HALMSTAD UNIVERSITY

MASTER THESIS

Master’s Program Information Technology

Prediction of Neurodegenerative Diseases using Brain Images

Aiman Mujtaba Khan and Abhishek Pulicherla

Halmstad University, March 11, 2020–version 4.0

Aiman Mujtaba Khan and Abhishek Pulicherla: *Prediction of Neurodegenerative Diseases using Brain Images*, , © March 2020

CONTENTS

LIST OF FIGURES

ACRONYMS

AD	Alzheimer's Disease
MCI	Mild Cognitive Impairment
PET	Positron Emission Tomography
ADNI	Alzheimer's Disease Neuroimaging Initiative
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic Curve

1 INTRODUCTION

Dementia is a term that is characterized by a mental decline. Every 3 seconds, someone in the world develops dementia. Dementia has many types, among these types, the most common form of dementia is Alzheimer's disease.[1]

What's difficult in Alzheimer is the diagnosis. Alzheimer's is a progressive disease. Initial stages include Normal cognition(which is how our brain normally is) to mild cognitive impairment(MCI), which involves small changes in a personâs behaviour and memory. For some people MCI might not lead to a more neurodegenerative disease and may revert back to normal cognition[3] but for the others MCI progressions to Alzheimer's. The main symptoms of AD are forgetfulness, confusion, disorientation, and loss of language abilities , the exact cause of this is not exactly known[20] Extensive and repetitive tests are performed in order to give a diagnosis. Most of the time, the diagnosis is too late or inaccurate[2].

The diagnosis of Alzheimer's disease can be very challenging for clinicians, especially in the early stages. Moreover, the use of particular medication and co-morbid conditions - the presence of more than one mental disorder, like anxiety, depression, and the likes. - can confound the assessment of clinicians. Extensive and repetitive tests are performed in order to give a diagnosis, but the results are more often not very definitive.

A high degree of certainty is needed before a diagnosis can be made, as it can have a significant effect on the patients, their families, and further medical treatment.[21] fragments in the brain. These tangles develop quickly, and in predictable patterns, this build-up blocks the communication between the nerve cells and leads to their death[14].

For this reason, the clinicians have started the use of imaging radioactive tracers in the workup of patients suspected of having AD or any other kind of dementia[17]The prime suspects of AD are plaques and tangle, which are the build-up of protein

Brain Positron Emission Tomography (PET) scans with 2-[fluorine-18] fluoro-2-deoxy-d-glucose as the radioactive tracer has improved the chances of timely diagnosis because they show the build-up of protein fragments in the brain. Figure 1 is an image of Brain PET, which shows the difference between a healthy brain and an Alzheimer's brain. Sometimes, the difference is not as apparent,

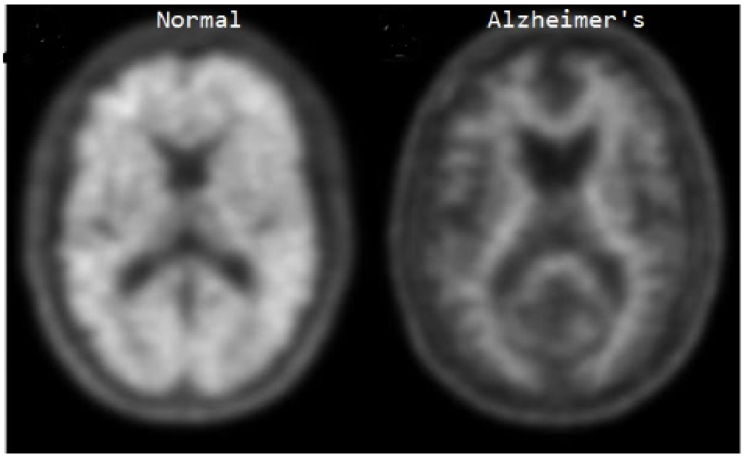


Figure 1: PET Brain Scan

Accurate interpretation of the PET scans requires knowledge and recognition of the standard and affected metabolic patterns that show up in the scans. Visual interpretations by the clinicians are essential components of the diagnosis, but they take a long time and have a level of uncertainty. Work has been done to develop machine learning and deep learning models that can make the diagnosis of a person having AD or not, using brain PET scans[23].

With the improvements of these models, the resulting diagnoses generated by these models are also improving. Numerous academic researches have been conducted concerning brain PET scans using 18F-FDG and learning models[27][10].The accuracy obtained from these models has increased significantly from the time these projects first started. We have reached the point where the performance of these models has exceeded the human levels.

However, the non-linear structure of the learning models makes them non-transparent; this means that we are not sure what information from the input data is being used by the model to make the diagnosis[26].This ambiguity is the reason that the deep learning models are regarded as black-box in nature. While in many applications of these learning models, it does not matter how a decision has been reached, in the medical applications, the impossibility of not being able to validate the decision process is a considerable drawback.

It would be irresponsible and dangerous to trust the diagnosis of a model, by default; this is the reason why a learning model needs to be interpretable. Every decision made by the model must be made accessible for verification by a human expert. Only the use of human interpretable models can guarantee that the decisions made by the model are reliable.

For our thesis, we are going to use brain PET scans to train a deep learning model to diagnose if a person has AD or not. To develop the model, we will be using a pre-trained network on our data and finetune it for our classification problem. Our main goal is not only the classification problem, although it is as much important. Our main contribution is an effort to open the black-box nature of the model that we are going to train.

Along with providing a diagnosis, we want the reasons for the diagnosis to be justified, thus making our results explainable for the general clinicians. We will achieve this through methods that deconstruct the logic used in deep learning networks and present the results to the clinicians in a format that they can comprehend.

2 LITERATURE REVIEW

Deep learning is rapidly becoming a standard technique for solving image analysis tasks related to the medical field. Uses of CNNs in the analysis of medical images can be traced back to the 1990s. Since then, CNNs have been used for breast cancer histology image classification[30],pulmonary nodule detection[18],liver lesion classification[15] and brain tumor classification[22].

CNNs consist of Convolution layers in their architectures and can detect certain local features in all locations of the input images. However, training a CNN from scratch comes with many complications. Some of which are that CNNs require a large amount of labeled training data, which may be challenging to meet in the medical domain as labeled data is limited, and training CNNs requires a lot of computational resources without which the training would be time-consuming.

An additional problem that people face, training on medical images, is that medical images mostly come in three dimensions, and CNNs accept only 2D images. The use of 3D-CNNs[11] solves this problem as they accept three-dimensional images, but computational resources needed for 3D-CNN are even higher than 2D CNN. In order to counteract the limited size of the medical datasets and the high level of computational resources needed for training from scratch, the transferability of knowledge embedded in the CNNs is being harnessed, so that the networks that are trained on other datasets are fined tuned to work with the specific medical tasks[29][9].

Research by Sullivan, A. Maki[9] and the team shows that the successful transfer of knowledge relies on the amount of difference between the dataset on which a CNN is trained and the data to which the knowledge is to be transferred.

Various networks like VGG, Inception, Xception, and Resnet have been trained on real-world images for object detection and classification purposes, and have proven to be very efficient too. However, there is a considerable difference between RGB images and grayscale images. It raises the question of how effective the transfer of knowledge will be from a pre-trained model, as it might have values and artifacts that are not intended for single-channel images.

This issue has been addressed in research by Yiting Xie[29] that used the pre-trained InceptionV3 network on ImageNet data after converting it to grayscale. No significant degradation was seen in the results, suggesting that color is not a critical feature in image classification using the Inception network. In another research by Nima Tajbakhsh[7], it has been shown that the transfer of knowledge from natural images to medical images is possible with the expectation of good results after considerable fine-tuning.

In the research by Ding[27],and the team, pre-trained network, InceptionV3, has been used on Positron Emission Tomography data to determine the performance, the accuracy was measured in terms of specificity and sensitivity. Specificity was 82 percent, and the sensitivity was a full 100percent, and the area under the ROC curve was 0.98, these results were quite admirable, and we are going to be using them as the baseline for our model.

In this research, 2109 images were used from 1002 unique patients taken from ADNI Database. 90 percent of the data was used for training, while 10 percent was used for testing. Model results were made explainable through saliency mapping. However, the interpretability by a human expert still lacked as the saliency map failed to provide significant markers that could help radiologists and neurologists to determine how a decision was made[27].

A pre-trained network, InceptionV3, had been used on the PET data to determine the performance, the accuracy was measured in terms of specificity and sensitivity, specificity was 82 percent and the sensitivity was a full 100 percent, and the area under the ROC curve was 0.98, these results were very good and these are going to be used as the baseline for our model. Though the results were great but a major drawback was that they were not interpretable by a human. Even the saliency map failed to provide significant markers that could help radiologists and neurologists to determine if a person had AD or not.[27]

The inception network was proposed as state of the art for classification and detection of images during the ImageNet Large-scale visual recognition challenge 2014(ILSVRC14). It was then called InceptionV1. The inception architecture moved from fully connected architecture to sparsely connected sophisticated architectures to manage the utilization of computational resources[6]. Later updated versions of this were released where the architecture was refined first by introducing batch normalization(InceptionV2) and then by introducing additional factorization ideas(inceptionV3).

The latest version Inception V4 introduced Residual connections and turned out to be more efficient in terms of top-1* and top-5* error[5]. We are going to be using one or more of the inception networks for dimensionality reduction and classification purposes. The reason for using more than one network is that we will have a better idea of which network produces better results.

The limited size of the dataset is also countered through data augmentation strategies. Data augmentation is the creation of warped versions of the data by applying methods like zooming, translation, shearing, stretching, flipping, and applying filters. This increases the training data as well as produces some variance in it. Although the resulting training images are slightly different, experts have argued that the augmentation causes the network to learn the general features in a better way rather than focusing too much on the individually specific features.

There are many data augmentation techniques, but we cannot use all of them at the same time. There has to be a specific strategy in place. In research by Zeeshan Hussain[31], it has been shown that the augmented training set must retain as many properties of the original medical images as possible, to get the optimal model performance. Flipping and Gaussian filters have proven to be the most suitable augmentation strategies for medical images for the best results[31].

To make the modeling more efficient, an additional method that we are going to use is autoencoders. Denoising Encoders using convolution layers can be used for efficient denoising of medical images. It was shown that denoising autoencoders could be stacked to form a deep network by feeding the output of one denoising autoencoder to the one below it[8]. An autoencoder first takes an input and maps(encoding) it to a hidden representation to ignore noise using deterministic mapping. This representation is then reconstructed(decoding) as close as possible to its original input. Convolutional autoencoders are based on standard autoencoder architecture but with convolutional encoding and decoding layers.

The inherent black box limitation is something humans are vary of especially after the implementation and biased results of Correctional Offender Management Profiling for Alternative Sanctions, which is a risk assessment tool that predicted whether a person would be committing a crime in the future or not. It was found that its predictions were unreliable and biased towards black people[13][19].

Along with that it has been demonstrated that trojaning attacks are possible on neural networks and that Deep Neural Networks can be fooled into misclassifying information without any relation to the right category.[28]

A definite example of the unreliability of neural networks in the medical domain is the Failure of IBM Watson: The AI Doctor[24]. There are numerous examples where AI fails to deliver, while it might be okay to have inaccurate results in many domains, but in the medical field, the unreliability of the models can lead to disastrous results.

As intelligent machines and black-box algorithms are making decisions that were previously entrusted to humans, it has become a necessity for these models to explain their decision-making process. There are three approaches to the explainability of deep neural networks and there are:

1. Explaining the representation of the data inside the network[12]

2. Explaining the processing of the data in the network[12]

3. Creating explanation producing systems[12]

The research by Ding Y[27],used Saliency Map as the explainability model which is a method that explains the processing of data inside the network. It visualized the parts of the image that were deemed to be important to the classification. Our main contribution will be the development of a hybrid explainability model. In another research by Benjamin Letham[4] a generative model was developed to make stroke prediction interpretable by experts. As we studied more, we found that saliency mapping was a tested approach for interpretability in many researches[32][16]

We are going to create a hybrid using the methods in the subfields of "Explaining the processing of the data inside the network" and "Explanation producing systems". The most interesting

The hybrid model will be compared with the singular methods to see which one is more explainable for the radiologists.These explanations will be evaluated by a professional radiologist and provide insights on how well he thinks the explainability is. We are going to try these methods and discern which one gives the best interpretability and explainability for the decision that the model makes.

3 CONTRIBUTION

We are going to be developing explainability models. All the researches that we have analyzed have used at the most one explainability model to try and explain the internal workings of the deep learning models. This approach has been successful to an extent, but something still lacked in these explanations.

In an assessment of these researches, our main contribution will be the development of the combination of two models; this is something that has not been done before. There has also not been a working comparison of the different explainability models in the domain of Radiology. We will implement both a stand-alone and hybrid explainability model on our classification model and record comparison of them.

The hybrid model that we will be making will be based on two kinds of explainability techniques, the processing of the data in the network, and verbal explanation producing system. From the processing of the data in the network, we will be focusing on the methods of Automatic Rule Generation and Saliency Mapping, and for the verbal explanations, we will be testing the methods of Attention network and disentangled representations.

4

EVALUATION METHODS

The quantitative evaluation for our deep learning model will be done by us. The baseline accuracy that will be used will is from the research by Ding Y[27], where an accuracy of 94% was achieved. We will be measuring our model performance in terms of accuracy and loss. After achieving a satisfactory accuracy, we will be developing the explainability models.

While interpretability is a substantial first step, it heavily relies on how well the decision-making process is understood and audited. The performance evaluation by the radiologist will demonstrate which one of our models is mode is more interpretable. The explainability of our model mainly involves explaining the working of the deep learning algorithm and interpretability is the human insight into the working and decision making of the algorithm. A certified radiologist would be called in for the evaluation of our explainability model. The complete evaluation for the model will be in the form of a report.

The report would include the qualitative analysis of the model, which would be a brief summary of questions like:

- How the decision making process is represented to im as an expert?
- Is the representation useful?
- How understandable it is?
- Is the presented information sufficient to understand the working of the classifications model?
- How improved the hybrid model is in comparison with the singular explainability model?
- Is it usable?
- How helpful can the model be in real world scenario?
- Is there any scope for improvement?

After the initial evaluation and feedback by the radiologist, further improvements will be made to the interpretability model and a revaluation would be done if necessary.

5 DESIGN OF EXPERIMENT

5.1 DATA

The data was acquired from the Alzheimer's Disease Neuroimaging Initiative(ADNI). The Alzheimerâs Disease Neuroimaging Initiative (ADNI) is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimerâs disease (AD). Results are shared by ADNI through the USC Laboratory of Neuro Imaging and Data Archive (IDA).

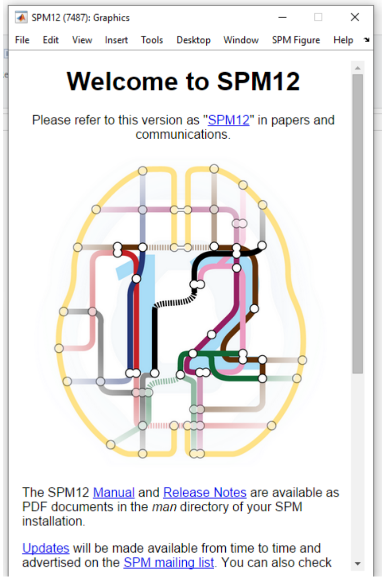
The data consists of about 3000 imaging studies collected from about 400 patients, most patients with multiple scans across ADNI-1, ADNI-GO, ADNI-2, and ADNI-3 which included scan images across different intervals ranging from October 2004 to December 2019. The imaging data consisted of PET scans of patients diagnosed with AD and the scans of patients who were diagnosed as Cognitively Normal.

5.2 PREPROCESSING

SPM software contains several tools mainly used in the preprocessing of PET and MRI Brain scan images. Most scanners produce data in DICOM format, which is the standard format for the communication and management of medical imaging information and related data. The images are converted into NIfTI -1 data format (Neuroimaging Informatics Technology Initiative), which can be done in the toolbox itself. These converted images have to be then Oriented manually to their origin to be Normalized later.2 shows the interface of the SPM Toolbox.

5.2.1 *Image Normalization*

There are two components to spatial normalization: There is the estimation part, whereby a deformation is estimated by deforming template data to match an individual scan; And there is the actual writing



Welcome to SPM12

Please refer to this version as "SPM12" in papers and communications.

The SPM12 Manual and Release Notes are available as PDF documents in the *man* directory of your SPM installation.

Updates will be made available from time to time and advertised on the SPM mailing list. You can also check

Figure 2: SPM Toolbox Interface

of the spatially normalized images, using the previously estimated deformation.

In spatial normalization, the estimated deformations in the image such as Image brightness, distortions, smoothness are normalized so as to ensure the image are evenly normalized. The writing of the spatially normalized images involves warping the images according to the given parameters. Figure 3 shows a normalized image.

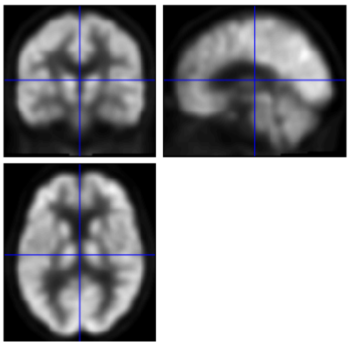


Figure 3: Image after Normalization as shown in SPM

5.2.2 *Image Segmentation*

Image segmentation involves separation of layers of the brain from the brain scan images. This function segments, bias corrects and spatially normalises - all in the same model. Basically, it involves tissue classification which can be useful to take in only the layers useful for us to feed in to the network. Typically, the order of tissues classified are Grey matter(C1)(figure 4), white matter(C2)(figure 5), Cerebro Spinal Fluid(CSF)(C3)(figure 6), bone(C4), soft tissue(C5) and air/background(C6).

5.2.3 *Image Calculation*

Image calculation involves the selection of the segmented layers of the brain scan images to produce a single nifti image. This step gets rid of the skull, hair, ears and any extra tissues that are not needed for classification. We have selected C1, C2 and C3 to produce the output image (figure 7). This image is a combination of the three segmented imaged.

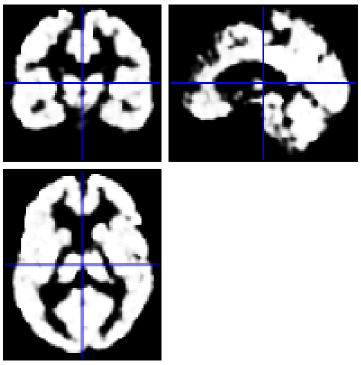


Figure 4: Grey Matter(C1)

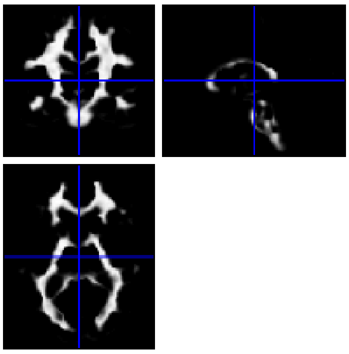


Figure 5: White Matter(c2)

After performing Image Calculation, we get a **2mm Isotropic voxel** output image with dimensions of **79x95x79**.

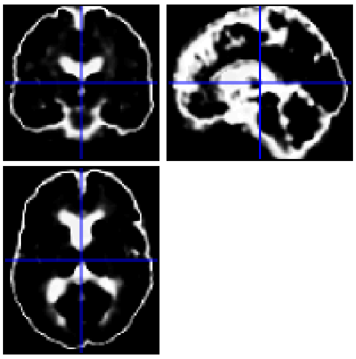


Figure 6: CSF(c3)

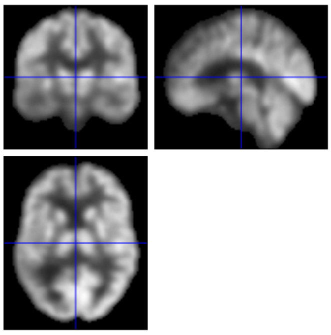


Figure 7: The Output Image as shown in SPM

5.2.4 *Data Representation for CNN*

After the initial pre-processing of the data, we had 3-dimensional - gray scale images in the nifti format. These images are complicated to work with, the two main problems that arose were

1. The gray scale image data is single-channel data. The pre-trained models are developed using RGB image data, which is three-channeled data.

2. The pre-trained models are trained on 2-Dimensional imaging data; that is why they do not accept 3D images.

We solved the first problem by stacking the gray scale image data in three layers on top of each other; this changed the data into three-channeled data. The three-channeled gray scale data is different from the RGB three-channeled data in the sense that all three layers have the same gray scale images instead of the three-channels representing the three colors of the r, g, and b.

Using this method, the same gray scale image is fed to the three channels of Convolutional Neural Network. For the representation of the data to a 2D format, we considered many strategies. We could have extracted single slices from the 3D image data and used it for training, or we could have created a grid of specific images as in the research by DingY, or the images could be converted from nifti to JPG or PNG. Using any one of the alternatives would have resulted in a considerable loss of information.

5.3 EXPLAINABILITY AND INTERPRETABILITY

As we have mentioned above, our Hybrid Explainability model would be our novel contribution. We are looking into using various methods such as Saliency Maps and Feature Maps for Visual Interpretation. Also, we are exploring interpretability via mathematical structure such as using Pre-defined models and feature extraction[25]. We are also looking at using Deep Belief Networks, a generative model which can be used to reconstruct features most used in a Neural Network.

5.4 PRELIMINARY RESULTS

We have some initial preliminary results from 3D-CNN and Inceptionv3.

5.4.1 *Modeling with 3D-CNN*

To get a new perspective on the problem of using 3D images, instead of starting with a pre-trained model off the bat, we first wanted to see how a simple Convolutional Neural Networks performed with the data. We used the 3D-CNN that can accept 3D imaging data. To down-sample the data, 3D-max-pooling was used to divide the data into cuboidal pools. The 3D-CNN was implemented successfully. We used 2-fold stratified validation, dropout of 0.6, learning late of 0.003, and a batch size of 5. We obtained an accuracy of 47% with 15.89 loss.

Tuning the CNN did nothing to improve the accuracy, adding more layers into the network could have improved the performance. However, it was also not an option because the 3D-CNN is computationally very expensive. Our servers could not provide the resources needed to build this classification model. Moreover, the data from 650 patients, was nowhere near enough for training a network from scratch. Because of the limited computational resources, we left 3D-CNN at a 47% accuracy.

Training a network from scratch on medical data is challenging. The accuracy we achieved proved to us that the reasons for using a pre-trained network are quite valid.

5.4.2 *Modeling with Inceptionv3*

After the results we got from 3D-CNN, we came back to the pre-trained network. The testing with the 3D-CNN gave us the idea to use 3D convolutional blocks in the inception architecture instead of the 2D blocks. The network worked and gave the initial accuracy of or 43% with a 12.53 loss. Tuning the learning rate and dropout gave us an accuracy of 49% with a 7.75 loss. A much more in-depth tuning would have to be done to improve accuracy.

5.5 PROJECT MANAGEMENT

As per our initial thesis plan, we have completed most of our tasks according to the time-frame we had designed. But the pre-processing of our data took more time than expected since we had to learn to work on new interfaces and the data itself was very large and complicated. Also, we had to figure out how to work with 3D images when using Convolutional Neural Networks and Inception Architecture. So this took us additional time since we had to do more literature review.

- For the next half time of our thesis we will be working on improving the accuracy of the model we got.
- The development for our explainability model will be started.
- The model will then be evaluated and validated by the radiologist.
- We will also leave some time to make modifications and a re-evaluation.
- We will start the thesis report a month before the final submission.

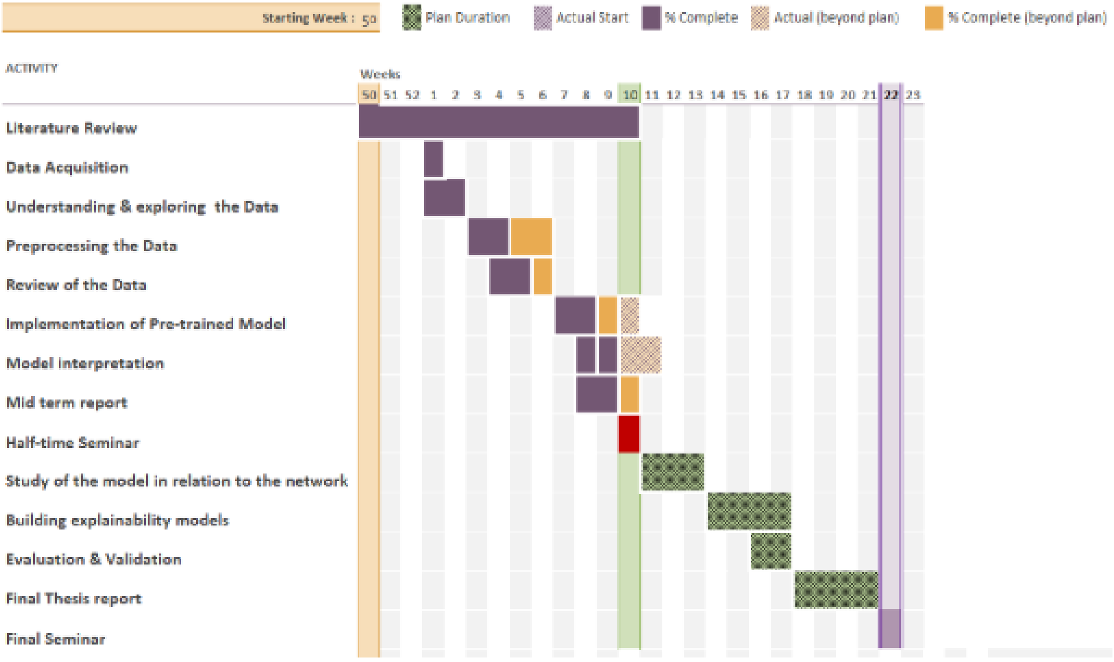


Figure 8: Updated Gantt chart of our Thesis Plan

BIBLIOGRAPHY

[1] Causes of dementia, . URL https://kids.alzheimersresearchuk.org/teens/what-is-dementia/causes-of-dementia/.

[2] What is alzheimerâs disease?, . URL https://www.alz.org/alzheimers-dementia/what-is-alzheimers.

[3] Mild cognitive impairment. URL https://www.alz.org/alzheimers-dementia/what-is-dementia/related_conditions/mild-cognitive-impairment.

[4] Tyler H. McCormick David Madigan Benjamin Letham, Cynthia Rudin. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction mode. *Annals of Applied Statistics 2015*, 5 Nov 2015.

[5] Vincent Vanhoucke Alex Alemi Christian Szegedy, Sergey Ioffe. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv:1409.4842*, 2016.

[6] Yangqing Jia Pierre Sermanet Scott Reed Dragomir Anguelov Dumitru Erhan Vincent Vanhoucke Andrew Rabinovich Christian Szegedy, Wei Liu. Going deeper with convolutions. *Computer Vision and Pattern Recognition*, 17 Sep 2014.

[7] N. Tajbakhsh et al. "convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299-1312*, May 2016.

[8] Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. *arXiv:1608.04667v2*.

[9] Josephine Sullivan Atsuto Maki Stefan Carlsson Hossein Azizpour, Ali Sharif Razavian. From generic to specific deep representations for visual recognition. *CVPR 2015*, 2015.

[10] Saykin AJ Jo T, Nho K. Deep learning in alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data. *Front Aging Neuroscience*, 2019 Aug 20.

[11] keras documentation. Convolutional layers. URL https://keras.io/layers/convolutional/.

[12] Ben Z. Yuan Ayesha Bajwa Michael Specter Lalana Kagal Leilani H. Gilpin, David Bau. Explaining explanations: An overview of interpretability of machine learning. *IEEE*, 2018.

[13] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang.

[14] Alzheimer's Association Logo. What is alzheimerâs disease? URL https://www.alz.org/alzheimers-dementia/what-is-alzheimers.

[15] Eyal Klang Michal Amitai Jacob Goldberger Hayit Greenspan Maayan Frid-Adar, Idit Diamant. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *arXiv:1803.01229*, 3 Mar 2018.

[16] Fernando Navarro Nassir Navab Magdalini Paschali, Sailesh Conjeti. Generalizability vs. robustness: Adversarial examples for medical imaging. *Computer Vision and Pattern Recognition*, 23 Mar 2018.

[17] Subramaniam RM. Marcus C, Mena E. Brain pet in the diagnosis of alzheimer's disease. *Clin Nucl Med*, 2015 Feb 18.

[18] Taco S. Cohen Marysia Winkels. 3d g-cnns for pulmonary nodule detection. *International conference on Medical Imaging with Deep Learning*, 12 Apr 2018.

[19] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv:1412.1897*.

[20] Ronald C. Petersen. Early diagnosis of alzheimer's disease: is mci too late? *Curr Alzheimer Res. 2009 Aug; 6(4): 324â330*, 2009.

[21] M.D. Ronald C. Petersen, Ph.D. Early diagnosis of alzheimerâs disease: Is mci too late? *Current Alzheimer Research*, 2009.

[22] P.M.Ameer S.Deepak. Brain tumor classification using deep cnn features via transfer learning. *Computers in Biology and Medicine*, August 2019.

[23] Liang Mi Richard J Caselli Kewei Chen Dhruman Goradia Eric M. Reiman Shibani Singh, Anant Srivastava and Yalin Wang. Deep learning based classification of fdg-pet data for alzheimers disease categories. *PMC*, 2017 Dec 18.

[24] Eliza Strickland. How ibm watson overpromised and underdelivered on ai health care. URL https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care.

[25] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): towards medical xai. *IEEE*, 2019.

[26] Klaus-Robert MÃŒller Wojciech Samek, Thomas Wiegand. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv:1708.08296v1*, 28 Aug 2017.

[27] Michael Kawczynski Hari Trivedi Roy Harnish Nathaniel Jenkins Dmytro S. Lituiev Timothy P. Copeland M S Aboian Carina Mari Aparici Spencer C. Behr Robert R Flavell Shih-Ying Huang Kelly A. Zalocusky Lorenzo Nardo Youngho Seo Randall A. Hawkins Miguel Hernandez Pampaloni Dexter Hadley Benjamin L. Franc Yiming Ding, Jae Ho Sohn. A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain. *Radiology*, 2018.

[28] Yousra Aafer Wen-Chuan Lee Juan Zhai Weihang Wang Xiangyu Zhang Yingqi Liu, Shiqing Ma. Trojaning attack on neural networks. *NDSS 2018*, 2018.

[29] David Richmond Yiting Xie. Pre-training on grayscale imagenet improves medical image classification. *European conference on computer vision*, 2018.

[30] Hai Zhang Xiao Xiao Yun Jiang, Li Chen. Breast cancer histopathological image classification using convolutional neural networks with small se-resnet module. *Plos One*, March 29, 2019.

[31] PhD Darvin Yi and Daniel Rubin MD MS Zeshan Hussain, Francisco Gimenez. Differential data augmentation techniques for medical imaging classification tasks. *AMIA Annu Symp Proceedings*, 2018 Apr 16.

[32] Charles DeCarli Lee-Way Jin Laurel Beckett Michael J. Keiser Brittany N. Dugger Ziqi Tang, Kangway V. Chuang. Interpretable classification of alzheimerâs disease pathologies with a convolutional neural network pipeline. *nature communications*, 15 May 2019.