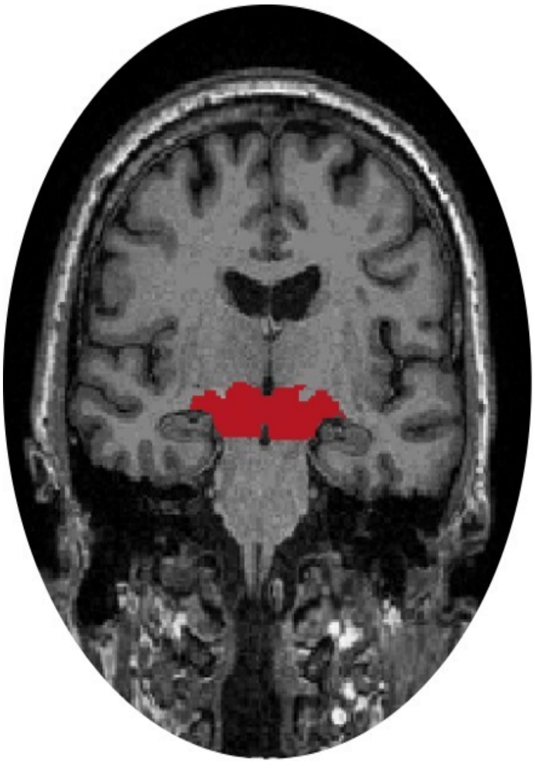
Predicting Parkinson's Disease based on 3D segmented ventral diencephalon using a Dense Convolutional Network



Group 417

Biomedical engineering and informatics
Aalborg University

Title sheet

THE FACULTY OF MEDICINE

AALBORG UNIVERSITY

Faculty of Health Science and Technology
Biomedical Engineering and Informatics
Fredrik Bajers Vej 7
9220 Aalborg Øst
http://www.hst.aau.dk/

Project title:

Predicting Parkinson's Disease based on 3D segmented ventral diencephalon using a Dense Convolutional Network

Project:

Master's Thesis

Project period:

February 2022 - June 2022

Project group:

417

Participants:

Gritt Olivia Asferg Læssøe Svendsen
Jens Tidemann Frøkjær
Trine Skjødt Thomsen

Supervisor:

Maciej Plocharski

Co-supervisor:

Lasse Riis Østergaard

Page numbers: 65

Finished: 1/6 - 2022

Abstract:

Parkinson's disease (PD) is a neurodegenerative disease that affects an increasingly large part of the elderly population. The symptoms are severe and get progressively worse over time, progressing from non-motor to motor symptoms. The symptoms are caused by neurons degenerating in brain structures such as Substantia Nigra. Diagnosis of PD remains a challenge as PD has symptom overlap with other neurodegenerative disorders. Additionally, the diagnosis is late in the progression of PD when the motor symptoms has set in. The aim of the study was to examine if a deep neural network could be trained to differentiate between PD and healthy controls based on ventral diencephalon. Magnetic Resonance Images were acquired from the Parkinson's Progression Markers Initiative database from 215 subjects with 156 diagnosed with PD and 59 healthy. The data was processed and used to train and test a densely connected convolutional neural network. The best trained network had a validation accuracy of 0.68, sensitivity of 0.82, specificity of 0.33 and an area under curve of the receiving operator characteristics of 0.58.

Titelblad

DET SUNDHEDSVIDENSKABELIGE FAKULTET

AALBORG UNIVERSITET

Det Sundhedsvidenskabelige Fakultet
Sundhedsteknologi
Fredrik Bajers Vej 7
9220 Aalborg Øst
http://www.hst.aau.dk/

Projekt titel:

Prædiktering af Parkinsons sygdom baseret på 3D segmenteret ventral diencephalon ved brug af et Dense Convolutional Netværk

Projekt:

Speciale

Projekt periode:

Februar 2022 - Juni 2022

Projekt gruppe:

417

Gruppemedlemmer:

Gritt Olivia Asferg Læssøe Svendsen
Jens Tidemann Frøkjær
Trine Skjødt Thomsen

Vejleder:

Maciej Plocharski

Bi-vejleder:

Lasse Riis Østergaard

Side antal: 65

Aflevering: 1/6 - 2022

Resumé:

Parkinsons sygdom (PS) er en neurodegenerativ lidelse, som rammer en tiltagende stor del af den ældre befolkning. Symptomerne starter som ikkemotoriske og ændres over tid til motoriske symptomer. Symptomerne er forårsaget af neuroner, der degenererer i hjernestrukturer såsom Substantia Nigra. Diagnosticering af PS er fortsat en udfordring, da PS har symptomoverlap med andre neurodegenerative lidelser. Diagnosticeringen sker sent i progressionen af PS, først efter de motoriske symptomer har sat ind. Formålet med studiet var at undersøge, om et dybt neuralt netværk kunne trænes til at skelne mellem PS og raske kontroller baseret på hjerneområdet ventral diencephalon. Magnetisk resonans billeder blev erhvervet af 215 forsøgspersoner 156 diagnosticeret med PS og 59 raske fra Parkinson's Progression Markers Initiative - databasen. Dataene blev behandlet og brugt til at træne og teste et dense convolutional netværk. Det bedst trænede netværk havde en valideringsnøjagtighed på 0,68, sensitivitet på 0,82, specificitet på 0,33 og et areal under kurve for receiver operating characteristic på 0,58.

Reading guidelines

The study is written in five parts: part I problem area, part II method, part III results, discussion and conclusion, part IV problem based learning and part V appendix. The study should be read in chronological order, however, the appendix is referred to elaborate relevant method theory. The theory included in appendix is relevant, but chosen to include in appendix instead of the method part due to it being theoretical information used to explain the concept of CNN for example and not describe the methods used in the study. In the study Harvard citation is used, meaning the sources are shown with a last name and year, [last name, year]. When a source is placed in front of a full stop, this means the source belongs to that specific sentence, whereas if the source is placed behind a full stop the source is used to the preceding paragraph.

Table of contents

Part I

Problem area

Background 1

Parkinsonism is a collection of symptoms that appear in several diseases. Symptoms of parkinsonism include, but are not limited to: rigidity, rest tremor, bradykinesia, axial, dysautonomia and cognitive deficits such as mild cognitive impairment [Armstrong and Okun, 2020; Prange et al., 2019; Goedert, 2015; Cleveland Clinic, 2019; Mayo Clinic, 2020; Dhakal and Bobrin, 2020]. Parkinson's Disease (PD) is the most common reason for a patient to suffer from parkinsonism [Armstrong and Okun, 2020]. PD is the most common movement disorder as well as the second most common neurodegenerative disorder [Goedert, 2015]. When a patient has developed PD the cause is often idiopathic, however, in some cases a patient's genes or geographic location can have an influence [Armstrong and Okun, 2020]. Clinical diagnosis of PD remains a challenge as PD symptoms overlap with other neurodegenerative disorders such as multiple system atrophy (MSA), progressive supranuclear palsy and dementia with Lewy bodies, making it hard to distinguish the earlier stages of PD from these diseases [Tang et al., 2017; Prange et al., 2019; Xiao et al., 2019; Haller et al., 2013].

1.1 Parkinson's Disease symptoms

The symptoms in PD patients can be split into motor and non-motor symptoms [Armstrong and Okun, 2020]. The most distinguishable symptoms of PD are the motor symptoms which include slow movements, stiffness in muscles, as well as a shaken and poor balance. [Armstrong and Okun, 2020; Titova et al., 2017; Xiao et al., 2019] The cause of the motor symptoms are still unknown [Hall and Guyton, 2011]. The non-motor symptoms can include: rapid eye movement sleep behaviour disorder, loss of smell, urinary dysfunction, excessive daytime sleepiness and depression. The non-motor symptoms appear before the motor symptoms, and can be harder to detect by the patient. Often the patient will first enter the healthcare system when the motor symptoms occur. [Armstrong and Okun, 2020; Titova et al., 2017]

1.2 Incidence and prevalence

PD is seen more often in men than women, but only slightly [Parkinson Foreningen, 2022]. The onset is seen between the ages of 60 to 70 and the typical age of people getting diagnosed are 60-62. PD is rarely diagnosed before the age of 50 but in about 5-10% of the cases, an onset before the age of 40 occur. [Parkinson Foreningen, 2022; World Health Organization, 2004; Dansk Selskab For Bevægeforstyrrelser, 2011] The worldwide incidence of PD is estimated to be around 16 to 19 cases per 100,000 people per year with an estimated prevalence of 160 per 100,000 people per year [World Health Organization, 2004]. This gives an estimate of around 12,5 million people worldwide living with PD. The prevalence of PD is increasing with an advancing age, where the occurrence is estimated to be 0.3% for the general population in the industrialised world and 1% for the ages of 65-75 and increasing to 3% for the ages 85 and over [Dansk Selskab For Bevægeforstyrrelser, 2011; World Health Organization, 2004; De Lau and Breteler, 2006]. In Denmark it is estimated that around 12.000 people lives with PD and an estimated incidence of 1.500 per year [Parkinson Foreningen, 2022].

1.3 Pathophysiology

PD is defined by the death of dopaminergic neurons and is therefore called a neurodegenerative disease [Armstrong and Okun, 2020; Titova et al., 2017; Jo and Oh, 2020; Xiao et al., 2019]. Before the motor symptoms sets in, approximately 70% degeneration of the dopaminergic neurons have already happened [Galvan and Wichmann, 2008]. One important brain structure affected by the degeneration of dopaminergic neurons is the Substantia Nigra (SN), sometimes defined as either a part of the basal ganglia or the related nuclei Ventral Diencephalon (VD). The basal ganglia and VD are two brain structures closely related. The basal ganglia and VD are involved in planning and controlling movement patterns and the reward system, which is possible due to a high production of dopamine. [Hall and Guyton, 2011; Vanderah and Gould, 2020; Sonne et al., 2019]

The decrease in dopaminergic neurons, will affect the use of dopamine as a neurotransmitter. Neuromelanin (NM) is a pigment found in the brain, the pigment is stronger in catecholaminergic neurons of SN and locus coeruleus. One of the neurotransmitters used in catecholamine neurons is dopamine. The location of NM in SN creates a darker appearance compared to other structures of the brain, meaning it is visible in Magnetic resonance imaging (MRI). Research suggest that dopaminergic neurons containing a higher amount of NM dies compared to neurons with lower amount of NM, this means that more neurons located in SN dies compared to neurons in the ventral tegmental area in PD patients. [Haining and Achat-Mendes, 2017] Within healthy subjects NM might have an influence of decreasing the possibility of oxidative stress caused by iron ions, by reducing the number of free hydroxyl radicals [Zecca et al., 2001; Haining and Achat-Mendes, 2017]. A hallmark of PD is damage to SN, specifically to the neuromelanin-containing dopaminergic neurons, which will reduce the size of SN seen in MRIs [Sasaki et al., 2006; Prange et al., 2019; Braak et al., 2003].

Besides PD being characterised by the death of dopaminergic neurons it is also characterised by intracellular deposits. These intracellular deposits are called Lewy bodies and consist of misfolded α-synuclein proteins [Goedert, 2015]. In PD patients Lewy bodies are located in SN and locus coeruleus, which are the brain structures with degeneration of the dopaminergic neurons. Furthermore, patients with severe PD have more Lewy bodies compared to patients with less severe PD. Therefore, it is assumed that Lewy bodies are connected to the death of dopaminergic neurons. However, the incidence of Lewy bodies also increase with ageing. [Schulz-Schaeffer, 2010] Inherited PD is caused by mutations in Synuclein Alpha, which is the gene for the α-synuclein protein. More than 95% of those diagnosed with PD have Lewy bodies. [Goedert, 2015] Dependent on the stage of PD certain brain structures will be affected [Armstrong and Okun, 2020; Titova et al., 2017].

1.3.1 Stages of disease

The disease progression of PD can be categorised in stages. There are multiple ways to categorise these stages, including the Hoehn and Yahr scale and the Braak hypothesis [Braak et al., 2003; Hoehn et al., 1967; Ogisu et al., 2013]. The latter is the most widely used [Armstrong and Okun, 2020]. The Braak hypothesis describes six stages of progression of PD. Neuropathology are based on topography of the changes, and the six stages are created to describe progressively worse neuropathology. Each stage include the pathology of the earlier stage with additional pathology. In the first and second stage, the pathology is confined to the medulla oblogenta, see figure 1.1. This can cause symptoms such as loss of smell, depression and rapid eye movement sleep behaviour disorder. In the second stage the location of Lewy bodies will expand to the caudal raphe nuclei. [Armstrong and Okun, 2020; Braak et al., 2003; Ziegler et al., 2013] In the third and fourth stage Lewy bodies are increasingly prominent in the midbrain including the basal ganglia, see figure 1.1. Pathology in these structures are linked with PD motor symptoms as specific motor skills are affected. PD is often diagnosed when patients have reached this stage. In stages five and six the pathology can be found in cerebral cortex as well, and the symptoms include cognitive impairment and hallucinations [Armstrong and Okun, 2020; Braak et al., 2003; Ziegler et al., 2013]. In stage five or six, of PD, almost the entire brain is affected, as seen on figure 1.1. The Hoehn and Yahr scale is not as finely divided as the Braak hypothesis as there only are three stages [Hoehn et al., 1967]. Importantly, the Hoehn and Yahr stages are used in the Unified Parkinson's Disease Rating Scale (UPDRS) which is the most used measurement of how serious the symptoms are in a patient [Dansk Selskab For Bevægeforstyrrelser, 2011]. Therefore, several studies are also using the Hoehn and Yahr and UPDRS when including subjects [Tang et al., 2017; Ogisu et al., 2013; Shinde et al., 2019].

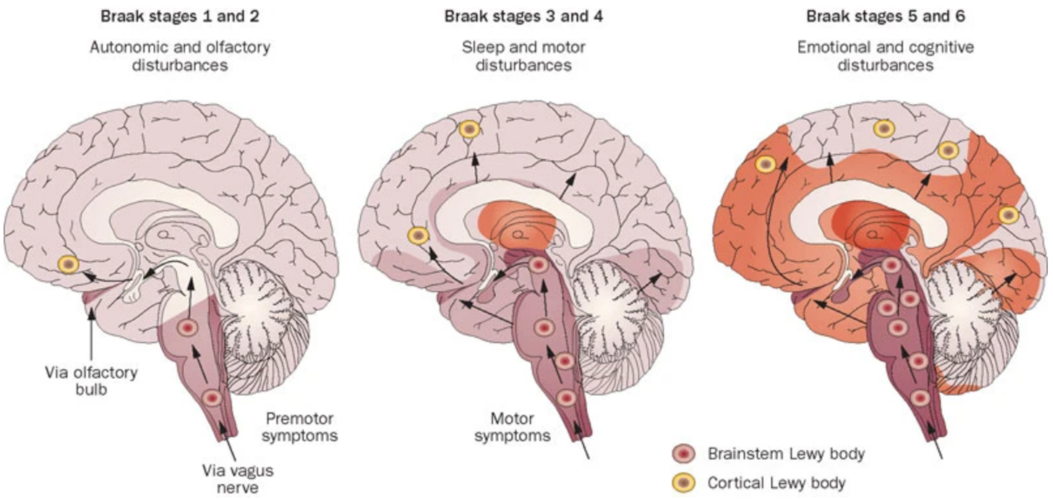


Figure 1.1. Affected brain structure in the six Braak stages. The coloured parts represents brain structures affected by Lewy bodies. [Doty, 2012]

1.4 Diagnosis

Diagnosing PD takes an examination of relevant symptoms, anamnesis and family history with the disease into account since PD can be inherited. A clinical diagnosis requires that the patient has bradykinesia and either rest tremor, rigidity or both for more than 10 years. [Armstrong and Okun, 2020; Xiao et al., 2019] Multiple criteria are used to support the diagnosis, such as improvement with levodopa treatment. Neurological imaging is not part of the criteria for the clinical diagnosis, but can be used to give a estimation of the dopaminergic neurons. Neurological images are also used when there is doubt regarding the diagnosis. [Dansk Selskab For Bevægeforstyrrelser, 2011] The type of brain images used are often Single Photon Emission Computed Tomography (SPECT)/Positron Emission Tomography (PET) and MRI/Computed Tomography (CT). SPECT and PET gives information about functional changes such as the number of dopaminergic neurons in specific structures of the brain. This helps to distinguish between PD and other diseases. The MR/CT images can give information regarding the structural changes of the anatomy. [Dansk Selskab For Bevægeforstyrrelser, 2011] Structural changes will be seen for a PD patient compared to a healthy control (HC), such as a smaller SN for the PD patient [Ogisu et al., 2013]. The structural changes visible on a MRI depends on which underlying condition the patient suffers from, such as PD, MSA or progressive supranuclear palsy which all causes variation in MRIs of the brain, see figure 1.2. [Dansk Selskab For Bevægeforstyrrelser, 2011] Misdiagnosis of PD are frequent as clinico-pathological studies found an incidence of 5-25% [Shin et al., 2021]. Therefore, it is of utmost importance to find a diagnostic tool that can visualise reduction of dopaminergic neurons and structural changes, such as MRIs [Shin et al., 2021; Raff et al., 2006].

Treatment

The treatment of PD focus on treating the symptoms caused by the dysfunctional produced cells rather than the underlying condition. The most common treatment supplements dopamine production. This will in turn only help on the symptoms caused by lacking dopamine, and not the symptoms caused by other neurotransmitter systems. The most used drug for this is levodopa. Patients will often take multiple types of medicine to combat all symptoms. Armstrong and Okun [2020] Furthermore, patients also receive physical treatment [Dansk Selskab For Bevægeforstyrrelser, 2011].

1.5 Previous work

1.5.1 Biomarkers used to examine PD

To improve the diagnosis of PD, which gives the opportunity of earlier treatment, a specific biomarker is needed [Tang et al., 2017]. Biomarkers are measures used to indicate a biological process. The information regarding the biological process can be examined using molecular or radiographic information [Califf, 2018].In SN dopaminergic neurons are clustered together, which is called nigrosomes, which can be examined using MRI. Nigrosome-1 is a focus within research, because it can be used to examine damage of dopaminergic neurons within SN and therefore, be used as a biomarker for PD. [Jo and Oh, 2020; Kim et al., 2021; Shin et al., 2021; Xiao et al., 2019; Prange et al., 2019]

The study by Ogisu et al. [2013] examines the volume of SN, using a region-growing technique, in PD patients and HCs and get the following mean volumes: 215.0 mm^3 and 370.3 mm^3. The patients with PD in this study has an 18% reduction in the volume of SN compared to the HCs.

In a MRI of SN, HCs would have a so called Swallow Tail Sign (STS), representing the concentration of iron-rich and -poor areas in SN. Irregular STS is an indication of PD. This can also be called Dorsal Nigral Hyperintensity [Prange et al., 2019]. The irregularity can be caused by an increasement in the amount of iron [Kim et al., 2021]. The meta-analysis by Mahlknecht et al. [2017] analysed 10 studies focusing on STS, the studies compared PD patients and HCs. A visual assessment of the MRIs lead to a sensitivity of 94.6% and a specificity of 94.4% [Mahlknecht et al., 2017].

PD versus parkinsonism

Studies also seek to discriminate PD from other types of parkinsonism, due to the difficulties differentiating between these [Haller et al., 2013; Ramli et al., 2015]. The study by Haller et al. [2013] examines PD and a group termed *other*, this group includes patients with other types of parkinsonism. The study wants to separate these groups by examining the volume and signal intensity of SN, red nucleus, putamen and dentate nucleus. Haller et al. [2013] discovered a significant difference between PD and *other* in the signal intensity of the left part of SN using the Susceptibility Weighted Imaging (SWI) sequence. However, no significant difference was found in volumes. This indicates that the signal intensity of SN can be applied as a biomarker when differentiating between PD and other types of parkinsonism. [Haller et al., 2013] Nevertheless, Ogisu et al. [2013] found a significant lower SN volume in PD patients compared to HCs.

The degeneration of dopaminergic neurons in SN result in loss of dopamine in the striatum which is a pathological symptom of PD. Besides striatum and SN, changes of the pons and cerebellum will be seen in MSA [Ramli et al., 2015; Prange et al., 2019]. Patterns of iron depositions in the brain can be seen for these neurodegenerative diseases which MRI has the ability to assess. In patients with MSA, higher iron deposition patterns can be observed in the posterior regions of the putamen, see figure 1.2. [Ramli et al., 2015; Haller et al., 2013]

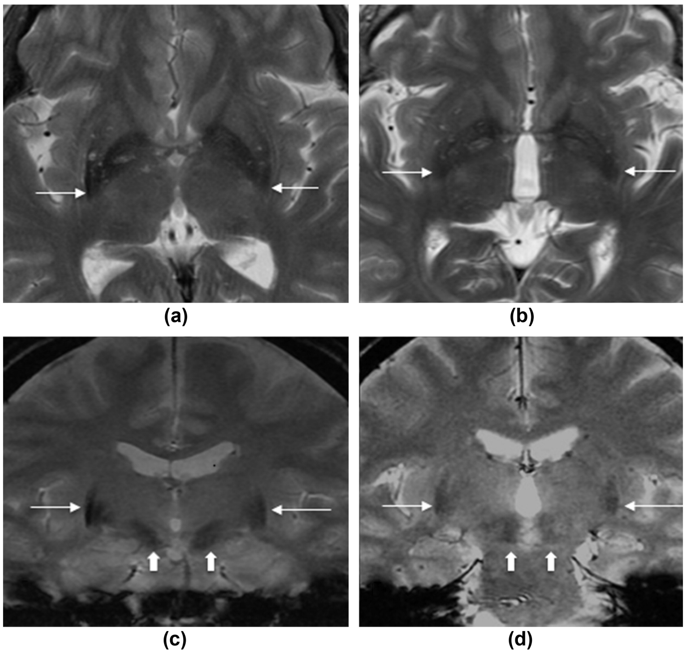


Figure 1.2. Iron deposition for MSA (a and c) and PD (b and d). Thin white arrows point to the iron depositions in the posterior putamen for MSA (a and c) where it can be distinctly seen, whereas, for PD it is not visible (b and d). Block arrows indicate the iron deposition in SN where a similar intensity can be seen for both MSA (c) and PD (d). The MRIs are T2 weighted and gradient-echo showing the brain in coronal and axial views. [Ramli et al., 2015]

Another biomarker used to examine MSA is the hot cross bun sign [Ramli et al., 2015; Prange et al., 2019]. It can be used as a biomarker due to loss of myelinated transverse pontocerebellar neurons in the pontine raphe [Ramli et al., 2015]. The most suitable type of biomarker depends on the types of parkinsonism wanted to differentiate between. Biomarkers as hot cross bun sign or STS can be examined using MRIs. [Prange et al., 2019]

1.5.2 MRI techniques

Both neuromelanin sensitive imaging and iron sensitive imaging can be used to examine PD, due to the decrease in NM in SN and the increase of iron in SN. [Prange et al., 2019]

Several studies, see table 1.1, choose to use Quantitative Susceptibility Mapping (QSM) or SWI as MRI techniques to examine biomarkers for PD. These techniques exploits that the amount of iron in SN in PD patients will increase compared to HCs [Kim et al., 2021; Liu et al., 2015; Wang et al., 2016]. The contrast between grey and white brain matter is caused by the iron in grey matter and myelin in the white matter. [Wang et al., 2016] Another MRI technique used in several studies is SWI see table 1.1, this technique can create a high spatial resolution. SWI is created by a combination of magnitude image and phase image. Due to a higher amount of iron in PD patients, this will lead to a shift in the phase image, this phase shift must be multiplied several times with the magnitude image to create a bigger phase shift contrast.

1.5.3 Segmentation of MRI

When analysing MRI, segmentation is often used as a tool to measure and visualise specific parts of the brains anatomical structures. The manual way of doing this is marking the images based on visual inspection. Manual segmentation is understood to be the most accurate segmentation method as it is difficult to accurately delineate structures in medical images. This method is still intensively used to define *ground truth*. However, it is an intensive and time-consuming task. Furthermore, it is not only tedious, due to the amount of images and resolution a modern MR scanner produces per image, but also especially prone to errors in the form of inter- and intra observer variation. The results of individual assessments are hard to reproduce due to this variation. [Despotović et al., 2015] Manual segmentation is used in studies such as Ziegler et al. [2013], where SN and additionally basal forebrain structures were manually labelled to examine the correlation between PD stage and degeneration of the specific brain structures. Compared to manually segmenting brain structures to examine PD, it can be advantageous to automatically segment these, due to the consistency and time efficiency [Xiao et al., 2014]. Besides manual segmentation there are multiple ways to automate the segmentation process one of these being atlas based segmentation methods. [Despotović et al., 2015]

Atlas based methods segment brain structures based on prior knowledge from a collection of MRIs from HCs. The image from the atlas containing *ground truth* is aligned with the image wanted to segment, and the *ground truth* of the atlas image is transferred to the new image. The advantage of using atlas based methods is a low cost and a possibility to segment all the structures included in the atlas. The disadvantage of the method is if the anatomy of the new image deviates too much from the atlas of *ground truth*. [Despotović et al., 2015]The study by Xiao et al. [2014] uses atlas segmentation to segment red nucleus, SN and subthalamic nucleus. The sizes and positions of the segmented brain structures were compared to the Talairach atlas. Xiao et al. [2014] showed significant differences between the sizes and positions of the segmented brain structures and the *ground truth* from the Talairach atlas.

Segmentation of images prior to using these in machine learning algorithms can lead to higher classification accuracy [Gao et al., 2012].

1.5.4 Classification performed using machine learning

Prange et al. [2019] notes that the use of machine learning may lead to further improvements in diagnosing PD.

Machine learning tools have been used to examine biomarkers, and generate classification based on these for PD. These tools include, among others: Support Vector Machine (SVM) and Convolutional Neural Network (CNN) [Shin et al., 2021; Tang et al., 2017; Xiao et al., 2019; Shinde et al., 2019; Haller et al., 2013].

The study by Xiao et al. [2019] aimed to differentiate between PD patients and HCs using SN as a biomarker. Xiao et al. [2019] examined the use of the QSM MRI technique to quantify the iron increase in vivo and the iron content of SN. Xiao et al. [2019] hypothesised that radiomics features and CNN can be used to evaluate the iron increase for discrimination between PD patients and HC. The radiomics features contained, among other, information about grey scale and texture. To extract the wanted features the QSM data was transformed into: original images, wavelet filtered images and Laplacian of Gaussian filtered images. The image data was evaluated in three ways with the use of three machine learning tools: Logistic Regression (LR), SVM and CNN.

1. LR and SVM tested using radiomics features
2. LR and SVM tested using 'hybrid features', the hybrid features used were a combination of the radiomics features and features used in the CNN
3. CNN with CNN features

These machine learning tools should predict if the subject had PD or was a HC based on the MRIs. The study included 87 PD patients and 53 HC. Xiao et al. [2019] used data augmentation for the CNN to help acquire a more robust model due to a limited amount of data, since CNN requires a certain amount of data. The best performance was found using SVM and hybrid features, however, these results could not have been obtained without creating the CNN and using CNN features in the hybrid features. So as a standalone classifier CNN had the best performance compared to using LR or SVM and radiomics features. The performance of CNN was the following: 0.93, 0.85, 0.86 and 0.83 for AUC, accuracy, sensitivity and specificity, respectively. [Xiao et al., 2019]

Shin et al. [2021] examined if a CNN could be used to detect abnormalities in Nigrosome-1 on SWI. Shin et al. [2021] showed no significant difference in diagnosis performance of the CNN and a visual assessment performed by a neuroradiologist.

1.5.5 Results of different CNN architectures

In segmentation and classification of hippocampus Liu et al. [2020] used a DenseNet, when classifying Alzheimer's disease (AD) an accuracy of 86.6% was obtained. The DenseNet obtained better results compared to LeNet and VGGNet. However, ResNet is also widely used within the field of medical image classification tasks, Shinde et al. [2019] used a ResNet to classify PD and obtained an accuracy of 80%. The study by Xiao et al. [2019] used a dense block in their CNN and obtained an accuracy of 85%. The advantages of DenseNet is that all the layers within a dense block are connected, which will alleviate the vanishing gradient problem. The connection of all layers also lead to a parameter efficiency, each new layer only contain a small amount of filters, so a small feature map from each layer will be added to the existing feature maps. The classification will be based on all feature maps from the network and not only from the last layer. [Huang et al., 2017] A comparison of ResNet and DenseNet was performed by Huang et al. [2017] due to ResNets high performance in image recognition. The DenseNet had similar accuracy using only 1/3 of the number of parameters ResNet used.

1.5.6 Previous work summary

Several studies have examined different biomarkers for PD, these studies were obtained in the block search explained in section 2. Table 1.1 shows an overview of eight studies and the methods used such as: the scanner settings, biomarkers and classifiers. Furthermore, the obtained results from the studies are also included in table 1.1.

Study	**Scanner setting/ techniques**	**Biomarker**	**Classification/ segmentation (s)**	**Results**

Ogisu et al. [2013]	T1 - 3D turbo field echo	SN
		Measure: volume
			Region-growing technique (s)	SE =0.83
				SP= 0.85

Jo and Oh [2020]	T2* - SMWI combined with QSM	Nigrosome-1	QSMnet - speed up the SMWI processing	ROI control =2.31 ROI PD =1.81
Shin et al. [2021]	SMWI	Nigrosome-1	CNN (YOLOv3)	AUC visual=0.9622 AUC CNN=0.9534
Haller et al. [2013]	SWI	SN, red nucleus, dentate nucleus and putamen Measure: Signal intensity and volume	SVM	ACC=0.87

Tang et al. [2017]	Resting-state fMRI	Lingual Gyrus, putamen and Cerebellum Posterior Lobe
		Measure: amplitude of low-frequency fluctuation (ALFF) and the fractional ALFF
			SVM	SE =0.92
				SP = 0.87

Xiao et al. [2014]	T2 - TSE	Subthalamic nucleus, SN and red nucleus Measure: Volume	Majority-voting label-fusion procedure	

Shinde et al. [2019]	Spectral pre-saturation with inversion recovery	SN
		Measure: Class Activation Maps
			CNN (ResNet50)	AUC=0.91
				ACC=0.80
				SE=0.86
				SP=0.70

Xiao et al. [2019]	QSM	SN	LR, SVM and CNN	*CNN* *AUC=0.93* *ACC=0.85* *SE=0.86* *SP=0.83*

Table 1.1. Overview of studies obtained in the block search. The scanner settings, biomarkers and classification tools are shown. Abbreviations: area under the curve (AUC), accuracy (ACC), sensitivity (SE), specificity (SP) and region of interest (ROI).

1.6 Research area

PD is a neurodegenerative disorder and defined by the death of dopaminergic neurons. The degeneration of dopaminergic neurons occur in SN, which is a brain structure involved in modulation of movements. PD patients will exhibit bradykinesia in combination with either one of three additional symptoms: resting tremor, rigidity or gait disturbance [Goedert, 2015]. The type of symptoms the patient experiences depends on the progression of the disease, this progression can be described in stages. In the first stages the symptoms are non-motor, whereas in the later stages the motor function will be affected. The clinical diagnosis of PD requires the patient to demonstrate bradykinesia and either rest tremor, rigidity or both for more than 10 years. Since PD can be inherited the diagnostic process of PD also take the family history into account [Armstrong and Okun, 2020]. Earlier stages of PD can be hard to distinguish since similar symptoms are seen with several other neurodegenerative diseases, making the clinical diagnosis a challenge. Therefore, biomarkers for PD needs to be examined. Being able to diagnose PD is important in relation to management and treatment and to prevent further progress of the disease. According to the review by Prange et al. [2019] examining structural imaging of PD patients, the use of machine learning on MRIs will be promising within future research of PD. The studies by Shin et al. [2021] and Xiao et al. [2019] obtain results above 90% in AUC using CNN. Whereas, the study by Xiao et al. [2019] uses radiomics features in a SVM and obtain an AUC of 89%. This indicates that CNN has a superior performance when examining SN as a biomarker for PD. A DenseNet has high performance and uses fewer parameters than ResNet, furthermore, DenseNet had high performance classifying AD. [Huang et al., 2017; Liu et al., 2020]

The aim of this study was to *examine the classification performance of a DenseNet based on 3D segmented MRIs of VD for the discrimination of PD patients from HCs.*

To the best of the authors knowledge the use of VD in combination with a DenseNet have not been examined in relation to classifying PD.

Part II

Methods

Literature search 2

The literature used in the current study was obtained by first performing an unstructured search and hereafter structured searches. In the beginning of the project period, an unstructured search was performed to gain knowledge within the field of using MRIs as an additional diagnostic tool when diagnosing PD. This search was also used to discover relevant keywords for the following structured literature searches. The first structured search consisted of the main terms *image diagnostic*, *disease* and *pattern recognition*. The block search and keywords used in the search can be seen in table 2.1. Other relevant keywords were also discussed, however these keywords were included in the MeSH terms.

	AND		
OR	**Image diagnostic**	**Disease**	**Pattern recognition**
	"Magnetic resonance imaging" [MeSH Terms]	“Parkinson disease" [MeSH Terms]	"Pattern recognition"
			"Machine learning" [MeSH Terms]

Table 2.1. Block search including MeSH terms performed in PubMed.

The block search was performed in PubMed and 13 articles were acquired. None were excluded due to irrelevant context. Furthermore, a chain search was also conducted based on the block search. Articles, books and other information not included in the block search or chain search were also included in the study. Relevant information and notes were documented in worksheets, based on these worksheets relevant background information were chosen. Additionally to the block search, a search for database specific studies was conducted.

2.1 Database specific search

This search was conducted when it was determined that data used in the current study would be data from the Parkinson’s Progression Markers Initiative (PPMI) database. The database search was conducted to examine what data other studies used and their methods. Therefore, the following search string was used: *"substantia nigra" [MeSH Terms] AND PPMI AND "Magnetic resonance imaging" [MeSH Terms].*
With this search, five articles were acquired, based on these articles the MR settings and methods were determined.

2.2 Literature search flowchart

The information included in the study was based on the *block search* in table 2.1, a database specific search, a *chain search* based on the block search and *other*. The number of included articles based on the four categories can be seen in flowchart 2.1.

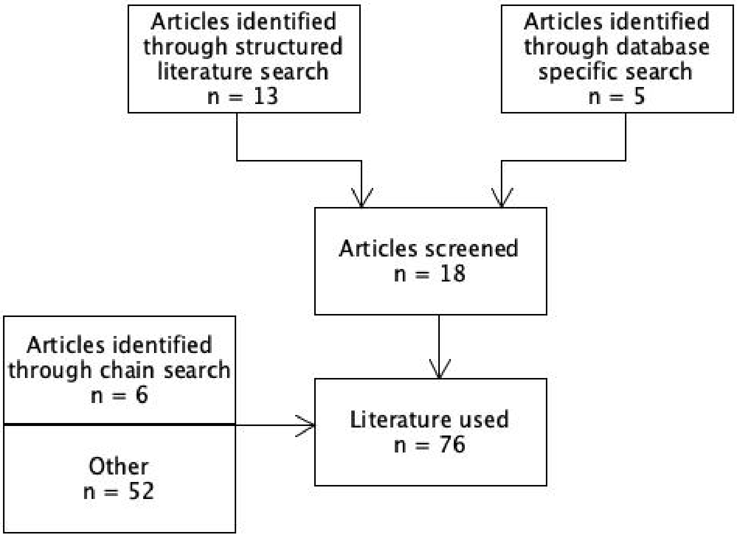

Articles identified through structured literature search n = 13
Articles identified through database specific search n = 5
Articles screened n = 18
Articles identified through chain search n = 6
Other n = 52
Literature used n = 76

Figure 2.1. Flowchart of the literature included in the study.

Pipeline 3

The pipeline consisted of multiple steps: MRI acquisition from PPMI, extraction of ROI using FreeSurfer, data processing, a DenseNet and evaluation of the network through performance metrics. The first step of the pipeline was MRI acquisition from PPMI. The second step was segmentation of the images using FreeSurfer and extraction of ROI. Each brain structure was given a label represented with a colour seen on images next to step two on figure 3.1. Step three was processing of the data, including cropping of data and data augmentation. In this step the data was transformed from NIfTI files, represented by a red ROI in step three on 3.1, to data arrays. The fourth step was training and optimisation of the DenseNet. On figure 3.1 the DenseNet is represented by several layers consisting of nodes and a classification of either class 0, Control, or class 1, Parkinson's Disease. Lastly, step five was validating the performance of the trained DenseNet from step four. This was performed on a validation set reserved for this purpose. Results of the last step was the performance metrics: sensitivity, specificity, accuracy and Area Under the Curve of the Receiver Operating Characteristic (AUC - ROC).

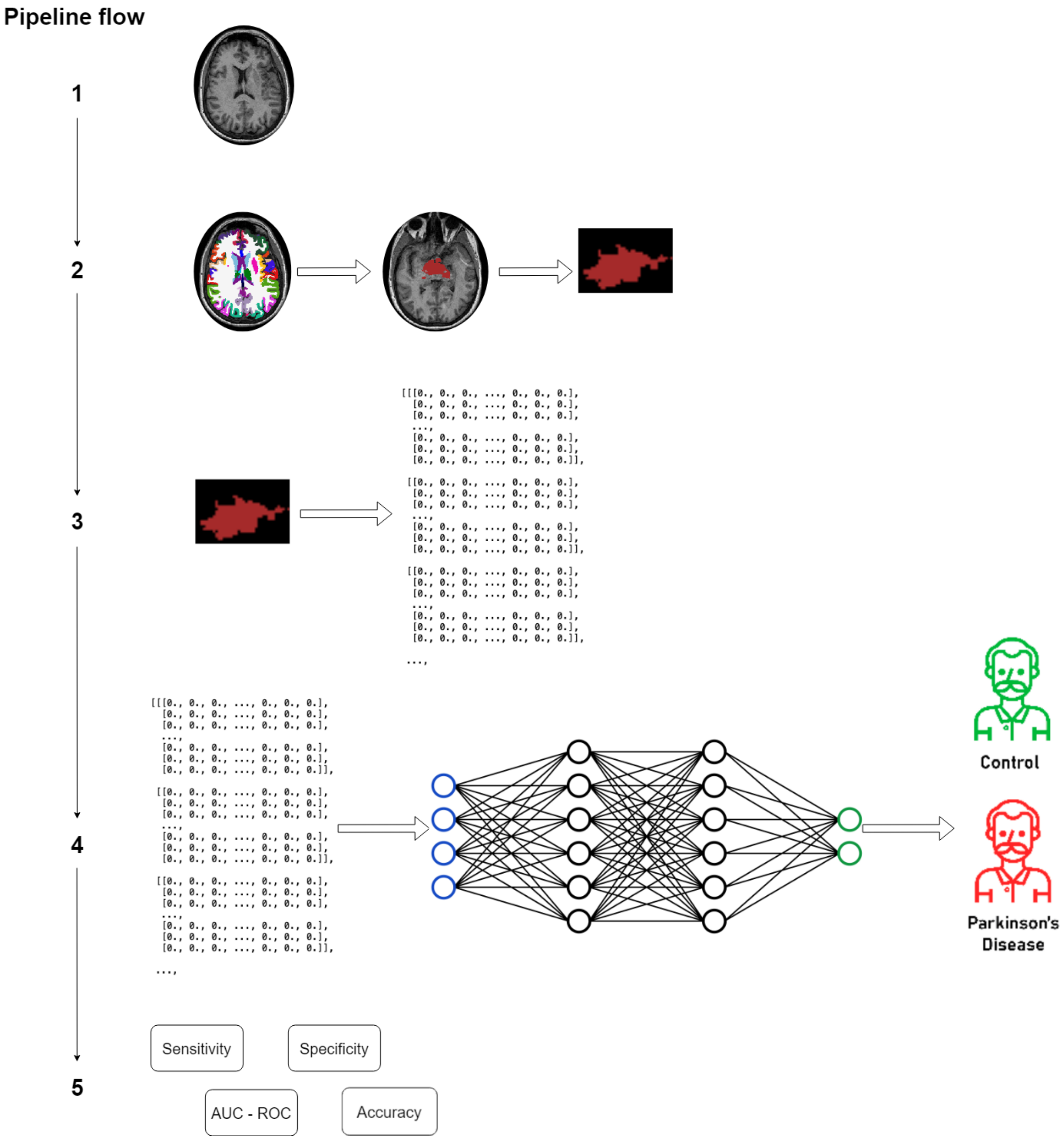


Figure 3.1. Pipeline of the steps included in the classification process. The steps included were: acquisition of data, subcortical segmentation of ROI, data processing, DenseNet and calculation of performance metrics.

Data acquisition, segmentation and data processing 4

4.1 Data collection

The data collected for the study were obtained from the PPMI database [PPMI initiative, 2022], which is a landmark study creating open-access data sets and biosample libraries for PD. Thousands of partners and study volunteers have been engaged to build a basis for Parkinson's research. The mission of PPMI is to profile clinical and biological changes through the spectrum of the disease to identify signals and intervention points of PD as early as possible. [PPMI initiative, 2022] The database contain thousands of images from multiple studies [Marek et al., 2011]. The images in PPMI are: SPECT, CT, fMRI, PET and MRI, where the latter is the neuroimages used in this study. The selected images were based on the specific parameters seen in table 4.1. All the images used in the study were obtained from a single type of scanner from Siemens, Munich, Germany. Additionally all the images were based on the Magnetisation Prepared—Rapid Gradient Echo (MP-RAGE) GeneRalised Autocalibrating Partial Parallel Acquisition (GRAPPA) sequence. This sequence type is used widely within both clinical practice and research when acquiring 3D images. Advantages of the sequence is a strong contrast between different tissue with a high spatial resolution and a very good image quality, due to a high signal-to-noise ratio. [Wang et al., 2014; Blaimer et al., 2004]

Imaging Protocol	Values
Research group	PD and HC
Visit	Baseline
Acquisition plane	Sagittal
Acquisition type	3D
Field strength	3.0 Tesla
Flip angle	9.0 Degree
Slice thickness	1.0 mm
Ecco time (TE)	3.0 ms
Inversion time (TI)	900.0 ms
Repetition time (TR)	2300.0 ms
Weighting	T1

Table 4.1. Overview of parameters used in the images from the PPMI database.

Based on this protocol, see table 4.1, MRIs were acquired for the subject pool.

4.2 Subjects

Subjects included were either HCs og PD patients. The images used were exclusively the baseline measurement from every subject. To segment brain structures, the software FreeSurfer was used for atlas segmentation. FreeSurfer requires 3D data, so the subjects included had such data. After applying the parameters mentioned in table 4.1, 221 MRIs were obtained. Due to data corruption (lack of ability to segment brain structures) a total of 215 segmented images were available. Out of 215 subjects, 80 were female and 135 were male with a mean age of 60.99 ± 9.91. The research groups were divided into different data sets.

4.2.1 Training, test and validation data

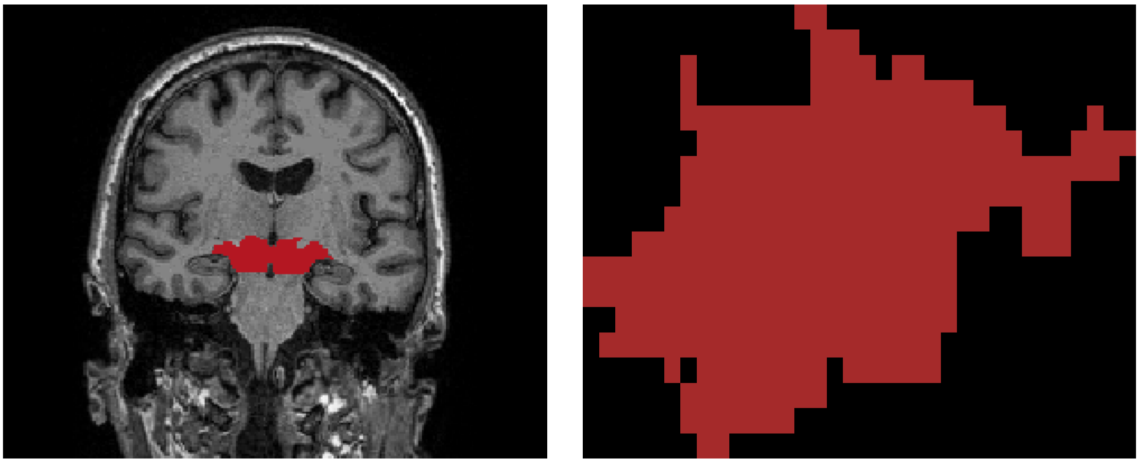
The available data were divided into three parts to train the network, test and validate the performance. The training data was used to find meaningful weights and biases for the network. The test set was used to test how generalisable the network was during training. Evaluation of the network configuration was influenced by the test set as the fine tuning of the hyperparameters depended on the training and test loss. The validation set was used, to provide an impartial generalised evaluation of the final network's fit. The validation set was used on the network with the best performance during hyperparameter optimisation. [Duda et al., 2001; Shah, 2017] 70% of the total data was used for the training set, with the remaining 30% divided evenly between the test set and the validation set. Data were distributed into the respective research groups as 59 for HC and 156 for PD, see table 4.2 for distribution of subjects.

	Training		**Test**		**Validation**	
	PD	*HC*	*PD*	*HC*	*PD*	*HC*
Subjects	110	41	23	9	22	9
Gender (M/F)	69/41	30/11	15/8	3/6	14/8	4/5
Age (mean ± std)	60.61 ± 9.25	59.90 ± 10.88	61.91 ± 9.39	53.11 ± 14.01	66.64 ± 6.73	62.22 ± 11.43

Table 4.2. Demographic overview of the two groups, PD and HC divided into training, test and validation sets.

4.3 Segmentation of region of interest

VD was segmented using FreeSurfer, the segmentation included the hypothalamus, mammillary body, subthalamic nuclei, substantia nigra and red nucleus [Abos et al., 2019]. The segmented VD can be seen in figure 4.1.



(a) Full brain image with VD

(b) Extracted VD

Figure 4.1. VD portrayed on a full brain image (a) and the extracted VD used in data processing (b).

4.3.1 FreeSurfer

In the segmentation process each voxel obtained a label indicating which brain structure the voxel belonged to. [Schmansky, 2022]

In FreeSurfers recon-all stream the input was 3D images (image one, figure 4.2). The skull was removed (image two, figure 4.2). Subcortical segmentation was performed, each subcortical structure was given a label (represented with different colours in image three on figure 4.2). The process of subcortical segmentation is further explained in the next section. Afterwards, the boundary between the white and grey brain matter was found (image four, figure 4.2). The boundary of the pial surface was found from the pial-cerebrospinal fluid intensity gradient (outer yellow mark, image five, figure 4.2). In the next two steps the right and left brain hemispheres were inflated, in the first step to even out the differences caused by the gyri (image six, figure 4.2) and in the second step it was inflated into a sphere. The sphere was normalised and afterwards the sphere was re-shaped back to having the shape of a pial surface. The cortical part of the brain was labelled, and afterwards statistical information like *"total grey matter volume"* was calculated. The extracted information from the recon-all stream were labels 28 and 60 representing the left and right VD (image seven, 4.2). [Fogarty, 2017; Jahn, 2019]

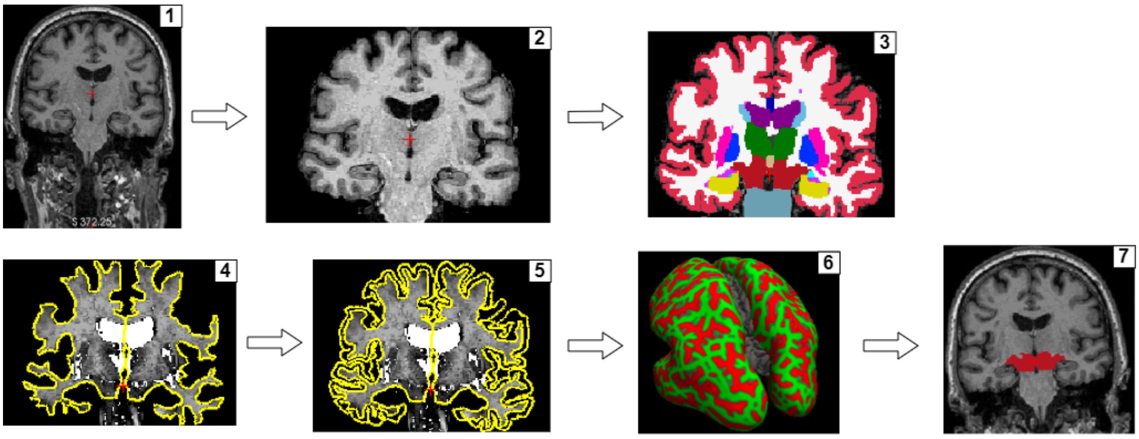


Figure 4.2. Seven steps in the recon-all stream and extraction of VD. The step of each image is marked in the right corner. The steps were as followed: 1. 3D input, 2. removal of skull, 3. subcortical segmentation, 4. determination of boundaries of white and grey brain matter, 5. determination of boundary of the pial surface, 6. brain inflation, 7. extraction of VD.

Subcortical segmentation

The subcortical segmentation step was included in the recon-all stream [Fogarty, 2017] and was used to extract labels indicating which brain structure a voxel belonged to. In the segmentation process a Gaussian Classifier Atlas (GCA) was used to estimate the probability of a voxel belonging to a specific brain structure.

The segmentation of the subcortical structures can be divided into six steps:

1. Transform file to coordinate with GCA
2. Normalise based on the GCA
3. A non-linear transformation was performed to adjust according to the GCA
4. The neck area of the brain was removed
5. Based on the GCA file including the skull a Talairach file also including the skull was created
6. Label each voxel [Fogarty, 2017]

After the segmentation and extraction of the VD, the data had to be processed before being used as an input to the network.

4.4 Data processing

The data processing involved converting the NIfTI files containing VD to arrays with the size 256x256x256. The non-zero values in the array were representing the area of VD, while all the rest of the brain structures were zeroed out. This resulted in an array for each image with mostly zero-values, therefore, the arrays were cropped to predominantly only contain the non-zero values. The cropping borders were decided based on the maximum and minimum placement of the VD in all dimensions, in all arrays. The cropping of the single array was based on extreme values of the whole data set, and therefore all arrays were cropped equally. After cropping, the size for all arrays were 76x73x50.

Afterwards, data augmentation was used to increase the amount of data the network could train on (see appendix A.5). An increase in training data should combat overfitting. Each array in the training set was augmented and added to the training set. This made the training set twice as large as the original set if one augmentation were used, and three times as large if two augmentations were used. One augmentation used was a two part rotation - one in the XY plane and one in the YZ plane. Both planes were rotated with the same degrees. The degree of rotation was chosen randomly between the values: -90°, -45°, -20°, 20°, 45°, 90°and 180°. Additionally, a two part shifting were included when using the two augmentations, the shifting was performed in the XY plane and the YZ plane. Both of the planes were shifted with the same degrees: -20°, -10°, -5°, 5°, 10°and 20°. These degrees of shifting was randomly chosen.

After processing the data was used as an input for a CNN (see appendix A.1) with a DenseNet architecture.

Network 5

5.1 DenseNet architecture

The DenseNet consisted of convolutional layers, pooling layers and activation functions. Additionally, the DenseNet had two dense blocks and a transition block. Compared to the feed-forward connections used in a CNN, the DenseNet uses additional input from the previous layers and passes it through to the following layers. The layers connected must have the same feature map size.

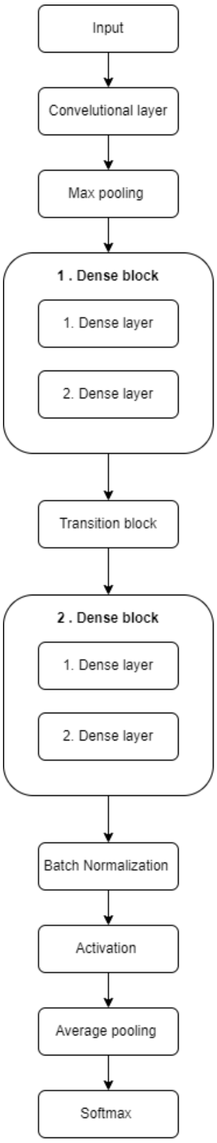


Figure 5.1. Architecture of the DenseNet. The network consisted of two dense blocks with two dense layers each. Between the dense blocks a transition block was placed.

The input size of the array was 76x73x50. Architecture of the network was a DenseNet, starting with a convolutional layer, an activation layer and a max pooling layer which halves the size in all dimensions continuing to a dense block. Based on the optimisation process (see section 5.3) the dense block consisted of two dense layers, shown in figure 5.1. Included in the dense layer was batch normalisation, activation, convolutional, dropout and concatenating layers. Following was a transition block consisting of batch normalisation, activation, convolutional and average pooling layers. Afterwards followed by the second dense block. The network consisted of two dense blocks containing two dense layers each. Following the last dense layer in the network was batch normalisation, activation as well as average pooling layers. A softmax activation function was used for the final classification. It gave the possibilities for each input to be contained in either the PD or HC class. The addition of the possibilities for the PD class and the HC class would have a sum of one. [Basta, 2020]

Hyperparameters were chosen for the network based on the initialised values in the code by Anwaar [2019] and parameters found in relevant literature (see appendix A.6).

5.1.1 Convolutional layer

A convolutional layer was implemented as the first step in the DenseNet. In a convolutional layer a kernel is used to find specific features within the image like edges or colours. The size of the feature maps can be reduced when applying a convolutional layer, depending on kernel, stride and padding, but this may not always be the case. [Saha, 2018] In DenseNet the number of filters were initially 256, while kernel size was set to 3x3x3 for depth, height and width. The padding was set to 'same' and stride of 1, so the output of the convolutional layer had the same size as the input. Following was a Rectified Linear Unit (ReLU) which is an activation function. ReLU does not allow negative values so if a value is $<$ 0 it will be returned as 0. For values $>$ 0 the function is linear, however, since negative values are returned as 0 the function is non-linear. [Brownlee, 2020]

5.1.2 Max pooling

Pooling was used to reduce the size of the feature maps. This was implemented to reduce the computational power required, while keeping the important information of the feature maps. In max pooling the maximum value of the portion of the image the kernel covers is found. [Saha, 2018]The max pooling implemented had a pool size of 2x2x2 and a stride of 2, while keeping padding 'same'. This reduced the feature maps to half the size.

5.1.3 Dense block

The dense block consisted of dense layers, which included a batch normalization layer, a ReLU activation layer, a convolutional layer and a concatenation layer. Within the dense block concatenation was used, meaning that each layer was connected to subsequent layers. This meant that dense layers within the dense block got feature maps from the previous dense layers. [Huang et al., 2017; Anwaar, 2019] This can be seen on figure 5.2, where the concatenation layer (red block) in the third dense layer receives feature maps from concatenation layers, from dense layers one and two. Furthermore, it receives feature maps from the preceding dropout layer within the same dense layer (blue block).

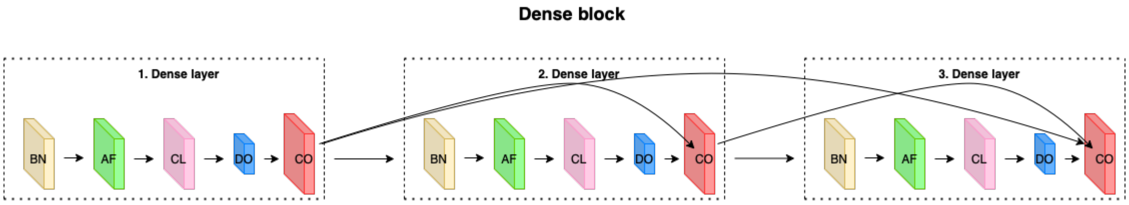


Figure 5.2. A dense block with three dense layers. Each layer contained batch normalisation (BN), a ReLU activation function (AF), convolutional layer (CL), dropout (DO) and concatenation (CO). The dropout reduced the size of the feature maps represented with a smaller block, whereas the concatenation added the feature maps from the BN layer and DO.

Batch normalisation is used to help prevent vanishing or exploding gradient as well as to improve the networks generalisation. In batch normalisation each feature map is normalised to have a mean of zero and a variance of one, for more information on batch normalisation (see appendix A.2).

Dropout was implemented in the dense block to reduce the risk of overfitting. [Ioffe and Szegedy, 2015] The dropout value was initialised to 0.2 but was tested in the optimisation process, see section 5.3.

5.1.4 Transition block

The transition block was placed between the dense blocks, and was used to down sample the feature maps. The transition block in the DenseNet consisted of batch normalisation, a convolutional layer and an average pooling layer. [Huang et al., 2017; Anwaar, 2019] Like max pooling, average pooling was used to reduce the size of the feature maps. In average pooling an average value for the kernel area was obtained and down sampled to this value. [Saha, 2018] The kernel size of the convolutional layer was set to 1x1x1. In the average pooling layer the pool size was 2x2x2 while the stride was set to 2, this reduced the size of dimensions of the feature maps to half.

After the creation of the DenseNet, the network was trained.

5.2 Training procedure

The DenseNet was trained to find the most optimal weights and biases as these can cause a change in the classification (see appendix A.1). The optimiser was Adam (see appendix A.4), whereas the loss function was binary cross entropy (see appendix A.3). The training procedure and update of weights and bias are shown in figure 5.3.

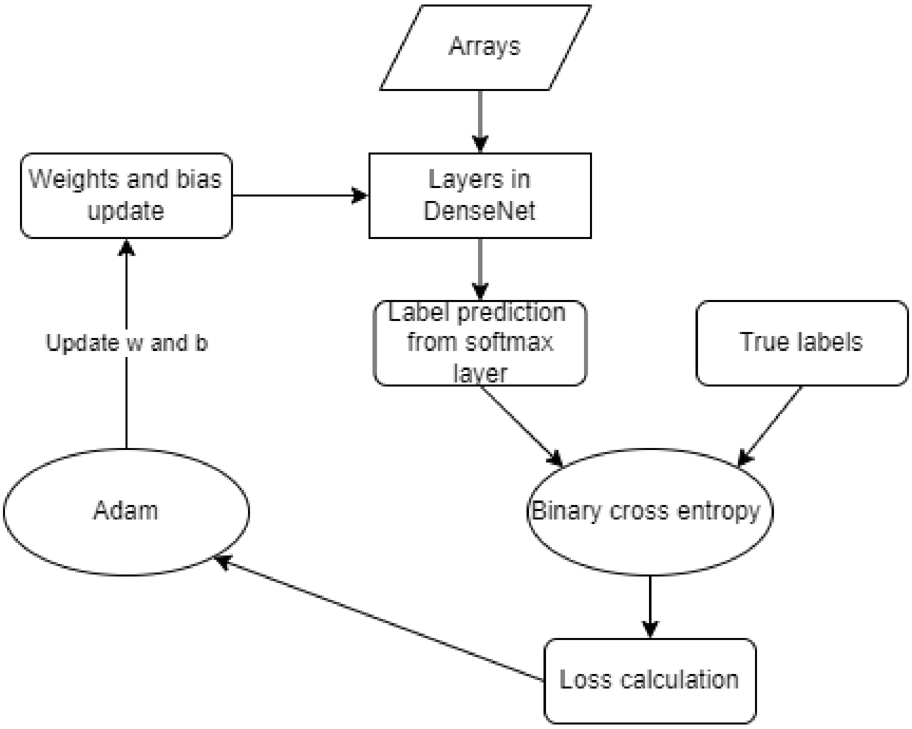


Figure 5.3. Weights and bias updated during training based on the Adam optimiser, which optimises based on the calculated loss.

Arrays containing the ROI was used as the input. The weights and bias was initialised, afterwards a batch of initially eight arrays were input to the DenseNet before the weights and bias were updated. Figure 5.3 show the arrays were used as input to the DenseNet. A comparison of the labels predicted by the network and the true labels was made and based on the comparison the binary cross entropy function calculated a loss, see figure 5.3. The calculated loss was input to Adam. Adam uses a moving average with bias correction to calculate the first and second moment of the batch, these moments are used to update weights and bias [Kingma and Ba, 2014]. After an update of weights and bias, eight new arrays was added into the DenseNet and the loop would start over. When no more arrays were available to create new batches from, a new epoch began.
Under the training procedure, optimisation of hyperparameters was performed to find the best combination of hyperparameters leading to the DenseNet having the most optimal weights and bias.

5.3 Optimisation process

An approach inspired by random search and grid search was used to optimise the chosen hyperparameters in the network, furthermore, different combinations of data augmentation were also tested. One parameter was tested at a time, and the best setting was chosen when testing the remaining parameters. When choosing the best parameter the representative networks learning curve was evaluated as a metric for performance (see appendix A.6). If the learning curves for the different parameters tested were to similar to differentiate the performance, the best test accuracy were used to decide the best parameter. The process was based on the assumption that each hyperparameters contribution to the performance improvement of the network was independent to each other. The optimisation process was performed in the order of the following sections, and can be seen on figure 5.4. The final network chosen for predictions was the best network in the full set of epochs.

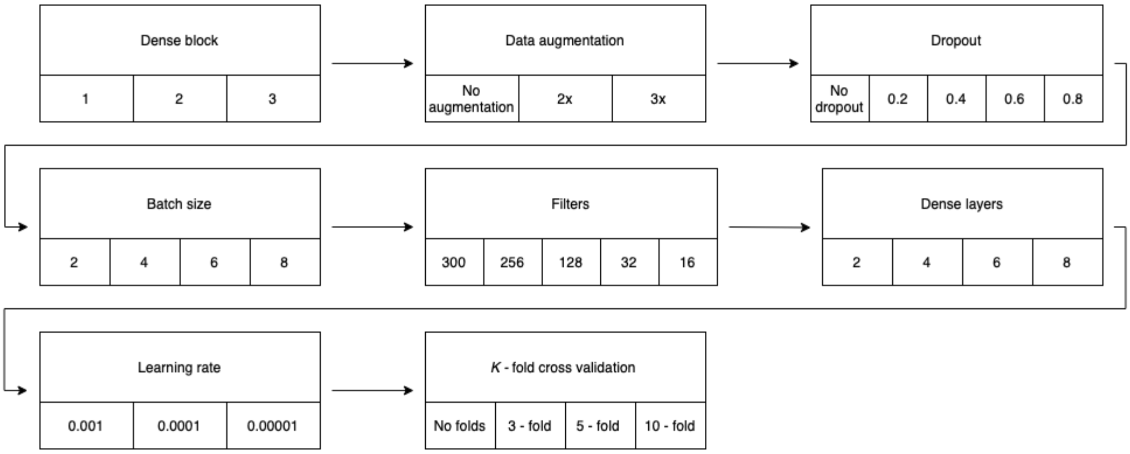


Figure 5.4. Pipeline illustrating the order of optimisation tests.

5.3.1 Computational memory

The first parameters examined was batch size and number of dense blocks. Batch size was initialised to 64, while number of dense blocks was four. However, an error occurred when making predictions on the test set due to lack of computational memory. Unstructured testing revealed the limits of the network size given the available hardware. The network could have a batch size of eight and a maximum of three dense blocks without lack of computational memory.

In some instances the computational memory limit caused training to stop prematurely. Therefore, not all plots of loss functions have the same number of epochs.

5.3.2 Dense blocks

The number of dense blocks tested were one, two and three.

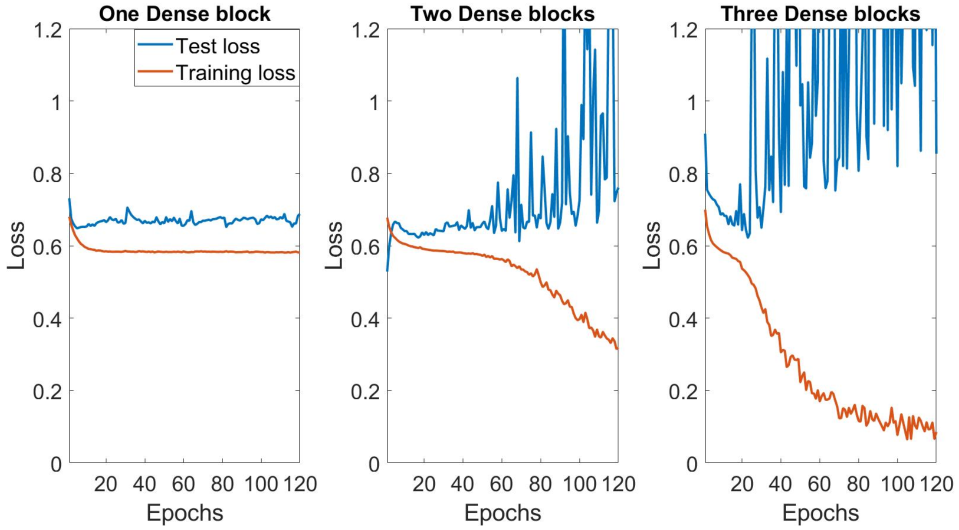


Figure 5.5. Loss functions of the network with one, two and three dense blocks. Epochs were initialised to 120, shown on the x-axis. Test loss is the blue graph while the training loss is the red graph.

On figure 5.5 three loss functions of the network trained and tested with one, two and three dense blocks with epochs initialised to 120 are shown. The network was overfitted in all three cases as the loss was low on training data while increasing on the test data, meaning the network lack ability to generalise. It was chosen to continue using two dense blocks based on the loss functions on figure 5.5, due to the training loss still decreasing after 120 epochs.
While testing dense blocks the loss function began converging at 300 epochs. Therefore, future tests were performed with 500 epochs.

5.3.3 Data augmentation

The data augmentation was included to test if this would reduce overfitting. The tested variation of data augmentation was: without data augmentation, rotation and rotation+shift. The amount of training data without data augmentation was 151 arrays, whereas with rotation it was 302 and with rotation+shift it was 453.

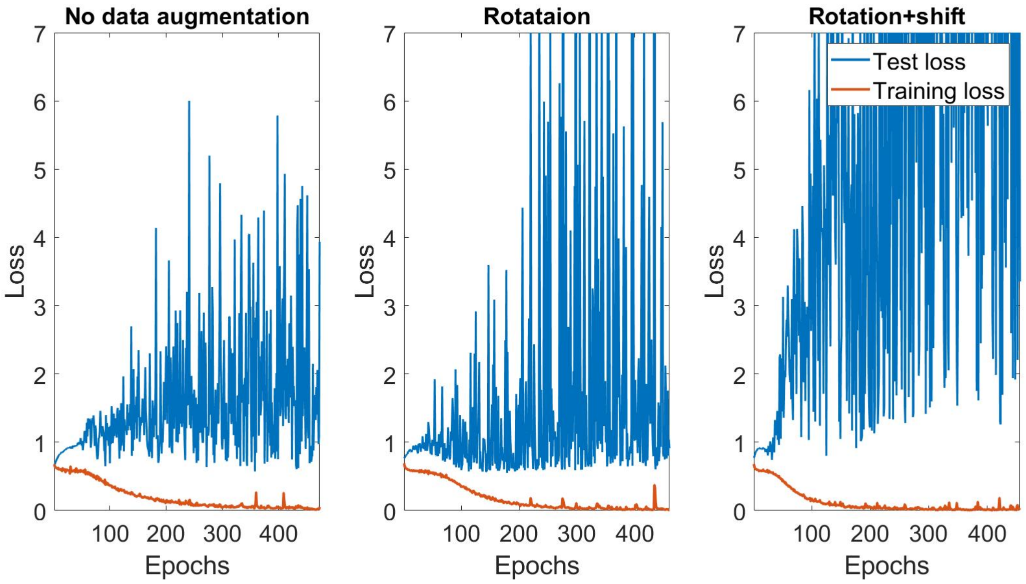


Figure 5.6. Loss functions representing the network trained using data augmentation: without data augmentation, rotation and rotation+shift. Number of epochs was 500 shown on the x-axis. Test loss is the blue graph while the training loss is the red graph.

The loss functions from the tested data augmentation can be seen in figure 5.6. Based on these loss functions the data augmentation chosen was rotation, as the test and training loss had the smallest gap.

5.3.4 Dropout

Dropout was also included in the network to prevent overfitting. The initial value of dropout was 0.2, in addition 0.4, 0.6 and 0.8 were tested. An extra test of the network without dropout was performed to examine if it was necessary to include dropout.

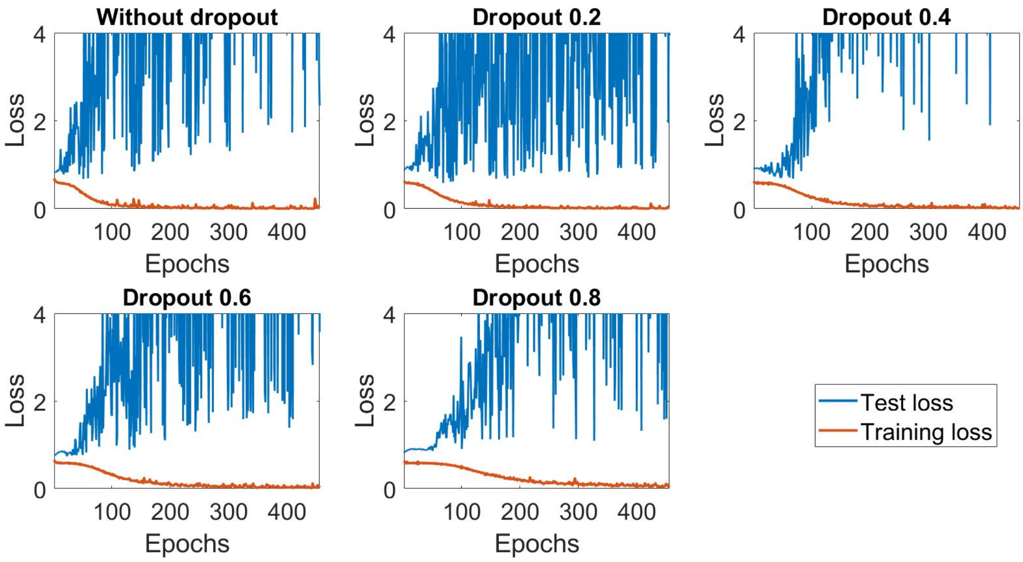


Figure 5.7. Loss functions representing the network trained using dropout values of: no dropout, 0.2, 0.4, 0.6 and 0.8. Number of epochs was 500 shown on the x-axis. Test loss is the blue graph while training loss is the red graph.

Figure 5.7 show five loss functions with five different dropout values. Based on these loss function the network was still overfitted, however, it was decided to use a dropout value of 0.2 as the training and test curve in this function had the smallest gap.

5.3.5 Batch size

Four different batch sizes was tested to examine which batch size was most optimal. The following batch sizes were tested: two, four, six and eight.

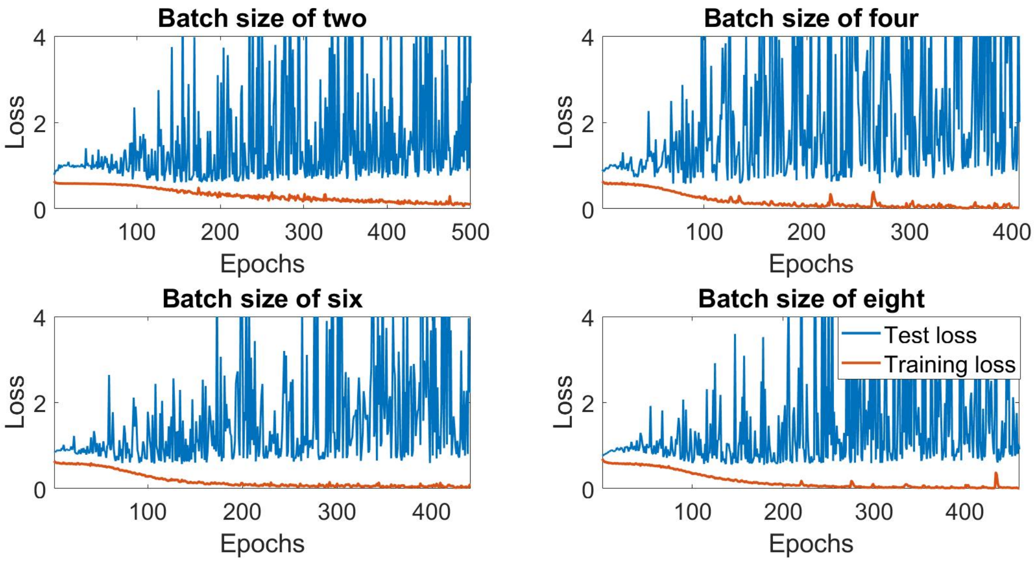


Figure 5.8. Loss functions representing the network trained using batch sizes of two, four, six and eight. Number of epochs was 500 shown on the x-axis. Test loss is the blue graph while training loss is the red graph.

The loss functions seen in figure 5.8 show the loss function using the four different batch sizes. The loss functions appear to be almost completely similar, therefore it was chosen to continue with a batch size of eight, because lager batch sizes often converge faster.

5.3.6 Filters

The number of filters tested were 300, 256, 128, 64, 32 and 16. These filters were placed in the first convolutional layer. In figure 5.9 the loss of the network trained and validated using 16 and 32 filters are shown. Number of filters per layer from 64 and above were not compared due to computational memory.

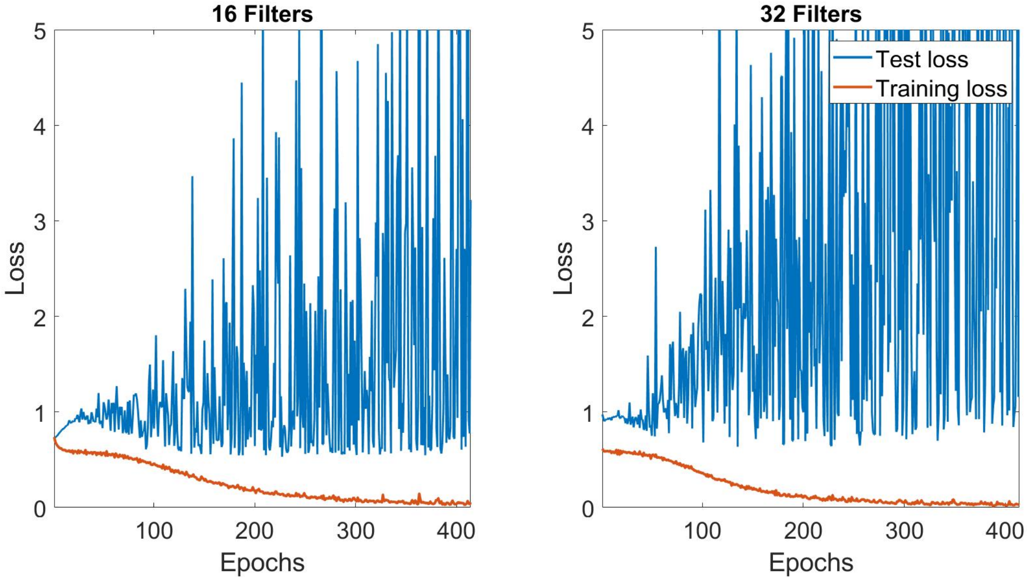


Figure 5.9. Loss functions representing the network trained and validated using 16 and 32 filters in the first convolutional layer. Number of epochs was 500 shown on the x-axis. Test loss is the blue graph while the training loss is the red graph.

Based on the loss functions in figure 5.9 it was determined to use 16 filters. In the loss function for the network trained and tested using 16 filters the test loss converge a bit more to the training loss compared to the network using 32 filters.

5.3.7 Dense layers

The number of dense layers tested were: two, four, six and eight. On figure 5.10 the loss of the network when trained and tested using two, four, six and eight dense layers, respectively, are shown.

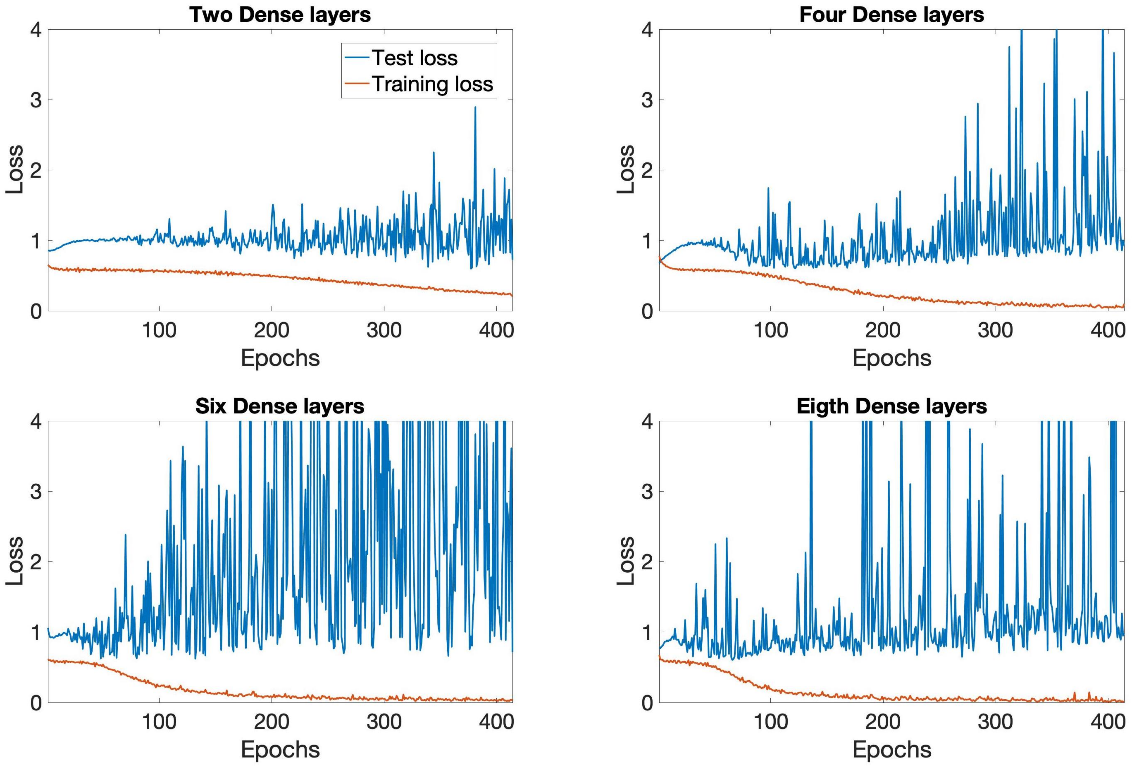


Figure 5.10. Loss functions representing the network trained and tested with two, four, six and eight dense layers. Number of epochs was 500 shown on the x-axis. Test loss is the blue graph while the training loss is the red graph.

Based on the loss functions on figure 5.10 it was decided to use two dense layers in the dense blocks. When observing the loss functions for the network trained and tested with the different dense layers, the test loss for only two dense layers converge a bit more to the training loss when compared to the loss functions for four, six and eight layers.

5.3.8 Learning rate

The learning rate was tested to examine if the network would obtain a higher performance or get a more stable test loss. The test loss used in the hyperparameter optimisation tests seen above are oscillating at a high frequency. Therefore, different learning rates of 0.00001, 0.0001 and 0.001 were tested to examine if these would have an impact on test loss.

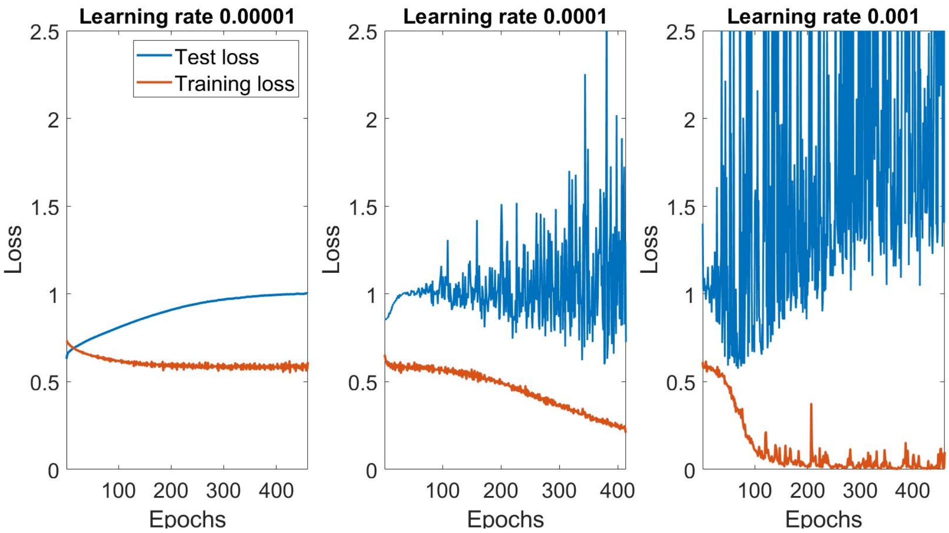


Figure 5.11. Loss functions representing the network trained and validated with learning rates of: 0.001, 0.0001 and 0.00001. Number of epochs was 500 shown on the x-axis. Test loss is the blue graph while training loss is the red graph.

Based on the loss functions on figure 5.11 it was clear, that the learning rate does impact the stability of both the test loss and training loss. A smaller learning rate will not cause big deviations in the values for the weights and bias, therefore, the performance of the network will not change as much as for a higher learning rate. However, based on the loss functions in figure 5.11, in the network with a learning rate of 0.001 the test loss is unstable and it increases over epochs. With a learning rate of 0.00001 the test loss is stable, nevertheless, this test loss also increases a bit over epochs. In the network with a learning rate of 0.0001 the training loss decreases over epochs as expected, and the test loss is unstable, but the test loss sometimes decreases to under 1, which the learning rate of 0.00001 does not. Therefore, a learning rate of 0.0001 was used in future tests.

5.3.9 K-fold cross validation

K-fold cross validation was implemented to test if it would reduce overfitting. When using k-fold cross validation the data set was resampled and split into a number of folds based on the specified number of folds. k-fold cross validation were tested with three, five and 10 folds on the training and test set. On figure 5.12 the loss of the network when trained and tested using no, three, five and 10 folds, respectively, are shown.

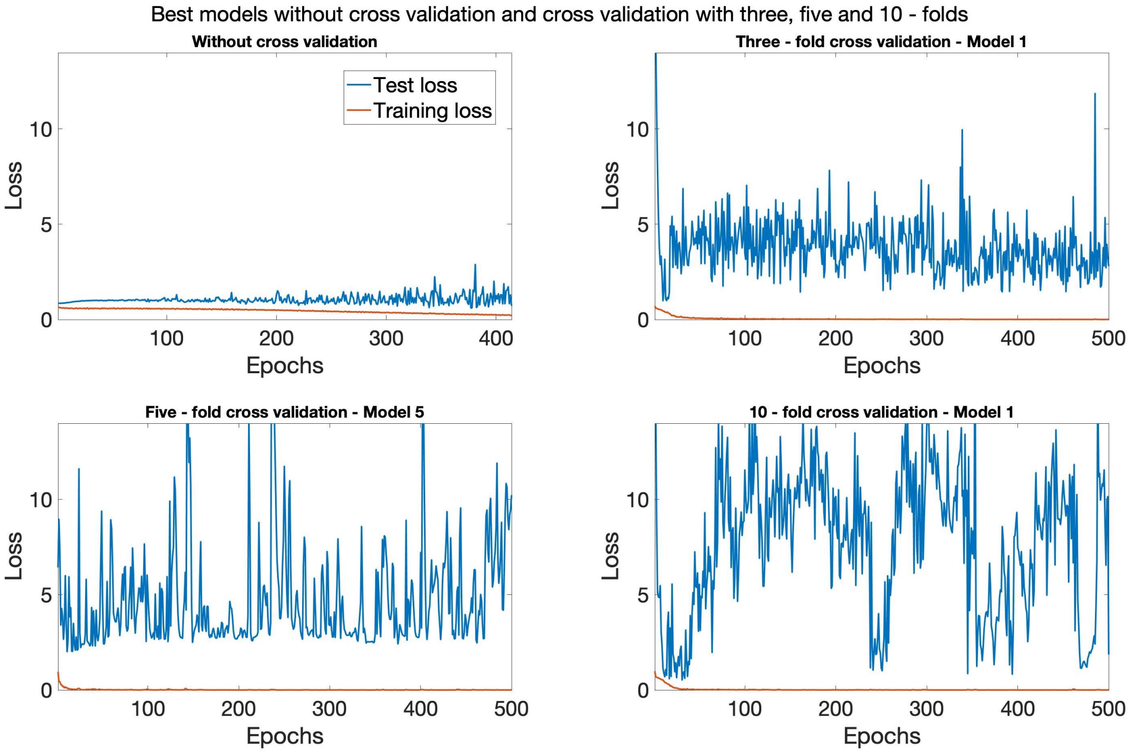


Figure 5.12. Loss functions representing the best network trained and tested when applying no, three, five and 10 folds, for the training of the network. Number of epochs was 500 shown on the x-axis. Test loss is the blue graph while training loss is the red graph.

Based on the loss functions on figure 5.12 it was decided to use no cross validation. When observing the loss functions for the network trained and tested with no, three, five and 10 folds, the test loss with no folds converge more to the training loss when compared to the loss functions using cross validation.

5.3.10 Summary of optimisation

A table summarising the values of the parameters tested can be seen in table 5.1. Both the initialised, tested and final values are listed.

Parameter	Initialised	Tested	Final
Dense block	3	1, 2, 3	2
Data augmentation	No augmentation	No augmentation, rotation, rotation + shift	Rotation
Dropout	0.2	No dropout, 0.2, 0.4, 0.6, 0.8	0.2
Batch size	8	2, 4, 6, 8	8
Filters	256	300, 256, 128, 64, 32, 16	16
Dense layers	4	2, 4, 6, 8	2
Learning rate	0.0001	0.001, 0.0001, 0.00001	0.0001
K - fold cross validation	No folds	No folds, 3 - folds, 5 - folds, 10 - folds	No folds

Table 5.1. Parameters initialised and tested as well as the final values.

After the parameter optimisation the performance was evaluated and compared to a CNN.

5.4 Performance metrics

Sensitivity, specificity, AUC and accuracy are metrics often used in the medical field, as seen in table 1.1, when performing tests such as screening tests [Trevethan, 2017]. These metrics were implemented in the current study. All these metrics are calculated based on the confusion matrix, see table 5.2, which is a table used to define the performance of an algorithm for classification. The table includes the percentage of correct classifications for each class as well as the error for various combinations in percentage. [Semmlow, 2008] A confusion matrix includes the predicted classes as columns and true classes as rows, see table 5.2. True negative (TN) and true positive (TP) is the off-diagonal whereas the diagonal is the false negative (FN) and false positive (FP).

	Predicted Negative	**Predictive Positive**
Actual Negative	TN	FP
Actual Positive	FN	TP

Table 5.2. Two-class confusion matrix presenting TP, TN, FP and FN.

Sensitivity referes to the ability to appropriately classify a subject as diseased when the subject has a disease [Parikh et al., 2008; Semmlow, 2008]. Therefore, sensitivity is a measure of how many subjects have been correctly classified as PD by the network. Sensitivity, expressed in terms of the confusion matrix, is given mathematically as:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity referes to the ability to adequately classify a subject as disease-free when the subject is without a disease [Parikh et al., 2008; Semmlow, 2008]. Meaning specificity is a measure of how many subjects the network has correctly classified as HC. Specificity, expressed in terms of the confusion matrix, is given mathematically as:

$$Specificity = \frac{TN}{TN + FP}$$

Accuracy, is characterised as the ratio of rightly classified patients to the total number of patients [Singh et al., 2021]. Meaning accuracy is the proportion of correct predictions of PD and HC over the total number of predictions, which is given mathematically as:

$$Accuracy = \frac{(TP + TN)}{All\ Predictions}$$

The information provided by a confusion matrix can be used to create ROC, for which the AUC can be calculated. The axes for the curve are TP rate (the y-axis) and FP rate (the x-axis). The points on the ROC curve are referring to the TP and FP rates for different thresholds of classification. For that reason, the AUC performance metric only works for binary classification. To calculate the metric, the area under the ROC curve is determined. [Kamarudin et al., 2017]

These performance metrics were calculated based on the DenseNet performance, furthermore, these performance metrics were compared to those of a conventional CNN.

5.5 Comparison of DenseNet and CNN

The performance of the DenseNet was compared to a conventional CNN to examine if the DenseNet had a better, worse or similar performance. The CNN had convolutional layers, max pooling layers and a softmax layer. Furthermore, the CNN used same optimiser and loss function as the DenseNet. A random search testing 200 different combinations was performed to find the most optimal number of filters in each layer (ranging from two to 50) and number of layers (ranging from two to 21). The final hyperparameters chosen were filters ranging from six to 49, and 18 convolutional layers all with max pooling. The best CNN found was compared to the DenseNet.

Part III

Results, discussion and conclusion

Results 6

6.1 DenseNet

The purpose of the study was to extract VD, and based on this brain structure classify the subjects using a DenseNet as either PD or HC. The performance of the classification was evaluated using accuracy, AUC, sensitivity and specificity as well as the confusion matrix. Furthermore, a ROC with the associated AUC was plotted. A loss curve for the final network and an accuracy curve were also included. The results were based on the best performing network chosen on the basis of parameter optimisation. The best network was trained using the training and test data. Performance of the network is showcased in the form of a confusion matrix in table 6.1.

	Predicted HC	Predicted PD
Actual HC	3	6
Actual PD	4	18

Table 6.1. Performance evaluation of the 3D DenseNet in the form of a confusion matrix based on the validation set. Green indicates correct classifications and red indicates wrong classifications.

Table 6.1 shows that the network most often predicted HC incorrectly versus PD, as 66% of HC were incorrectly classified and only 19% for PD were incorrectly classified.

Table 6.2 showcases the performance of the network on the test set and the validation set.

Metric	**Test**	**Validation**
Accuracy	0.84	0.68
AUC - ROC	0.76	0.58
Sensitivity	0.91	0.82
Specificity	0.67	0.33

Table 6.2. Performance evaluation of the 3D DenseNet.

Table 6.2 shows that the network performed better on the test set than the validation set.

Results arranged as a ROC curve can be seen on figure 6.1.

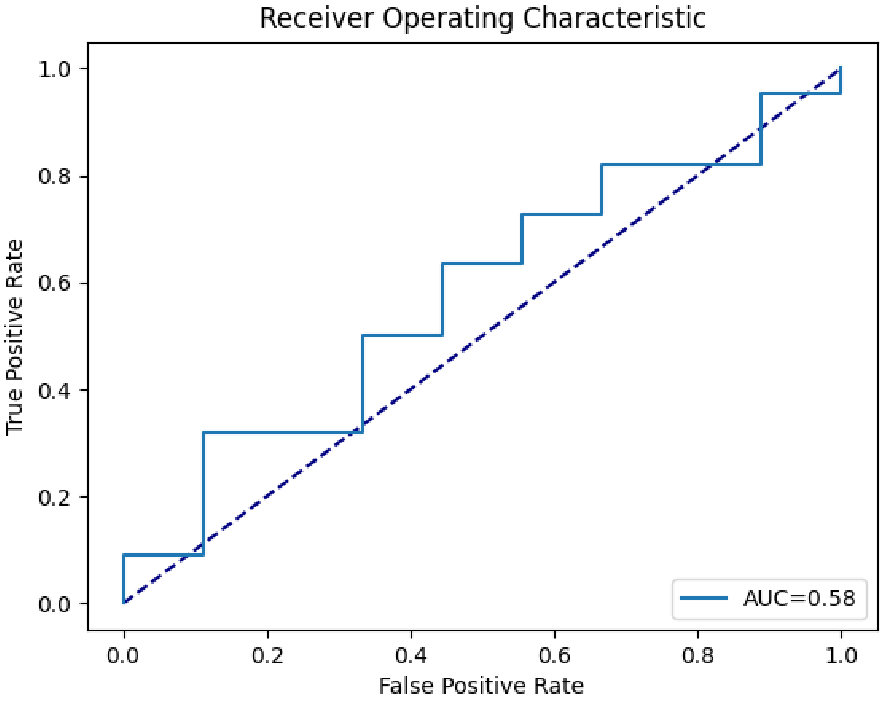


Figure 6.1. AUC - ROC of the 3D DenseNet. False Positive Rate on the x-axis and True Positive Rate on the y-axis. The diagonal line illustrates a random guess.

Figure 6.1 illustrates the diagnostic ability of the network. The more the curve is above the diagonal line, the better the relationship between true and false positives are as the class prediction threshold changes. Oppositely, the more the curve is under the diagonal line the worse the relationship are between true and false positives. Figure 6.1 shows the network classifications were around the diagonal line so the prediction fluctuated between being better than random guessing or worse than random guessing, yet primarily above the diagonal line.

Figure 6.2 shows the final loss of the network on the validation set.

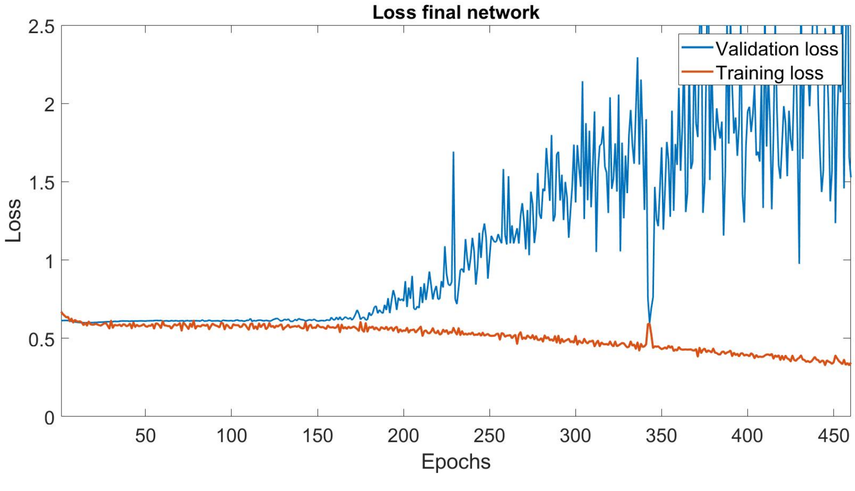


Figure 6.2. The loss function of the final DenseNet. Validation loss is the blue graph while training loss is the red graph.

Figure 6.2 shows the loss of the final DenseNet, as the figure shows the fitting of the network to the training data was not very good. However, the training loss decreased, meaning the network was learning over epochs. Contrary to the training loss, the validation loss increased over epochs.

Figure 6.3 illustrates the accuracy of the final network.

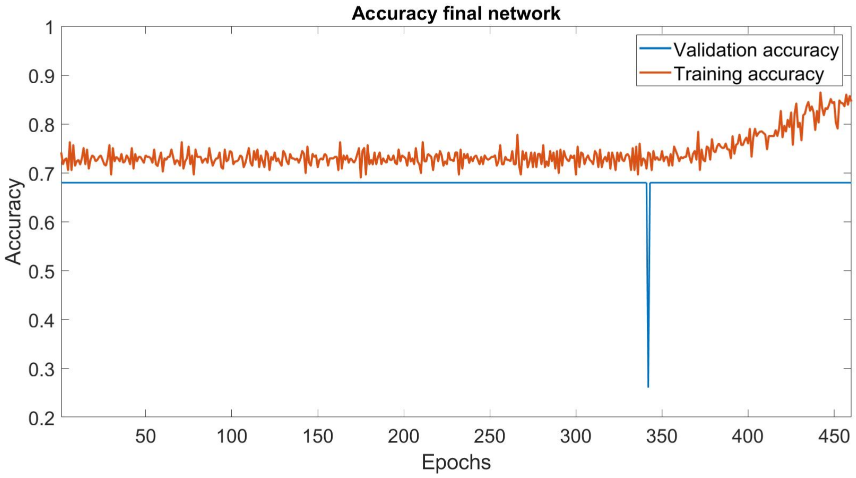


Figure 6.3. The accuracy of the final DenseNet. The validation accuracy is the blue graph while the training accuracy is the red graph

Figure 6.3 shows the accuracy of the network. The accuracy, for the training, was at a plateau until 350 epochs whereafter the accuracy increased. The validation accuracy was stagnant meaning the network was not able to learn to a better degree. After 350 epochs the gap between validation and training became larger which was an indication of overfitting.

6.2 CNN

Additionally to the DenseNet a conventional CNN was tested, which was evaluated using the same performance metrics as for the DenseNet.

The confusion matrix is shown in 6.3.

	Predicted HC	Predicted PD
Actual HC	2	7
Actual PD	8	14

Table 6.3. Performance evaluation of the 3D CNN in the form of a confusion matrix. Green indicates correct classifications and red indicates wrong classifications.

Table 6.1 shows that the network most often predicted HC incorrectly versus PD, as 78% of HC were incorrectly classified and only 36% for PD were incorrectly classified.

Metric	**Validation**
Accuracy	0.52
AUC - ROC	0.39
Sensitivity	0.64
Specificity	0.22

Table 6.4. Performance evaluation of the 3D CNN.

As shown in the metrics in table 6.4 the CNN had worse performance in accuracy, AUC-ROC, sensitivity and specificity compared to the DenseNet.

Results of the CNN arranged as a ROC curve can be seen on figure 6.4.

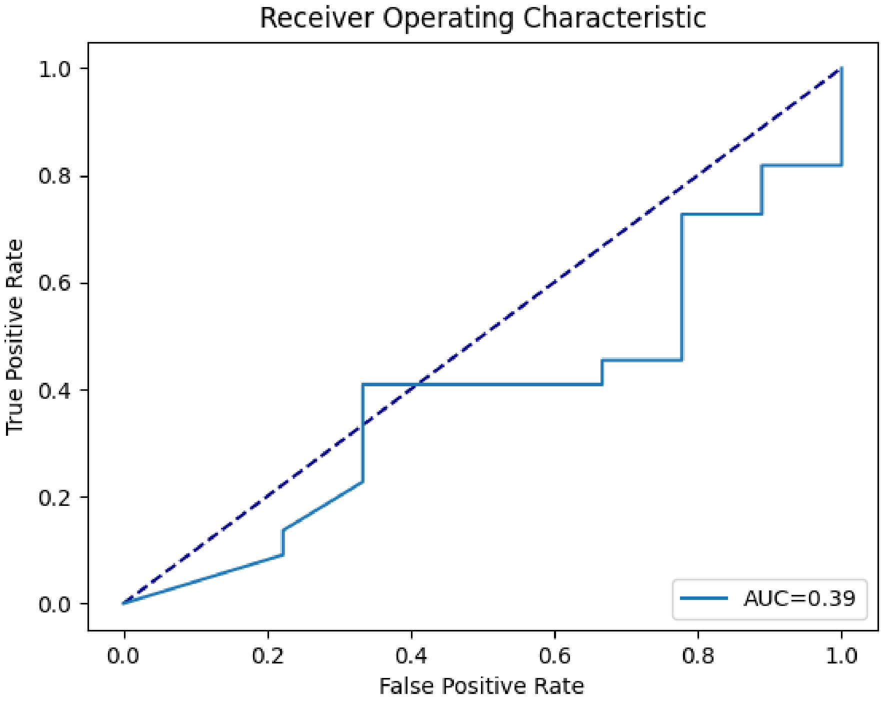


Figure 6.4. AUC - ROC of the 3D CNN. False Positive Rate on the x-axis and True Positive Rate on the y-axis. The diagonal line illustrates a random guess.

Figure 6.4 shows the network classifications were under the diagonal line, except at one point, meaning the predictions were generally worse than random guessing.

Figure 6.5 shows the loss of the final CNN using the validation set.

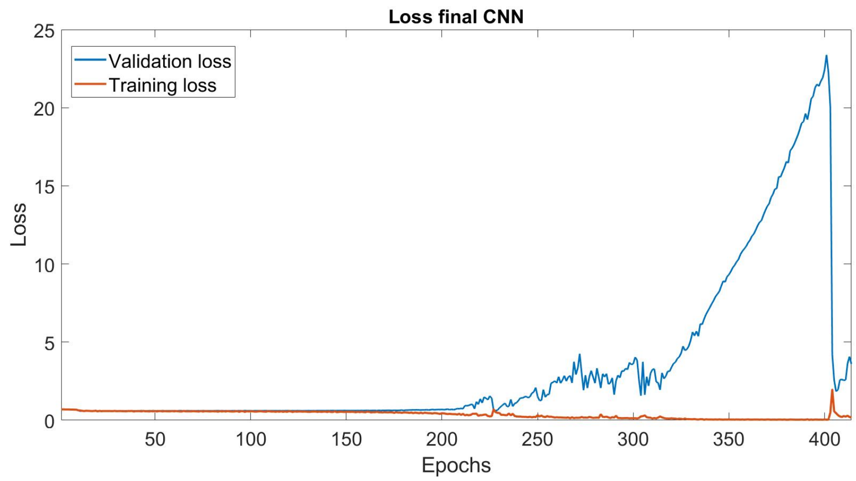


Figure 6.5. The loss function of the final CNN. Validation loss is the blue graph while training loss is the red graph.

Figure 6.5 shows the loss of the final CNN, the training and validation loss were lower in the first 200 epochs. After 200 epochs the validation loss increases extremely, this indicates the fit of the validation data only get worse over epochs. An increase in the validation data indicates the network was overfitted to the training data.

Figure 6.6 illustrates the accuracy of the final CNN.

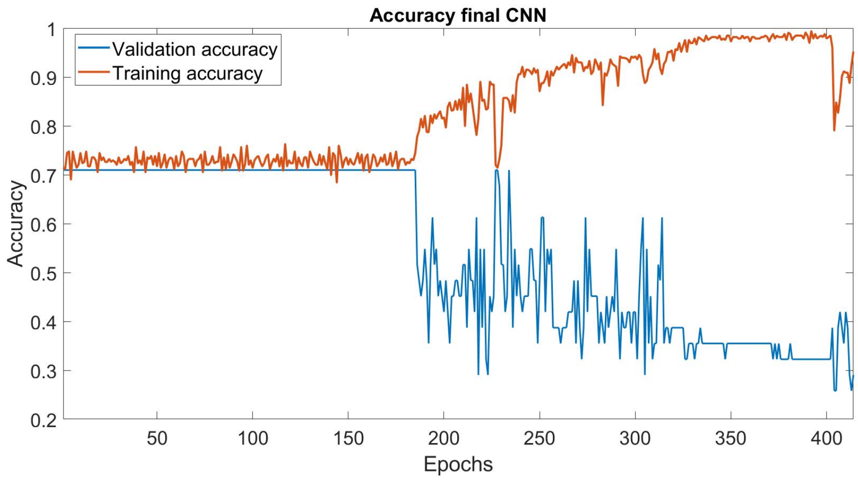


Figure 6.6. The accuracy of the final CNN. The validation accuracy is the blue graph while the training accuracy is the red graph

Figure 6.6 shows the accuracy of the CNN. In the first 200 epochs the training and validation accuracy was a stable horizontal line, meaning the accuracy did not improve or get worse. After the 200 epochs the training accuracy increased while the validation accuracy decreased, this indicates the network was overfitted to the training data.

Discussion 7

7.1 Results

Two studies who used a CNN to discriminate PD patients from HCs were Shinde et al. [2019] and Xiao et al. [2019]. Shinde et al. [2019] used a ResNet50 to classify PD patients and obtained an accuracy of 0.80, a sensitivity of 0.86, specificity of 0.70 and an AUC - ROC of 0.91, whereas Xiao et al. [2019] obtained an accuracy of 0.85, a sensitivity of 0.86 and a specificity of 0.83 as well as an AUC - ROC of 0.93. The accuracy obtained in the current study is 0.68, a sensitivity of 0.82 while the specificity is 0.33 and AUC-ROC of 0.58. It is clear that the specificity and AUC-ROC are less accurate compared to those by Shinde et al. [2019] and Xiao et al. [2019]. However, in general PD is hard to diagnose as misdiagnosis of PD have an incidence of 5-25% [Shin et al., 2021; Raff et al., 2006]. Given the incidence of misdiagnoses is up to 25%, and the DenseNet having an accuracy of 68%, the DenseNet is not far of from clinicians when predicting PD. Leading to the DenseNet having some potential, if optimised when predicting PD. Howbeit, when clinicians misdiagnose PD it is often mistaken for other Parkinsonian Syndromes where the DenseNet only have to distinguish between PD and HCs [Raff et al., 2006].

A conventional CNN architecture was designed to examine if there would be a difference in performance. As seen in results, the CNN performed worse than the DenseNet when evaluated on accuracy, sensitivity, specificity and ROC-AUC. This express that the network architecture had an impact on the output, and the output was not solely dependent on the data used as input.

The results obtained in the current study can not classify PD patients to the same extent as similar studies, see table 1.1.

The gap between the current study's results and those found by similar literature might be caused by a number of factors: different brain structures, possibility of bad segmentation of brain structure, network architecture and optimisation process.

7.2 Ventral diencephalon used as a biomarker for PD

The novelty of the study is to use the segmented structure VD and a DenseNet as classifier for PD patients and HCs. Therefore, the results obtained can be used to examine if VD can be used as a biomarker for PD. The segmented VD contain the following brain structures: hypothalamus, mammillary body, subthalamic nuclei, substantia nigra and red nucleus. Similar literature uses substantia nigra or Nigrosome-1 as a biomarker. It is well documented that changes in the amount of neurons within these brain structures are reduced as well as the volume of the structure. [Jo and Oh, 2020; Prange et al., 2019; Ogisu et al., 2013]

Xiao et al. [2014] segmented the structures: subthalamic nuclei, substantia nigra and red nucleus in 33 PD patients. A PCA-Based Variability Analysis was performed to examine if all three brain structures had equally big variation in PD patients. Xiao et al. [2014] found that the structures subthalamic nuclei and substantia nigra had significantly higher variability in PD patients compared to red nucleus. Xiao et al. [2014] also performed a correlation of the UPDRS III score and the shape variations in the brain structures and indicate that these might be related. The three brain structures examined by Xiao et al. [2014] are included in VD. The subthalamic nuclei possibly includes additional relevant information in the classification process of PD. However, the difference of shape variations in brain structures dependent on the disease might have impacted the classification. Therefore, it would have been an advantage to examine if there were a higher mismatch in the classifications of PD patients in the mild end of the UPDRS scale compared to patients in the high end.

It could also have helped the classification if only data from patients in the higher stages of PD were included, as more clear structural changes would have been ensured. The volume of SN decreases as the disease progress and the patients reaches a higher stage of PD [Ziegler et al., 2013]. The patients included in the study only had baseline images, therefore, it can be assumed that a part of the included patients will not have an advanced enough progression of PD to make it distinguishable enough from the HCs for the classifier.

7.3 Optimisation process

Choosing fitting hyperparameters is a complicated task, which can be improved by testing many different combinations of hyperparameters. The approach used to select the hyperparameters in the study might be improved by testing different combinations of hyperparameters. In the approach used the hyperparameters were tested one at a time, while the other hyperparameters were kept at the initialised value. However, the best combination of hyperparameters was possibly not found using this approach. A grid search testing multiple values for multiple hyperparameters to determine the best combination of hyperparameters might have provided an architecture capable of better performance. However, the values of some hyperparameters was also limited by the computational memory, one being the batch size. Radiuk [2017] used CIFAR-10 and MNIST data sets, with 60.000 and 70.000 images, respectively, and CNN as classifier and examined if using different batch sizes would affect the accuracy. Radiuk [2017] used batch sizes varying from 16 to 1024. The highest accuracies for the CIFAR-10 and MNIST data sets were obtained using the highest batch sizes of 512 and 1024. In the current study it was planned to initialise the batch size to 64, based on the study by Liu et al. [2020] including 97 AD patients and 119 HCs. A batch size of 64 might have been to high for the current study, however, this remains unclear if a more fitting batch size existed due to the computational memory. Furthermore, the learning rate can also be complicated to determine.Smith [2018] tested different learning rates on the CIFAR-10 dataset, based on a learning rate of 0.001 the loss function showed an underfitted curve whereas a learning rate of 0.004 was overfitted, therefore, Smith [2018] suggest to use a grid search to find the optimal learning rate. Using a smaller learning rate in the current study lead to a more stable test loss, however, the test loss did not converge to the training loss it was an almost horizontal and slightly increasing loss. The most fitting learning rate was probably not tested in this study, however, it might have been discovered using a grid search.

7.4 Performance metrics

The results is a collection of various performance metrics based on a networks prediction on a validation data set. One metric used for assessing the quality of a network while hyperparameter optimising were accuracy. Accuracy was chosen for the intuitiveness and inclusiveness of both positive and negative predictions. While the inclusiveness made it easily understandable, it could falsely show low predictive power as "good". If a network blindly predicts the class which is most present in the data set, the accuracy would be falsely high as the network have learnt it will get a high accuracy simply by guessing the predominant class.As seen in section 4.2.1, the validation split is 70.96/29.04 in PD's favor. Multiple trained networks were confirmed to only predict PD. For a more nuanced look into the networks performance, sensitivity and specificity were reported alongside accuracy. Whether importance is highest on specificity or sensitivity depends on the usage of the network, and the consequences of the prediction. Especially the consequence of false negatives versus false positives should be taken into consideration. For many early predictive methods in the healthcare system false negatives may have graver consequences than false positives. In case of a Parkinson's diagnosis, a positive prediction means that the patient will receive treatment, often in the form of levodopa medicine which have many known side effects [Armstrong and Okun, 2020]. These side effects should be compared to the consequence of a false negative - that the disease will go untreated and thereby progress faster than a treated one [Jankovic and Tan, 2020].

7.5 Practical implementation of system

A practical implementation of the DenseNet used to classify PD patients would require the segmentation process to be less time consuming and implemented directly as a part of the network. Furthermore, it would also require the network to perform better than the clinician with no more than 5% of misclassifications [Shin et al., 2021]. Shin et al. [2021] classifies PD patients based on Nigrosome-1, this area of SN is extracted using manual segmentation, which is really time consuming. Nevertheless, Despotović et al. [2015] states that manual segmentation is the most accurate segmentation method. Even though the segmentation was more time consuming compared to Shin et al. [2021], it would not require a professional neuroradiologist or neurologist to segment the brain structure manually. The diagnosis performance of the CNN created by Shin et al. [2021] had similar performance as visual assessment performed by a neuroradiologist. The segmentation was performed using the FreeSurfer software, with the segmentation process being approximately seven hours for each image. Liu et al. [2020] created a multi-task CNN used to both segment hippocampus and based on this segmentation classify either AD or HC. The segmentation of one image in the study by Liu et al. [2020] lasted 0.29s whereas the classification was 0.85s. The CNN used by Liu et al. [2020] trained the ability to segment hippocampus based on images segmented using a software library, afterwards the initial segmentation was edited by three neuroradiologists. As it is necessary to include a neuroradiologist or neurologist to mark the ROI, it would also be a comprehensive process to include segmentation in the DenseNet created in this study. Nevertheless, it would be beneficial for the practical implementation of the system not to use two different software programs and to make the process of classification less time consuming.

Conclusion 8

Literature and research examine the classification of PD using different machine learning tools, with the aim of potentially improving the diagnosis process of PD in the future. The study examined the classification performance of a DenseNet based on 3D segmented MRIs obtained from the PPMI database. The use of VD have not been examined in relation to classifying PD, therefore, this study can help evaluate the use of VD as a biomarker for PD. The best architecture of the DenseNet were found through manual hyperparameter optimisation. To evaluate the final network, the performance metrics: accuracy, specificity, sensitivity and AUC-ROC were chosen. For comparison, a sequential CNN architecture were found using random search and trained, tested and validated on the same data as the DenseNet. The networks had similar results in all performance metrics.

The performance metrics from the best DenseNet and CNN showed worse performance compared to previous studies. Therefore, the use of VD as a biomarker and DenseNet as a classifier does not seem beneficial compared to other studies, as it showed worse performance.

Part IV

Problem Based Learning

Time management 9

Time management was controlled using different schedules, with daily agendas being the most detailed and Gantt chart giving an overview of the entire project period. All levels were used for the project management.

9.1 Gantt chart

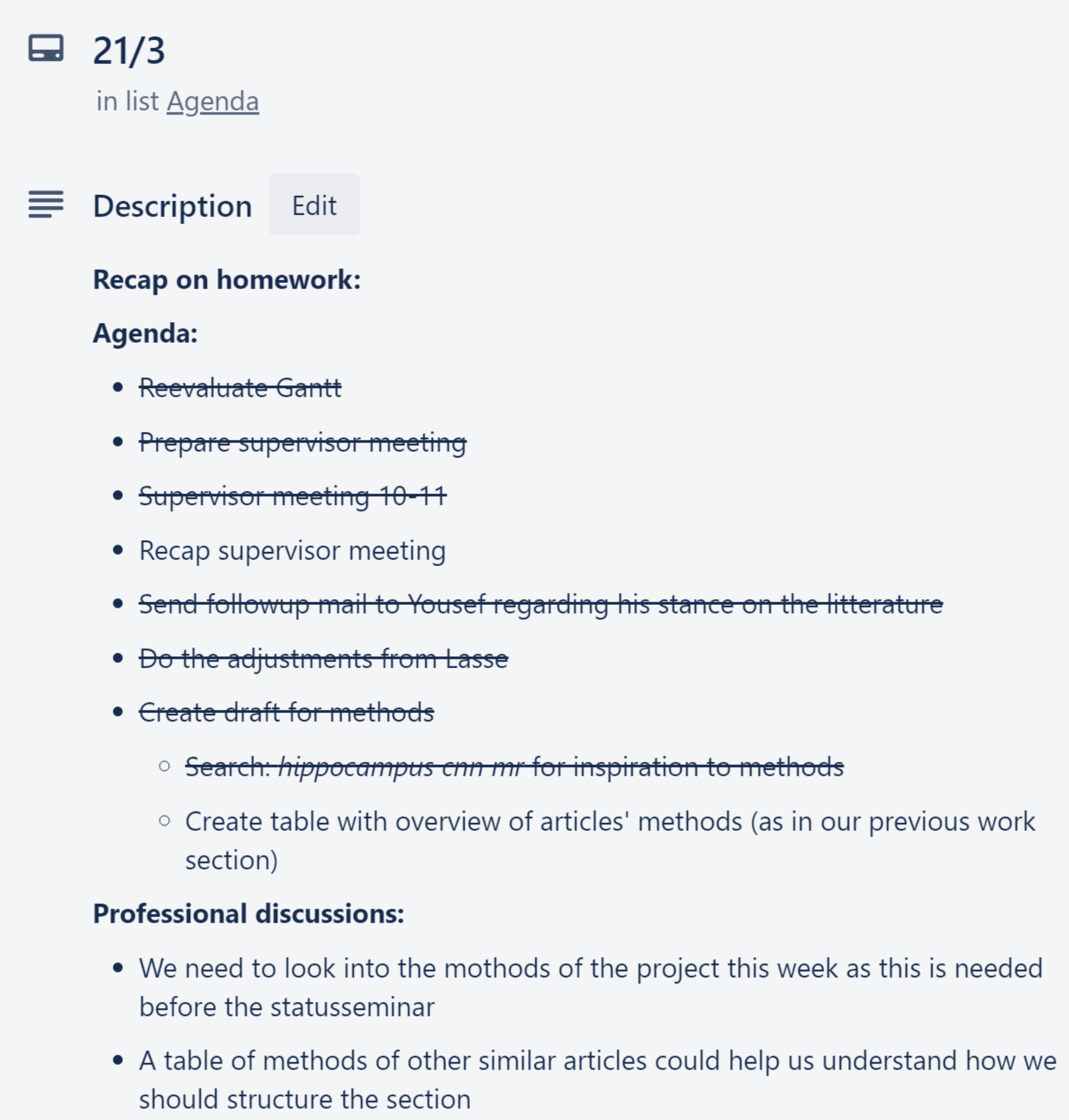
A Gantt chart was made for the whole project period at the beginning. It included all the tasks of the project the group members could think of in the beginning, and the time frame for each task in weeks. The process of creating Gantt chart consisted of brainstorming the different tasks, creating deadlines for the tasks using backcasting. Throughout the project period the Gantt chart was revisited in the beginning of each week for adjustments, to create subtasks for the week to accommodate the time frames of the tasks. This was an effort to ensure that the deadlines given by the Gantt chart chart were upheld.

9.2 Week plan

The Gantt chart gave an overall view of the work and goals for each week, whereas a week plan was used to give a better overview of the tasks for a week. The week plan was made each monday based on the Gantt chart. The purpose of the week plan was to plan some general activities each day which made it easier to accommodate the tasks of the week. Additionally, the week plan was the foundation of the daily agenda. The purpose of the week planning at the beginning of the period was to specify the task at hand whereas from midway-through to the end of the period it was more detailed with goals of the week. A more detailed week plan gave the group members a better overview of the week as well as the tasks of the week became more palpable.

9.3 Daily agenda

Everyday the group members started by making a daily agenda of the tasks based on the week plan and dividing the tasks between the group members. The daily agenda was based on the week plan. A card showing the daily agenda was created each day to keep track of the tasks. Besides the daily agenda, recap of homework from the day before was also included, a recap of professional discussion and homework, if any was noted. An example of a daily agenda card can be seen on figure 9.1.



21/3

in list Agenda

Description

Recap on homework:

Agenda:

- ~~Reevaluate Gantt~~
- ~~Prepare supervisor meeting~~
- ~~Supervisor meeting 10-11~~
- Recap supervisor meeting
- ~~Send followup mail to Yousef regarding his stance on the litterature~~
- ~~Do the adjustments from Lasse~~
- ~~Create draft for methods~~
  - ~~Search: *hippocampus cnn mr* for inspiration to methods~~
  - Create table with overview of articles' methods (as in our previous work section)

Professional discussions:

- We need to look into the mothods of the project this week as this is needed before the statusseminar
- A table of methods of other similar articles could help us understand how we should structure the section

Figure 9.1. Example of an agenda with the agenda of the day as well as the professional discussion.

The daily agenda helped the group keep track of the progression made on the weekly goals, and ensure that none were lacking. If any were lacking behind, extra focus would be on completing these goals. At the end of the workday the group members recapped what they had worked on and it was noted in the *"professional discussions"* point. This ensured peer-learning, and it was planned to document it each day to be sure everyone had an understanding of all tasks. However, this was not documented every day. The days it was documented the group members agreed that it helped with the overview of progression in the different tasks, and it would have proven useful to do it more rigorously. The documentation of "professional discussions" also helped if the group members swapped tasks and a group member was assigned a task for the first time, an introduction were given by the last group member who had worked on the task.

9.4 Goal of the week

Goal of the week were goals that the group members strove to achieve by the end of the week through tasks, and were based on the tasks in Gantt chart. *Goal of the week* was created because of missing clarity for the tasks in Gantt. There was uncertainty for the task when planning the week, therefore, *goal of the week* was a newly introduced tool used to provide clarity as well as help optimise the time management. Before the week plan was made, *goal of the week* was determined based on Gantt chart. When the *goal of the week* was determined, cascading was used to determine sub goals necessary to reach the *goal of the week*.

When this tool was taken into use, the goal became more tangible and it gave the clarity that otherwise would have been lacking for some of the tasks in the Gantt chart. The sub goals provided a better overview of how the goal was going to be accomplished. At first the final goal of the week was discussed by all group members, here the big decisions were discussed, this could for example be which type of network to use. When the primary decision had been discussed and agreed on, sub goals were divided between group members. Using *Goal of the week* gave a more specific and structured overview, of how to obtain the goal, this was also the intention since this overview was lacking when using Gantt chart.

9.4.1 The end period

Furthermore, within the last five weeks of the project period a timescale was made. The purpose of the timescale was to prioritise the remaining time. The timescale had three subcategories: Code for network, written work and proofreading. Each subcategory had different tasks given a number representing the prioritisation, additionally the three subcategories were marked with the specific weeks used to work on the subcategories. This method was deemed more manageable than the larger Gantt chart chart given the limited amount of tasks left and given that the tasks generally did not have many levels of subtasks, contrary to the tasks earlier in the project. It allowed the group members to have a clear view of the remaining tasks, but would not have worked with too many larger tasks, as earlier in the project period.

9.5 Unforeseen situations and challenges to the time schedule

In the beginning of the project it was the clear collective understanding that the data the project should use was to be available from the Department of Radiology at Aalborg University Hospital. Therefore, communication through e-mails and meetings was included in the daily agendas. However, the cooperation was challenged due to data not arriving earlier in the project period. This delayed several processes in the time schedule, and required major revisions of Gantt. Therefore, two plans were created plan A and B, where plan A was to use the MRIs from the hospital, and plan B was to examine if any data was available in online databases. Data from the PPMI database was available after filling in a formula on the purpose of the study. Obtaining the data from the hospital was delayed to a point were it caused serious challenge to the initial time schedule. Eventually, due to the delays, the plan B was decided upon. This complication regarding plan A caused a net loss of time, as resources were spent on specific tasks related to data from the hospital. As the original Gantt chart was planned with some weeks worth of buffer time, the change in data supplier did not have as grave an impact on the schedule as it could have had without the buffer. However, some changes in the Gantt chart schedule had to be made. The consequence of all this were a timeschedule with less time allocated for tasks regarding network design, testing architectures, hyperparameter optimisation, the writing process and proofreading. The writing tasks and network design were completed without any substantial compromises, as they were tasks with a specific end. Hyperparameter optimisation and proofreading could be performed in a near infinite manner, therefore, clear boundaries were established before either process began. More time for testing architectures and different data sets was wanted by the end of the project period. Both processes could have been continued with potential gains from the extra resources spent, but those resources were spent on getting the data from the hospital as mentioned.

Bibliography

A. Abos, H. C. Baggio, B. Segura, A. Campabadal, C. Uribe, D. M. Giraldo, A. Perez-Soriano, E. Munoz, Y. Compta, C. Junque, et al. Differentiation of multiple system atrophy from parkinson's disease by structural connectivity derived from probabilistic tractography. *Scientific reports*, 9(1):1–12, 2019.

S. Anwaar. *Dense Net Image Classification*, 2019. URL `https://www.kaggle.com/code/sohaibanwaar1203/dense-net-image-classification/notebook`.

M. J. Armstrong and M. S. Okun. Diagnosis and treatment of parkinson disease: a review. *Jama*, 323(6):548–560, 2020.

D. Balduzzi, M. Frean, L. Leary, J. Lewis, K. W.-D. Ma, and B. McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning*, pages 342–350. PMLR, 2017.

N. Basta. *The Differences between Sigmoid and Softmax Activation Functions*, 2020. URL `https://medium.com/arteos-ai/the-differences-between-sigmoid-and-softmax-activation-function-12adee8cf322`.

M. Blaimer, F. Breuer, M. Mueller, R. M. Heidemann, M. A. Griswold, and P. M. Jakob. Smash, sense, pils, grappa: how to choose the optimal method. *Topics in Magnetic Resonance Imaging*, 15(4):223–236, 2004.

H. Braak, K. Del Tredici, U. Rüb, R. A. De Vos, E. N. J. Steur, and E. Braak. Staging of brain pathology related to sporadic parkinson's disease. *Neurobiology of aging*, 24(2): 197–211, 2003.

J. Brownlee. *How to use Learning Curves to Diagnose Machine Learning Model Performance*, 2019. URL `https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/`.

J. Brownlee. *A Gentle Introduction to the Rectified Linear Unit (ReLU)*, 2020. URL `https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/`.

R. M. Califf. Biomarker definitions and their applications. *Experimental Biology and Medicine*, 243(3):213–221, 2018.

N. S. Chauhan. *Loss Functions in Neural Networks*, 2021. URL `https://www.theaidream.com/post/loss-functions-in-neural-networks`.

Cleveland Clinic. *Mild Cognitive Impairment*, 2019. URL `https://my.clevelandclinic.org/health/diseases/17990-mild-cognitive-impairment`.

Dansk Selskab For Bevægeforstyrrelser. Parkinsons sygdom klinisk vejledning - diagnose, forløb og behandling fra et tværfagligt perspektiv. 2. udgave:13, 2011.

L. M. De Lau and M. M. Breteler. Epidemiology of parkinson's disease. *The Lancet Neurology*, 5(6):525–535, 2006.

I. Despotović, B. Goossens, and W. Philips. Mri segmentation of the human brain: challenges, methods, and applications. *Computational and mathematical methods in medicine*, 2015, 2015.

A. Dhakal and B. D. Bobrin. Cognitive deficits. *StatPearls [Internet]*, 2020.

R. L. Doty. Olfactory dysfunction in parkinson disease. *Nature Reviews Neurology*, 8(6): 329–339, 2012.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. 2001. ISBN 978-0-471-05669-0.

V. Flovik. *How to use machine learning for production optimization - Utilizing data to improve performance*, 2018. URL `https://towardsdatascience.com/machine-learning-for-production-optimization-e460a0b82237`.

M. Fogarty. *Recon All*, 2017. URL `https://surfer.nmr.mgh.harvard.edu/fswiki/recon-all`.

A. Galvan and T. Wichmann. Pathophysiology of parkinsonism. *Clinical neurophysiology*, 119(7):1459–1474, 2008.

Y. Gao, N. Kerle, J. Mas, J. Pacheco, and I. Niemeyer. Optimized image segmentation and its effect on classification accuracy. 05 2012.

M. Goedert. Alzheimer's and parkinson's diseases: The prion concept in relation to assembled aβ, tau, and α-synuclein. *Science*, 349(6248):1255555, 2015.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*, chapter 3. MIT Press, 2016. `http://www.deeplearningbook.org`.

R. L. Haining and C. Achat-Mendes. Neuromelanin, one of the most overlooked molecules in modern medicine, is not a spectator. *Neural regeneration research*, 12(3):372, 2017.

J. E. Hall and A. C. Guyton. *Guyton and Hall textbook of medical physiology*, pages 552–554;681–694. Elsevier Health Sciences, 2011.

S. Haller, S. Badoud, D. Nguyen, I. Barnaure, M. Montandon, K. Lovblad, and P. Burkhard. Differentiation between parkinson disease and other forms of parkinsonism using support vector machine analysis of susceptibility-weighted imaging (swi): initial results. *European radiology*, 23(1):12–19, 2013.

M. M. Hoehn, M. D. Yahr, et al. Parkinsonism: onset, progression, and mortality. *Neurology*, 17(5):318–318, 1967.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

A. Jahn. *FreeSurfer Tutorial 3: Recon-all*, 2019. URL `https://andysbrainbook.readthedocs.io/en/latest/FreeSurfer/FS_ShortCourse/FS_03_ReconAll.html`.

J. Jankovic and E. K. Tan. Parkinson's disease: Etiopathogenesis and treatment. *Journal of Neurology, Neurosurgery & Psychiatry*, 91(8):795–808 (800), 2020.

M. Jo and S.-H. Oh. A preliminary attempt to visualize nigrosome 1 in the substantia nigra for parkinson's disease at 3t: An efficient susceptibility map-weighted imaging (smwi) with quantitative susceptibility mapping using deep neural network (qsmnet). *Medical physics.*, 47(3), 2020. ISSN 0094-2405.

A. N. Kamarudin, T. Cox, and R. Kolamunnage-Dona. Time-dependent roc curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17(1):1–19, 2017.

D. Kim, G. Tung, U. Akbar, and J. Friedman. The evaluation of the swallow tail sign in patients with parkinsonism or gait disorders. In *Movement Disorders*, volume 36, pages S360–S361, 2021.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

T. K. Larsen. Lecture in machine learning. The spring semester, 2021.

C. Liu, H. Wei, N.-J. Gong, M. Cronin, R. Dibb, and K. Decker. Quantitative susceptibility mapping: contrast mechanisms and clinical applications. *Tomography*, 1 (1):3–17, 2015.

M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, M. Xu, A. D. N. Initiative, et al. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease. *Neuroimage*, 208:116459, 2020.

P. Mahlknecht, F. Krismer, W. Poewe, and K. Seppi. Meta-analysis of dorsolateral nigral hyperintensity on magnetic resonance imaging as a marker for parkinson's disease. *Movement Disorders*, 32(4):619–623, 2017.

K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.

Mayo Clinic. *Mild Cognitive Impairment (MCI)*, 2020. URL `https://www.mayoclinic.org/diseases-conditions/mild-cognitive-impairment/symptoms-causes/syc-20354578`.

A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick. On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*, 2018.

Musstafa. *Optimizers in Deep Learning*, 2021. URL https://medium.com/mlearning-ai/optimizers-in-deep-learning-7bf81fed78a0.

K. Ogisu, K. Kudo, M. Sasaki, K. Sakushima, I. Yabe, H. Sasaki, S. Terae, M. Nakanishi, and H. Shirato. 3d neuromelanin-sensitive magnetic resonance imaging with semi-automated volume measurement of the substantia nigra pars compacta for diagnosis of parkinson's disease. *Neuroradiology*, 55(6):719–724, 2013.

R. Parikh, A. Mathai, S. Parikh, G. C. Sekhar, and R. Thomas. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56 (1):45, 2008.

Parkinson Foreningen. *Hvad er parkinson?*, 2022. URL https://www.parkinson.dk/viden-forskning/om-parkinson/hvad-er-parkinson/#section-3.

PPMI initiative. *Data at a Glance*, 2022. URL https://www.ppmi-info.org/access-data-specimens/data.

S. Prange, E. Metereau, and S. Thobois. Structural imaging in parkinson's disease: new developments. *Current Neurology and Neuroscience Reports*, 19(8):1–13, 2019.

P. M. Radiuk. Impact of training set batch size on the performance of convolutional neural networks for diverse datasets. 2017.

U. Raff, M. Hutchinson, G. M. Rojas, and I. Huete. Inversion recovery mri in idiopathic parkinson disease is a very sensitive tool to assess neurodegeneration in the substantia nigra: preliminary investigation. *Academic radiology*, 13(6):721–727, 2006.

N. Ramli, S. Nair, N. Ramli, and S. Lim. Differentiating multiple-system atrophy from parkinson's disease. *Clinical radiology*, 70(5):555–564, 2015.

P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. *Encyclopedia of database systems*, 5:532–538, 2009.

S. Saha. *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*, 2018. URL https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

M. Sasaki, E. Shibata, K. Tohyama, J. Takahashi, K. Otsuka, K. Tsuchiya, S. Takahashi, S. Ehara, Y. Terayama, and A. Sakai. Neuromelanin magnetic resonance imaging of locus ceruleus and substantia nigra in parkinsons disease. *Neuroreport*, 17(11): 1215–1218, 2006. ISSN 0959-4965.

N. Schmansky. *FreeSurfer*, 2022. URL https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferMethodsCitation.

W. J. Schulz-Schaeffer. The synaptic pathology of α-synuclein aggregation in dementia with lewy bodies, parkinson's disease and parkinson's disease dementia. *Acta neuropathologica*, 120(2):131–143, 2010.

J. L. Semmlow. *Biosignal and medical image processing*, page 545. CRC press, 2008.

T. Shah. *About Train, Validation and Test Sets in Machine Learning*, 2017. URL `https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7`.

D. H. Shin, H. Heo, S. Song, N.-Y. Shin, Y. Nam, S.-W. Yoo, J.-S. Kim, J. H. Yoon, S. H. Lee, Y. H. Sung, et al. Automated assessment of the substantia nigra on susceptibility map-weighted imaging using deep convolutional neural networks for diagnosis of idiopathic parkinson's disease. *Parkinsonism & Related Disorders*, 85:84–90, 2021.

S. Shinde, S. Prasad, Y. Saboo, R. Kaushick, J. Saini, P. K. Pal, and M. Ingalhalikar. Predictive markers for parkinson's disease using deep neural nets on neuromelanin sensitive mri. *NeuroImage: Clinical*, 22:101748, 2019.

C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

K. K. Singh, M. Elhoseny, A. Singh, and A. A. Elngar. *Machine Learning and the Internet of Medical Things in Healthcare*. Academic Press, 2021.

L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

J. Sonne, V. Reddy, and M. R. Beato. Neuroanatomy, substantia nigra. 2019.

Y. Tang, L. Meng, C.-m. Wan, Z.-h. Liu, W.-h. Liao, X.-x. Yan, X.-y. Wang, B.-s. Tang, and J.-f. Guo. Identifying the presence of parkinson's disease using low-frequency fluctuations in bold signals. *Neuroscience letters*, 645:1–6, 2017.

N. Titova, C. Padmakumar, S. J. Lewis, and K. Chaudhuri. Parkinson's: a syndrome rather than a disease? *Journal of Neural Transmission*, 124(8):907–914, 2017.

R. Trevethan. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Frontiers in public health*, 5:307, 2017.

T. W. Vanderah and D. J. Gould. The thalamus and internal capsule : Getting to and from the cerebral cortex - nolte's the human brain. In *Nolte's The Human Brain*, pages 372–394, 2020.

J. Wang, L. He, H. Zheng, and Z.-L. Lu. Optimizing the magnetization-prepared rapid gradient-echo (mp-rage) sequence. *PloS one*, 9(5):e96899, 2014.

Z. Wang, X.-G. Luo, and C. Gao. Utility of susceptibility-weighted imaging in parkinson's disease and atypical parkinsonian disorders. *Translational neurodegeneration*, 5(1):1–8, 2016.

World Health Organization. *Brief review of selected topics - Neurology Atlas*, 2004. URL `https://www.who.int/mental_health/neurology/neurogy_atlas_review_references.pdf`.

B. Xiao, N. He, Q. Wang, Z. Cheng, Y. Jiao, E. M. Haacke, F. Yan, and F. Shi. Quantitative susceptibility mapping based hybrid feature extraction for diagnosis of parkinson's disease. *NeuroImage: Clinical*, 24:102070, 2019.

Y. Xiao, P. Jannin, T. D'Albis, N. Guizard, C. Haegelen, F. Lalys, M. Vérin, and D. L. Collins. Investigation of morphometric variability of subthalamic nucleus, red nucleus, and substantia nigra in advanced parkinson's disease patients using automatic segmentation and pca-based analysis. *Human brain mapping.*, 35(9), 2014. ISSN 1065-9471.

L. Zecca, D. Tampellini, M. Gerlach, P. Riederer, R. Fariello, and D. Sulzer. Substantia nigra neuromelanin: structure, synthesis, and molecular behaviour. *Molecular Pathology*, 54(6):414, 2001.

D. A. Ziegler, J. S. Wonderlick, P. Ashourian, L. A. Hansen, J. C. Young, A. J. Murphy, C. K. Koppuzha, J. H. Growdon, and S. Corkin. Substantia nigra volume loss before basal forebrain degeneration in early parkinson disease. *JAMA neurology*, 70(2): 241–247, 2013.

Part V

Appendix

Network theory A

The appendix contains further information of topics in the study, and is referred to if the reader wants more information regarding a topic.

A.1 Convolutional Neural Network

CNN can be used as a classification tool using automatic feature extraction [Shinde et al., 2019]. CNN is often used in image classification due to its ability to understand spatial and temporal information obtained through training. A CNN contain neurons with learnable weights and biases. [Saha, 2018] The weights and biases are updated during training, a small change in a weight or a bias will cause a small change in the output. A CNN consists of convolutional layers, pooling layers and activation functions. The number of convolutional layers and pooling layers depends on the complexity of the image. [Saha, 2018; Duda et al., 2001] A CNN uses a kernel to perform a convolution operation, the kernel consists of a specific pattern of values. The kernel slides, also called strides, over an image to examine if this pattern is present. The kernel is placed in the top left corner of the image and will slide towards the right corner using the stride. The kernel multiplies the value of the kernel with the image value contained within a pixel or voxel. When the kernel is placed in the right corner and has performed the multiplication it will be moved down and to the left of the image, this will continue until the kernel has performed the convolution of the entire image. The convolution of the image can either down sample the image, keep the same size or increase the size. Valid padding is used to down sample the size of the image, whereas same padding is used to keep the same size or increase the size. [Saha, 2018]

A.2 Batch normalisation

Batch normalisation aspires to improve the training of the network by stabilising the distribution of layer inputs. Therefore, further layers are introduced to the network that control the first two moments, mean and variance. The batch normalisation layers controls the mean and variance, of the distribution of each activation to zero, for mean, and one, for variance, this will achieve the stabilisation of the distribution. [Santurkar et al., 2018] Batch normalisation makes the optimisation problem more smooth ensuring the gradients are more predictive hence allowing the use of larger learning rates and faster network convergence. The optimisation problem seeks to find the best combination of all parameters to maximise the performance of a network [Flovik, 2018]. Properties of batch normalisation

include: robustness to different settings of learning rate to prevent exploding or vanishing gradients. [Santurkar et al., 2018] Valuable impact of batch normalisation on the training process is the reparametrisation of the underlying optimisation problem. [Santurkar et al., 2018]. This will lead to a decrease of the loss, due to the loss changing at a smaller rate as well as the magnitude of the gradients are smaller too. [Santurkar et al., 2018]
Balduzzi et al. [2017] found that networks without batch normalisation indicate to suffer from small correlation between different gradient coordinates, which is reported to be behaviour abstruse in deeper network. Morcos et al. [2018] observed that network with batch normalisation rely less on single directions in the activation space, which they report is connected to the generalisation properties of the network.

A.3 Loss function

A loss function is used in a neural network to compare the actual labels to the predicted values. Therefore, loss is an indication of how good the performance of the network is, a lower loss represent a better performance. The cross entropy loss function was used in the current study, it speeds up the learning of the network. Entropy is a measure describing disorder, a low entropy means a small degree of disorder which makes it easier to separate the two classes. [Chauhan, 2021] Each neuron in the network learns weights and bias when training based on the input to the neuron. The cross entropy will be positive and go towards zero as the neurons train and improve in computing the needed result. A decrease in the loss will occur faster in the beginning of the training due to a higher error, as the error becomes smaller the neurons will not learn as fast. The advantage of using cross entropy as loss function is the faster learning of the network, compared to other loss functions like the quadratic loss function. [Goodfellow et al., 2016; Brownlee, 2019]

A.4 Optimiser

In neural networks optimisers are used in the training process to minimise the loss, and hereby optimise the network's performance. Optimisers are algorithms used to change the weights and learning rate in a neural network. Stochastic gradient descent (SGD) is an optimisation algorithm often used within research and engineering, due to it being an effective optimisation method. Adam was created to produce a more effective stochastic descend method. Compared to SGD, Adam updates the learning rate (α) during training. The learning rate determines how large step the gradient takes trying to find the local minimum. The name Adam comes from adaptive moment estimation, due to the learning rate being updated based on estimates of first and second order moments, mean and variance, of the parameters. Adam uses a moving average to update the gradient, the decay rates of the moving averages are managed by the two parameters: $\beta1$ and $\beta2$.[Kingma and Ba, 2014; Musstafa, 2021]

To examine the performance of Adam Kingma and Ba [2014] tested Adam, SGD and Adagrad. Kingma and Ba [2014] showed that the cost was reduced further and faster when using Adam compared to the other optimisers. The advantages of using Adam is that it is effective in obtaining convergence, it works well with a large amount of parameters, furthermore, it does not require a lot of memory. In addition, the studies by Xiao et al. [2019]; Shinde et al. [2019] used the Adam optimiser and obtained high accuracy when classifying PD using a CNN.

A.5 Data augmentation

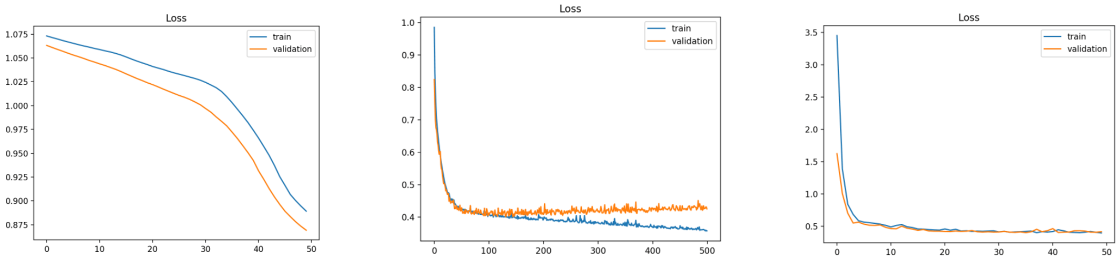
Data augmentation is a method to combat eventual overfitting by superficially increasing the data size used for training. It can be performed in multiple domains, but one way is by geometrically changing the training data, and adding this changed data to the training set. Examples of geometric transformations can be: flipping (the horizontal axis gets flipped), rotation (rotating the image right or left in some degree) or noise injection (add Gaussian noise to the image). [Shorten and Khoshgoftaar, 2019] Data augmentation can be necessary if the amount of data in a study is limited. It differentiates whether the studies obtained from the literature search described in section 2 use data augmentation or not. This also depend on the amount of data the studies required. [Xiao et al., 2019] Shinde et al. [2019] included 80 subjects (35 HCs and 45 PD) in their study and used data augmentation to obtain 10 times the amount of data. The data augmentation used in the study by Shinde et al. [2019] were translations, minor shifting and flipping. Whereas, Shin et al. [2021] included 427 subjects (267 PD and 160 HCs) and did not use data augmentation. In the study by Xiao et al. [2019] 140 subjects (87 PD and 53 HCs) was included, who also chose to use data augmentation to reduce the possibility of overfitting.

A.6 Hyperparameter initialisation and optimisation

Shinde et al. [2019] used a ReLU activation function, and implemented a learning rate of 0.0001, therefore, the same learning rate would be implemented in the network created in the current study. Liu et al. [2020] used an almost similar 3D DenseNet structure to segment and classify AD. They obtained an accuracy of 88.9% classifying AD subjects and HCs. Liu et al. [2020] used the Adam optimiser with a learning rate of 10^{-4}. The number of epochs used was 120, whereas a batch size of 64 was chosen. In the DenseNet four dense blocks and three transitions blocks was initialised. [Liu et al., 2020; Anwaar, 2019]

The size of the convolution layer was set to 3x3x3 convolutions, whereas the average pooling in the transition layer was set to 2x2x2 and the convolutional layers in the dense block was 3x3x3 based on Liu et al. [2020] and Anwaar [2019]. The max pooling was set to 2x2x2. Initialisation of parameters, besides the learning rate, for the optimiser was: $\beta 1$ of 0.9, $\beta 2$ of 0.999 and epsilon of $1e^{-08}$. Epsilon was used as a parameter to prevent any division by zero.

To optimise the hyperparameters the training and test loss would be plotted as a learning curve. The learning curve would show the performance of the network over time [Brownlee, 2019]. The x-axis would show the time illustrated using epochs, whereas the y-axis would represent the learning and show the loss. When examining the learning curve, it was important to check if the network was under-/overfitted or if the fit was suitable. From the learning curve it was clear that a network was underfitted if the training loss decreased over all epochs, see figure A.1a, furthermore, the validation loss was lower than the training loss, which also implies underfitting. In this case the network can obtain lower loss with more training. In a learning curve of an overfitted network the training loss will keep decreasing, whereas the validation loss will start to increase at some point, due to the models lack of ability to generalise. An overfitted network can be seen in figure A.1b. When the network have a good fit, the training and validation loss will decrease and stabilise at almost same point, furthermore, the space between the two graphs will be small, this can be seen in figure A.1c. [Brownlee, 2019]



(a) Underfitted network. ***(b)*** Overfitted network. ***(c)*** Network with a good fit.

Figure A.1. Learning curves representing an under fitted, overfitted and a good fitted network. [Brownlee, 2019]

A.7 K-fold cross validation

K-fold cross validation is a method often used on small sample sizes to combat eventual overfitting. In the k-fold cross validation, the entire data set is split in k different groups. One group is used as testing data and the remaining as training data, see figure A.2. K different iterations of which fold was used as test were made. This gives k differently trained models, which are compared. [Refaeilzadeh et al., 2009]

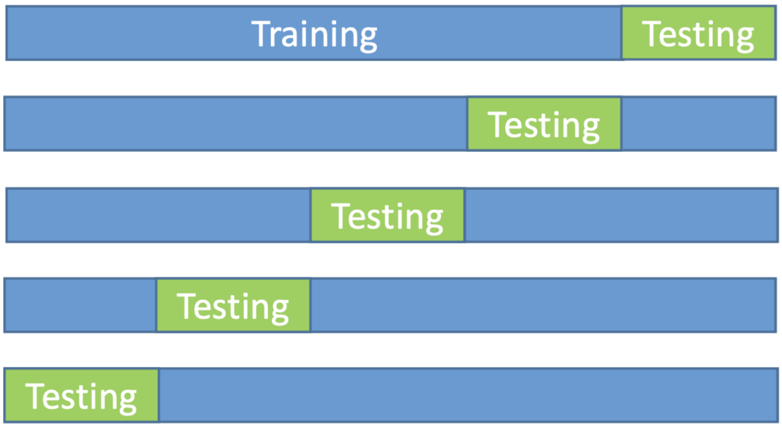


Figure A.2. Data set split into training and test data using k-fold cross validation. [Larsen, 2021]