zamaskovaných obrázkov tiež pozostáva z niekoľkých krov, pričom niektoré z nich môžu byť vykonávané paralelne. Tento krok sme znázornili diagramom (Obr. 4.2). Masky sa vytvárajú paralelne, pretože ”čierna” maska ma jemné hrany a na dokreslenie potrebujeme naopak masku s ostrými hranami.

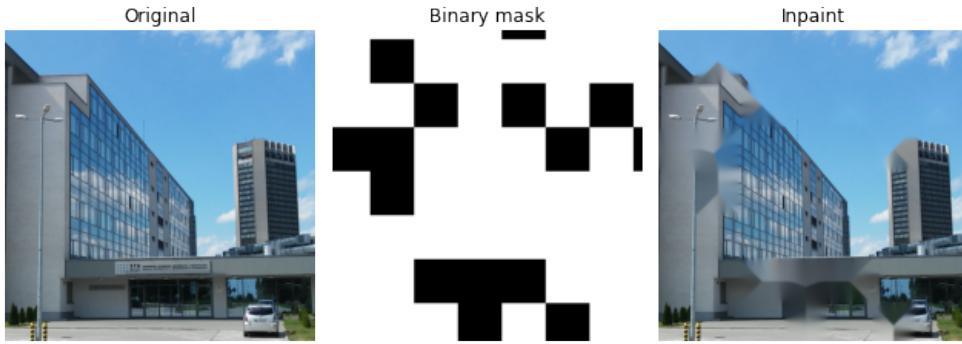


Obr. 4.2: BPMN diagram generovania jedného obrázka prekrytého maskou

Oproti metóde *Rise* vytvárame o jednu masku naviac, a teda je originálny obrázok prekrytý s viacerými maskami. Jednotlivé masky cez seba prekryjeme, pričom každej z nich nastavíme určité množstvo priehľadnosti. S týmto pomerom môžeme ďalej experimentovať a výsledky porovnávať. Môžeme porovnať použitie iba dokreslenej masky s iba ”čiernej” maskou a tiež s použitím oboch v rôznych pomeroch.

Vytvorenie ”čiernej” masky je rovnaké, ako pri metóde *Rise*. Dokreslená maska vznikne dokreslením zakrytych (zamaskovaných) častí obrázka pomocou jedného z algoritmov na dokreslovanie (angl. inpainting). Tieto algoritmy sme popísali v sekcií 2.4 Spracovanie obrazu. Obrázok 4.3 je príkladom dokreslenia častí vzorového obrázka na základe masky náhodne vygenerovanej masky (tentotýž príklad je v 2D, naša metóda bude pracovať s 3D). V našej metóde budeme experimentovať s

rôznymi hodnotami prekrytie (priemer, maximum, minimum, medián).



Obr. 4.3: Niektoré časti vzorového obrázka (vľavo) boli dokreslené podľa náhodne vygenerovanej binárnej masky (v strede). Výsledný obrázok (vpravo) môže byť ešte prekrytý ”čiernou” maskou s určitou priehľadnosťou.

2. Vloženie zamaskovaných obrázkov do modelu a následné získanie aktivácie pre triedu T . Tento krok je identický s originálnou metódou *Rise*.

3. Vytvorenie a vizualizácia vysvetlenia pomocou tepelnej mapy. Tento krok je identický s originálnou metódou *Rise*. Nasledovný vzorec 4.1 vyjadruje výpočet dôležitosti I pre každý voxel $[x, y, z]$ snímky, kde n je počet všetkých zamaskovaných snímok. Funkcia $p(k, x, y, z)$ vracia vracia predikciu (tj. aktiváciu v kontexte neurónových sietí) pre predikovanú triedu (v prípade binárnej klasiifikácie) z modelu pre zamaskovanú snímku k . Funkcia $c(k, x, y, z)$ vracia mieru zakrytie/dokreslenia maskou, pričom $H(c) = <0, 1>$, kde 1 znamená úplné prekrytie/dokreslenie a 0 žiadne prekrytie/dokreslenie. Rovnako, ako metóde *Rise*, počítame vážený priemer.

$$I_{x,y,z} = \frac{\sum_k^n p(k, x, y, z) * (1 - c(k, x, y, z))}{\sum_k^n p(k, x, y, z)} \quad (4.1)$$

Navrhovaná metóda do originálnej metódy pridáva niekoľko parametrov a najmä výpočtovo náročné dokreslovanie, preto bude nutné nájsť vhodné nastavenie pa-

rametrov, aby výpočet vysvetlenia neboli príliš časovo náročný. Práve výpočtová náročnosť môže byť jednou zo slabín tejto metódy. Takisto aj samotná dokreslená časť obrázka môže byť príčinou zmätenia neurónovej siete.

4.2 Overenie riešenia

Našu metódu v prvom rade porovnávame s originálnou metódou RISE (tj. či sa nám podarilo vytvoriť lepšiu metódu) a následne s inou existujúcou metódou (LRP, GradCAM, Guided Backprop alebo Guided GradCAM). Z týchto metód je najviac vhodná metóda LRP, keďže už bolo jej použitie pri vysvetľovaní rozhodnutí neurónových sietí detegujúcich Alzheimerovu chorobu (Sekcia 2.3.1) skúmané. Tieto experimenty vykonávame na CN a AD vzorkách. Sledujeme kvalitu navrhnutej metódy (oproti ostatným metódam) a na základe týchto tepelných máp vyhodnocujeme mieru správnosti modelu.

4.2.1 Dátová sada

Experimenty budeme vykonávať na dátovej sade ADNI, ktorá obsahuje MRI snímky AD pacientov. Táto dátová sada bola použitá aj na trénovanie state-of-the-art modelu na diagnostiku Alzheimerovej choroby [23], ale aj pri vysvetľovaní rozhodnutí neurónovej siete pomocou LRP [28]. Na tejto dátovej sade budeme musieť vykonať rovnaké predspracovanie ako Böhle et al., aby sme sa s ich výsledkami mohli porovnať. Prípadne môžeme vykonať vlastné predspracovanie, ale budeme musieť vykonať aj experimenty s metódou LRP.

4.3 Model

Na vytváranie tepelných máp je nevyhnutný model. Navrhnutú metódu porovnávame na niekoľkých modeloch - architektúrach neurónových sietí. Z nich vyberieme

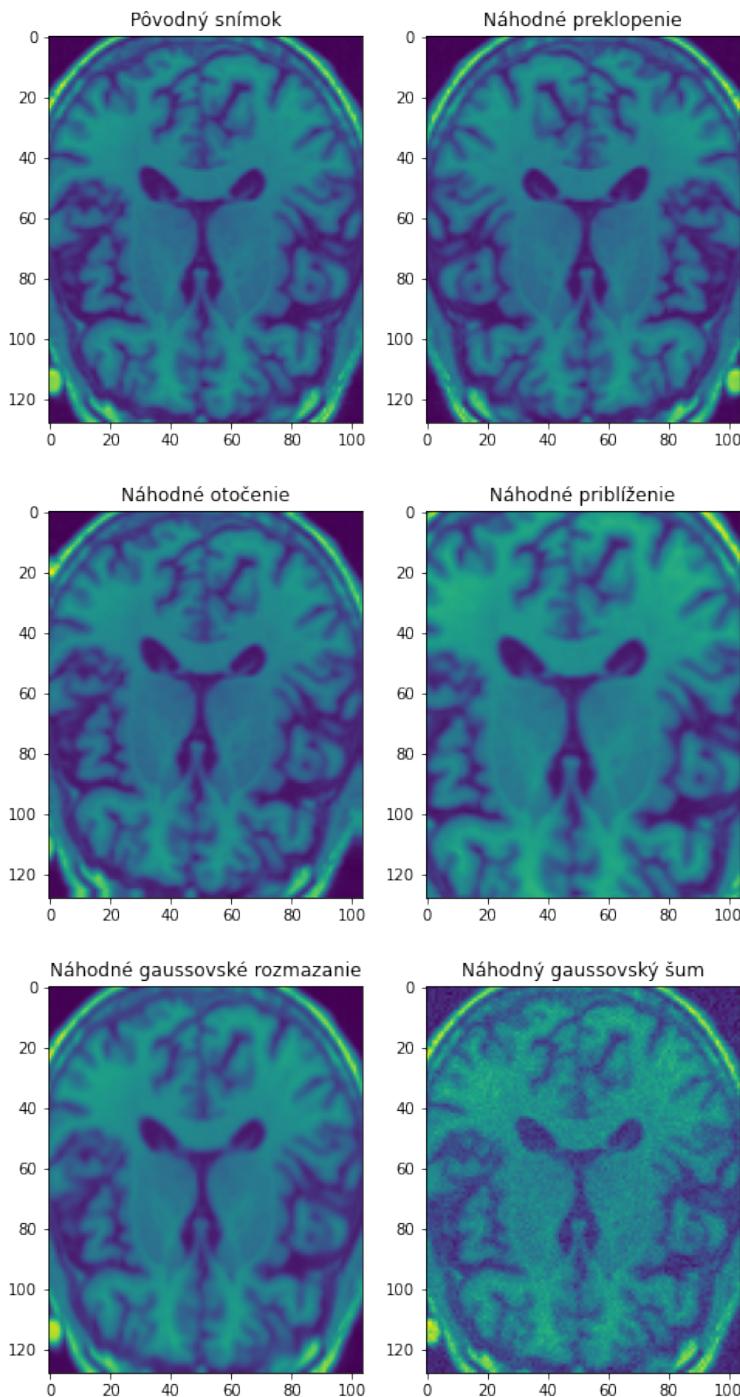
najvhodnejší pre ďalšie experimenty.

- **3D konvolučná neurónová sieť od Esmaeilzadeh et al.** Túto neurónovú sieť sme vybrali, pretože jej autori pomocou nej dosiahli veľmi dobré výsledky (94.1% presnosť).
- **2D ResNet a 3D ResNet** Keďže reziduálne neúronové siete dosahujú pri klasifikačných úlohách nad obrazovými dátami veľmi dobre výsledky použijeme aj tieto architektúry.

Do týchto neurónových sietí sme ešte pridávame dropout a dávkovú normalizáciu (angl. batch normalization). Dropout pridávame pred plne prepojené vrstvy. Dávkovú normalizáciu pridávame v konvolučných vrstvách pred aplikovaním nelinearity, tak ako je to odporučené od Ioffe et al. v *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. Na poslednej vrstve používame dva neuróny s aktivačnou funkciou *softmax*.

Snímky s dátovej sady sme sa rozhodli augmentovať (Obr. 4.4) s cieľom zväčšenia počtu rôznych pozorovaní. Dáta náhodne augmentujeme v každej dávke (angl. batch). Používame nasledovné augmentácie:

- Vymenenie hemisfér mozgu (Esmaeilzadeh et al. [23]) s pravdepodobnosťou 50%
- Náhodná rotácia o 0 až 5 stupňov s pravdepodobnosťou 20%
- Náhodné priblíženie do 80% veľkosti snímku s pravdepodobnosťou 20%
- Náhodné gaussovské rozmazanie ($\max \sigma = <0.85, 1>$) s pravdepodobnosťou 20%
- Náhodný gaussovský šum pravdepodobnosťou 20%



Obr. 4.4: Príklady aplikácie augmentácií.

4.3.1 Experimenty

Najskôr vyhodnocujeme nami navrhnutú metódu pomocou sledovania kvality tepelných máp. Následne overujeme správnosť modelu pomocou nami navrhнутej metódy, avšak je nutné, aby metóda generovala kavlitné tepelné mapy.

4.3.1.1 Určenie kvality metódy vysvetľovania rozhodnutí modelu

Kvalitu metódy vysvetľovania rozhodnutí modelu sledujeme určovaním kvality tepnej mapy. Tá v kontexte našej práce hovorí o tom, do akej miery táto mapa odzrkadľuje to, na základe čoho sa model rozhoduje. Toto budeme merať metrikami *insertion (AUC)* a *deletion (AUC)*, ktoré sme bližšie popísali v sekciu 2.2.6.6. Tieto metriky nám povedia, aká dobrá je naša metóda na vysvetľovanie.

Keďže naša metóda generuje tepelné mapy pomocou vygenerovania veľkého množstva náhodných masiek, je vhodné skúmať, ako sú tieto tepelné mapy konzistentné pri niekoľkých použitiach metódy rovnakom MRI snímku. Konzistentnosť máp môžeme merať pomocou podobnosti medzi jednotlivými tepelnými mapami vygenerovaními pre tú istú snímku a ten istý model (napr. ako súčet absolútlných hodnôt rozdielov medzi voxelmi v oboch tepelných mapách). Čím je táto podobnosť väčšia, tým je metóda pri generovaní máp viac konzistentná.

4.3.1.2 Určenie správnosti modelu

Správnosť modelu sme určujeme na základe tepelných máp vytvorených pomocou metódy na vysvetľovanie predikcií modelu. Overujeme do akej miery dávajú tepelné mapy zmysel v kontexte skutočnej anatómie mozgu, tj. či tepelná mapa pre správnu predikciu ukazuje na klinicky relevantné oblasti mozgu. Sledujeme, či tepelná mapa nehovorí o tom, že sa model rozhodol na základe takej oblasti mozgu, z ktorej sa Alzheimerova choroba nedá zistiť. Veľkú úlohu pri určovaní správnosti modelu zohráva aj kvalita natréновaného modelu, tú môžeme merať pomocou metrík z práce od Böhle et al. v ktorej sa autori zaoberali vyhodnocovaním tepelných

máp vypočítaných pomocou metódy LRP. Tieto metriky sú nasledovné (relevancia je v našom prípade teplota na tepelnej mape):

- súčet relevancie v jednotlivých častiach mozgu (podľa segmentačných mapek) pre AD a CN
- hustota relevancie v jednotlivých častiach mozgu (podľa segmentačných mapek) pre AD a CN, berie ohľad na veľkosť danej časti mozgu
- prírastok relevancie v jednotlivých častiach mozgu (podľa segmentačných mapek) vypočítaný ako pomer priemernej relevancie každej triedy v danej časti mozgu

4.4 Zhrnutie

V tejto kapitole sme navrhli metódu na vysvetľovanie rozhodnutí modelov strojového učenia a spôsob jej implementácie. Navrhnutú metódu budeme overovať na neurónových sieťach detegujúcich Alzheimerovu chorobu s cieľom odhaľovania nesprávnych rozhodnutí.

5. Implementácia

5.1 Metóda RISEI

Metódu RISEI sme sa rozhodli implementovať v jazyku Python, keďže plánujeme používať knižnice pre strojové učenie akými sú *tensorflow* či *scikit-learn*.

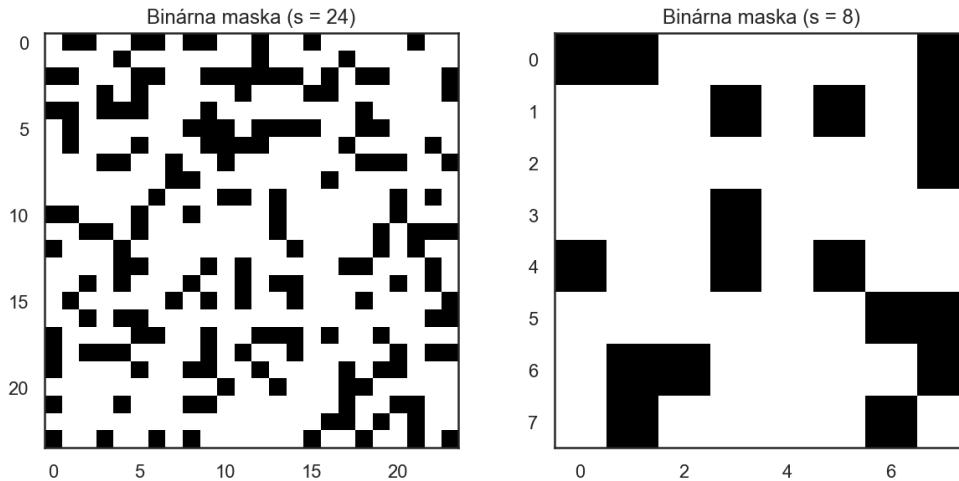
5.1.1 Generovanie masiek

Na základe BPMN diagramu (Obr. 4.2) sme implementovali proces generovania masiek. Generovanie masiek prebieha paralelne vo viacerých procesoch použitím knižnice *multiprocessing*. Metóda RISEI síce pracuje s trojrozmernými dátami, avšak diagramy v tejto sekcii zobrazujú snímky a masky v 2D (konkrétnie určitú vrstvu z 3D snímku) kvôli jednoduchšej vizualizácii. V tejto sekcii popíšeme jednotlivé kroky generovania masiek.

Vytvorenie náhodnej binárnej masky Náhodné binárne masky generujeme pomocou knižnice *numpy*. Pomocou nasledovného kódu vygenerujeme N náhodných masiek 3D binárnych matice. Obr. zobrazuje takúto binárnu maticu, ale v 2D. *size* (veľkosť) a *probability* (pravdepodobnosť) sú hyper-parametrami RISEI metódy. *size* hovorí o veľkosti generovanej masky, čím je toto číslo väčšie tým bude výsledná maska viac fragmentovaná na malé plochy. *probability* hovorí o tom, s akou pravdepodobnosťou daná plocha neprekrytá maskou. Metóda RISE, z ktor-

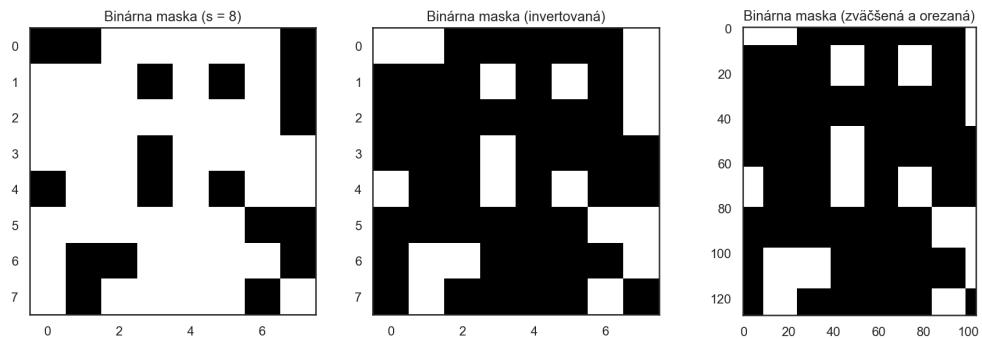
rej vychádzame používa predvolenú hodnotu $size = 8$, preto sme použili rovnakú hodnotu.

```
binary_masks = np.random.rand(N, size, size, size) < probability
```



Obr. 5.1: Porovnanie dvoch binárnych masiek s rôznou veľkosťou ($size$), čím väčšia veľkosť, tým je obrázok viac fragmentovaný. Keďže fragmentácia vyššia, zakrývame menšie časti mozgu, predpokladáme, že takto sa nám nepodarí zakryť relevantné časti z čo zapríčiní nižšiu kvalitu tepelnej mapy (predpokladáme, že v takomto prípade bude "teplo" rovnomerne rozmiestnené po celej snínke).

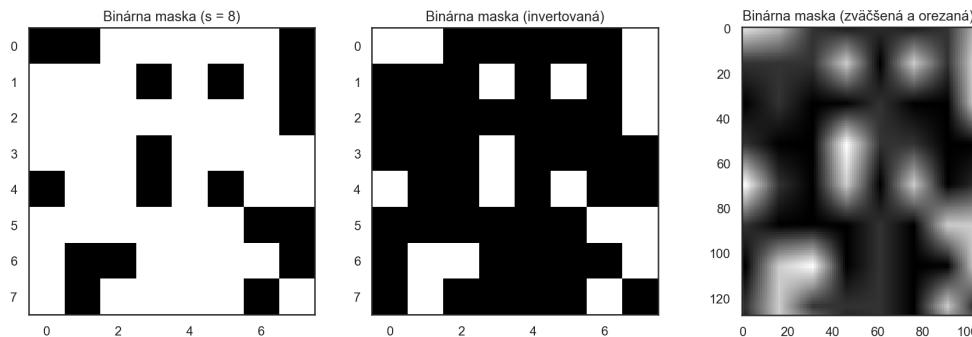
Náhodné nastavenie pozície vyrezania, zväčšenie binárnej masky a orezenie na veľkosť obrázka Binárnu masku zväčšíme na veľkosť vstupnej snímky plus menší offset (o veľkosti $size$). Následne zo zväčšenej masky na náhodnej pozícii vyrežeme masku o veľkosťi vstupnej snímky (Obr. 5.2). Táto maska určuje, ktoré miesta na snímke je potrebné dokresliť (biele miesta sú určené na dokreslenie). Tento krok v pôvodnej implementácii RISE nie je.



Obr. 5.2: Vygenerovaná maska je zväčšená a orezaná na veľkosť vstupnej snímky (o veľkosti [104, 128, 104] pričom na obrázkoch je vizualizovaná druhá a tretia dimenzia). Úplne vľavo je binárna maska o veľkosti 8. V strede je invertovaná binárna maska (kvôli ďalšiemu pracovanou s ňou) a vpravo je orezaná binárna maska o veľkosti vstupnej snímky.

Zväčšenie pomocou bilineárnej interpolácie a orezanie masky na veľkosť obrázka

Tak ako v poôvodnej implementácii RISE, vytárame "čiernu" masku na zakrytie častí obrázku. Pôvodnú binárnu masku pomocou bilineárnej interpolácie (funkcia `resize` z knižnice `scikit-learn`) zväčšujeme na veľkosť o niečo väčšiu ako je vstupná snímka (aby sme mohli vykonať náhodný posun), následne vyrežeme na náhodnej pozícii masku o veľkosti vstupnej snímky (táto náhodná pozícia je rovnaká ako pri orezávaní binárnej masky bez interpolácie, preto je v BPMN diagrame v samostatnom kroku).



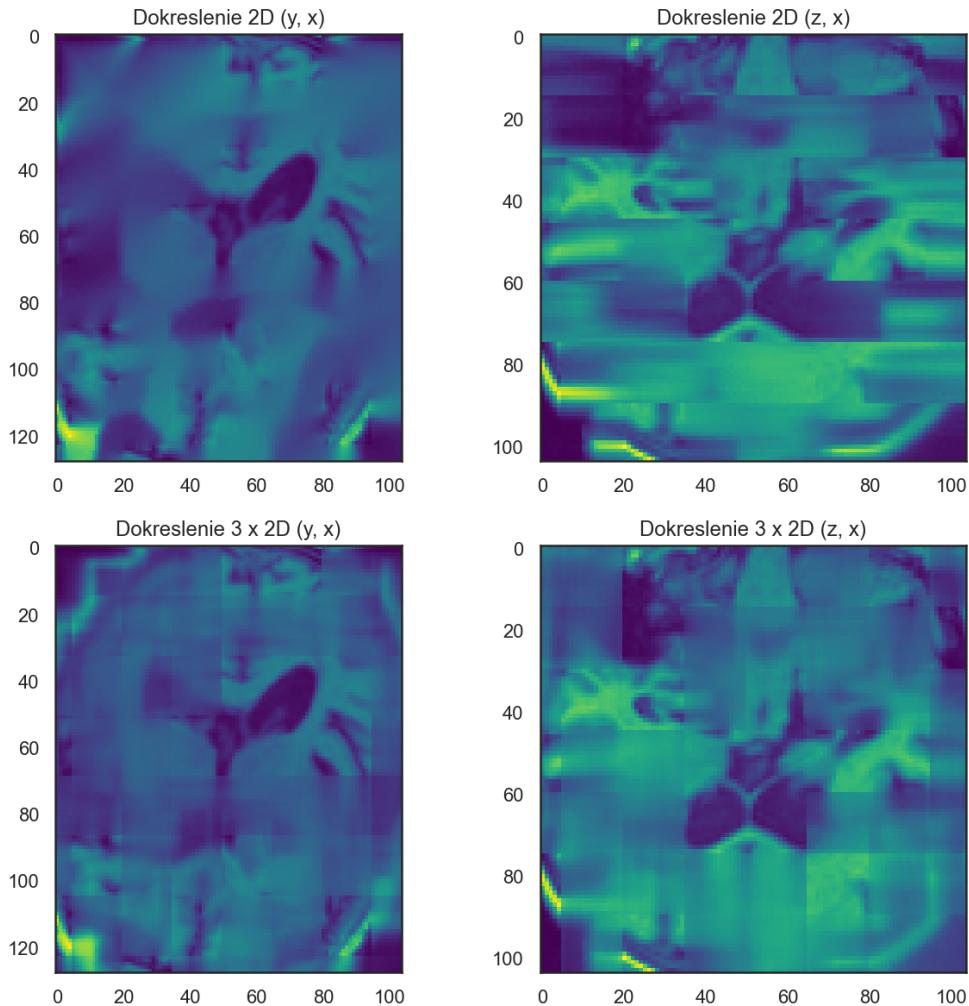
Obr. 5.3: Vygenerovaná maska je zväčšená pomocou bilineárnej interpolácie a orezaná na veľkosť vstupnej snímky (tá je o veľkosti [104, 128, 104] pričom na obrázkoch je vizualizovaná druhá a tretia dimenzia). Úplne vľavo je binárna maska o veľkosti 8. V strede je invertovaná binárna maska (kvôli ďalšej práci s ňou) a vpravo je orezaná interpolovaná ”čierna” maska o veľkosti vstupnej snímky.

Prekrytie masky s obrázkom a dokreslenie zamaskovaných častí obrázka

Keďže pracujeme nad trojrozmernými dátami, pokúsili sme sa použiť dokreslovanie obrázka v 3D. Na to sme sa pokúsili použiť funkciu *inpaint* s knižnicou *scikit-image*, avšak dokreslenie jednej masky bolo veľmi časovo náročné (trvanie bolo až v minútach kde dokreslenie v 2D je v sekundách) pričom v rámci navrhovanej metódy je ich potrebné generovať tisíce, preto sme od trojrozmerného dokreslovania upustili.

Dokreslovanie dvojrozmerných snímok z 3D snímku má avšak svoje nevýhody. Nech máme snímky o veľkosti $[z, y, x]$, pri 2D dokreslení musíme dokreslovať z snímok o veľkosti $[y, x]$ (alebo y snímok o veľkosti $[y, x]$, alebo x snímok o veľkosti $[y, z]$). Pri takomto dokreslovaní, dokreslenie z pohľadu $[y, x]$ vyzerajú byť správne, avšak z iného pohľadu, napr. $[z, x]$ sa javí byť dokreslenie nesprávne, najmä kvôli vzniknutým ostrým hranám (Obr. 5.4). Tento problém sme adresovali tak, že dokreslovanie vykonávame vo všetkých troch rovinách a následne počítame priemer pre každý voxel zo všetkých troch dokreslení. Takto je výsledok o niečo lepší, tj. z každej strany je dokreslenie lepšie ako nesprávne dokreslenie z 2D ale o niečo horšie ako správne dokreslenie z 2D. Na označenie miest, ktoré treba do-

kresliť sme použili zváčšenú binárnu masku (Obr. 5.2). Dokreslenie vykonávame funkciou *inpaint* z knižnice *cv2* (*Open CV*). Používame dokreslovací algoritmus *cv2.INPAINT_TELEA*, keďže pomocou neho sme dosahovali vizuálne najlepšie výsledky. Funkcia *cv2.inpaint* vyžaduje ako parameter *inpaint_radius* (Obr. 5.6), čo je jedným z hyper parametrov našej metódy.

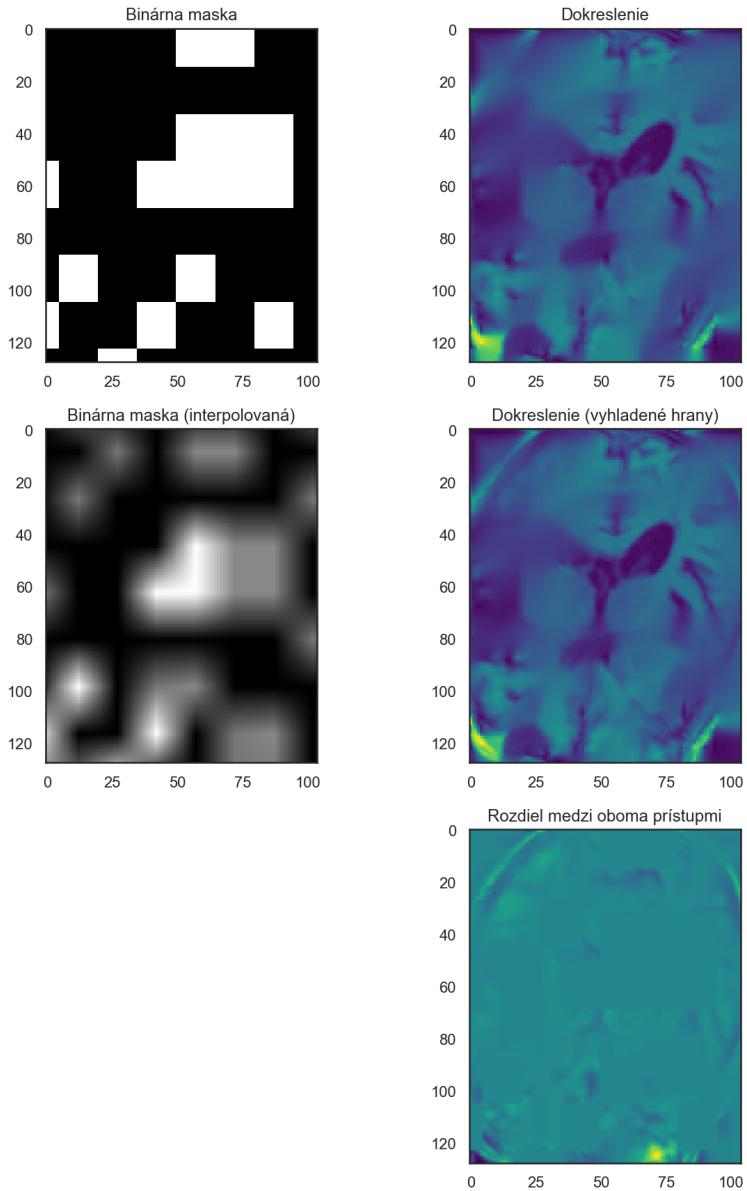


Obr. 5.4: Porovnanie 2D dokreslenia (iba v jednej dimenzii) a spriemerovaného 3x 2D dokreslenia (v každej dimenzii). Použitie iba 2D dokreslenia je kvalitné iba v jednej dimenzii a v ostatných je deštruktívne - vytvára ostré hrany. Použitie 3x 2D dokreslenia a spriemerovanie pre každý voxel produkuje primerane dobré dokreslenia po pohľade z každej dimenzie.

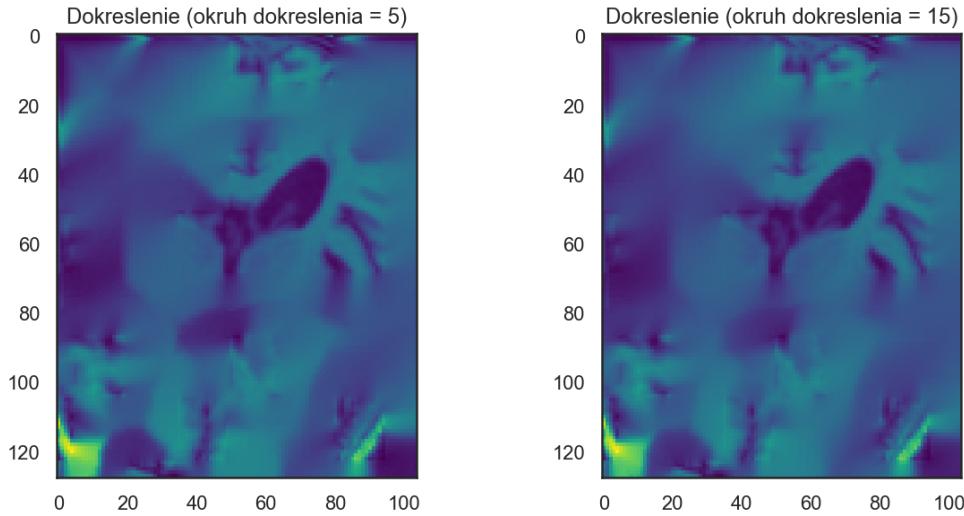
Kapitola 5. Implementácia

Kedžže sa pôvodná implementácia RISE prekrýva miesta tak, aby nevznikali ostré hrany medzi zakrytým miestom a pôvodným obrázkom, a teda vznikol plynulý prechod, aj pri dokreslení vytvárame plynulý prechod medzi dokresleným a pôvodným obrázkom (Obr. 5.5). Tento prechod je implementovaný nasledovne.

```
# binary_mask int[z, x, y] - upsized binary mask
# image float[z, x, y] - original image
# mask float[z, x, i] - upsized and interpolated binary mask
# inpaint_radius int
inpainted = cv.inpaint(image, binary_mask, inpaint_radius,
                       cv2.INPAINT_TELEA)
inpainted_blend = image * mask + inpainted * (1 - mask)
```



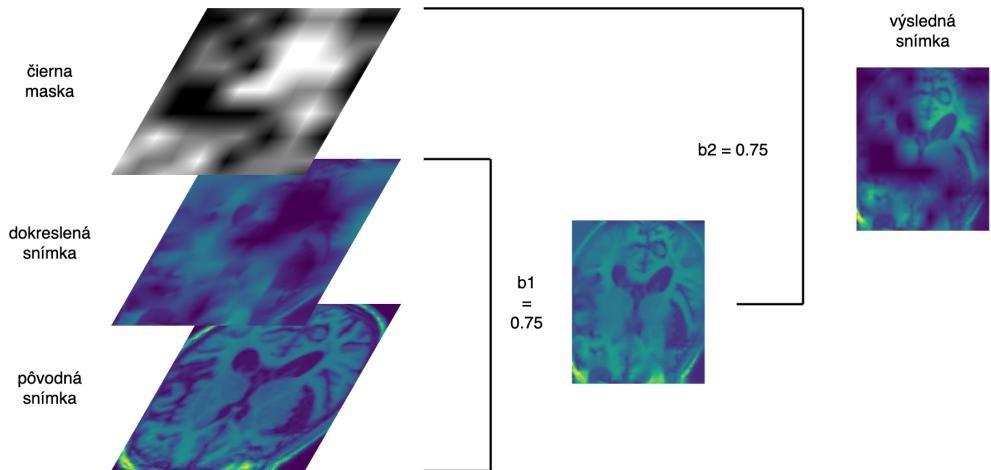
Obr. 5.5: Príklad vyhladzovania hrán dokreslenia - splynutie dokreslenia s pôvodným snímkom (štvrťá snímka). Druhá snímka zobrazuje ostré hrany po dokreslení - bez splývania s obrázkom. Piata snímka zobrazuje rozdielový obrázok medzi oboma prístupmi.
Môžeme si všimnúť, že na obrázku sú viditeľné miesta, kde sa nachádza prechod na interpolovanéj binárnej maske. O tieto miesta (informácie) je dokreslenie s vyhladenými hranami "bohatšie".



Obr. 5.6: Porovnanie okruhov dokreslenia (parameter *inpaint_radius*), rozdiel vo výsledku nie je veľmi viditeľný, avšak s väčším oruhom dokreslenia je generovanie rádovo pomalsie. (pri generovaní bolo vypnuté splynutie dokreslenia so snímkom aby bol rozdiel aspoň trochu viditeľný)

Prekrytie dokreslenej masky a čiernej masky s obrázkom Kedže v rámci metódy sa prekrývajú tri rôzne vrstvy - originálna snímka, čierna maska a dokreslená snímka, môžeme tieto vrstvy skombinovať v rôznom pomere a tým vytvoriť novú snímku.

Toto sme implementovali zavedením parametrov $b1$ a $b2$ (skratka od slova prechod, angl. blend), ktoré hovoria o pomere medzi originálnou snímkou a dokreslenou snímkou, a originálnou snímkou spojenou s dokreslením a čiernou maskou (Obr. 5.7). Pri týchto parametroch platí, že $0 \leq b1, b2 \leq 1$. Takto zadefinované parametre mi umožňujú vytvoriť zakaskovaný snímok iba s čiernou maskou ($b1 = 0, b2 = 1$) či iba s dokreslením ($b1 = 1, b2 = 0$).



Obr. 5.7: Príklad, ako vyzerá spojenie originálnej snímky, dokreslenej snímky a čiernej masky. V diagrame je zobrazený aj výsledok medzikroku spojenia dokreslenej snímky a pôvodnej snímky.

Parametre boli nastavené na $b1 = 0.75$ a $b2 = 0.75$.

Názov "čierna" maska pochádza z pôvodnej implementácie RISE, kde sa obrázok prekrýval čierňou maskou. V našej implementácii neprekrývame farbou, ale hodnotou, tj. "čierna" je hodnota 0 (minimum). Okrem použitia hodnoty 0, môžeme použiť aj 1, *priemer* či *medián* (toto je ďalším hyper-parametrom našej metódy). Zjednodušená (a menej efektívna, v produkčnej implementácii sa niektoré inštrukcie nevykonávajú keď $b1$ je 0 alebo $b2$ je 0) implementácia spojenia jednotlivých vrstiev vyzerá nasledovne.

```

# image float[z, x, y] - original image
# inpainted_blend float[z, x, y] - inpainted image
# mask float[z, x, i] - upsized and interpolated binary mask
# b1 float <0, 1>
# b2 float <0, 1>
# b2_value string - what value use in "black" mask
(min/max/mean/median)

# merge with inpainted image
new_image = (1 - b1) * original_image + b1 * inpainted_blend

```

```

value = 0 # black
if b2_value == 'max':
    value = 1 # white
elif b2_value == 'mean':
    value = np.mean(original_image)
elif b2_value == 'median':
    value = np.median(original_image)
# merge with "black" mask
new_image = b2 * mask * new_image + (b2 * (1 - mask) * value)

```

Kompletný zoznam parametrov metódy RISEI sa nachádza v tabuľke 5.1.

Názov	Dátový typ	Popis
s	int	Veľkosť strany binárnej 3D matice.
p	float	Pravdepodobnosť, že plocha nebude prekrytá maskou.
b1	float	Miera prekrytie medzi originálnym snímkom a dokresleným snímkom.
b2	float	Miera prekrytie s "čiernej" maskou.
b2_value	string	Hodnota "čiernej" masky, môže to byť minimum, maximum, medián, priemer.
in_paint_radius	float	Polomer dokreslenia algoritmom z knižnice OpenCV.

Tabuľka 5.1: Zoznam parametrov metódy RISEI.

5.1.2 Vytvorenie tepelných máp

Na základe návrhu (Sekcia 4.1) sme implementovali vytváranie tepelných máp. Keďže generovanie tepelnej mapy si vyžaduje vygenerovať veľký počet zamaskovaných snímok, ktoré v istom momente musia byť všetky uložené v pamäti, generujeme a vyhodnocujeme zamaskované snímky v dávkach (angl. batch). Zdrojový kód nižšie, implementuje vytvorenie jednej tepelnej mapy. Príklad vytvorenjej tepelnej mapy uvádzame na obrázku 5.8.

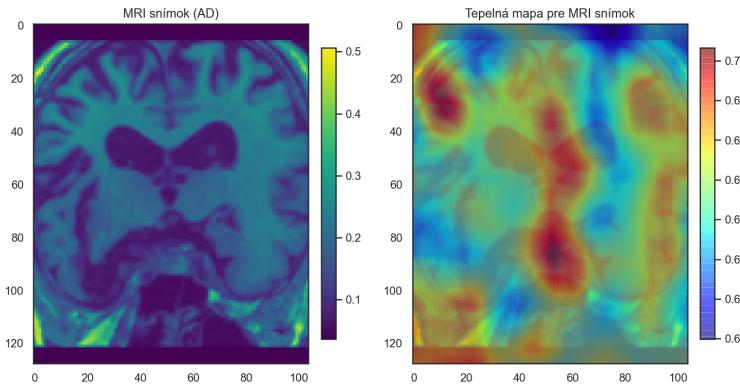
```
# image_x float[z, x, y, 1] - original image
# masks_count int - how many masks are generated to create a heatmap
# batch_size - how many masks to evaluate on model
# risei_batch_size int - how many masks to generate in one batch
# seed int int - seed for mask generation
# cls_idx int - index of target class in model output vector
# model tf.keras.Model - instance of tensorflow model

risei = RISEI(s=8, p=0.5, b1=0.5, b2=0.5, b2_value='median',
               in_paint_radius=5)
heatmap = np.zeros(shape=image_x.shape[:3])
batch_count = math.ceil(masks_count / risei_batch_size)
weights = 0

for batch_idx in range(batch_count):
    batch_masks_count = min(risei_batch_size, masks_count - batch_idx *
                           risei_batch_size)
    # reshape input for RISEI since it works with [z, y, x] shape
    # batch_x float[z, x, y] - images to evaluate with masks already
    # applied
    # masks float[z, x, y] - interpolated binary masks (so we know which
    # places we inpainted or masked)
    batch_x, masks = risei.generate_masks(batch_masks_count,
                                           image_x.reshape(image_x.shape[:3]), seed=seed)
    y_pred_batch_x = model.predict(batch_x.reshape((-1, *image_x.shape)),
                                   batch_size=batch_size)

    for mask, y_pred in zip(masks, y_pred_batch_x):
        # invert the mask, since 1 is for no masking
        # y_pred is the activation for the input masked image on last
        # layer (softmax)
        heatmap = heatmap + y_pred[cls_idx] * (1 - mask)
        weights += y_pred[cls_idx]
```

```
heatmap = heatmap / weights
```



Obr. 5.8: Príklad vytvorennej tepelnej mapy (vpravo) k MRI snímke (vľavo). Mierka vujadruje priemernú mieru aktivácie pre daný voxel.

5.1.3 Vyhodnotenie tepelných máp

Zatiaľ sme implementovali, podľa návrhu riešenia (Sekcia 4.3.1.1), iba metriky *insertion* a *deletion*.

5.1.3.1 Metriky insertion & deletion

Tieto metriky fungujú tak, že postupne odstraňujeme/pridávame pixely z obrázku a tieto obrázky vkladáme do modelu a zaznamenávame si aktiváciu na poslednej vrstve pre predikovanú triedu. V prípade obrázkov, a teda dvojrozmerných dát je to ešte výpočtovo zvládnutelné, avšak v prípade trojdimenzionálnych rádiologických simkov to už môže byť problém. Naše vstupné snímky majú po zmenšení rozmer [104, 128, 104], čiže ak aby sme odstraňovali zo snímku po jednom voxelu, museli by sme vykonať 1 384 448 evaluácií pomocou nášho modelu (čo trvá niekoľko hodín, aj pri evaluovaní v maximálnych možných dávkach vzhľadom na pamäť grafickej karty). Preto sme sa rozhodli pridávať po n (~ 100) voxeloch v každom kroku. V

prípade metódy insertion vkladáme do snímku plného núl (môžeme prípadne aj jednotiek). Keďže kód je rozsiahlejší, uvedieme len pseudokód.

```
method = 'insertion'

step_size = 150 # how many voxels to insert/delete in one evaluation
image_x, image_y = get_image()
image_y_pred = model.predict(image_x)
heatmap = get_heatmap()
voxels = get_ordered_voxels_by_heat(heatmaps)
sequence = get_images_sequence(voxels, step_size) # create a sequence
    from images where each next image has n inserted/deleted voxels
y_pred = []

for batch_x, batch_y in sequence:
    batch_y_pred = model.predict(batch_x)
    for y in batch_y_pred:
        y_pred.append(y)

auc = metrics.auc([i * step_size for i in range(len(y_pred))], y_pred) /
    get_voxels_count(image_x)
```

5.2 Model na detekciu Alzheimerovej choroby na základe MRI snímok

V tejto sekcii popíšeme implementáciu modelu, z ktorého predikcií vytvárame teplné mapy. Náš model - neuónovú sieť sme sa rozdihodli implementovať v knižnici Tensorflow (v2.3.0). Naším cieľom nie je natrénovať najlepší model na dekodovanie Alzheimerovej choroby, ale model ktorý je použiteľný na overenie nami narvhnutej metódy. Preto nevykonáme komplexnejšie prístupy k detekcii Alzheimerovej choroby, ktoré sme popísali v analýze (Sekcia 2.3), ako je napríklad učenie prenosom pomocou autoenkodéra.

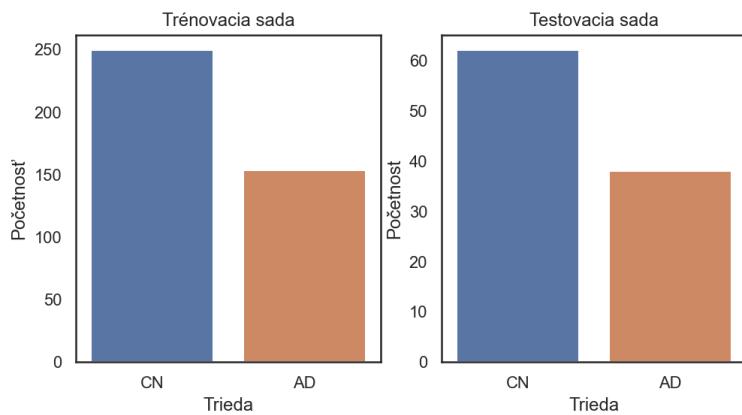
5.2.1 Dátová sada

Použili sme dátovú sadu ADNI. Ako vstup modelu je celý MRI snímok (tj. všetky tri dimenzie), nepoužívame žiadné ine údaje z dátovej sady ADNI, ako napríklad demografické údaje a pod. keďže model plánujeme používať iba na vytváranie tepelných máp pre vstupné snímky.

V dátovej sade sa nachádza celkom 502 MRI snímok, z toho 311 pacientov s Alzheimerovou chorobou (AD) a 191 bez (CN). Dátovú sadu máme teda nevyváženú a model môže začať preferovať jednu triedu. Na zabránenie tomuto javu existuje niekoľko techník, napríklad nadvzorkovanie (angl. oversampling) alebo podvzorkovanie (angl. undersampling) kedy sa doplní synetetickými minoritnou triedou, alebo sa odstránia nejaké pozorovania z majoritnej triedy. My sme sa však rozhodli nastaviť predikovaným triedam váhy, ktoré sú zohľadnené v chybovej funkcií, taktiež sme nainicializovali chybu¹ pre neuróny na poslednej vrstve aby reflektovala to, že triedy sú nevyvážené.

Dátovú sadu MRI snímkov pacientov sme náhodným výberom rozdelili na trénovaciu a testovaciu v pomere 80/20. Validačnú sadu sme nevytvárali, z dôvodu malého množstva dát, ktoré máme k dispozícii a taktiež neplánujeme prehľadávať priestor hyper parametrov za účelom nájsť ich najoptimálnejšiu kombináciu. Alternatívne by bolo vykonanie krížovej validácie (angl. cross validation) pri trénovaní. Aj po rozdelení sa nám podarilo zachovať pôvodný pomer medzi triedami – 62/38 (Obr. 5.9).

¹https://www.tensorflow.org/tutorials/structured_data/imbalanced_data#optional_set_the_correct_initial_bias



Obr. 5.9: Početnosť tried medzi trénovacou a testovacou sadou - je zrejmá prevaha triedy AD.

5.2.1.1 Predspracovanie

MRI snímky boli predspracované štandardnou postupnosťou nástroja Freesurfer², avšak nevykonali sme odstránenie lebky z MRI snímok.

Ďalej sme vykonali:

- Upravenie vstupných snímok zmenšením na rovnakú veľkosť 104 x 128 x 104 voxelov. Esmaeilzadeh et al. upravili vstupné snímky na veľkosť 116 x 130 x 83, k týmto číslam sme sa pokúsili priblížiť. Pomer veľkostí dimenzií ale nemáme rovnaký, aj z dôvodu, že sme nevykonali odstránenie lebky zo vstupných snímok.
- Štandardizáciu vstupných dát (preškálovanie na rozsah $<0,1>$) nasledovným vzorcom: $\frac{(image_x - images_min)}{(images_max - images_min)}$.

Snímky z dátovej sady sme augmentovali na základe návrhu.

²<https://surfer.nmr.mgh.harvard.edu/>

5.2.2 Model

Implementujeme neurónové siete na základe návrhu (Sekcia 4.3).

3D konvolučná neurónová sieť od Esmaeilzadeh et al. Implementovali sme jej jednoduchšiu verziu, ktorá dosahovala lepšie výsledky, opísali sme ju v sekcií 2.3. Táto neurónová sieť má celkovo $2\ 899\ 778$ parametrov.

2D ResNet a 3D ResNet V prípade 2D ResNet-u používame 2D konvolúcie, tie nám budú fungovať aj napriek tomu, že máme 3D dátu. Vstup do 2D ResNet-u je tiež 3D matica, ktorej tretia dimenzia býva o obvykle o dĺžky 1 alebo 3 (RGB), v našom prípade je o veľkosti poslednej dimenzie snímky. Rozmery vstupných dát pre 2D ResNet sú [104, 128, 104] a pre 3D ResNet [104, 128, 104, 1]. Za konvolučné vrstvy a globálnu združovaciu vrstvu sme pripojili dve plne prepojené vrstvy s 512, 256 a 128 neurónami a s aktivačnou funkciou *ReLU*, následne už nasleduje iba posledná vrstva s aktiváciou *softmax*. Tieto neurónové siete majú celkovo $12\ 689\ 602$, resp. $34\ 356\ 354$ parametrov.

5.2.3 Trénovanie

Pri trénovaní sme použili:

- kategorickú entropiu (angl. categorical crossentropy) ako chybovú funkciu s podporou pre nevyvážené triedy (tj. táto funkcia brala ohľad na váhy tried, ktoré sme nastavili nepriamo úmerne ich veľkosti),
- optimalizačný algoritmus Adam s prevolenými nastaveniami,
- exponenciálne tlmenie rýchlosťi učenia (angl. learning rate decay), s hodnotou 0,96 každých 25 epoch,
- skoré zastavenie trénovania ak sa metrika AUC (plocha pod krivkou) nezlepšila za posledných 50 epoch,

- veľkosť dávky (angl. batch size) – 10 (vždy tak, aby sme naplno využili pamäť grafickej karty),
- l_2 regularizáciu (rovnako ako Esmaeilzadeh et al.).

Trénovanie sme začali s modelom bez augmentácie, dávkovej normalizácie, dropoutu a regularizácie pričom sme ich postupne pridávali, dolaďovali a sledovali zmeny v úspešnosti modelu. Najlepšie výsledky sme dosiahli s architektúrou 3D ResNet s presnosťou 80% (Tabuľka 5.2). Nepodarilo sa nám nám teda priblížiť k výsledkom analyzovaných prác, čo však v konečnom dôsledku ani nie je cieľom tejto práce.

		Baseline	+ Augmentácie	+ Dávková normalizácia	+ Dropout	+ Regularizácia (l_2)
3D CNN	Acc.	0.71	0.67	0.75	0.75	0.71
	Sens.	0.76	0.68	0.77	0.76	0.70
	Spec.	0.63	0.66	0.71	0.74	0.71
3D ResNet	Acc.	0.71	0.69	0.80	0.74	0.79
	Sens.	0.79	0.84	0.85	0.94	0.87
	Spec.	0.57	0.45	0.71	0.42	0.66
2D ResNet	Acc.	0.67	0.76	0.77	0.77	0.78
	Sens.	0.77	0.90	0.89	0.85	0.89
	Spec	0.50	0.52	0.58	0.63	0.61

Tabuľka 5.2: **Výsledky trénovania.** Acc. = presnosť (angl. Accuracy), Sens. = senzitivita (angl. Sensitivity), Spec. = Špecifita (angl. Specificity)

Oproti Esmaeilzadeh et al., ktorí dosiahli presnosť až 94%, sme dosiahli presnosť len 72% avšak sme mali menej dát (o 339 pozorovaní menej), neodstraňovali sme zo snímok lebku a nepoužili sme pri klasifikácii vek pacienta. Avšak robili sme viac augmentácií, no po ich pridaní sa úspešnosť modelu zhoršila (Obr. 5.10) (ale následne sa už iba zlepšovala), je teda možné, že niektoré augmentácie sú nekorektné a deštruktívne voči vstupným snímkam a vedú k zhoršeniu výkonnosti modelu. Aj po pridaní veľmi slabej regularizácie, sa úspešnosť modelu zhoršila. V prípade 2D a 3D ResNet architektúr sa nám podarilo dosiahnuť lepšie výsledky, avšak v

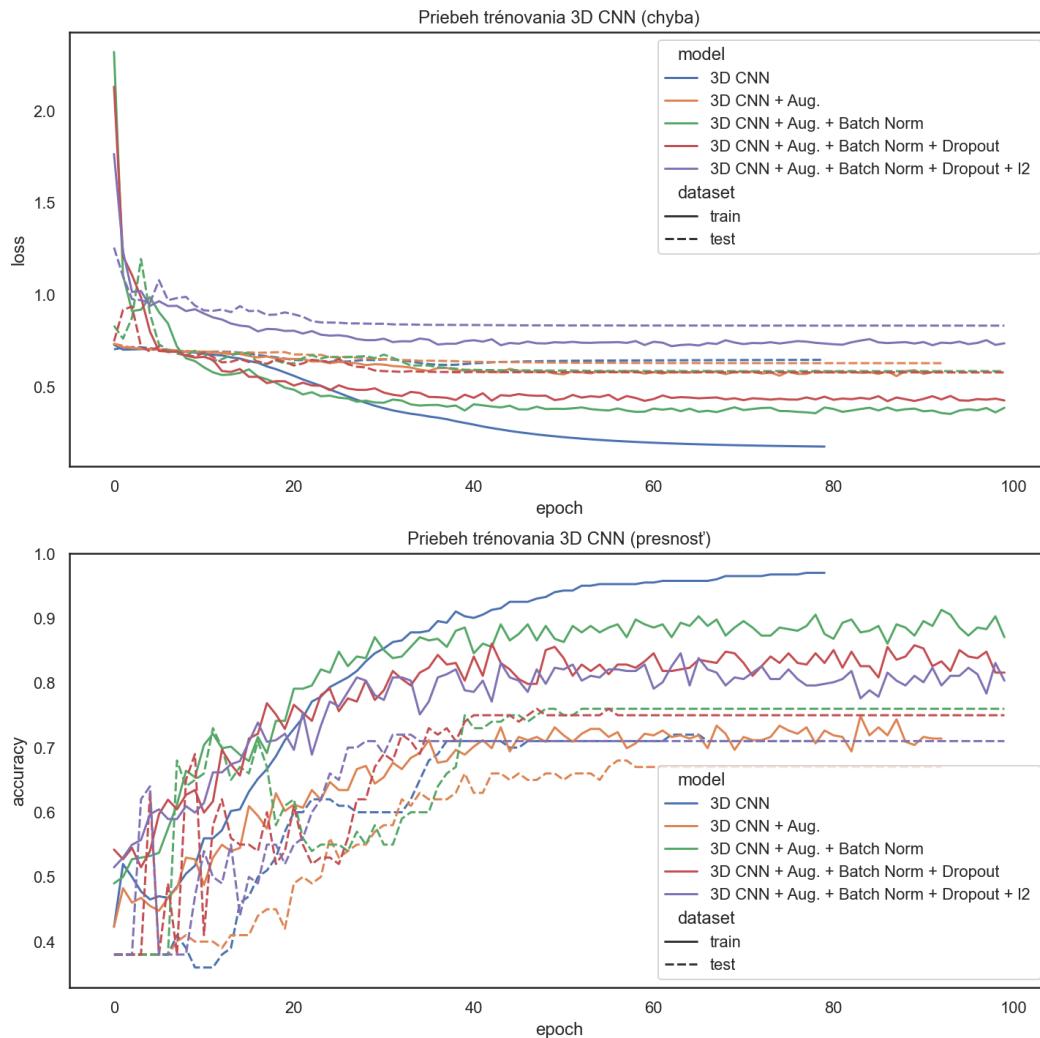
Kapitola 5. Implementácia

ich prípade sa sieť neskôr začala pretrénovať (Obr. 5.11) aj napriek použitej regularizácii, čo môže naznačovať, že sú tieto architektúry na náš problém príliš komplexné.

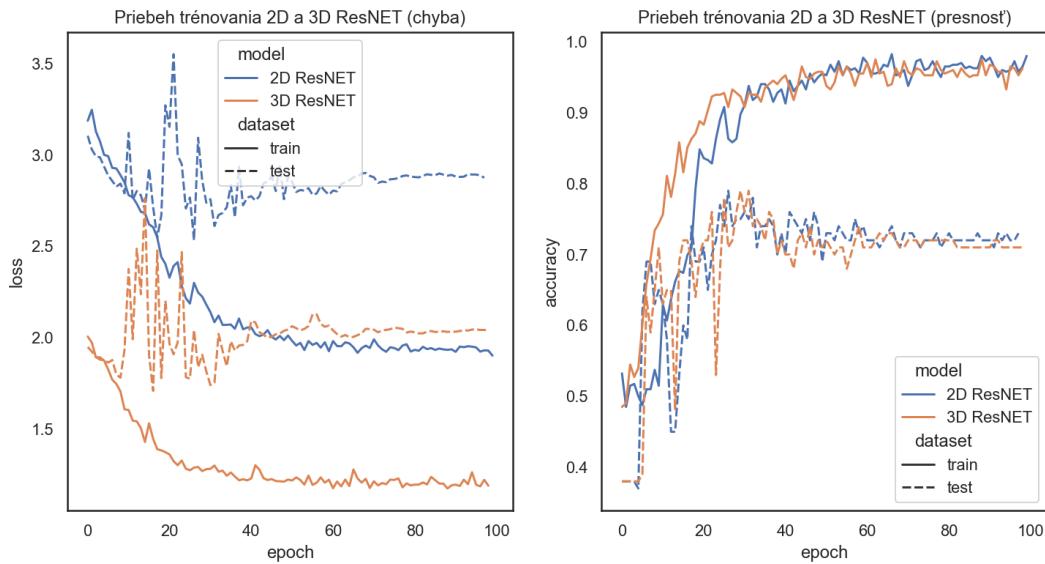
Identifikovali sme nasledovné možné vylepšenia (zoradené podľa subjektívneho pomeru úsilie/vplyv):

- Porovnať jednotlivé augmentácie a vyhodiť tie deštruktívne.
- Odstránenie lebky zo snímok.
- Nájsť a použiť viac dát.
- Krížová validácia v prípade väčšieho nastavovania hyperparametrov.
- Učenie prenosom pomocou autoenkodéra [26].

Vzhľadom na to, že našim cieľom nie je natrénovať najlepšiu neurónovú sieť, nevykonali sme všetky identifikované možné vylepšenia. Vykonali sme porovnanie jednotlivých augmenácií na architektúre 3D CNN a to tak, že sme pre každú augmentáciu natrénovali samostatný model (bez dropout-u a regularizácie) a augmentovali sme vstupné snímky s pravdepodobnosťou 50%. Sledovali sme, či sa aj s augmentovanými snímkami netrénované modely bez akejkoľvek regularizácie dokážu preučiť. Model sa nepreučil iba pri augmentácii *náhodné priblíženie*. Následne sme natrénovali modely pre jednotlivé architektúry znova, lepší model sa nám podarilo natrénovať iba pre architektúru 3D CNN, presnosť - **0.78**, senzitivita - **0.81**, špecificka - **0.74**.



Obr. 5.10: Priebeh trénovalia 3D konvolučnej neurónovej siete, čím viac sme pridali regularizácie (l2 alebo dropout) tým sme dosiahli horšie výsledky. Po pridaní augmentácií sa úspešnosť modelu zhoršila, avšak následne sa už iba zlepšovala.



Obr. 5.11: Priebeh trénovania 2D a 3D ResNet neurónovej siete. V oboch prípadoch sme použili dropout aj regularizáciu. Od približne 30-tej epochy sa neurónová sieť začala pretrénovať, aj napriek regularizácii. Ako vylepšenie je možné skúsiť silnejšiu regularizáciu (väčšiu hodnotu l2 a dropout), ak ani to nepomôže, je možné, že architektúra je príliš komplexná náš problém (neurónová sieť je príliš hlboká). Môžeme ďalej skúsiť odstrániť jednu plne prepojenú vrstvu alebo znížiť počet neurónov v týchto vrstvách.

5.3 Zhrnutie

V tejto kapitole sme opísali, ako sme implementovali nami navrhovanú metódu a spôsob jej overenie. Opísali sme implemetáciu kľúčových prvkov (generovanie masiek a vytvorenie tepelnej mapy) navrhovanej metódy RISEI a jej hyper parametre. Rovnako sme opísali aj implementáciu neurónovej siete, architektúru, spôsob trénovania a použitú dátovú sadu. Metódu RISEI sa nám podarilo implementovať a je ju možné použiť v experimentoch. Podarilo sa nám natrénovať niekoľko architektúr neurónových sietí, pričom sme identifikovali možné príčiny výsledkov, ktoré dosahujú (a navrhli spôsob ako ich riešiť). Nie všetky identifikované príčiny sa nám podarilo vyriesiť, keďže cieľom tejto práce nie je natrénovať najlepší model, ale

natrénovať taký model, ktorý je postačujúci na overenie navrhnutej metódy.

6. Overenie riešenia

Kedže sme natrénovali viacero modelov, a metóda RISEI má veľké množstvo parametrov zvolili sme nasledovnú stratégiu pri overovaní riešenia, tak aby sme nemuseli overovať všetky možné kombinácie parametrov/architektúr a dokázali vykonať všetky navrhnuté experimenty (Sekcia 4.3.1) v stanovenom čase.

1. Výber architektúry neurónovej siete na ktorej budeme overovať metódu RISEI a porovnávať ju z ostatnými. *Ktorý z natrénovaných modelov je najvhodnejší pre navrhnutú metódu?*
2. Overenie metódy RISEI. *Dávajú výsledky z navrhнутej metódy zmysel a nie sú náhodné?:*
 - Overenie stability tepelných máp.
 - Overenie kvality tepelných máp (metriky *insertion* a *deletion*).
3. Porovnanie kvality tepelných máp s metódou RISE (s doimplementovanou podporou pre 3D snímky) - *je navrhnutá metóda lepšia ako metóda, z ktorej vychádza?*
4. Overenie správnosti tepelných máp a porovnanie s inými metódami - GradCAM, Guided Backprop, Guided GradCAM. *Sú tepelné mapy nie len kvalitné, ale dávajú zmysel v kontexte anatómie mozgu? Ako sú na tom iné metódy?*

Adresovanie náhodnosti metódy RISE pri porovnávaní dvoch roznych modelov Kedže vygenerované masky sú náhodné, pri porovnávaní dvoch metód (kombináciu rôznych parametrov metódy RISEI) generujeme rovnaké binárne masky pre každý i-ty snímok, tj. zakryté pozície sú rovnaké, rôzna je len hodnota zakrytia. Vygenerované binárne masky pre jednotlivé snímky v testovacej sade sú stále náhodné (a teda aj medzi sebou rôzne). Rovnaké sú len binárne masky pre dve rôzne generovania masiek pre ten istý snímok v testovacej sade. Takto dosiahneme presnejšie výsledky, pretože nebudeme porovnávať miesto prekrytia ale spôsob - hodnotu prekrytia.

6.1 Experiments

6.1.1 Výber architektúry neurónovej siete pre ďalšie experimenty

V tomto experimente sme porovnali metódu RISE na nami natrénovaných modeloch s cieľom vybrať jeden z nich pre ďalšie experimenty (kedže experimenty sú časovo náročné, nechceme ich robiť na všetkých modeloch). Vybrali sme najlepšie natrénované modely pre každú architektúru (Tabuľka 5.2). Použili sme nami implementovanú metódu RISEI pričom sme nastavili jej parametre tak, aby fungovala ako metóda RISE. Parametre sme nastavili nasledovne: $s = 8$, $p = 1/3$, $b1 = 0$, $b2 = 1$ (opis parametrov sa nachádza v tabuľke 5.1). Na vytvorenie tepelnej mapy sme vygenerovali 1024 masiek. Pri vyhodnocovaní metrík insertion a deletion sme nastavili veľkosť kroku na 2500 voxelov (takto trvalo vyhodnotenie tepelnej mapy 3 minúty). Vybrali sme 25 náhodných snímok z testovacej sady (13 AD, 12 CN), generovanie a vyhodnotenie tepelných máp k nim trvalo približne 1 hodinu.

Najlepšie výsledky sme dosiahli na architektúre 3D CNN (Tabuľka 6.1). Je dôležité si ale uvedomiť, že na vyhodnotenie metrík *insertion* a *deletion* z vytvorennej tepelnej masky sa používa model samotný - zo snímky sú pridávané/odoberané voxely pričom sa sleduje sa zmena predikcie modelu. Môže nastať teda situácia, že

		3D CNN	3D ResNET	2D ResNET
Insertion (AUC)	priemer	0.50	0.46	0.38
	medián	0.53	0.13	0.30
Deletion (AUC)	priemer	0.53	0.81	0.62
	medián	0.60	0.83	0.43

Tabuľka 6.1: **Porovnanie metódy RISE na rôznych architektúrach.** Vybrali sme najlepšie natrénované modely pre každú architektúru (Tabuľka 5.2). Pre insertion sú lepšie vyššie hodnoty (očakávame, že keď vložíme najpodstatnejšie voxely aktivácia bude stúpať), pre deletion sú lepšie nižšie hodnoty (očakávame, že keď ostráníme najdôležitejšie voxely, aktivácia bude klesať).

dva rôzne modely vytvoria dve identické tepelné mapy, pričom výsledná hodnota metriky bude rozdielna. Na základe týchto metrik teda nemôžeme tvrdiť, že jeden model vytvára lepšie tepelné mapy ako druhý. Metrika *insertion* a *deletion* nie je vhodná na takéto porovnanie dvoch rôznych modelov medzi sebou. Dobrým signálom by bolo ak by hodnota metriky *insertion* bola väčšia ako metriky *deletion*, takto to nie je ani u jedného z modelov (najbližšie má k tomu 3D CNN). Aj kvôli tomuto sme vybrali do ďalších experimentov model 3D CNN, zároveň na základe jeho specificity a senzitivity môžeme tvrdiť, že nepreferuje ani jednu z tried - čo u iných modeloch tak nie je (tie preferujú AD pozorovania). Taktiež je tento model jednoduchší.

6.1.2 Overenie metódy RISEI

6.1.2.1 Stabilita tepelných máp

Kedže metóda RISEI (aj metóda RISE z ktorej vychádzame) používa na generovanie tepelným máp náhodné masky, výsledná tepelná mapa je touto náhodnosťou ovplyvnená a je teda do určitej miery náhodná. Očakávame, že čím viac masiek vygenerujeme, tým bude vplyv náhodny na výslednú tepelnú mapu nižší. Keď teda pre tú istú snímku vygenerujeme niekoľko tepelných máp, tieto tepelné mapy sa budú lísiť čo najmenej = budú stabilné.

Metóde RISEI sme nastavili nasledovné parametre $s = 8$, $p = 1/3$, $b1 = 0$, $b2 = 1$ a $b2_value = 1$ - nepoužívame dokreslenie, keďže je časovo náročné a vykonávame veľké množstvo experimentov. Uvažujeme, že nezáleží na tom, čo je hodnota zakrytie vo vygenerovaných maskách, pokial tá hodnota nie je náhodná.

Použili sme model 3D CNN so senzitivitou - 0.81 a špecifítou - 0.74.

Porovnanie vytvorených tepelných máp N vytvorených tepelných máp porovnávame tak, že počítame štandardnú odchíľku pre každý voxel medzi vytvorenými tepelnými mapami. Tak z tepelných máp o rozmere $[N, z, y, x]$ vznikne 3D matica štandardných odchíľok $[z, y, x]$. Ak má voxel rovnakú/blízku hodnotu tepla medzi tepelnými mapami, štandardná odchýlka nula/blízka nule. Z 3D matice štandardných odchíľok vypočítame priemernú/strednú hodnotu - táto hodnota reprezentuje vzniknutú chybu medzi tepelnými mapami plynúcu z náhodnosti tepelných masiek.

6.1.2.2 Experiment 1 (jedna snímka)

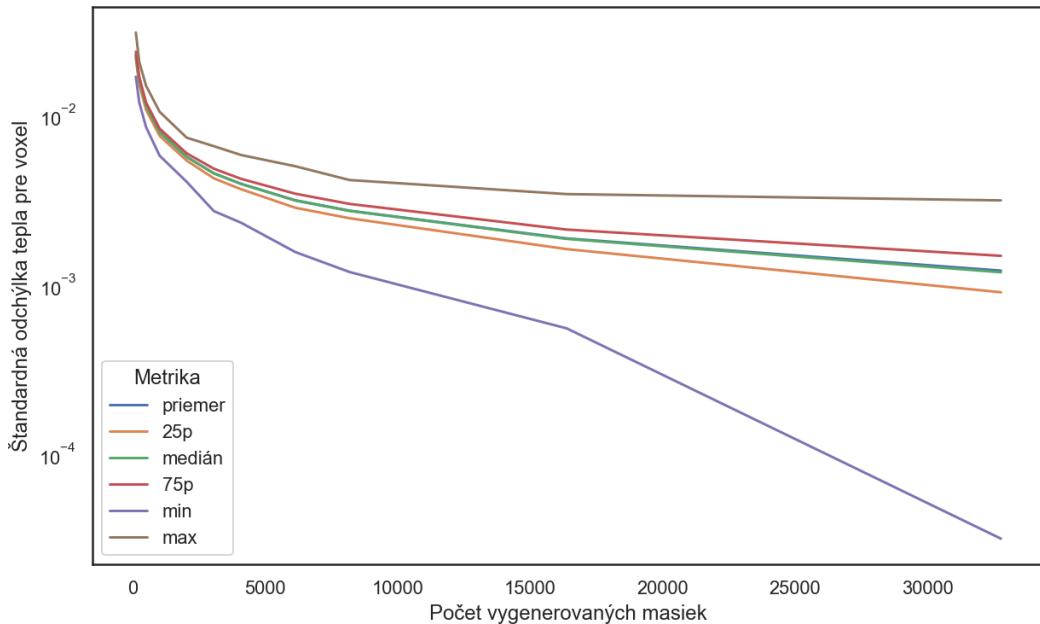
Pre náhodnú snímku vytvoríme K tepelných máp, pričom tepelné mapy vytvárame z 16, 128, 256, 512, 1024, 2048, 3072, 4096, 6144, 8192, 16384 alebo 32768 masiek. Kvôli vysokej pamäťovej náročnosti, v experimentoch, v ktorých vytvárame tepelné mapy z vysokého počtu masiek vytvoríme menej tepelných máp (Tabuľka 6.2). Zistili sme, že s vyšším počtom vygenerovaných masiek chyba klesá logaritmicky (Obr. 6.1). Zároveň, chyba sa javí byť náhodná a nie systematická (Obrázok 6.2).

Počet vygenerovaných masiek	Počet vytvorených tepelných máp	Medián štandardnej odchýlky pre voxel (chyba)
16	100	0.0640
128	100	0.0225
256	100	0.0160
512	100	0.0113
1024	100	0.0080
2048	100	0.0057
3072	50	0.0045
4096	50	0.0039
6144	25	0.0031
8192	25	0.0027
16384	15	0.0019
32768	5	0.0011

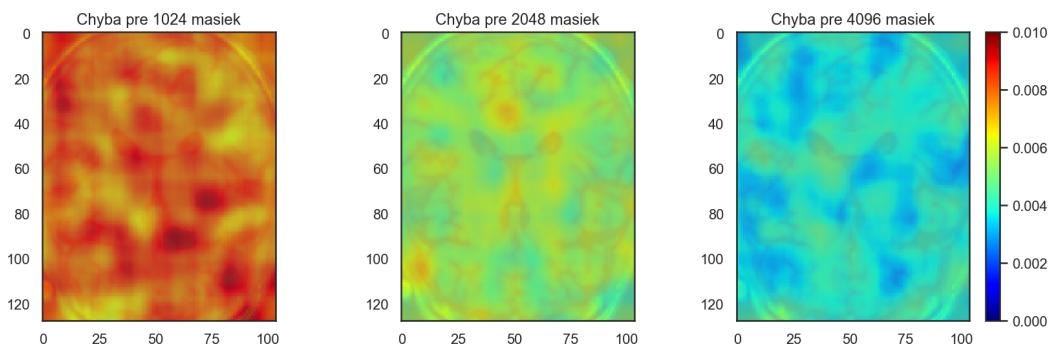
Tabuľka 6.2: Stabilita vytvorených tepelných máp podľa počtu vygenerovaných masiek pre jeden snímok. S vyšším počtom vygenerovaných máp chyba výrazne klesá. Už pri 2048 maskách je chyba zanedbateľná, keďže hodnoty voxelov v tepelných mapách sú z intervalu $<0, 1>$. Väčšie množstvo náhodných masiek nemá zmysel generovať, pretože výpočtová náročnosť stúpa rýchlejšie ako stabilita. V porovnaní 2048 a 32768 chyba klesla 5 krát, avšak sme museli vygenerovať 15 krát viac masiek.

6.1.2.3 Experiment 2 (viacero snímok)

Kedže sme v predchádzajúcim experimente overovali stabilitu iba na jednej snímke, v tomto experimente overíme stabilitu na viacero snímkach. Z testovacej sady sme vybrali 5 TP (viď. Zoznam použitých skratiek), 5 TN, 5 FP a 5 FN pozorovaní (tj. celkovo 20 pozorovaní), podľa toho ako ich neurónová sieť označila. Takto zabezpečíme vyváženosť tried pozorovaní v experimente. Kvôli časovej aj pamäťovej náročnosti tohto experimentu sme vytvárali 10 tepelných máp pre každé jedno pozorovanie. Výsledky boli takmer identické s predchádzajúcim experimentom (Sekcia 6.1.2.2) a trend poklesu chyby pri zvyšujúcim sa počte masiek bol zachovaný (6.3). Z oboch experimentov vyplýva, že je vhodné použiť vyšší počet



Obr. 6.1: Stabilita vytvorených tepelných máp podľa počtu vygenerovaných masiek. Os y je v logaritmickej škále a reprezentuje chybu. Táto chyba klesá logaritmicky s vyšším počtom vygenerovaných masiek. Priemer sa veľmi blíži mediánu, preto ho na diagrame nie je takmer vôbec vidieť.



Obr. 6.2: Vizualizácia chyby z generovania 1024, 2048 a 4096. Z vizualizácie je zdrejmé, že chyba je skôr náhodná ako systematická, keďže sa nachádza na rôznych častiach medzi snímkami. Škála tepla má výrazne znížené maximum oproti maximálnej možnej chybe (1) aby boli rozdiely viditeľné.

masiek pri vytváraní tepelných máp, aby sa odstránil vplyv náhody. Ako vhodným počet považujem 2048 masiek, pri tomto počte je chyba vzhľadom na hodnoty v tejepnej mape minimálna (Tabuľka 6.2, 6.3).

Počet vygenerovaných masiek	Medián štandardnej odchýlky pre voxel
16	0.0594
128	0.0207
256	0.0160
512	0.0105
1024	0.0074
2048	0.0052
3072	0.0043
4096	0.0037

Tabuľka 6.3: Stabilita vytvorených tepelných máp podľa počtu vygenerovaných masiek pre 20 snímiel. Trend poklesu chyby, rovnako ako v prvom experimente, (Tabuľka 6.2) ostal zachovaný.

Na stabilitu masiek môže mať vplyv aj parameter p , v tomto experimente sme použili konštantnú hodnotu $p = 1/3$. Jeho zmena môže stabilitu zvýšiť už pri menšom/až pri väčšom počte generovaných masiek. Vplyv parametra p na stabilitu tepelných máp podľa počtu masiek je vhodným predmetom ďalšieho skúmania.

6.1.3 RISE vs RISEI (s rôznymi parametrami)

V tomto experimente sme porovnali RISE a rôzne nastavenia metódy RISEI. Použili sme model 3D CNN so senzitivitou 73% a špecifitou 71%.

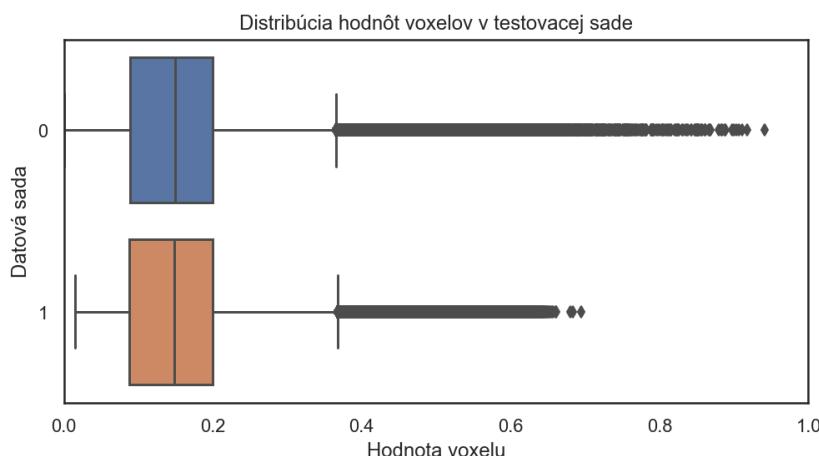
Použili sme rovnaké nastavenie parametrov, a rovnaký počet pozorovaní ako pri výbere architektúry neurónovej siete (Sekcia 6.1.1) a menili sme len parametre $b1$, $b2$ a $b2_value$. Parametre in_paint_radius sme nastavili na hodnotu 5.

Vyhodnocovali sme len metriku *insertion* (aby sme čo najrýchlejšie získali prvotné výsledky), najlepšie výsledky sme dosiahli bez použitia dokreslenia ale s prekrytím

hodnotou jedna (Tabuľka 6.4). To si vysvetľujeme tým, že hodnota voxelov blízka jednej je v trénovacej sade veľmi ojedinelá (Obr. 6.3) a neurónová siet na základe nich nerozhoduje, takéto voxelové neurónovú siet nepomýlia. Naopak, voxelové s hodnotou/blízke hodnote nula sú pomerne časté, ak na základe nich neurónová siet rozhoduje, možno to byť dôvod prečo prekrytie zaznamenalo horšie výsledky.

Metóda s dokreslením si počínala horšie ako prekrytie hodnotou jedna, ale stále lepšie ako prekrytie s hodnotou nula (hodnota pôvodnej metódy RISE). Zároveň dosiahla takmer identický výsledok ako prekrytie mediánom hodnôt voxelov snímky.

Obrázok 6.4 zobrazuje príklad vygenerovanej tepelnej mapy pre MRI snímok a výsledný graf zmeny aktivácie pre metriku *insertion*. Na diagrame je vidieť, že s postupným pridávaním voxelov stúpa aktivácia pre skutočnú triedu pozorovania, takto to však nie je u všetkých pozorovaní.



Obr. 6.3: Distribúcia hodnôt voxelov v trénovacej (0) a testovacej (1) sade. Testovacie dátá boli štandardizované (pred zmenšením) do intervalu $<0, 1>$ podľa maximálnych hodnôt v trénovacej sade.

Kedže sme neskôr natrénovali lepší model pre 3D CNN architektúru so senzitivitou 81% a špecifitou 74%, overlili sme ho v tomto experimente tiež, ale len na najlepšej zistenej kombinácii parametrov. Výsledok bola nižšia hodnota v metrike *insertion* (priemer - 0.60, medián - 0.63). Vytvorené tepelné mapy medzi obomi

	Insertion	
	Priemer	Medián
b1 = 0, b2 = 1, b2_value = 0 (RISE)	0.43	0.37
b1 = 0, b2 = 1, b2_value = 1	0.65	0.67
b1 = 0, b2 = 1, b2_value = medián	0.52	0.48
b1 = 1, b2 = 0	0.53	0.47
b1 = 1, b2 = 0.25, b2_value = 0	0.49	0.42
b1 = 1, b2 = 0.50, b2_value = 0	0.44	0.37
b1 = 1, b2 = 0.75, b2_value = 0	0.39	0.30

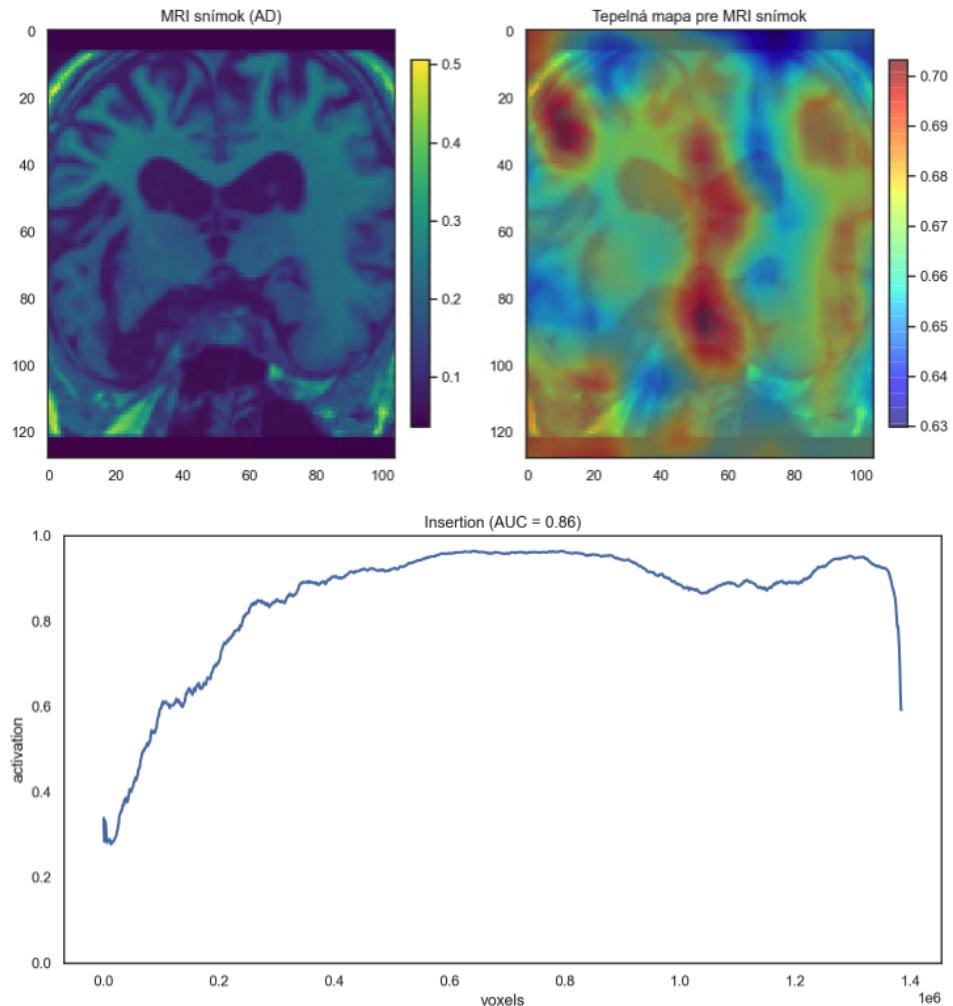
Tabuľka 6.4: Porovnanie rôznych nastavení metódy RISEI.

Najlepšie výsledky dosiahla metóda bez použitia dokreslenia ale s prekrytím hodnotou jedna.

modelmi boli veľmi podobné, no každý model ich vyhodnotil inak, tento problém sme načrtli v sekciu 6.1.1. Obrázok 6.5 zobrazuje takmer identicky vygenerované tepelné mapy, avšak lepší model, ktorý označil dané pozorovanie s vyššou istotou dosiahol nižšie skóre v metrike insertion. Tu sa črtá ďalší problém metrík *insertion* a *deletion*, ak model vykazuje nižšiu mieru istoty pre pozorovanie - hodnota aktivácie, plocha pod krivkou (metrika AUC), ktorá k nej smeruje nemôže byť jedna. V našom prípade sú tieto aktivácie pomerne nízke (autori RISE uvádzali také príklady, kde aktivácie pre predikované triedy boli viac ako 0.9). Toto avšak nemusí byť nutne problém, keďže sa v snímke môžu nachádzať voxely, ktoré sú proti predikovanej triede, a teda mali nastavené teplo správne, a boli správne vložené až na konci vyhodnotenia.

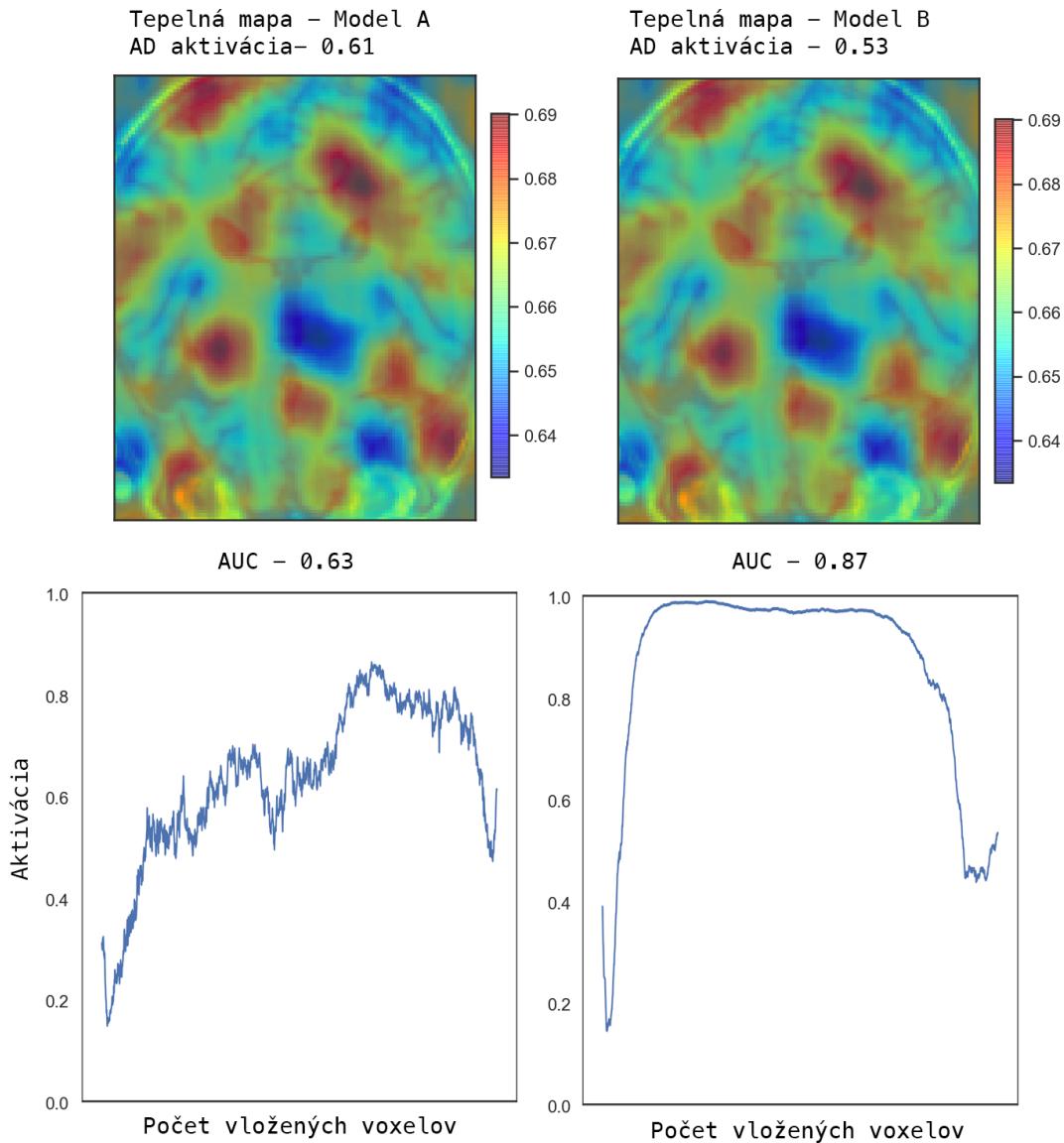
Ďaľším problémom týchto metrík môžu byť hodnoty voxelov, ktoré sú na miestach kde sme už zmazali/doposiaľ nepridalí voxely, tie by mali byť neutrálne a nemali by hovoriť o žiadnej triede. My používame hodnotu nula (tá môže skôr hovoriť o AD pacientoch - chýbajúce tkanovo), avšak je tiež možné použiť opačný extrém, hodnotu jedna.

Aj kvôli zisteným problémom vyššie budeme tepelné mapy overovať voči segmentačným maskám aby sme získali lepšiu predstavu o správnosti tepelnej mapy. Aj napriek horšej metrike *insertion* budeme ďalej používať lepšie natrénovaný mo-



Obr. 6.4: Vytvorená tepelná mapa a graf zmeny aktivácie po pridávaní voxelov pre vybranú MRI snímku. Metrika AUC je pomerne vysoká, avšak je potrebné tepelnú mapu ešte vyhodnotiť z pohľadu segmentačných masiek. Tepelná mapa bola vytvorená s parametrami $b1 = 1$, $b2 = 0$ (RISEI s dokreslením).

del.



Obr. 6.5: Porovnanie tepelných máp vygenerovanými dvoma rôznymi modelmi. Model A) je model 3D CNN so senzitivitou 81% a špecifitou 74%. Model B) je model 3D CNN so senzitivitou 73% a špecifitou 71%. Oba modely vytvorili takmer identické tepelné mapy, avšak každý ich vyhodnotil inak. Kvalitatívne vyhodnotenie tepelnej mapy lepšie natrénovaného modelu dosiahlo horší výsledok.

6.1.3.1 Nastavenie parametrov RISEI

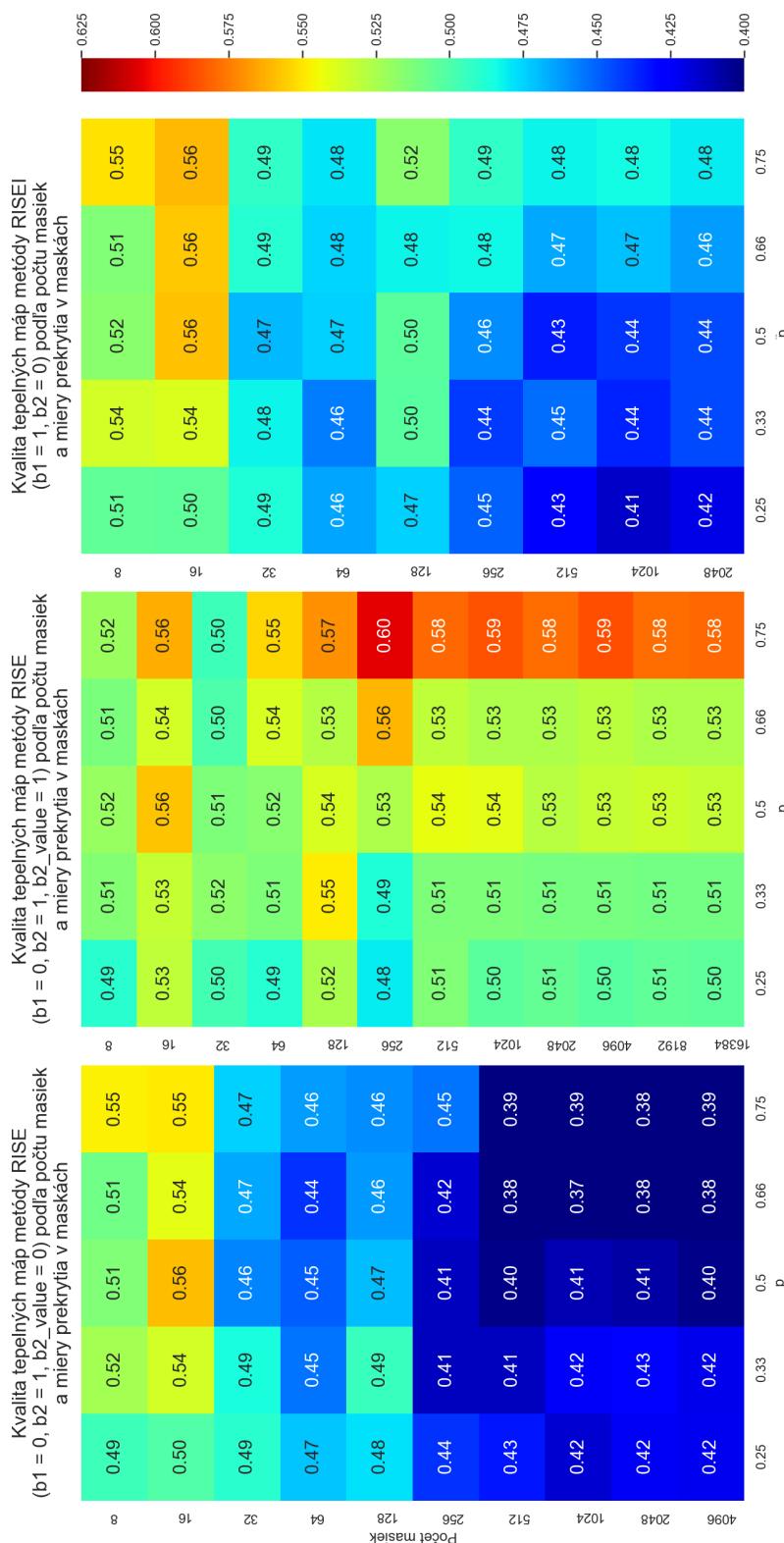
Doposiaľ sme menili parametre $b1$, $b2$ a $b2_value$, ktoré nastavujú hodnotu prekrytia v generovaných maskách. Okrem týchto parametrov má metóda RISEI ďalšie parametre, ktoré nastavujú veľkosť prekrytia a počet generovaných masiek a významne ovplyvňujú výslednú tepelnú mapu. Vybrali sme 3 kombinácie parametrov, z predchádzajúcich experimentov, ktoré menia hodnotu prekrytia a to:

- A $b1 = 0$, $b2 = 1$ a $b2_value = 0$ (RISE),
- B $b1 = 0$, $b2 = 1$ a $b2_value = 1$ (RISE s prekrytím hodnotou 1)
- C $b1 = 1$, $b2 = 0$ (RISEI s dokreslením)

čím sme zahrnuli pôvodnú metódu RISE, metódu RISEI s dokreslením a najlepšiu kombináciu parametrov. Tak ako sme uviedli v predchádzajúcim experimente, používame v tomto aj v ďalších eksperimentoch model 3D CNN so senzitivitou 81% a špecifitou 74%. Kvôli časovej náročnosti týchto experimentov sme použili menšiu dátovu vzorku o veľkosti 10 pozorovaní (5 AD + 5 CN).

Aj tieto experimenty potvrdili, že metóda s prekryvom hodnotou jedna dosahuje najlepšie výsledky. Ďalej sme zistili, že nastavením menšieho prekryvu (parameter p má vyššiu hodnotu) sme v kombinácii s vyšším počtom masiek dosiahli lepšie výsledky, pričom od 256 masiek boli tieto hodnoty takmer rovnaké (Obr. 6.6). Od 2048 masiek boli dosiahnuté výsledky takmer rovnaké, toto môže súvisieť aj so stabilitou tepelných máp pri vyšších hodnotách (Sekcia 6.1.2.1). Metódy RISEI a RISE s prekryvom nula dosiahli najlepšie výsledky s nízkym počtom masiek. Toto je prekvapivé, vzhľadom na to, že tepelné masky z nízkym počtom masiek niesú stabilné. Je preto možné, že sa jedná o chybu alebo náhodu. Preto je vhodné ďalšie skúmanie týchto výsledkov.

Parameter $inpaint_radius$ sme za rozhodli netestovať, keďže z testov vyplynulo, že najvhodnejšie je prekrývať extrémnymi hodnotami, ktoré sú ojedinelé v dátach ($b2_value = 1$). Zmena tohto parametra by znamenala, že by algoritmus použil iné okolité hodnoty - z nich by nevytvoril ojedinelé hodnoty.



Obr. 6.6: Porovnanie kvality nastavených parametrov počtu masiek a veľkosti prekrytia (p) na rôznych nastaveniach metódy RISEI. Kvalitu sme merali ako $(insertion + (1 - deletion))/2$, z toho vypĺýva, že čím vyššia hodnota, tým je vytvorená tepelná mapa kvalitnejšia. Keďže po 2048 generovaných maskách sa hodnoty výrazne nemenili (diagram v strede), tieto počty masiek sme vyniechali v zvyšných experimentov výnemchal (digaramy vľavo a vpravo). Metóda RISEI s prekrytím 1 dosiahla najlepšie výsledky, zároveň výšie hodnoty parametra p sa ukázali ako lepšie.

6.1.4 Porovnanie s existujúcimi metódami

Všetky tri nastavenia metódy RISEI (vrátane metódy RISE) z predchádzajúceho experimentu (Sekcia 6.1.3.1) sme sa rozhodli porovnať s inými existujúcimi metodami. Na základe predchádzajúceho experimentu sme nastavili počet masiek na 2048 a parameter p na 0.75 , keďže pri týchto parametroch sme dosiahli najlepšie výsledky a so zvoleným počtom masiek sú tepelné mapy dostatočne stabilné (Sekcia 6.1.2.1).

Z existujúcich metód sme vybrali metódy GradCAM, Guided Backprop a Guided GradCAM. Keďže sme sa rozhodli, že v našej práci venujeme čo najviac času experimentom a nie implementácií iných, už existujúcich metód, vybrali sme také metódy, ktoré nám ponúkala použitá knižnica *captum* (preto sa neporovnávame s metódou LRP). V prípade metódy GradCAM a Guided GradCAM sme pri zväčšovaní konvolučnej vrstvy použili trilineárnu interpoláciu, keďže pracujeme s trojrozmernými dátami.

Zoznam porovnávaných metód je teda nasledovný:

- A RISE: $b1 = 0$, $b2 = 1$ a $b2_value = 0$,
- B RISE s prekrytím hodnotou 1: $b1 = 0$, $b2 = 1$ a $b2_value = 1$,
- C RISEI: $b1 = 1$, $b2 = 0$,
- D GradCAM,
- E Guided Backprop,
- F Guided GradCAM,
- G Guided RISE (s prekrytím hodnotou 1) - vypočítaný rovnako ako Guided GradCAM a to ako súčin tepelných máp po prvkoch medzi RISE a Guided Backprop.

Ako testovaciu vzorku sme použili 40 pozorovaní (20 AD + 20 CN).

Vytvorené tepelné mapy sme porovnali aj oproti segmentačným maskám, v kto-

rých sme mali vyznačené nasledovné oblasti - šedá hmota, biela hmota, komory a hipokampus. V rámci týchto oblastí nás bude zaujímať koľko zachytili "tepla" z tepelných máp. Na základe ich veľkosti budeme následne počítať hustotu tepla podľa vzorca (6.1), kde o je oblasť, pre ktorú počítame hustotu, N je počet voxellov v oblasti, a t_{no} je hodnota tepla n-tého voxely v oblasti o . Zároveň jednotlivé oblasti dáme do pomeru medzi sebou.

$$h(o) = \frac{\sum_{n=1}^N t_{no}}{N_o} \quad (6.1)$$

6.1.4.1 Metriky

Na vyhodnotenie tepelných máp pre každú snímku sme použili nasledovné metriky, pričom sme vychádzali z existujúcich prác popísaných v návrhu (4.3.1.2).

Kedže tepelné mapy rôznych metód vyjadrujú teplo na rôznych škálach, tepelné mapy normalizujeme a preškálujeme ich do intervalu $< 0, 1 >$.

- A *insertion x deletion* $((insertion + (1 - deletion)) / 2)$ - vyššia hodnota je lepšia,
- B hustota tepla v jednotlivých častiach mozgu - šedá hmota, biela hmota, komory a hipokampus (Vzorec 6.1) - vyššia hodnota je lepšia,
- C pomer hustoty tepla mimo mozgu oproti hustote tepla v mozgu,
 $\frac{h(oblasc_mimo_mozgu)}{h(oblasc_v_mozgu)}$ kde $h(o)$ je (6.1) - nižšia hodnota je lepšia,
- D pomer hustoty tepla zvyšnej časti snímky oproti významným časťam mozgu (biela hmota, komory a hipokampus), $\frac{h(oblasc_mimo_mozgu) + h(seda_hmota)}{h(biela_hmota) + h(komory) + h(hipokampus)}$ kde $h(o)$ je (6.1) - nižšia hodnota je lepšia,
- E pomer hustoty tepla mimo mozgu oproti významným časťam mozgu (biela hmota, komory a hipokampus), $\frac{h(oblasc_mimo_mozgu)}{h(biela_hmota) + h(komory) + h(hipokampus)}$ kde $h(o)$ je (6.1) - nižšia hodnota je lepsia.

6.1.4.2 Výsledky

Najlepšie výsledky takmer vo všetkých metrikách dosiahla metóda Guided Backprop (Tabuľka 6.5).

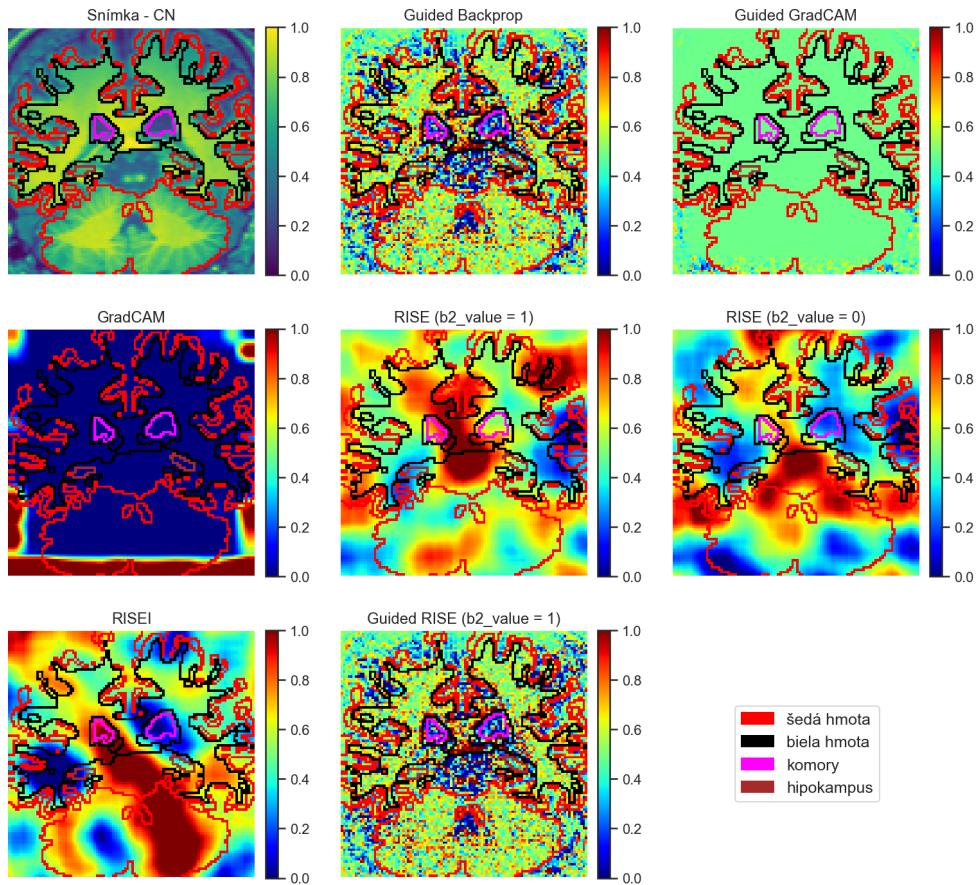
Metriky B nám dávajú predstavu o tom ako jednotlivé metódy rozdeľovali teplo a pomocou nich vieme koľko tepla bolo prideleného mimo mozog. Všetky metódy rozdeľovali teplo pomerne rovnomerne medzi jednotlivé časti mozgu. U metódy GradCAM je na základe metriky C vidieť, že CN pozorovaniam rozdeľovala viac tepla mimo mozog, toto je viditeľné aj na obrázku 6.7. Metrika C je prvotným ukazovateľom ako dobre metóda funguje, najlepší výsledok v nej dosiahla metóda Guided Backprop, a to 0.808. Táto hodnota je avšak pomerne vysoká, keďže hovorí, že hustota tepla mimo mozgu je až 80.8% hustoty tepla v mozgu. Toto môže signalizovať zlý model. Podobne je tomu aj u metrík D a E . Metódy RISE nedosiahli dobré výsledky, najlepšie výsledky dosiahla metóda RISE s $b2_value = 1$, pričom prekonala aj GradCAM.

Samotná metóda RISE dosiahla zo všetkých metód najhoršie výsledky, pričom ju prekonala aj metóda RISEI. Metóda RISE s prekrytím hodnotou jedna dosiahla porovnatelné výsledky s metódou GradCAM, pričom ju v niekoľkých metrikách (A, C) prekonala. Metóda Guided RISE nedokázala vylepšiť metódu Guided Backprop a bola s ňou v niekoľkých metrikách (C, D, E) takmer totožná, v týchto metrikách dokonca prekonala metódu Guided GradCAM.

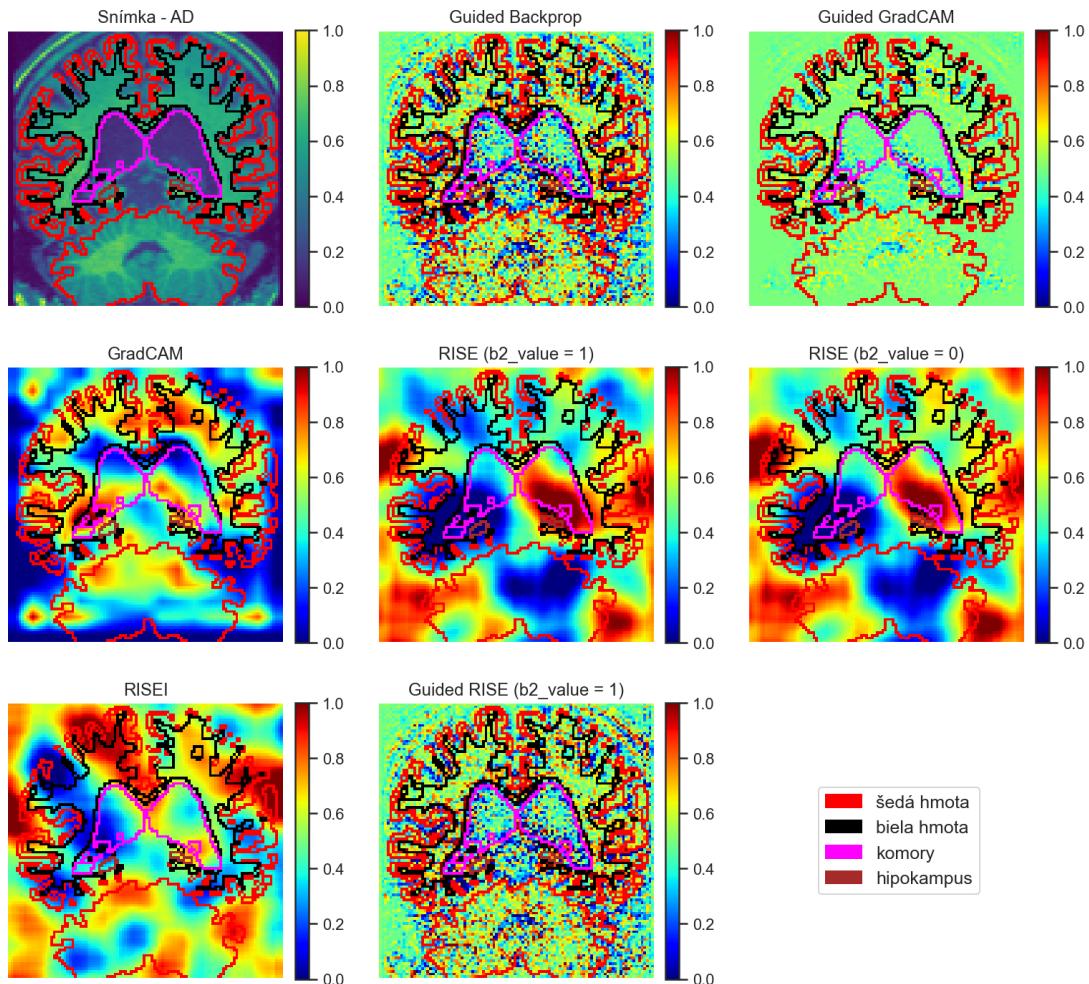
Zo sady pozorovaní sme vybrali sme dva snímky, jeden s najlepšou tepelnou mapou podľa metriky C najlepšej metódy, druhú s najhoršou tepelnou mapou (Obr. 6.8, 6.7). Na základe vizualizácií nemôžeme tvrdiť, že by sa model rozhodoval na základe relevantných častí mozgu (Obr. 6.7). Metóda Guided Backprop zachytila pomerne veľké množstvo tepla na lebke mozgu, čo môže naznačovať, že model môže v niektorých prípadoch vykazovať chybné správanie, tj. rozhodovať sa na základe náhodných, resp. nerelevantných častí mozgu.

Metrika / Metóda	Grad-CAM	Guided Back-prop	Guided Grad-CAM	Guided RISE (b2_value = 0)	RISE (b2_value = 0)	RISE (b2_value = 1)	RISEI
A (AD + CN)	0.572	0.678	0.673	0.547	0.401	0.610	0.443
A (AD)	0.606	0.803	0.810	0.552	0.372	0.550	0.423
A (CN)	0.541	0.564	0.538	0.473	0.426	0.638	0.462
B - biela hmota (AD + CN)	0.466	0.607	0.507	0.606	0.497	0.512	0.498
B - hipokampus (AD + CN)	0.400	0.559	0.500	0.556	0.491	0.435	0.501
B - komory (AD + CN)	0.417	0.333	0.491	0.331	0.494	0.537	0.484
B - nie mozog AD + CN)	0.503	0.464	0.498	0.464	0.502	0.494	0.500
B - šedá hmota (AD + CN)	0.452	0.572	0.504	0.571	0.499	0.504	0.503
C (AD + CN)	1.108	0.808	0.986	0.809	1.014	0.970	1.002
C (AD)	0.854	0.810	0.942	0.810	0.985	0.996	1.004
C (CN)	1.234	0.802	0.999	0.803	1.015	0.968	0.999
D (AD + CN)	1.118	0.875	0.991	0.875	1.000	0.989	0.999
D (AD)	0.948	0.882	0.970	0.882	0.991	1.005	0.990
D (CN)	1.171	0.868	0.998	0.869	1.004	0.983	1.005
E (AD + CN)	1.130	0.836	0.989	0.837	1.004	0.982	0.995
E (AD)	0.910	0.840	0.957	0.841	0.984	1.002	0.993
E (CN)	1.206	0.828	0.998	0.830	1.006	0.975	1.002

Tabuľka 6.5: Porovnanie kvality a správnosti tepelných máp RISEI oproti iným existujúcim metódam. Hodnoty jednotlivých metrik sú strednými hodnotami metrik pre jednotlivé pozorovania v testovacej sade. Popis metrik sa nachadza v sekcií 6.1.4.1



Obr. 6.7: Porovnanie tepelných máp vytvorených porovnávaními metódami. Vybrali sme snímku s najlepšiou tepelnú mapu pre metódu Guided Backprop podľa metriky $C = 0.781$. Tak ako aj na obrázku 6.8, aj tu si môžeme všimnúť výrazné rozdiely medzi jednotlivými metódami. Metóda GradCAM pridelila veľmi málo tepla do mozgu a veľmi málo tepla na krajoch snímky.



Obr. 6.8: Porovnanie tepelných máp vytvorených porovnávaními metódami. Vybrali sme snímku s najlepšou tepelnú mapu pre metódu Guided Backprop podľa metriky $C = 0.839$. Môžeme si všimnúť výrazné rozdiely medzi jednotlivými metódami. Teplo z metódy Guided Backprop, ktoré sa nachádza mimo mozgu, lemuje časť lebky, v ľavej a aj v pravej časti snímky. Metóda GradCAM oproti metódam RISE a RISEI rozdelila pomerne veľa tepla v rámci mozgu. Metóda RISE s prekrytím hodnotou nula aj jedna rozdelila pomerne veľké množstvo tepla v oblasti komôr.

6.2 Zhrnutie

Metódu RISE sme otestovali na niekoľkých architektúrach neurónových sietí pričom architektúrou, ktorú sme vyhodnotili ako navhodnejšiu pre ďalšie experimenty sme použili v ďalších experimentoch. Nevyhodnocovali sme, ktorý model je najlepší, pretože použité metriky *insertion* a *deletion* nie sú na takúto úlohu vhodné.

Podarilo sa nám vyhodnotiť stabilitu vytváraných tepelných máp. Z výsledkov sme usúdili, že so stúpajúcim počtom generovaných masiek stúpa stabilita vytvárania tepelných máp pričim chyba po určitem počte je už zanedbateľná.

Vyhodnotili sme rôzne hodnoty prekrytie pre metódu RISEI. Následne sme hľadali optimálny počet masiek a mieru prekrytie.

Metódu RISEI (a jej rôzne nastavenia na základe doterajších experimentov) sa nám podarilo porovnať s niekoľkými existujúcimi metódami. V sledovaných metrikách sme dosiahli horšie výsledky ako metóda Guided Backprop, ktorá vo väčšine metrič dominovala. Metóda Guided Backprop na rozdiel od metódy RISE(I) musí poznať model, čo môže byť v niektorých prípadoch nevýhodou. Je teda uplatniteľná na menšiu množinu modelov ako metóda RISE(I).

Metóda RISEI dosiahla lepšie výsledky ako pôvodná metóda RISE. Avšak oveľa jednoduchšou úpravou metódy RISE, a to použitím prekrytie hodnotou jedna sme dosiahli lepšie výsledky ako použitím dokreslenia, pričom sme dosiahli rýchlejšie generovanie masiek.

Dosiahnuté výsledky naznačujú, že máme model, ktorý sa nerozhoduje na základe relevantných častí mozgu - rozhoduje napr. aj na základe lebky. Toto má vo veľkej miere vyplýv na kvalitu tepelných máp metódy RISE, keďže vytvára tepelné mapy z veľkého počtu predikcií k zamaskovaným snímkam. Ďaľšie experimenty by bolo vhodné realizovať na modeli s lepšou úspešnosťou a na pozorovaniach s odstránenou lebkou zo snímek v predspracovaní (kedže náš model sa rozhodoval na základe tej a tie najlepšie modely boli trénované na snímkoch bez tej).

7. Zhodnotenie

V našej práci sme sa venovali uplatneniu interpretovateľnosti a vysvetliteľnosti neurónových sietí pri vyhodnocovaní medicínskych obrazových dát.

Skúmanú doménu a problém sme si naštudovali, a zistené informácie uviedli v analýze práce. Na základe toho sme navrhli modifikáciu existujúcej metódy na vyhodnocovanie rozhodnutí neurónových sietí. Výhodou navrhnutej metódy je, že nemusí poznať použitý model, a tak je ju možné použiť aj pri komplikovanejších modeloch (napr. kombinácia neurónovej siete a inej metódy strojového učenia, model so špecifickým predspracovaním do vektoru črt a pod.). Zároveň sme narvhlí spôsob jej overenia - vyhodnotenie a porovnanie výsledkov.

Na základe návrhu sme naimplementovali modifikáciu existujúcej metódy RISE s dokreslením pre 3D volumetrické dáta. Vytvorenú metódu sme overili v niekoľkých experimentoch na nami natrénovanom modeli. Vytvorená metóda disponuje viačerími parametrami ovplyvňujúcimi jej správanie (napr. hodnota prekrytie, miera prekrytie), preto sme v experimentoch overovali zvolené kombinácie parametrom, pričom u jednej dvojice parametrov sme prehľadávali mriežku parametrov. Keďže metóda používa na vytváranie tepelných máp masky s náhodným prekrytím, overili sme vplyv počtu vygenerovaných masiek na stabilitu tepelných máp. Metóda sa u vyššieho počtu vygenerovaných masiek ukázala ako stabilná.

Metódu sme porovnali s inými existujúcimi metódami - GradCAM, Guided Backprop a Guided GradCAM. Metóda v sledovaných metrikách dosiahla horšie výsledky ako metóda Guided Backprop, pričom dosiahla porovnateľné výsledky s metódou

GradCAM. Kombináciou s Guided Backprop sme vytvorili metódu Guided RISE, ktorá dosiahla výsledky blízke Guided GradCAM a v niektorých ohľadoch takmer rovnaké s Guided Backprop. Použitie dokreslenia ako prekrycia sa neukázalo ako vhodný prístup v doméne rádiologických obrazových dát.

Výsledné tepelné mapy sme vizualizovali. Tepelné mapy z metódy Guided Backprop naznačujú, že sa model nerozhoduje na základe relevantných častí mozgu. V ďalších experimentoch by bolo vhodné použiť lepší model.

7.0.1 Zhodnotenie cieľov práce

Vytvorenie novej, alebo vylepšenie existujúcej metódy pre vysvetľovanie rozhodnutí neurónových sietí (Sekcia 3.1) Vytvorili sme modifikáciu už existujúcej metódy RISE, do ktorej sme priniesli niekoľko funkcionálnych vylepšení - podporu pre 3D volumetrické dáta a možnosť nastavenia rôznych hodnôt preprvy (vrátane dokreslením). Ukázalo sa, že rôzne hodnoty prekryvu majú vplyv na kvalitu tepelných máp. Navrhnutý prekryv dokreslením bol lepší oproti hodnote nula (pôvodná metóda RISE), avšak prekryv hodnotou jedna prekonal dokreslenie pričom je výpočtovo jednoduchší (Sekcia 6.1.4.2. V doméne rádiologických dát sa dokreslenie neukázalo ako navhodnejší spôsob prekrycia.

Využitie vytvorenej metódy na určenie miery správnosti modelu neurónovej siete detegujúcej Alzheimerovu chorobu (Sekcia 3.2) Vytvorenú metódu sme využili na určenie správnosti modelu tak, že vytvorené tepelné mapy sme vyhodnocovali na základe nami zadefinovaných metrík (Sekcia 6.1.4.1). Tieto metriky využívali segmentačné masky, ktoré overovali, či vytvorená tepelná mapa dáva zmysel z pohľadu anatómie mozgu. Výsledky ukázali, že natrénovaný model sa nerozhodoval na základe relevantných častí mozgu, čo sa odzrkadlilo v sledovaných metrikách, napr. pomer tepla mimo mozgu voči teple v mozgu.

7.0.2 Limitácie

Oproti metódam ako je GradCAM alebo Guided Backprop vytvorená metóda vyžaduje viac času na vytvorenie tepelnej mapy. Ten je variabilný a závisí od počtu masiek, veľkosti pamäte, počtu jadier a GPU. Viac výpočtových zdrojov umožňuje generovanie masiek paralelizovať a generovať vo väčších dávkach.

7.0.3 Možné rozšírenia

Ako možné rozšírenia tejto práce sme identifikovali:

- Preskúmať metódu na lepšie natrénovaných modeloch (s úspešnosťami blížiacim sa k state-of-the-art), keďže vytvorená metóda bola (detailnejšie) oveřená iba na jednom modeli, u ktorého sme identifikovali, že sa rozhoduje na základe nie relevantných častí snímky. Takou časťou je napr. lebka, ktorý by bolo preto vhodné v predspracovaní odstrániť.
- Preskúmať vplyv počtu masiek na stabilitu tepelných máp, pretože to môže zredukovať ich potrebný počet a celkový čas generovania tepelných máp.
- Porovnať metódu s inými perturbačnými metódami, keďže sa navrhnutá metóda medzi ne zaraďuje. Navrhujeme porovnanie tepelných máp z hľadiska rýchlosťi ich vytvárania a ich správnosti.

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. GILPIN, Leilani H; BAU, David; YUAN, Ben Z; BAJWA, Ayesha; SPECTER, Michael; KAGAL, Lalana. Explaining explanations: An overview of interpretability of machine learning. In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. 2018, s. 80–89.
3. 2019. Dostupné tiež z: <http://www.alzheimer.sk/informacie/alzheimerovachoroba.aspx>.
4. DUTHEY, Béatrice. Background paper 6.11: Alzheimer disease and other dementias. *A Public Health Approach to Innovation*. 2013, s. 1–74.
5. KHAN, Tapan. *Biomarkers in Alzheimer's Disease*. Academic Press, 2016.
6. 2017. Dostupné tiež z: <https://www.alz.org/alzheimers-dementia/facts-figures>.
7. WORKING, G Biomarkers Definitions. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*. 2001, roč. 69, č. 3, s. 89–95.
8. HAYKIN, Simon S et al. *Neural networks and learning machines/Simon Haykin*. New York: Prentice Hall, 2009.
9. 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).

10. O'SHEA, Keiron; NASH, Ryan. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*. 2015.
11. SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2014.
12. 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.
13. SZEGEDY, Christian; LIU, Wei; JIA, Yangqing; SERMANET, Pierre; REED, Scott; ANGUELOV, Dragomir; ERHAN, Dumitru; VANHOUCKE, Vincent; RABINOVICH, Andrew. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, s. 1–9.
14. MONTAVON, Grégoire; SAMEK, Wojciech; MÜLLER, Klaus-Robert. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*. 2018, roč. 73, s. 1–15.
15. SIMONYAN, Karen; VEDALDI, Andrea; ZISSERMAN, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*. 2013.
16. 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.
17. 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.
18. PETSIUK, Vitali; DAS, Abir; SAENKO, Kate. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*. 2018.

19. 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.
20. 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.
21. 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.
22. 2017. Dostupné tiež z: <http://adni.loni.usc.edu/>.
23. ESMAEILZADEH, Soheil; BELIVANIS, Dimitrios Ioannis; POHL, Kilian M; ADELI, Ehsan. End-to-end Alzheimer's disease diagnosis and biomarker identification. In: *International Workshop on Machine Learning in Medical Imaging*. 2018, s. 337–345.
24. SMITH, Stephen M. Fast robust automated brain extraction. *Human brain mapping*. 2002, roč. 17, č. 3, s. 143–155.
25. SUK, Heung-Il; LEE, Seong-Whan; SHEN, Dinggang; INITIATIVE, Alzheimer's Disease Neuroimaging et al. Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. *Brain Structure and Function*. 2016, roč. 221, č. 5, s. 2569–2587.
26. HOSSEINI-ASL, Ehsan; KEYNTON, Robert; EL-BAZ, Ayman. Alzheimer's disease diagnostics by adaptation of 3D convolutional network. In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, s. 126–130.
27. SUK, Heung-Il; LEE, Seong-Whan; SHEN, Dinggang; INITIATIVE, Alzheimer's Disease Neuroimaging et al. Deep ensemble learning of sparse regression models for brain disease diagnosis. *Medical image analysis*. 2017, roč. 37, s. 101–113.
28. BÖHLE, Moritz; EITEL, Fabian; WEYGANDT, Martin; RITTER, Kerstin. Layer-wise relevance propagation for explaining deep neural network decisions.

- ons in MRI-based Alzheimer's disease classification. *Frontiers in aging neuroscience*. 2019, roč. 11, s. 194.
- 29. CHEN, Wai Kai. *The electrical engineering handbook*. Elsevier, 2004.
 - 30. RAVI, S.; PASUPATHI, P.; MUTHUKUMAR., S.; KRISHNAN, N. Image in-painting techniques - A survey and analysis. In: *2013 9th International Conference on Innovations in Information Technology (IIT)*. 2013, s. 36–41.
 - 31. IOFFE, Sergey; SZEGEDY, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*. 2015.

A. Plán práce

A.1 Letný semester - DP1

V tomto semestri plánujem pracovať na analýze domény, návrhu metódy a jej implementácií.

A.2 Zimný semester - DP2

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.2.1 Vyjadrenie k plneniu plánu

V tomto semestri sa nám podarilo splniť všetky stanovené ciele. Natrénovali sme niekoľko modelov detekujúcich Alzheimerovu chorobu z MRI snímkov. Čo sa týka úspešnosti týchto modelov, bohužiaľ sa nám nepodarilo dosiahnuť tak dobré výsledky ako u iných prác. Avšak, naším cieľom nie je natrénovať najlepší model, takže táto úspešnosť vyzerá byť zatial pre nás postačujúca.

Metódu sme implementovali, tak, ako sme ju navrhli, pričom sme pridali vylepšenia ako multiprocessing - paralelné generovanie masiek.

S hyper-parametrami navrhnutej metódy sme experimentovali (ale nie so všetkými, pretože ich je veľa), avšak sme nerobili žiadne prehľadávanie optimálnych parametrov.

Dosiahnuté výsledky sme skúmali a diskutovali ich v závere overenia riešenia pričom sme navrhli ďalšie kroky.

A.3 Letný semester - DP3

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,
- vykonať úpravy v navrhnutej metóde na základe doterajších výsledkov experimentov,
- vyhodnotiť stabilitu tepelných máp,
- optimalizovať vstupné parametre do RISEI metódy,
- porovnať navrhnutú metódu s existujúcimi metódami,

- vyhodnotiť a porovnať vykonané experimenty,
- odovzdať prácu.

A.3.1 Vyjadrenie k plneniu plánu

Plán práce sa nám v tomto semestri podarilo dodržať. Z experimentov nevzišli žiadne potrebné úpravy metódy, preto žiadne úpravy metódy neboli v tomto semestri vykonané. Taktiež sme vyhodnotili stabilitu tepelných máp. Rovnako sme hľadali aj optimálnu kombináciu vstupných parametrov metódy RISEI tak, že sme si vytvorili možné kombinácie parametrov, ktoré sme vyhodnotili. Neoptimalizovali sme ale všetky vstupné parametre, pri niektorých sme uznali, že to u nich nedáva zmysel. Vytvorenú metódu sme taktiež porovnali s existujúcimi metódami GradCAM, Guided Backprop a Guided GradCAM. Vykonali sme veľké množstvo experimentov (rôzne parametre RISEI, kombinácia RISEI a guided backprop atď.), ktoré sme vyhodnotili (napr. vyhodnotenie na segmentačných maskách) a porovnali.

Dodatok A. Plán práce

B. Technická dokumentácia

Metóda RISEI je implementovaná v jazyku Python, rovnako ako aj jej vyhodnotenie a porovnanie s ostatnými metódami.

B.1 Príprava vývojového prostredia

Na správu python-ovských balíkov a vývojového prostredia je použitá conda, ktorú je nutné nainštalovať. Condu je možné nainštalovať cez distribúciu Anaconda (Anaconda obsahuje grafické rozhranie a množstvo nástrojov/programov) alebo menšiu distribúciu Miniconda. V prípade, že potrebujete šetriť miesto na disku, odporúčam menšiu distribúciu Miniconda.

Po inštalácii condy zadajte nasledovný príkaz v koreňovom adresári repozitára. Tento príkaz vytvorí nové conda prostredie v ktorom nainštaluje potrebné python balíky.

```
1 $ conda env create -f environment.yml
```

Následne pre aktiváciu conda prostredia zadajte nasledovný príkaz.

```
1 $ conda activate dp-timzatko
```

Teraz je možné používať shell, v ktorom bolo aktivované conda prostredie *dp-timzatko*, na spúšťanie Python skriptov a Jupyter notebookov.

Následne spustite Jupyter notebook klienta.

```
1 $ jupyter-notebook
```

Teraz je možné, prezerať, spúštať a upravovať jupyter notebooky v repozitári.

B.2 Závislosti (použité knižnice)

Na implementáciu riešenia sme použili nasledovné Python knižnice (uvádzame len tie najvýznamnejšie). Kompletný zoznam sa nachádza v súbore */REPOZITÁRJ/environment.yml*.

- numpy - na prácu s vektormi, maticami a matematickými operáciami nad nimi.
- pandas - na vytváranie, a ukladanie tabuľkami.
- seaborn - vykreslovanie grafov.
- matplotlib - vykreslovanie rádiologických snímkov, tepelných mám a segmentačných masiek.
- opencv - na dokreslenie v RISEI.
- tensorflow (v2.3.1), tensorboard - implementácia, trénovanie, evaluácia modelu na predikciu alzheimerovej choroby.
- torch, torchvision - evaluácia modelu na predikciu alzheimerovej choroby. Pytorch je potrebný, pretože je závislosťou knižnice *captum*, ktorú používame na vytváranie tepelných masiek existujúcimi metódami (GradCAM a pod.).
- SimpleITK - načítanie volumetrických dát z disku.
- scikit-image - práca s vizuálnymi dátami (augmentácie, zmena veľkosti).

B.3 Technické riešenie

Implementácia riešenia (RISEI, model, evaluácia atď.) sa nachádza v adresári *[REPOZITÁR]/src*. Funkcionalita z tohto adresára je následne importovaná jupyter notebookmi v adresári *[REPOZITÁR]/conda_notebooks*. Každý z notebookov má iný účel - trénovanie modelu, vyhodnotenie metódy RISEI, porovnanie metód a pod (detailný opis k týmto notebookom sa nachádza v prílohe C Opis digitálnej časti práce).

B.4 Moduly

Adresár *[REPOZITÁR]/src* obashuje nasledovné Python moduly, ktoré majú uvedené zodpovednosti.

- **src.risei** - implementácia metódy RISEI (exportuje triedu RISEI) - generovanie masiek.
- **src.model** - obsahuje pomocné funkcie na prácu s tensorflow modelom (načítanie checkpointu atď.).
- **src.model.cnn_3D** - implementácia 3D konvolučnej neurónovej siete v tensorflow-e.
- **src.model.res_net** - implementácia 3D a 2D siete ResNet v tensorflow-e.
- **src.model.compile_model** - komplilácia tensorflow modelu a nastavenie metrík, optimizéru, chybovej funkcie a predvolených nastavení.
- **src.model.create_model** - vytvorenie modelu.
- **src.model.training** - spustenie trénovania modelu, vrátane vytvorenie tensorflow datasetu, nastavenia augmentácií, pripojenia k tensorboardu a pod. na základe vstupných parametrov.

- **src.model.mri_tensorboard_callback** - výpis rádiologických snímkov po epochách/iteráciách do tensorboard-u pri trénovaní.
- **src.model.evaluation** - vyhodnotenie modelu (matica zmätenia, klasifikačné metriky) a priebehu jeho trénovalia.
- **src.model.torch.cnn_3D** - implementácia 3D konvolučnej neurónovej siete v pytorch-y.
- **src.data** - práca s volumetrickými dátami, konverzia tensorflow sequence do numpy a opačne.
- **src.data.description** - popisné štatistiky o dátovej sade.
- **src.data.augmentations** - augmentácie.
- **src.data.mri_sequence** - načítanie MRI snímkov, štítkov a segmentačných masiek z disku. Zmena veľkosti snímkov, orezanie snímkov, šandardizácia dát, rozdelenie dát do dávok.
- **src.data.train_test_split** - rozdelenie dátovej sady na trénovaciu, testovaciu a validačnú.
- **src.data.selector** - výber záznamov z dátovej sadny na základe triedy AD/CN a správnosti klasifikácie modelom.
- **src.data.evaluation.segmentation_masks** - evaluácia tepelných máp podľa segmentačných masiek.
- **src.heatmaps** - generovanie tepelných máp.
- **src.heatmaps.evaluation** - evaluácia kvality tepelnej mapy podľa metrík *insertion* a *deletion*, perzistencia histórie - tepelných máp a ich evaluácie, načítanie histórie evaluácie, vizualizácie v diagramoch.

Funkcie a triedy v moduloch sú v primeranom rozsahu dokumentované pomocou komentárov. Ďalej bližšie opíšeme najviac dôležitý modul implementujúci navrhnutnú metódu *RISEI*.

B.4.1 Modul: src.risei

Tento modul poskytuje triedu RISEI ktorá slúži na generovanie masiek, z ktorých sa vytvárajú tepelné mapy.

B.4.1.1 Trieda: RISEI

Trieda RISEI slúži na generovanie masiek.

```
class src.risei.RISEI(input_size, s=8, p1=0.5, b1=0.8, b2=0.5,
                      b2_value=0,
                      in_paint='2d',
                      in_paint_radius=20,
                      in_paint_algorithm=cv2.INPAINT_NS,
                      in_paint_blending=True,
                      in_paint_2d_to_3d=False,
                      processes=4,
                      debug=False,
)
```

Parametre

- **s** - veľkosť mriežky, z ktorej sa vytvára binárna maska.
- **p1** - pravdepodobnosť, že pixel v mriežke bude biely.
- **b1** - miera prekryvu medzi pôvodným obrázkom a dokreslením. Ak je θ tak sa dokreslenie vôbec nevykoná.
- **b2** - miera prekryvu medzi pôvodným obrázkom s dokreslením a "čierrou" maskou.
- **b2_value** - hodnoty v "čiernej" maske.

Dodatok B. Technická dokumentácia

- **in_paint** - typ dokreslenia. Môže byť *2d* alebo *3d*. V prípade *2d* je dokreslenie realizované iba v prvej dimenzií.
- **in_paint_radius** - rádius dokreslenia (posúva sa ďalej do knižnice *opencv*).
- **in_paint_algorithm** - algoritmus dokreslenia, môže byť *cv2.INPAINT_NS* alebo *cv2.INPAINT_TELEA*.
- **in_paint_blending** - ak *True* dokreslenie bude prekryté s pôvodným snímkom podľa interpolovanej čiernej masky (tak nevzniknú žiadne ostré hrany).
- **processes** - počet procesov v ktorých sa budú generovať masky.
- **debug** - ak *True* budú do pamäte ukladané medzivýsledky z generovania masiek (binárna maska, interpolovaná maska atď.).

Metódy Metódy triedy RISEI.

```
generate_masks(n, image, log=True, seed=None)
```

Metóda vygeneruje masky k poskytnutej snímke (*image*).

Parametre:

- **n** - počet masiek na vygenerovanie,
- **image** - 3D snímka (*x, y, z*) (k nej budú vygenerované masky),
- **log** - ak *True* na štandardnom výstupe bude zobrazený aktuálny stav generovania masiek,
- **seed** - seed pre náhodu.

```
show_from_last_run(i, z, figsize=(12, 8), dim=0):
```

Parametre:

- **i** - i-ta maska na zobrazenie,
- **z**,
- **figsize** - veľkosť vykresleného diagarmu,
- **dim** - reprezentuje rozmer - 0, 1, 2.

B.4.2 Modul: src.heatmaps.evaluation

Temto modul poskytuje generovanie a evaluáciu kvality tepelnej mapy podľa metrík *insertion* a *deletion*, perzistenciaj histórie - tepelných máp a ich evaluácie, náčitanie histórie a vizualizácie v diagramoch.

B.4.2.1 Trieda: HeatmapEvaluationV3

Táto trieda vytvorí a vyhodnotí tepelné mapy, následne vráti históriu evaluácie.

```
class src.risei.HeatmapEvaluationV3(  
    predict_fn,  
    heatmap_fn,  
    sequence,  
    evaluation_step_size=1000,  
    evaluation_max_steps=-1,  
    evaluation_batch_size=32)
```

Parametre

- **predict_fn** - funkcia *def predict_fn(batch_x)*, ktorá pre dávku snímkov vráti pravdepodobnosti predikovaných pre jednotlivé triedy.
- **heatmap_fn** - funkcia *heatmap_fn(image_x, image_y, evaluation_idx, seed, log)*, ktorá vráti tepelnú mapu pre snímku *image_x*.
- **sequence** - sekvencia snímkov - musí byť *src.data.MRISequence*.

- **evaluation_step_size** - kol'ko voxelov bude vložených v jednom kroku pri evaluácii tepelnej mapy.
- **evaluation_batch_size** - koľko snímkov vkladať pri evaluácii do modelu.

Metódy Metódy triedy HeatmapEvaluationV3.

```
evaluate(log=False, verbose=0, seed=None)
```

Spustí generovanie a evaluáciu tepelných máp. Metóda vráti *src.heatmaps.evaluation.HeatmapEvaluationHistory* s históriou evaluácie a s metrikami.

Parametre:

- **log** - zapne alebo vypne výpisu.
- **verbose** - úroveň výpisov 0, 1, 2 (najmenej až najviac).
- **seed** - seed pre náhodu (posunutý do *heatmap_fn*).

B.4.2.2 Trieda: HeatmapEvaluationHistory

Táto uchováva históriu evaulácie tepelných máp. Umožňuje ich načítavať a ukladať na disk.

```
class src.risei.HeatmapEvaluationHistory(  
    method, auc, arr_auc, arr_heatmap, arr_x, arr_y, arr_y_pred,  
    arr_y_pred_heatmap,  
    arr_voxels, arr_max voxels, arr_step_size)
```

Túto triedu inicializuje knižnica (v *src.heatmaps.evaluation.HeatmapEvaluationV3*), nemal by ju inicializovať klientský kód.

Metódy Metódy triedy HeatmapEvaluationHistory.

```
save(path, filename)
```

Uloží históriu na disk.

Parametre:

- **path** - cesta k súboru.
- **filename** - názov súboru.

```
@staticmethod  
def load(path, filename):
```

Načíta históriu z disku.

Parametre:

- **path** - cesta k súboru.
- **filename** - názov súboru.

```
description(percentage=True, cls_index=None)
```

Vypíše popisné štatistiky (min, max, priemer, štandardná odchýlka) pre plochy pod krivkou jednotlivých evaluácií.

Parametre:

- **percentage** - ak *False* vráti AUC v absolútnych číslach vzhľadom na počet voxelov a nie hodnotu z intervalu $<0, 1>$.
- **cls_index** - ak *None* vráti štatistiky pre všetky pozorovanie, ak číslo, vráti štatistiky pre index danej triedy.

Táto trieda poskytuje aj ďalšie metódy pre vizualizáciu plochy pod krivkou pre

jednotlivé pozorovania a pod., popísali sme iba najdôležitejšie z nich.

Atribúty

- **arr _ heatmap** - vygenerované tepelné mapy pre všetky snímky.
- **arr _ x** - snímky.
- **arr _ y** - skutočne triedy ku snímkam.
- **arr _ y _ pred** - predikované triedy ku snímkam.
- **arr _ auc** - hodnoty AUC pre jednotlivé snímky.

Dodatok B. Technická dokumentácia

Dodatok B. Technická dokumentácia

C. Obsah priloženého digitálneho média

Evidenčné číslo práce v informačnom systéme FIIT-182905-86077.

Obsah digitálnej časti práce (archív ZIP):

- `/DP_TimotejZatko.pdf` — Práca vo formáte PDF.
- `/DP_prilohy_TimotejZatko.pdf` — Prílohy vo formáte PDF.
- `/repo.zip` — Zdrojové súbory a dát vo formáte ZIP.

Obsah súboru `repo.zip`:

- `/repo/thesis/` — Zdrojové súbory práce vo formáte L^AT_EX.
- `/repo/tmp/` — Dátová sada, zoserializované natrénované modely, záznamy z trénovania modelov, zoserializované výsledky experimentov.
- `/repo/src/` — Všetky zdrojové súbory, vrátane metódy RISEI, modelov, generovania tepelných máp. Obsahuje Python balíky.
- `/repo/scripts/` — Pomocný skript na stiahnutie dát z Google Drive.
- `/repo/assets/` — Zdrojové súbory diagramov použitých v práci.
- `/repo/colab_notebooks/` — Obsahuje Jupyter notebooky z trénovania modelov na platforme Google Colab s využitím GPU aj TPU (obsahuje prvé natrénované modely)

Dodatok C. Obsah priloženého digitálneho média

- `/repo/conda_notebooks/` — Obsahuje Jupyter notebooky.
- `/repo/conda_notebooks/dataset.ipynb` — Rozdelenie dátovej sady.
- `/repo/conda_notebooks/tensorboard.ipynb` — Tensorboard.
- `/repo/conda_notebooks/augmentations.ipynb` — Vizualizácia implementovaných augmentácií.
- `/repo/conda_notebooks/training/training_history.ipynb` — Vizualizácia priebehu trénovania.
- `/repo/conda_notebooks/training/augmentations/` — Porovnanie vplyvu augmentácií na výsledný model.
- `/repo/conda_notebooks/training/2d_ResNet18/` — Trénovanie modelu 2D ResNet18 (viacero jupyter notebookov).
- `/repo/conda_notebooks/training/3d_ResNet18/` — Trénovanie modelu 2D ResNet18 (viacero jupyter notebookov).
- `/repo/conda_notebooks/training/3d_cnn/` — Trénovanie modelu 3D CNN (viacero jupyter notebookov).
- `/repo/conda_notebooks/risei/model_comparison_v1/` — Porovnanie metódy RISEI a jej parametrov na natrénovaných modeloch (prvá iterácia, staré modely, viacero jupyter notebookov).
- `/repo/conda_notebooks/risei/model_comparison_v2/` — Porovnanie metódy RISEI a jej parametrov na natrénovaných modeloch (druhá iterácia, nové modely, viacero jupyter notebookov).
- `/repo/conda_notebooks/risei/evaluation_history_ins_del.ipynb` — Vyplňanie štatistik pre vybranú evaluáciu metódy.
- `/repo/conda_notebooks/risei/evaluation_history_ins_del_comparison.ipynb` — Porovnanie vybraných dvoch evaluácií metód.

Dodatok C. Obsah priloženého digitálneho média

- `/repo/conda_notebooks/risei/evaluation_history_segmentation_masks-.ipynb` — Vyhodnotenie metódy voči segmentačným maskám.
- `/repo/conda_notebooks/risei.evalution_all.ipynb` — Porovnanie všetkých experimentov.
- `/repo/conda_notebooks/risei/risei.ipynb` — RISEI - zobrazenie generovaných masiek podľa nastavených parametrov.
- `/repo/conda_notebooks/risei/experiments/methods/` — Vyhodnotenie tepelných máp z iných, existujúcich metód (viacero jupyter notebookov).
- `/repo/conda_notebooks/risei/experiments/parameters/` — Hľadanie optimálneho počtu parametrov (viacero jupyter notebookov).
- `/repo/conda_notebooks/risei/experiments/stability/` — Vyhodnotenie kvality stability tepelných máp podľa počtu vygenerovaných masiek (viacero jupyter notebookov).