

Deep Learning and Automated MRI Analysis in Idiopathic Normal Pressure Hydrocephalus

**Methodological developments for outcome
prediction and quantitative DESH assessment**

Klara Mogensen



UMEÅ UNIVERSITY

This work is protected by the Swedish Copyright Act (Act 1960:729)

ISBN: 978-91-8070-969-9 (print)

ISBN: 978-91-8070-970-5 (pdf)

ISSN: 0346-6612

Umeå University Medical Dissertations New Series no. 2419

The cover is created by Britta Söderkvist in collaboration with Lovis Heggenberger, Tuva Heggenberger, Stina Thorsén, Johanna Thorsén, Jonas Mogensen and Karin Bengtsson

Electronic version available at: <http://umu.diva-portal.org/>

Printed by: Scandinavian Print Group, Skarpnäck, 2026

Till Lill, Sten, Bengt och Lena

Contents

Abstract.....	iii
Populärvetenskaplig sammanfattning på svenska	v
List of publications.....	ix
List of abbreviations.....	xi
1 Introduction	1
2 Background.....	3
2.1 The cerebrospinal fluid system.....	3
2.2 Idiopathic normal pressure hydrocephalus	4
2.2.1 <i>Higher-level gait disorder</i>	5
2.2.2 <i>Guidelines</i>	5
2.2.3 <i>Radiological markers</i>	6
2.2.4 <i>qDESH</i>	7
2.3 Magnetic resonance imaging.....	8
2.4 Medical image analysis	9
2.5 Machine learning	11
2.5.1 <i>Neural networks</i>	11
2.5.2 <i>Deep learning and convolutional neural networks</i>	11
2.5.3 <i>Model training and evaluation</i>	12
2.5.4 <i>Transfer learning</i>	13
2.5.5 <i>Model generalization</i>	14
2.5.6 <i>Ensemble learning</i>	14
2.5.7 <i>Multimodal learning</i>	15
3 Research Rationale	17
4 Aims.....	19
5 Materials and Methods.....	21
5.1 Ethical approval	21
5.2 Cohorts	22
5.3 MRI data.....	24
5.3.1 <i>Spatial preprocessing</i>	25
5.3.2 <i>Tissue extraction and segmentation</i>	25
5.3.3 <i>Intensity normalization</i>	26
5.3.4 <i>Radiomics and machine learning models</i>	26

5.4	Neural networks	27
5.5	Ensemble search	29
5.6	Sequential forward search	31
5.7	Auto-qDESH pipeline	32
5.8	Training of networks and algorithms	33
5.8.1	<i>Competitors and baselines</i>	34
5.9	Statistics and validation	35
6	Results	37
6.1	HLGD classification	37
6.1.1	<i>Ensemble search</i>	37
6.2	Multimodal fusion.....	39
6.3	Shunt outcome prediction	40
6.4	Auto-qDESH	42
7	Discussion.....	45
7.1	Assessing brain changes with AI.....	45
7.1.1	<i>HLGD</i>	45
7.1.2	<i>Shunt outcome prediction in INPH</i>	46
7.2	Auto-qDESH	48
7.3	Development of search algorithms	49
7.3.1	<i>Ensemble search</i>	49
7.3.2	<i>Sequential Fusion Search Algorithm</i>	50
7.4	Methodological considerations.....	52
7.5	Future outlook.....	53
8	Conclusions	55
9	Acknowledgements	56
10	Funding.....	58
11	References	59

Abstract

Idiopathic normal pressure hydrocephalus (INPH) is a neurological disorder characterized by impaired gait and balance, cognitive decline and incontinence, in combination with enlarged lateral ventricles. Although symptoms can often be alleviated through surgical insertion of a cerebrospinal fluid (CSF) shunt, a substantial proportion of patients do not improve after surgery. There is therefore a need for new analytical methods that can extract more informative features from MRI to improve diagnostic and prognostic accuracy.

This thesis consists of the work from four papers with the overall aim to develop and assess artificial intelligence (AI)-based and fully automated MRI-based methods, to improve objective assessment and shunt decision support in INPH.

Several well-known convolutional neural networks (CNNs) were applied to 3D brain magnetic resonance imaging (MRI) data to distinguish between participants with an INPH-typical gait pattern and controls. An ensemble model search was developed to find the optimal ensemble for the task at hand, by optimizing combinations of diverse models. A fusion search strategy was also developed, to determine the optimal fusion points for information fusion between different MRI sequences. Shunt outcome prediction was evaluated with both deep learning approaches, using the two search algorithms, as well as with radiomics-based machine learning models. Finally, a fully automated pipeline was developed for assessment of disproportionally enlarged subarachnoid space hydrocephalus (DESH), utilising image segmentation and image analysis techniques to determine a quantitative DESH metric (qDESH). The work was conducted on brain MRI from one population-based cohort (Paper I), two open access datasets (Paper II), a clinical cohort of shunted INPH patients (Paper III) and a retrospective cohort of INPH patients and controls (Paper IV).

All CNNs distinguished between gait-impaired and controls, in terms of a chi-square test of independence. The optimized ensemble model achieved the highest classification performance, exceeding that of the individual networks and conventional radiological measures. The results support the presence of detectable structural differences in brain MRI between the groups. The sequential search of multimodal fusion points improved classification performance compared with unimodal and

conventional fusion strategies, while reducing computational cost. However, when applying these methodologies to predict shunt outcome, no model achieved clinically sufficient performance. These findings indicate that structural MRI alone is not yet reliable for shunt prediction in INPH. The fully automated qDESH pipeline demonstrated high agreement with the established semi-automatic qDESH method, although the agreement was lower than between two raters of the semi-automatic method. The automated measure of qDESH aligned well with visual assessment of DESH.

In conclusion, this thesis advances methodologies for AI-based and automated brain MRI analysis, particularly for INPH. Introducing and evaluating an optimized ensemble strategy, a systematic multimodal fusion approach, and a fully automated quantitative imaging pipeline, the work demonstrates both the potential and the current limitations of advanced and automated MRI analysis in INPH. The fully automated qDESH pipeline showed good agreement with both the semi-automatic method and visual DESH ratings, although further refinement is required before it can be applied in clinical practice. While CNNs can capture differences in brain MRI beyond conventional linear measures, they cannot yet predict shunt response in a clinically useful way. Structural MRI data alone might be insufficient, and additional non-imaging data might be required. The findings highlight the importance of diversity across models and imaging sequences to improve data-driven image assessment. The need for large clinical datasets is a limiting factor, making collaboration among multiple centres necessary to enable further methodological developments. The methodological approaches and insights presented here may also be transferable to other neurological disorders in which MRI plays a central diagnostic role, thereby contributing more broadly to the neuroimaging field.

Populärvetenskaplig sammanfattning på svenska

Idiopatisk normaltryckshydrocefalus (INPH) är en sjukdom som drabbar äldre personer, och blir vanligare ju äldre man blir. Symtomen påminner mycket om vanliga tecken på åldrande; man får sämre gång och balans, problem med minnet och inkontinensbesvär. Men det finns behandling som kan lindra symptomen. Behandlingen består av att en shunt sätts in, en slang som leder bort vätska från hjärnan. Omkring 50–90 procent av patienterna som opereras blir bättre i sina symptom, men man vet inte varför behandlingen inte fungerar för alla. Vi behöver förstå sjukdomen bättre, för att bättre välja ut vilka som ska få en shuntoperation.

Vid en magnetkameraundersökning av huvudet kan man se att hjärnan hos INPH-patienter har fått en annan form. Framför allt är det vätskeutrymmena i och kring hjärnan som förändras. Det viktigaste tecknet är att det stora hålrummet mitt i hjärnan är förstorat. Ofta ser man också att vätskeutrymmena ovanför hjärnan är mindre, samtidigt som det har blivit mer plats på hjärnans sidor. Detta mönster kallas DESH. Det har utvecklats många testmetoder, både för att ställa diagnos och för att uppskatta om patienten kommer att bli bättre av en shuntoperation, som bygger på mätningar av hjärnan och vätskeutrymmena. Dessa tester tar tid att utföra och bygger på att läkaren har erfarenhet eftersom flera av testerna är subjektiva. Detta gör det svårare att jämföra patientgrupper mellan olika sjukhus. Det är inte heller klarlagt hur hjärnans förändrade form hänger ihop med symptomen, eller förväntad förbättring av en operation.

Artificiell intelligens (AI) används i allt större utsträckning för att analysera data. För medicinska bilder används ofta faltningsnätverk, som är en typ av djupa neurala nätverk. Genom att testa om flera, välkända faltningsnätverk kunde skilja mellan hjärnor hos personer med typiska INPH-symtom, i det här fallet gångstörning, och personer utan gångstörning, kunde vi undersöka om hjärnförändringarna hänger ihop med symptomen. Med den här metoden kunde vi också undersöka om hjärnans förändrade form kunde säga något om vilka som blir bättre av en shuntoperation, med målet att förbättra urvalet av patienter till operation. När vi undersökte detta använde vi också tabeller med olika

kvantitativa mått som beräknats från bilderna (radiomics), och tränade enklare maskininlärningsmodeller på samma uppgift.

Vi använde de tränade falttningsnätverken, inte bara var och ett för sig, utan också tillsammans i en ensemble. För att förhoppningsvis förbättra resultaten från de enskilda nätverken skapade vi en algoritm som sökte efter den bästa kombinationen av nätverk. Algoritmen optimerades för kombinationer av nätverk som gjorde olika fel, men samtidigt hade bra resultat på uppgiften.

Vi utvecklade också en algoritm som optimerade när man ska väga samman information från olika bilder, om man använder mer än en bild från varje patient i sitt nätverk. Genom en stegvis sökning analyserade algoritmen hur djupt i nätverket det är optimalt att blanda information.

För att bättre kunna jämföra patienter mellan varandra är det bra om måtten som tas på hjärnorna är kvantifierbara. qDESH är ett kvantitativt mått på DESH, som kan beräknas genom en halvautomatisk programvara. Vi utvecklade ett program som helt automatiserar beräkningen av qDESH, för att spara tid för läkarna och underlätta spridningen av detta mått.

Våra resultat visade att falttningsnätverken kunde skilja mellan hjärnorna hos personer med och utan gångstörning. Bäst resultat fick ensemblen som skapades av sökalgoritmen. Det här kan betyda att det finns flera olika anledningar till gångstörningen som syns i hjärnan, och att olika nätverk hittar olika strukturer för att förklara detta. Resultaten visar att det finns något att undersöka vidare, för att förstå vad nätverken baserar sin uppdelning på.

Algoritmen som stegvis sökte efter bra punkter att blanda olika bilder hittade en modell som presterade bättre än alla andra jämförda modeller, och framför allt bättre än om man bara använder en enda bildtyp. Genom att använda den här algoritmen kan man få hjälp med när man ska blanda information, och spara mycket beräkningskraft eftersom man inte behöver testa alla möjliga varianter.

När vi undersökte hjärnbilderna för att hitta de patienter som förbättrades av shuntoperation, hittade visserligen några nätverk en skillnad, men den var inte så tydlig att resultaten kan förbättra arbetsgången på sjukhusen idag. Det här kan bero på att det dataset vi använde var för litet för att modellen skulle lära sig att hitta skillnader, men det kan också betyda att det inte finns tillräcklig information i

bilderna för att helt förklara om en patient kommer förbättras av en shuntoperation. Det händer troligen mycket i hjärnan som inte syns på bilderna. En fortsättning på denna studie skulle kunna vara att ta med annan typ av data när man tränar nätverken, till exempel patientens ålder, hur länge patienten har haft symptom, vilka symptom det gäller, och vilket tryck det är i hjärnan.

Den automatiska qDESH-beräkningen stämde bra överens med den halvautomatiska, men programmet hade svårigheter att avgöra vad som var ben och vad som var vätska. Detta gjorde att de två metoderna inte hade lika överensstämmande resultat som när två olika personer använde den halvautomatiska metoden, framför allt när värdena var höga. När den automatiska metodens resultat jämfördes med två läkares manuella bedömning av DESH stämde det överens bra. I dagsläget kan programmet användas i forskningssyfte när man främst tittar på grupper av patienter, men det behöver bli bättre för att kunna användas på sjukhus.

Den här avhandlingen visar olika sätt att använda AI och bildbehandling för att undersöka magnetkamerabilder, framför allt vid INPH. Vi har visat att det finns en koppling mellan hjärnförändringar och den gångstörning som är vanlig hos patienter med INPH, men att AI idag inte kan förutse vilka som förbättras av en shuntoperation, enbart genom att analysera hjärnbilder. Det finns många fördelar med att ha kvantitativa mått på hjärnförändringarna, och genom att utveckla en helt automatisk metod underlättar vi spridningen av qDESH och minskar det manuella arbetet. För att med AI bättre kunna analysera INPH och vilka som förbättras av en shuntoperation behövs större, standardiserade dataset med mer information än bara bilder. Den här typen av modeller skulle då kunna bli ett tidseffektivt och objektiva verktyg för läkare som arbetar med äldre patienter.

List of publications

This thesis is based on the work presented in the following papers and manuscripts.

Paper I:

Klara Mogensen*, Valerio Guarrasi*, Jenny Larsson, William Hansson, Anders Wåhlin, Lars-Owe Koskinen, Jan Malm, Anders Eklund, Paolo Soda, Sara Qvarlander. An optimized ensemble search approach for classification of higher-level gait disorder using brain magnetic resonance images. *Computers in Biology and Medicine*, 2025; 184: 109457 **

Paper II:

Valerio Guarrasi, **Klara Mogensen**, Sara Tassinari, Sara Qvarlander, Paolo Soda. Timing Is Everything: Finding the Optimal Fusion Points in Multimodal Medical Imaging. *Proceedings of 2025 International Joint Conference on Neural Networks*, 2025 **

Paper III:

Klara Mogensen, Valerio Guarrasi, Sofia Behndig, Johan Eriksson De Ryst, Paolo Soda, Anders Eklund, Jan Malm, Sara Qvarlander. Can AI applied on MRI reliably predict shunt response in INPH? A comprehensive exploration of deep learning and radiomics approaches using preoperative MRI. *In manuscript*.

Paper IV:

Klara Mogensen, Sofia Behndig, Afroditi Lalou, Anders Wåhlin, Anders Eklund, Jan Malm, Sara Qvarlander. Automated computation of a quantitative DESH score in Brain MRI for reproducible radiological assessment of hydrocephalus. *In manuscript*.

* The authors contributed equally to this work.

** Reprinted with permission.

List of abbreviations

AC	Anterior commissure
AD	Alzheimer's disease
AI	Artificial intelligence
AUROC	Area under the receiver operating characteristic curve
B-accuracy	Balanced accuracy
CI	Confidence interval
CNN	Convolutional neural network
CSF	Cerebrospinal fluid
CT	Computed tomography
DESH	Disproportionately enlarged subarachnoid-space hydrocephalus
DL	Deep learning
DP	Decision profile
EI	Evans' index
FLAIR	Fluid-attenuated inversion recovery
F-score	F1-score
HLGD	Higher-level gait disorder
ICC	Intraclass correlation coefficient
INPH	Idiopathic normal pressure hydrocephalus
ML	Machine learning
MMTM	Multimodal transfer module
MNI	Montreal Neurological Institute
MRI	Magnetic resonance imaging
PC	Posterior commissure
qDESH	Quantitative DESH
SEM	Standard error of the mean
SFSA	Sequential Fusion Search Algorithm
SD	Standard deviation
T1-w	T1-weighted
T2-w	T2-weighted
XAI	Explainable artificial intelligence

1 Introduction

Gait disturbance, cognitive decline, and urinary incontinence in older adults are often dismissed as normal consequences of ageing. Yet for some individuals, these symptoms arise from the potentially reversible disorder idiopathic normal pressure hydrocephalus (INPH). Many INPH patients experience symptom relief by insertion of a ventriculoperitoneal shunt [1,2]. INPH is therefore sometimes described as a treatable form of dementia. INPH has an estimated prevalence of 1.5% in the population above 65 years old, increasing to 7.7 % among those over 80 [3–5]. With an ageing population, the number of patients requiring accurate diagnosis and treatment is thus expected to rise. However, electing patients for surgery remains challenging, with improvement rates around 50-90% [6]. The current radiological and physiological assessments appear to capture only part of the underlying pathology, motivating the development of more objective and data-driven approaches. Artificial intelligence (AI) methods may be particularly suitable for this purpose, as they can identify complex patterns in imaging data that may not be captured by conventional visual or linear measurements [7].

As the name implies, the cause of INPH remains unknown, and the intracranial pressure is typically near normal or slightly increased [8]. A characteristic finding in cerebrospinal fluid (CSF) dynamics, however, is an increased resistance to CSF outflow [9]. INPH patients present with enlarged lateral ventricles together with a characteristic triad of symptoms: impaired gait and balance, cognitive decline, and urinary incontinence [10,11]. Since the pathophysiology of INPH is not fully understood, the diagnosis and selection for shunt surgery relies on a combination of clinical and physiological assessments together with neuroimaging assessments, often based on magnetic resonance imaging (MRI), to rule out overlapping differential diagnoses and estimate the prognosis for improvement from shunting [10,11]. However, there is yet no combination of assessments that provides optimal results in this regard [12–14].

Radiological assessments for INPH involve measures that reflect altered CSF distribution, and several rely on subjective interpretation. Established quantitative metrics require expertise and manual effort [10,11], which limits scalability and comparability across centres. Brain MRI provides information beyond what these measures can capture, and

more than one imaging sequence is often acquired per patient. This motivates the development of more advanced methods that can uncover imaging patterns beyond traditional assessments, and automated pipelines that can standardize quantitative measures to enable more objective and scalable tools for evaluation.

AI offer opportunities to analyse the MR images in new ways. Neural networks can process large numbers of medical images together with clinical outcomes, such as symptoms or level of improvement, and may detect patterns or relationships beyond current radiological knowledge [15]. Automated and quantitative methods also have potential to reduce subjectivity and manual workload for the clinician, increasing reproducibility and facilitating analyses across larger and more diverse cohorts.

This thesis explores how AI methods can enhance our understanding and assessment of INPH. It investigates whether brain imaging features are related to the cardinal feature of the disorder, i.e., the gait disorder, and to improvement after shunt surgery. To leverage complementary information across networks and imaging sequences, this thesis aimed to develop and evaluate ensemble-based and multimodal neural network models. In addition, it aimed to establish a fully automated pipeline for obtaining a quantitative measure of disproportionately enlarged subarachnoid space hydrocephalus (DESH), which is used in INPH investigations to assess how the CSF is distributed around the brain. This pipeline is based on an established manual method but removes the need for subjective interpretation and manual labour, enabling more consistent measurement and potentially supporting future clinical and research applications. Together, these approaches aimed to improve both the clinical work-up of INPH and our overall understanding of this complex condition.

2 Background

This chapter provides a background to the research presented in this thesis. It begins with a brief explanation of the cerebrospinal fluid (CSF) system, which is important in patients with INPH. INPH is then described, including the characteristic gait disturbance often seen in these patients and how patients are radiologically assessed today, as well as a newly developed quantitative radiological measure. The chapter then continues with an introduction to MRI, including how images can be pre-processed and analysed, and concludes with an overview of machine learning and deep learning, as well as how such models are trained and evaluated.

2.1 The cerebrospinal fluid system

CSF surrounds the brain and spinal cord. It consists mainly of water and functions as a shock absorber, reduces the effective weight of the brain, and helps maintain a stable extracellular environment. In the classical view, most CSF is produced by the choroid plexus, which is located in the ventricles, see Figure 1. From there, it circulates through the ventricular system and into the subarachnoid space around the brain and spinal cord before being reabsorbed into the bloodstream via arachnoid granulations in the dural venous sinuses [16].

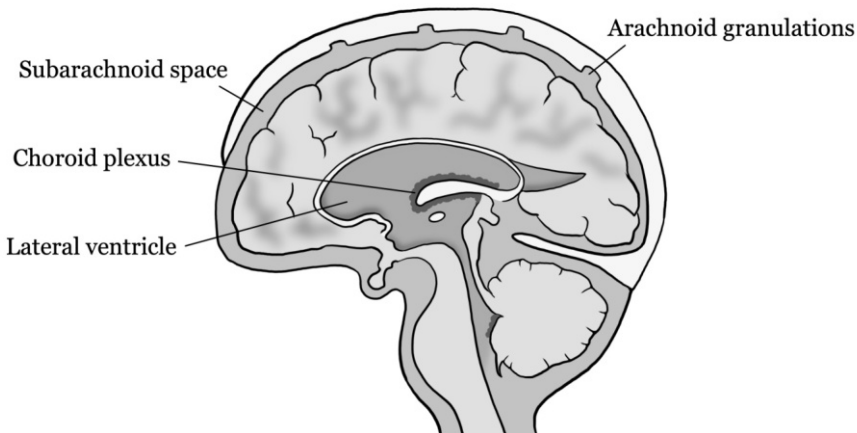


Illustration by Britta Söderkvist, 2026.

Figure 1, Schematic of the brain and cerebrospinal system.

More recent research has expanded this view by describing the glymphatic system, which acts as a waste-clearance pathway in the brain [17]. In this theory, the CSF flows not only in the subarachnoid space, but also enters the brain parenchyma, facilitating the removal of metabolic waste products such as amyloid- β . In addition to production in choroid plexus, a substantial portion of CSF is exchanged with interstitial fluid along perivascular spaces, and clearance occurs partly through perivenous outflow pathways that connect to cervical lymph nodes or recirculate back to the subarachnoid space.

CSF dynamics refers to the processes regulating CSF production, flow, and absorption within the brain and spinal canal. The total intracranial CSF volume in adults is approximately 300 mL [18], and the daily CSF production is around 500 mL [19,20], meaning the CSF volume is renewed every day. CSF movement is influenced primarily by cardiac and respiratory pulsations [21], and the glymphatic clearance appears to be more active during sleep [22]. Normal mean intracranial pressure ranges from about 7.5 to 15 mmHg [23] and reflects the balance between CSF production, circulation, and absorption.

2.2 Idiopathic normal pressure hydrocephalus

Of disorders involving altered CSF dynamics, INPH is one of the most common [5]. It affects older adults, and becomes more common with age, with a reported prevalence of 1.5 % in individuals aged 65 years or older and 7.7% among those older than 80 years [3–5]. The classic symptom triad consists of impaired balance and gait, sometimes combined with cognitive impairment and/or incontinence. The insidious onset and overlap with other diseases can complicate diagnosis. Many patients experience symptom improvement after shunt surgery, with gait often showing the most pronounced response [2]. Cognition and incontinence may also improve, although typically to a lesser extent [24].

A major clinical challenge is, apart from identifying which patients have INPH, determining which patients are likely to benefit from shunt treatment. Because INPH is a diagnosis that requires exclusion of other conditions, and comorbidities are common in this patient group, determining suitability for shunt surgery can be difficult [25]. Reported shunt improvement rates vary widely, typically ranging from approximately 50% to 90% depending on study design and patient selection [1,2,6,26]. However, shunt surgery is associated with a risk of complications. Among patients undergoing shunt surgery for INPH,

approximately 10% develop subdural hematomas [27]. A number of clinical tests and radiological assessments are therefore used to help identify patients who are most likely to benefit from surgery [28]. Nevertheless, many of the commonly used assessments have the limitation that patients with radiologically or physiologically normal findings may still improve after surgery [29,30]. This uncertainty highlights an important gap in current clinical practice and underscores the need for alternative methods for evaluating INPH.

2.2.1 Higher-level gait disorder

Gait disturbance is the most characteristic feature in INPH and is often the first component of the clinical triad to appear [31]. It is also the symptom that most frequently improves after shunt surgery [2]. The gait pattern in INPH is typically symmetrical and broad-based, with reduced step height and step length. It is often described as shuffling or magnetic, and freezing may occur [11].

This gait pattern reflects dysfunction in the higher-level nervous system and is therefore categorized as a higher-level gait disorder (HLGD) [32]. HLGD is defined as a gait impairment that cannot be explained by dysfunction in lower- or mid-level motor systems, such as peripheral motor or sensory pathways, pyramidal tracts, in the cerebellum, or the basal ganglia. Although characteristic in INPH, HLGD can also occur in other conditions, including cerebrovascular disease and neurodegenerative disorders [33].

By investigating the relationship between gait disturbance and brain structure, a deeper understanding of the underlying mechanisms in INPH may be achieved, and potentially more accurate tools for patient assessment can be developed.

2.2.2 Guidelines

Two sets of international guidelines, which are currently being revised, outline current recommendations for the diagnostic evaluation of INPH; the American-European guidelines [11], and the Japanese guidelines [10]. Both incorporate clinical findings and radiological markers, although they emphasize certain aspects differently, with the Japanese guidelines generally applying stricter criteria [3]. Both guidelines distinguish between probable INPH and possible INPH, the latter representing a less certain diagnosis

2.2.3 Radiological markers

The most characteristic radiological finding associated with INPH is enlarged lateral ventricles, which is required before the diagnosis can be considered [11]. This is commonly determined using Evans' Index (EI) [34], defined as the ratio between the maximum width of the frontal horns and the maximum intracranial width on the same axial slice, depicted in Figure 2. An EI greater than 0.30 is often used as a threshold suggesting ventriculomegaly. A related measure is the z-EI, which uses the ratio between the height of the frontal horns and the height of the intracranial space above the frontal horns on the same coronal slice instead. A z-EI above 0.42 has been proposed as a complementary cutoff value in the evaluation of INPH [35]. Another way to assess ventriculomegaly is to measure the ventricular volume, where the limit for enlarged ventricles has been determined to >77 mL [36]. However, ventricular enlargement can also occur in cerebral atrophy, which is one of the reasons why additional radiological markers are used to differentiate INPH from other conditions.

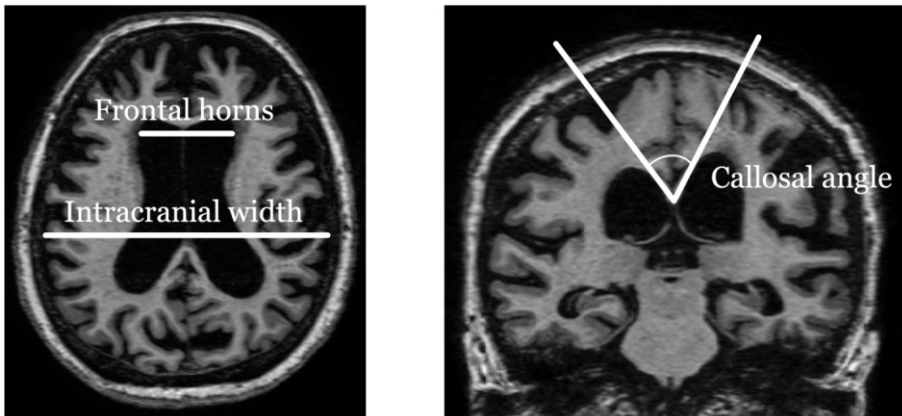


Figure 2, Two important radiological measures for INPH. The Evans' index is computed as the ratio of the width of the frontal horns and the intracranial width, depicted to the left. The callosal angle is presented to the right.

One commonly used measure is the callosal angle [37], defined as the angle between the medial walls of the lateral ventricles on a coronal image through the posterior commissure, see Figure 2. In INPH the angle is often below 90° , and this can assist in differentiating INPH from disorders such as Alzheimer's disease [37]. Another key feature, especially in the Japanese guidelines, is disproportionately enlarged subarachnoid-space hydrocephalus (DESH) [13], which describes

ventriculomegaly with the combination of enlarged Sylvian fissures and tight subarachnoid spaces over the high convexities. Enlargement of the temporal horns of the lateral ventricles is also considered in the American-European guidelines as an additional supportive marker. To support more standardized radiological assessment, the INPH Radscale [38] was introduced in 2017, integrating imaging features described in both guidelines into a structured diagnostic scoring tool.

In addition to these structural measures, several physiological imaging markers have been described. Periventricular signal changes on MRI may reflect altered water content in the adjacent brain tissue and have been proposed as a marker in both diagnostic evaluation and shunt selection [11,30], although such changes alone are nonspecific and are also common in small vessel disease and cerebral atrophy. Perfusion methods such as single-photon emission computed tomography (SPECT) or perfusion MRI have shown regional blood-flow alterations in some INPH cohorts, but these findings have limited diagnostic value and are not recommended for routine assessment [10]. CSF flow-related MRI features, including aqueductal flow voids and altered CSF stroke volumes have also been investigated, but the available evidence does not support their use as predictors of treatment response [39,40].

Overall, the radiological structural measures of INPH rely mainly on a few predefined landmarks that describe the relationship between the brain and the CSF spaces, while physiological imaging techniques provide additional but currently limited clinical insight. Much of the information contained in the brain images likely remains unexplored by these conventional radiological measures, and the interpretation often depends on the experience of the rater. This motivates the exploration of other approaches, such as AI-based image analysis, which may reveal additional patterns or shapes associated with INPH and improve diagnostic evaluation.

2.2.4 qDESH

qDESH is a quantitative method to assess DESH [41]. In a brain volume aligned to the anterior commissure-posterior commissure (AC-PC) line and the midsagittal plane, qDESH is the ratio between the combined CSF volumes of the two Sylvian fissures and the CSF volume at the high convexity. Both measures are evaluated at the level of the PC and 20 mm anteriorly. The high-convexity CSF volume is restricted to the uppermost 30 ml of this search volume within the dura. For the Sylvian fissures, CSF volumes lateral to the cerebellar border are constrained to lie within

10 mm of the Sylvian fissure midline, whereas no distance restriction is applied medially.

qDESH is implemented as a semi-automatic tool that requires manual input to define anatomical regions of interest. While this approach improves objectivity compared to purely visual assessment, the need for manual interaction limits scalability and clinical utility, motivating the development of a fully automated method.

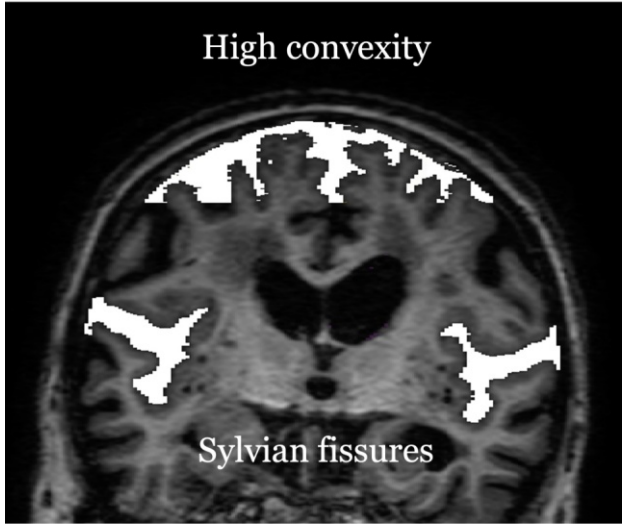


Figure 3, qDESH. The three regions that compose the qDESH metric, computed in the semi-automatic software.

2.3 Magnetic resonance imaging

Both computed tomography (CT) and MRI are used in the evaluation of INPH. CT is often used as an initial screening tool to detect ventricular enlargement, whereas MRI is typically preferred as part of the diagnostic work-up [10]. MRI provides superior soft tissue contrast and a wider range of imaging sequences that can highlight structural and physiological brain features.

MRI covers a broad range of sequences which can highlight different features of the imaged tissue. The technique works by measuring the behaviour of hydrogen nuclear spins. When placed in the strong static magnetic field of an MR scanner, the magnetic moments associated with hydrogen nuclei spins partially align with the main magnetic field,

producing a net longitudinal magnetization. Individual magnetic moments precess around the magnetic field at the so-called Larmor frequency, but their phases are randomly distributed, resulting in no net transverse magnetization. An applied radiofrequency pulse tips the magnetization away from the longitudinal axis and brings the spins into phase, creating a transverse component of the net magnetization that precesses coherently and induces a voltage in the receiver coils. After the radiofrequency pulse, the signal decays as the spins lose phase coherence (T₂-relaxation) and as the longitudinal magnetization recovers back towards equilibrium (T₁-relaxation). Different tissues have different T₁ and T₂ relaxation times, which produce the contrast seen in MRI images depending on the sequence used.

To determine where the MR signal originates in the body, the MR scanner applies gradient magnetic fields in addition to the strong static field. These gradients introduce small, controlled variations in the magnetic field strength across space, causing the Larmor frequency and phase of the nuclear spins to differ depending on their location. By applying gradients in three orthogonal directions, three-dimensional images can be reconstructed.

T₁-weighted (T₁-w) images are one of the most common sequence types for brain imaging. These show white matter as bright, grey matter as intermediate, and CSF as dark, which makes them suitable for structural assessments. T₂-weighted (T₂-w) images are another very common sequence type, where CSF appears bright, while white matter is darker than grey matter. T₂-w sequences are useful for detecting edema or inflammatory changes as regions with increased water content appear bright on these images. A commonly used variant of the T₂-w sequence is T₂ FLAIR (fluid-attenuated inversion recovery), which suppresses the CSF signal while maintaining the T₂-w contrast in the brain tissue. This makes periventricular white matter hyperintensities more easily distinguishable which is particularly relevant in the assessment of INPH [42].

2.4 Medical image analysis

In clinical practice, brain images are typically assessed by a neuroradiologist, who visually inspects the images for previously described radiological signs in order to support a diagnosis and exclude other potential causes of the patient's symptoms. While this approach benefits from expert clinical judgement and contextual interpretation, it

is resource-intensive and subject to observer variability. In a research context, medical image analysis aims to extract meaningful information from imaging data, often in a reproducible and quantitative manner. To ensure that images from different MR scanners, sequences, and subjects can be compared, several preprocessing steps are often required.

Standard structural MRI does not use an absolute intensity scale, so identical tissues can appear with different voxel intensities depending on scanner hardware, sequence type, and acquisition parameters. Pixel intensity normalization can be applied to bring the images to a common range. Intensity variations can also occur within a single MRI volume, due to receiver coil sensitivity, why bias-field correction is often applied.

For many applications, it is desirable that the brain image has the same orientation and coordinate system across subjects. For this, neuroimaging reference spaces exist. A widely used example is the Montreal Neurological Institute (MNI) space [43], which provides population-derived template brains and coordinate systems. Aligning images to such a template enables group-level comparisons and supports the use of predefined anatomical labels.

A central task in medical image analysis is the segmentation of anatomical structures, i.e., identifying and labelling tissues or regions of interest such as ventricles and cortical areas. Although manual segmentation is considered a reference standard, it is time-consuming and requires anatomical expertise. These limitations restrict the number of images that can be reliably processed and may introduce subjectivity into quantitative measurements. To overcome these challenges, various automated and semi-automated software packages have been developed. Tools such as FreeSurfer [44] and SPM12 [45,46] use predefined anatomical probabilistic atlases, statistical models and intensity-based methods to perform tasks including skull stripping, tissue classification, and cortical parcellation for brain regions. These pipelines are widely adopted and are reproducible, enabling large-scale studies that would not be feasible with manual segmentations alone. Automated segmentation also facilitates the use of established labels for specific regions, which support consistent reporting across studies. More recent approaches extend these ideas using machine learning or deep learning to further reduce manual intervention and potentially capture more complex patterns in the data [47,48].

2.5 Machine learning

In AI, machine learning (ML) refers to a class of algorithms that learn patterns from data and use this knowledge to make predictions or decisions on unseen data. ML is therefore a broad term, encompassing methods ranging from linear regression and decision trees to more computationally demanding models such as large neural network models [49]. ML models can be trained using different learning paradigms, most commonly supervised and unsupervised learning.

This thesis focuses on supervised classification tasks. In supervised learning, the model is provided with input data together with corresponding labels, such as a disease diagnosis or a segmented image. The objective is to learn a function that maps inputs to outputs by minimizing a predefined loss function, and typical tasks include classification and regression.

In medical image analysis, many clinical tasks can be formulated using labelled data, such as diagnoses, treatment outcomes, or expert-defined segmentations, although the need for expert annotation makes data collection time-consuming and costly. As a result, supervised learning is central to many medical AI applications, despite growing interest in semi-supervised and self-supervised methods [50].

2.5.1 Neural networks

Neural networks are a class of ML models consisting of interconnected computational units, often referred to as neurons. The neurons are organized in layers, with an input layer, one or more hidden layers, and an output layer. Each connection between two neurons also has a weight, which determines how strongly the signal from one neuron influences the next and is adjusted during training to minimize the loss function. Deep learning (DL) is a subset of ML, characterized by neural networks with multiple hidden layers. In this thesis, however, the term *machine learning* will be used to refer to traditional, non-deep learning methods, while *deep learning* refers specifically to models based on deep neural networks.

2.5.2 Deep learning and convolutional neural networks

DL models use stacked computational layers that create increasingly abstract representations of the data. Different layer types perform distinct operations within the network, such as feature extraction or improving model generalization. Although the internal representations

within these models are often difficult to interpret, their ability to model complex, non-linear relationships has led to widespread adoption, in tasks such as image recognition, speech recognition, and text classification [51].

For image analysis, a widely used type of deep neural network is the convolutional neural network (CNN) [52]. CNNs take advantage of the spatial structure of images, where neighbouring pixels tend to contain related information, and meaningful patterns often appear as clusters rather than isolated pixels. By applying convolutional filters that scan across the image, CNNs can automatically learn features such as edges, textures, and shapes, and combine them into more complex representations. Convolutional layers are computationally efficient because they apply small kernels whose weights are shared across the image, instead of using one weight per input unit as in fully connected neural networks. These shared kernel weights are updated during training, meaning that the number of learnable parameters does not increase with the size of the input image. Stacking multiple convolutional layers allows the network to learn increasingly more complex patterns in the data.

2.5.3 Model training and evaluation

During network training, it is essential to reserve part of the data for model evaluation. A common approach is to divide the dataset into a training set, a validation set, and a test set, which is used only once after training to provide an unbiased estimate of the final model performance. Ideally, model performance should also be evaluated on an external dataset acquired independently of the training process, as this provides a stronger assessment of generalizability and reduces the risk of dataset-specific bias.

Neural networks learn by iteratively minimizing a loss function that corresponds to the difference between predictions and the expected labels or values. The data are typically processed in mini-batches over multiple epochs, where one epoch corresponds to training one on all mini-batches in the training set. Model performance is monitored throughout training, often using the validation set, to guide decisions such as hyperparameter tuning or early stopping. Hyperparameters are model settings defined before training, such as the learning rate or batch size. Early stopping ends training when validation performance no longer improves, helping to prevent the model from adapting too closely to the training data.

For smaller datasets, the split into training, validation, and test sets is especially critical, as individual subjects can have a substantial influence on the data distribution. To obtain more reliable performance estimates under such conditions, cross-validation is often employed. In cross-validation, the model is trained multiple times using different data splits, and performance is averaged across these runs. When the splits are constructed so that each fold preserves the class distribution of the full dataset, the procedure is referred to as stratified cross-validation.

When using cross-validation, care must be taken to avoid data leakage, meaning that the test set is no longer independent from the training set. Leakage can occur if parts of the preprocessing are derived from the distribution of the entire dataset rather than being computed using only the training data within each split. In datasets that include multiple images from the same individual, it is also essential to ensure that all images from a given subject are kept within the same split.

After training is completed, the model performance is typically assessed using evaluation metrics such as area under the receiver operating characteristic curve (AUROC). For imbalanced datasets, the use of accuracy can be misleading, therefore balanced accuracy (B-Accuracy) is a better choice. B-accuracy is defined as the mean of sensitivity and specificity. Another common metric is the F1-score (here denoted F-score), which is the harmonic mean of precision and sensitivity.

2.5.4 Transfer learning

A large number of neural network architectures have already been developed, and for many tasks it is therefore impractical to design a model entirely from scratch. Instead, existing networks are often adapted to the task at hand by training them on task-specific data.

Many low-level features image learned by neural networks are shared across tasks, such as the detection of edges, corners, or simple textures in images. For this reason, models can be initialized using pre-trained weights obtained from a different but related task rather than starting from randomly initialized weights, as is done when training a model from scratch. This strategy, known as transfer learning, is widely used in deep learning and is particularly valuable in medical imaging applications, where the amount of labelled data is often limited. During training, all the pre-trained weights may then be fine-tuned for the new task, although in some cases the weights in the earlier layers are kept fixed while only later, task-specific layers are updated.

2.5.5 Model generalization

A common challenge in medical imaging is overfitting, which occurs when a model adapts too closely to the training data and fails to generalize to unseen data. This can occur when the dataset is small relative to the model complexity, allowing the model to learn patterns that are specific to the training data rather than generalizable features. As a result, an overfitted model typically shows substantially better performance on the training set than on a test set. Overfitting is particularly problematic in medical imaging, where datasets are often limited in size and models can accidentally learn scanner- or subject-specific characteristics.

There are several ways to counteract overfitting and improve a model's ability to generalize to unseen data. One approach is data augmentation, in which the input images are modified during training in ways that preserve the relevant information. Examples include flipping, rotating, or adding small amounts of noise to the images. Data augmentation effectively increases the variability of the training data and reduces the risk of the model learning irrelevant patterns.

Another common strategy is the use of dropout layers, where a fraction of the activations (e.g. layer outputs) is randomly set to zero during training. This prevents the network from relying too heavily on individual features and encourages more stable representations. Pooling layers, which reduce the spatial resolution of feature maps, can also limit model complexity and may therefore contribute to improved generalization.

2.5.6 Ensemble learning

Selecting the optimal neural network architecture for a given task can be challenging, different models may capture different aspects of the data. Ensemble learning is therefore a strategy in which multiple models are combined to produce a final prediction, and it has repeatedly been shown to improve performance compared to using a single model [53]. A key reason for this improvement is that ensembles can reduce model variance and mitigate the impact of individual model errors.

For an ensemble to be effective, the included models should not only perform well individually but also be diverse [54], meaning that they do not make the same errors on the same data. Diversity can be achieved by training models with different architectures. By combining such models,

the ensemble can leverage complementary strengths, where one network may misclassify a case, other networks classify correctly.

The final output of an ensemble can be obtained using different fusion strategies, or aggregation functions, such as majority voting or averaging predicted probabilities. Probability averaging can lead to more stable predictions because it incorporates confidence information from the models. On the other hand, if the predicted probabilities are poorly calibrated, majority voting may be more robust [55]. Overall, ensemble learning provides a practical approach to improve generalizability and reliability, which is especially important in medical imaging applications where datasets are often small and inter-subject variability is high.

2.5.7 Multimodal learning

Multimodal learning refers to methods that integrate information from multiple data sources to improve predictive performance. In medical imaging applications, combining complementary modalities can provide a more complete representation of a patient's condition than relying on a single input type alone. For example, multimodal models may integrate different imaging modalities such as CT and MRI, combine structural images with derived representations such as segmentation maps, or use multiple MRI sequences. A multimodal approach resembles clinical radiological assessments, where several sequences are typically reviewed together since they highlight different tissue properties and pathological features.

A key design choice in multimodal DL is to choose how and when information from different modalities is fused. In early fusion, modalities are combined at the input level, for example by concatenating different MRI sequences as separate channels of one input image before training. In late fusion, each modality is processed by a separate model or network stream, and the outputs are combined at the decision level, for example by averaging predicted probabilities or using voting strategies. Intermediate fusion refers to approaches where modality-specific feature extraction is performed in the early stages of the network, followed by sharing of features in later stages that enables interactions between modalities.

The optimal fusion strategy can depend on the task, dataset size, and practical constraints. Early or intermediate fusion can utilize the multimodal interactions more directly but often requires careful alignment between modalities and can increase computational cost. Late

fusion is simpler to implement and may be more stable when modalities differ in resolution or are occasionally missing.

3 Research Rationale

Diagnosis and prediction of shunt responsiveness in INPH is challenging due to overlapping symptoms with other age-related disorders and incomplete understanding of the underlying pathophysiology. Brain MRI is central to clinical evaluation, but current radiological assessment is largely based on a limited set of predefined structural markers, several of which rely on subjective interpretation and require manual effort, limiting reproducibility and scalability across centres. At the same time, MRI data contains high-dimensional anatomical information that is likely not fully captured by conventional measures.

Advances in AI, particularly DL, provide an opportunity to extract clinically relevant patterns from imaging data in a more data-driven manner. In this thesis, we adopt a whole brain approach, enabling models to learn from the entire image, without restricting the analysis to a predefined set of regions or measurements. In addition, for the existing imaging markers there is also a need for methods that are stable, quantitative and computationally feasible.

This thesis therefore investigates deep learning-based classification using ensembles and multimodal fusion strategies to classify HLGD and shunt response, as well as an automated pipeline for the radiological measure qDESH. Together, these papers aim to contribute toward more reproducible and scalable, clinically applicable imaging-based tools for INPH assessment and decision support.

4 Aims

The overall aim of this thesis was to develop and assess AI-based and fully automated MRI-based methods, to improve objective assessment and shunt decision support in INPH.

The specific aims of this thesis were:

- ◆ To investigate whether whole-brain MRI contains sufficient morphological information to distinguish older individuals with HLGD from controls using CNN classifiers.
- ◆ To develop and evaluate an ensemble search algorithm that systematically identifies networks to include in an optimal ensemble in order to improve classification performance.
- ◆ To develop and evaluate a search algorithm for identifying optimal fusion points in multimodal DL architectures for medical imaging classification tasks.
- ◆ To assess whether the developed DL-based methods, and radiomics-based machine learning models, can provide clinically relevant prediction of shunt responsiveness in patients with INPH, using multi-sequence structural MRI.
- ◆ To develop and validate a fully automated quantitative DESH assessment pipeline (auto-qDESH) to enable stable and scalable analysis of DESH without manual intervention.

5 Materials and Methods

This thesis includes four papers whose methods and materials are summarised in Table 1. In Paper I, an ensemble search algorithm was developed for selecting individual CNNs to include in an ensemble model for classification of individuals with HLGD using T1-w brain MRI. Paper II proposes a sequential fusion search algorithm (SFSA) to determine the optimal stage for fusing multi-sequence CNN streams in brain MRI-based classification tasks. Paper III applied the models from the previous papers, together with conventional ML models using radiomic features to predict shunt response in a cohort of shunted INPH patients. Finally, Paper IV developed a fully automated image analysis pipeline for qDESH estimation.

Table 1, Overview of the methods and materials in the papers.

Method\Paper	I	II	III	IV
<i>Method development</i>	✓	✓		✓
<i>CNN</i>	✓	✓	✓	
<i>Radiomics based ML</i>			✓	
<i>Ensemble models</i>	✓		✓	
<i>Multimodal models</i>		✓	✓	
<i>Classification task</i>	✓	✓	✓	
<i>Automatization task</i>				✓
<i>Patient group</i>	HLGD	Epilepsy, Alzheimer's disease	INPH	INPH
<i>Control group</i>	✓	✓		✓
<i>MRI sequence</i>	T1-w	T1-w, T2-w, T2 FLAIR	T1-w, T2-w, T2 FLAIR	T1-w

5.1 Ethical approval

The datasets used in this thesis were obtained from five different cohorts. The study in Paper I was approved by the Regional Ethical Review Board in Umeå (Dnr. 2017-335-31M; 2017-528-32M) and all participants provided written and oral informed consent.

Paper II included two publicly available datasets: an epilepsy cohort and the OASIS-3 dataset. The epilepsy cohort is openly available for anyone whereas OASIS-3 can be accessed by the research community after accepting their data use agreement and include a research statement.

The study in Paper III was approved by the Swedish Ethical Review Authority (no. 2020-04469), and patients were informed by letter with the possibility to opt out, deceased individuals were included.

The study in Paper IV was approved by the Regional Ethical Review Board in Umeå (Dnr. 2017/252-31) and all participants provided written and oral informed consent.

All data were analysed in de-identified form and handled in accordance with applicable ethical and data protection requirements.

5.2 Cohorts

The cohort used in Paper I is a subset of the participants in the VESPR (Ventriculomegaly and Gait Disturbance in the Senior Population in the Region of Västerbotten) cohort [56]. The total cohort consists of 1047 participants living in Umeå municipality, aged 65-86. It is a population-based cohort, collected to investigate gait disorders in the older population. The inclusion was based on self-reported gait problems through a survey. The participants underwent clinical assessments including tests of gait, balance, cognition, and brain imaging. Based on clinical assessment, the participants were diagnosed according to gait disorder, resulting in four groups: HLGD, (other) neurological gait disorder, non-neurological gait disorder and no (objective) gait disorder. Four age and gender matched controls were invited per participant with HLGD, chosen among the survey responses with no subjective gait disorder. From this cohort, all participants with HLGD and T1-w MRI ($n = 73$) and twice the number of controls ($n = 146$) were included.

Two openly available datasets have been used in this thesis for Paper II: Epilepsy [34] and OASIS-3 [35]. The Epilepsy dataset consists of 85 patients with epilepsy and 85 healthy adult controls, age range 11–60 years. The patients were treated at University Hospital Bonn between 2006 and 2021, with histologically confirmed or radiologically suspected focal cortical dysplasia type II. Only participants who were adults at the time of data preparation were included. The dataset also includes regions of interest of the lesions, histopathologic grading, as well as the

patients' improvement post-surgery. However, in this thesis only the group labels Epilepsy and Control and the corresponding MRI data (T1-w and T2 FLAIR) were used. The OASIS-3 dataset regards Alzheimer's disease (AD) and the subset used in this study consists of 1098 participants, with different stages of AD and healthy controls, age range 42–95 years. The subset used in this thesis consists of the 339 patients with AD and 508 controls, that had both T1-w and T2-w brain MRI available.

The cohort used in Paper III (Shunted INPH) is part of a cohort of patients investigated for INPH at Umeå University Hospital in 2007–2019. All participants underwent clinical investigation due to suspected INPH. The ones manifesting clinical and radiological signs of INPH together with a positive CSF tap test and/or an elevated CSF outflow resistance were considered eligible for shunt treatment. The subset used in this thesis were the ones that underwent shunt surgery, had available outcome data and preoperative MRI including T1-w, T2-w, and T2 FLAIR sequences. Patients were excluded if the time between symptom assessment and surgery surpassed 1 year, if the follow-up assessments were more than 2 years after surgery, or if the patient had a non-functional shunt, or other relevant adverse event, at the follow-up assessments. The shunt response was determined based on postoperative gait speed improvement, measured as the average of up to 6 repetitions of a 10 m maximum gait speed test. Patients were classified as responders ($n = 113$) if their gait speed was improved 0.16 m/s or more, or non-responders ($n = 36$) if the improvement was less than 0.1 m/s. The cutoff is based on a previous study estimating minimal clinically important difference in gait speed to be 0.16 m/s [57]. Patients with improvement between 0.1–0.16 m/s, or clear improvement in balance but not in gait speed, were excluded to provide a clear distinction between the groups.

For Paper IV a retrospective cohort was used (qDESH). It consists of 35 patients diagnosed with INPH at Umeå University hospital and 45 age and gender matched healthy controls. All participants underwent T1-w MRI. The INPH diagnosis was based on patient history and clinical and radiological findings in coherence with the American-European INPH guidelines. In the INPH group, 31 patients were categorized as probable INPH and 4 as possible INPH. The controls were volunteers with normal neurological exams and no major illnesses or MRI contraindications.

The demographics of the datasets collected at our department are presented in Table 2.

Table 2, Demographics of the three cohorts from our department used in this thesis.

Paper	Cohort	Number of participants	Female (n)	Age (year \pm SD)
<i>I</i>	VESPR cohort			
	HLGD	73	27	77.2 \pm 4.8
	Control	146	64	75.0 \pm 4.0
<i>III</i>	Shunted INPH cohort			
	Non-responder	36	10	74.1 \pm 5.9
	Responder	113	38	74.3 \pm 5.6
<i>IV</i>	qDESH cohort			
	INPH	35	10	74.9 \pm 7.1
	control	45	13	72.2 \pm 5.7
<i>Abbreviations; SD: standard deviation.</i>				

5.3 MRI data

The MRI data used in this thesis are T1-w, T2-w and T2 FLAIR brain volumes.

In Paper I, MRI data were acquired using a 3D T1-w fast spoiled gradient echo sequence (BRAVO) on a 3T scanner (Discovery MR 750, GE Healthcare), with a 32 channel head coil. Images were obtained with 1 mm isotropic resolution and underwent intensity correction (PURE).

In Paper II, for the OASIS-3 cohort, MRI was primarily acquired on 3T Siemens scanners, most commonly the TrioTim system. T1-w MPRAGE sequences with approximately 1 mm isotropic resolution and mainly two T2-w acquisition protocols were used, a 3D SPACE sequence with approximately 1 mm isotropic resolution and a 2D turbo spin echo (TSE) sequence with 4 mm slice thickness. Because the dataset spans multiple studies and acquisition protocols, imaging parameters vary between sessions. For the epilepsy cohort, MRI was performed on a 3T scanner (Magnetom Trio, Siemens Healthineers) including T1-w MPRAGE and T2 FLAIR sequences. Two T1-w acquisition protocols were used, resulting in differences in spatial resolution (1.0 mm vs 0.8 mm isotropic). Most participants received isotropic 1 mm T2 FLAIR imaging, while a small subset underwent a 2D FLAIR sequence.

In Paper III, MRI was acquired across multiple scanners from Philips, GE, and Siemens, with the majority obtained on Philips Achieva/Achieva dStream scanners. Examinations were performed at 3T for 134 patients and at 1.5T for 15 patients. The T1-w images consisted primarily of 3D gradient echo sequences with near-isotropic voxel sizes ranging from approximately 0.7-1.0 mm. T2-w and FLAIR sequences were mainly acquired as 2D multi-slice turbo/fast spin echo acquisitions, with in-plane resolution ranging from approximately 0.6-1.6 mm and slice thickness typically between 3-5 mm.

In Paper IV, 3D T1-w fast spoiled gradient echo sequences were acquired using four different scanners. Most MRI were conducted with 3T scanners (Discovery MR 750, GE Healthcare, Philips Achieva) but three were acquired using 1.5T systems, (Philips Achieva and Achieva dStream), with reconstructed voxel sizes ranging from approximately 0.5–1.0 mm³.

5.3.1 Spatial preprocessing

To reduce inter-subject variability and differences in image orientation, rigid-body alignment was applied to all brain MRI volumes prior to analysis in Papers I, III, and IV. In Papers I and III, the volumes were rigidly aligned to a MNI template brain using SPM12. The ICBM-152 template corresponding to each MRI sequence was used as the reference. During this procedure, the images were resliced to match the grid of the reference template. In Paper IV, the rigid-body alignment was instead performed using the *acpcdetect* software [58], which aligns the brain volume according to the AC-PC line and the mid-sagittal plane, with the AC positioned at the centre of the image volume. This strategy was selected to mimic the workup for the qDESH method.

5.3.2 Tissue extraction and segmentation

To reduce irrelevant information and for anonymization purposes, prior to network training, brain extraction or segmentation methods were applied in all papers. In Papers I and II, a publicly available U-Net-based brain extraction tool [59] was used. In Paper III, SynthStrip was used for the same purpose. In Paper IV, the CNN-based segmentation software SynthSeg [47] was used to segment the brain into cortical regions, CSF spaces, and white matter. This segmentation formed the basis for subsequent analyses.

5.3.3 Intensity normalization

Prior to use as input to CNNs, in Papers I and II, voxel intensity distributions were examined and truncated to reduce the influence of extreme values. The intensities were then scaled to the appropriate range for each network. In Paper III, where data were acquired from multiple scanners, voxel intensities were clipped at three standard deviations above the mean before scaling, to mitigate scanner-related intensity variations. In all three papers, background voxels were set to the minimum value of the scaled intensity range. In Paper IV, bias field correction was applied to the MRI volumes, in order to improve intensity homogeneity and facilitate threshold-based analyses.

5.3.4 Radiomics and machine learning models

In Paper III, conventional ML models were used, in addition to CNNs. Radiomic features were used as input to these. To extract these features, Laplacian of Gaussian filters and wavelet transforms were applied to the pre-processed MRI volumes generating several representations of each image. Feature extraction was performed in three dimensions, treating the segmented brain as one region of interest, resulting in 1030 radiomic features. The final features were standardized using z-score normalization.

Within each cross-validation fold, dimensionality reduction was applied using the training set to limit redundancy and reduce the risk of overfitting. Features with low variance and highly correlated features were excluded. A random forest classifier ranked the remaining features by feature importance, and the top 20 were selected for model training. The same features were used for the corresponding validation and test sets.

25 conventional ML classifiers were evaluated, including linear models (e.g., logistic regression, ridge classifier), support vector machines, discriminant analysis methods, probabilistic classifiers (e.g., Naïve Bayes), instance-based methods (e.g., k -nearest neighbors), and multiple tree-based ensemble methods such as random forests and gradient boosting. Shallow neural networks were also included.

5.4 Neural networks

All CNNs in this work were implemented as the 3D versions to fit the input MRI volumes. The different CNNs used in Papers I–III are presented in Table 3, together with their pretraining. The specific networks were chosen because they are well established or had demonstrated good performance on similar tasks [60–64]. The pretraining was based on models trained on data most similar to brain MRI. In paper I, a modification of the ResNet-18 was also included, which has an additional fully connected layer in the end, before the classification output. Similarly, in Paper III, SFCN-FC is a simple fully convolutional network (SFCN), with an additional fully connected layer in the end. These modifications were added to investigate whether they would influence classification performance. ResNet-18-MM is the multimodal model used in Papers II and III, based on two (Paper II) and three (Paper III) ResNet-18 streams, one for each sequence type.

Table 3, Overview of the 3D convolutional neural networks used in this work (Papers I, II and III), together with their corresponding pretraining datasets. Resnet-18-MM is the multimodal network, with two streams in Paper II and three streams in Paper III.

Architecture family	Network	Pretraining	I	II	III
<i>Dense Convolutional Network</i>	DenseNet-121		✓		✓
	DenseNet-169		✓		✓
	DenseNet-201				✓
	DenseNet-264				✓
<i>EfficientNet</i>	EfficientNet-Bo	ImageNet [65]	✓		
	EfficientNet-B1	ImageNet	✓		
<i>Residual Network</i>	ResNet-10	Med3D [66]	✓		✓
	ResNet-18	Med3D	✓		✓
	ResNet-18-FC	Med3D	✓		
	ResNet-18-MM	Med3D		✓	✓
	ResNet-34	Med3D	✓		✓
	ResNet-50	Med3D	✓		✓
	ResNet-101	Med3D	✓		✓
	ResNet-152	Med3D			✓
	ResNet-200	Med3D			✓
<i>Squeeze and Excitation Network</i>	SENet-154	ImageNet	✓		✓
	SEResNet-50	ImageNet	✓		✓
	SEResNet-101	ImageNet	✓		✓
	SEResNet-152	ImageNet			✓
	SEResNeXt-50	ImageNet	✓		✓
	SEResNeXt-101	ImageNet	✓		✓
<i>Simple Fully Convolutional Network</i>	SFCN	UK Biobank [67]	✓		✓
	SFCN-FC	UK Biobank			✓

5.5 Ensemble search

To improve performance beyond single CNNs, we developed an ensemble search algorithm (Paper I). All CNNs were trained separately, and their predictions were combined using late fusion. The goal of the ensemble search was to construct the optimal ensemble, K^* , consisting of classifiers that are both accurate and diverse. To achieve this, we defined following loss function

$$F(K) = \frac{1}{|E|} \sum_{i=1}^{|E|} (1 - e_i(K)) + \frac{1}{|D|} \sum_{j=1}^{|D|} (1 - d_j(K)),$$

where K denotes a subset of classifiers $C = \{c_1, c_2, \dots, c_{|C|}\}$, $E = \{e_1, e_2, \dots, e_{|E|}\}$ is the set of evaluation metrics and $D = \{d_1, d_2, \dots, d_{|D|}\}$ is the set of diversity metrics. In this thesis, the evaluation metrics were B-accuracy and F-score and the diversity metrics were Yule’s Q -statistic, Pearson correlation and Cohen’s kappa.

The backwards search algorithm starts with the full set of trained networks C . At each iteration, the current subset K is evaluated by temporarily excluding one network at a time and calculating the resulting value of the loss function F . The subset that results in the lowest value of F is selected for the next iteration. This iterative process is repeated until removing an additional network no longer decreases F .

To compute $F(K)$, a decision profile DP is created for each input, consisting of the posterior class probabilities from all separate networks. These probabilities are then combined, using an aggregation function, to determine the joint predicted probability. In this thesis, maximum probability, product of class probabilities, and averaging the class probabilities were evaluated as aggregation functions. The final predicted label was determined using the maximum membership rule. After performing this for the whole validation set, the evaluation metrics are computed, by comparing the ensemble predictions to the true labels. The diversity metrics are calculated pairwise between the single networks’ classifications. For consistency, all evaluation and diversity metrics are normalized to the range $[0,1]$, with 1 being the most accurate, or most diverse, respectively. The computation of $F(K)$ is illustrated in Figure 4.

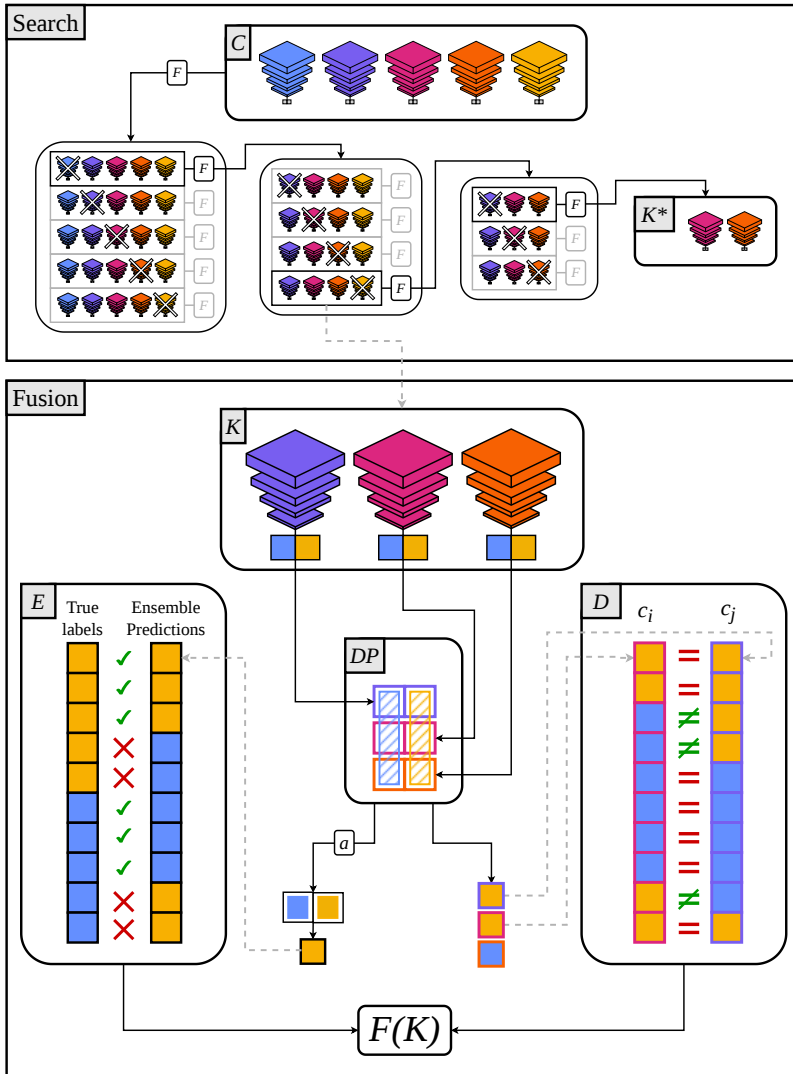


Figure 4. The ensemble search. The algorithm starts with the full set of networks and iteratively removes one network at a time until the cost function F is no longer decreased, yielding the optimal ensemble K^* . In the illustrated example for the fusion, the ensemble K consists of three networks. Their posterior probabilities are combined into a decision profile (DP) and an aggregation function (a) is applied to produce the ensemble prediction. From this prediction, the evaluation metrics E are computed. The individual networks' predictions (here the networks c_i and c_j) are used pairwise to compute the diversity metrics D . The evaluation and diversity are then combined to compute $F(K)$. Adapted from Paper I [68], © Elsevier.

5.6 Sequential forward search

To make use of different MRI sequences for the same patient, we developed another search algorithm (Paper II), to optimize where to fuse the data in an intermediate fusion architecture. The base architecture was ResNet-18, with one modality specific stream per sequence type, interconnected with n fusion modules, see Figure 5. In our case $n = 4$, implemented as Multimodal Transfer Modules (MMTM) [69], placed after each residual block. Each fusion module can be activated (allowing interaction between modalities) or deactivated (keeping the modality streams separate). The posterior class probabilities are combined at prediction stage.

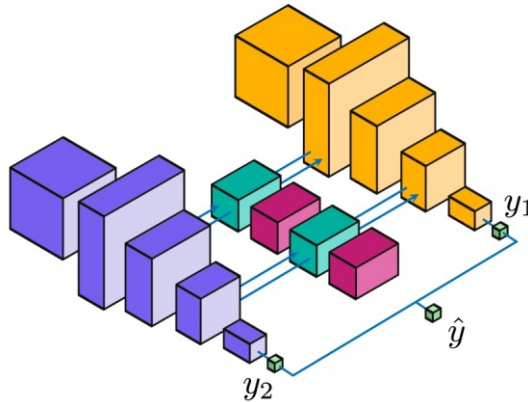


Figure 5, The sequential forward search setup. Schematic drawing of two ResNet-18 (yellow and blue), each processing one MRI sequence from the same patient. The streams are interconnected by four multimodal transfer modules (MMTMs). When an MMTM is activated (turquoise), the modality-specific feature representations can interact across the two streams. When all MMTMs are deactivated, the streams remain independent until prediction stage. The configuration of activated MMTMs is determined by the sequential forward search algorithm.

To determine which fusion modules should be activated, we developed a Sequential Forward Search Algorithm (SFSA). For the baseline configuration, no fusion modules are activated. After training, the loss is calculated on the validation set. Next, the algorithm evaluates the configurations where one fusion module is activated at a time. For each candidate configuration, training is initialized from the final weights of the baseline configuration, and the validation loss is computed. The

fusion module that yields the largest decrease in validation loss is selected. This procedure is repeated iteratively by activating one additional fusion module per stage. At each stage, the weights are initialized from the current best configuration from the previous stage. The search terminates when activating an additional fusion module no longer decreases the validation loss, and the final model is chosen as the configuration with the lowest validation loss during the search.

5.7 Auto-qDESH pipeline

For the computation of auto-qDESH (Paper IV), the automated qDESH, the brain segmentation was used together with the MRI aligned to the AC-PC line. The segmentation maps were used to identify the corresponding voxels in the MRI volume belonging to the lateral ventricles. For coherence with the semiautomatic qDESH method, this ventricular volume was dilated by 7 mm, and a voxel intensity histogram was created from the dilated region. The midpoint between the grey matter peak and the CSF peak was used as the intensity threshold for CSF for the rest of the algorithm. The overall search volume was defined as a slab extending 20 mm anterior to the PC (along the AC-PC line).

For the quantification of the CSF volume in the high convexity, the outer boundary of the brain segmentation was used instead of the manually defined dural boundary in the semiautomatic qDESH method. The upper 30 ml of the search volume was filtered using the CSF intensity threshold to compute the high-convexity CSF volume.

For quantification of the CSF volumes in the Sylvian fissures, the segmentation map was used. The lateral borders of the cerebellum were first identified to determine the appropriate sagittal slices for defining the Sylvian fissure midline. Within these two slices, cortical labels were used to identify the inferior boundary of the Sylvian fissure, defined by the transition from superiotemporal or transversetemporal cortex labels to CSF. The superior boundary of the Sylvian fissure was determined by detecting the transition from CSF back to non-CSF tissue labels. Based on these two boundaries, the midline of the Sylvian fissure was estimated. A restricted search region was then defined by extending 10 mm superiorly and inferiorly perpendicular from this midline, as in the original qDESH method. This restriction was applied to all sagittal slices lateral to the cerebellar border.

Within the resulting search region, defined by the overall anterior-posterior limits and the lateral restriction relative to the cerebellum, voxels labelled as CSF were grouped into connected three-dimensional objects. These objects were dilated using a spherical kernel with a diameter of 4 mm. Finally, the corresponding voxels in the MRI volume were filtered using the previously determined CSF intensity cutoff to compute the Sylvian fissure CSF volume. The process for determining the CSF spaces is illustrated in Figure 6.

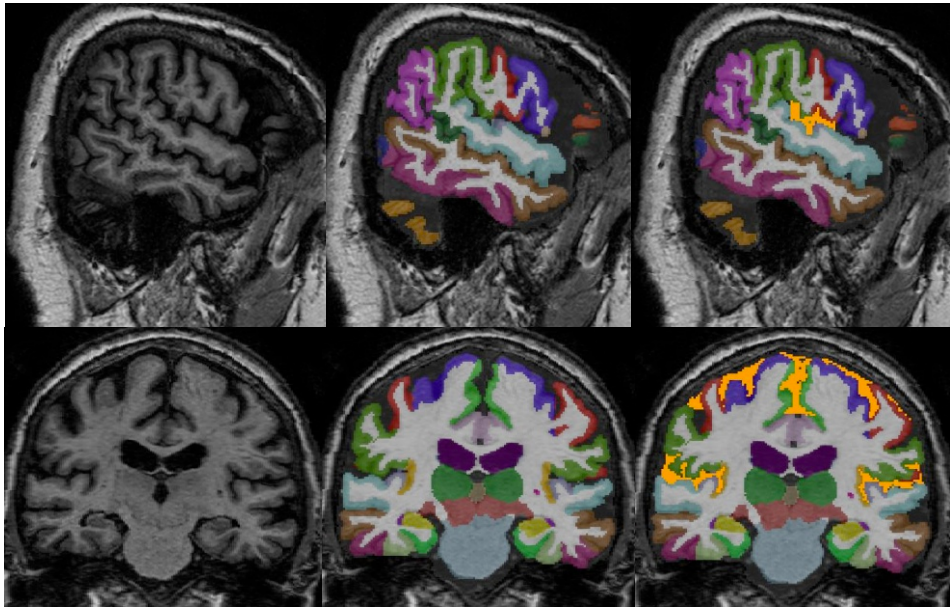


Figure 6, Different steps of the qDESH algorithm. Top row: Sagittal view of a brain in the same sagittal slice as the lateral border of the cerebellum. Bottom row: Coronal view of the same brain. Left: The rigid body rotated T1-weighted brain MRI, where the mid-sagittal section is aligned with the anterior commissure-posterior commissure line. Middle: The segmentation map overlaid on the brain MRI, segmenting the cortex and cerebrospinal fluid (CSF) spaces into different regions. Right: The final, segmented volumes for the Sylvian fissures' and high convexities CSF spaces, here marked in orange.

5.8 Training of networks and algorithms

All CNNs were trained in a tenfold stratified cross-validation, for Paper I and III with eight sets for training, one for validation and one for testing, and in Paper II, with seven sets for training, two for validation and one for testing.

Data augmentation was applied, consisting of random shifts (± 3 voxels) along each spatial axis and random left-right flipping (reflection across the sagittal plane) and was applied on the training set. In paper I and III the minority class (HLGD and shunt non-responders) were duplicated in the training and validation sets.

All CNNs were trained with a cross-entropy loss function and the Adam optimizer with an initial learning rate of 10^{-5} (10^{-4} in Study II). The learning rate was reduced by one order of magnitude if the validation loss did not reduce for 10 consecutive epochs. The maximum number of epochs was 300, and early stopping was applied if the validation loss did not decrease for 50 consecutive epochs. Minibatch size was 16 and default hyper parameters were used. All the computations were performed using PyTorch as the main deep learning library.

For the shunt outcome prediction, the single CNNs obtaining a specificity and sensitivity above 60% on the validation set were included in the ensemble search. This was decided to reduce the number of networks at starting point and avoid early local minima.

For the conventional ML models, the same cross-validation folds were used as for the CNNs. Default hyperparameters from the Python libraries scikit-learn and XGBoost were used.

For the auto-qDESH development, a smaller group of 14 participants, selected based on age, gender, image quality and diagnosis, were used during testing and development of the algorithm, whereas the remaining 66 were used to evaluate the final algorithm.

5.8.1 Competitors and baselines

To validate our method choices, we introduced competing methodologies or baselines.

In Paper I, an ablation study was conducted to investigate the importance of the diversity metrics together with the evaluation metrics in the ensemble search cost function. In this, all possible combinations of the available evaluation and diversity metrics were used in the cost function F , and through a brute force search, the ensemble of networks minimising each F the most, on the validation set, were chosen as competing ensembles. The ensemble found when F was computed based only on evaluation metrics is denoted K_E .

To be put in perspective, the performance of the networks were also compared to the best linear radiological measure to distinguish between the two groups in the cohort, previously detected to be z-EI [70].

In Paper II, several competitors were evaluated. The first, referred to as brute force, corresponded to the best configuration obtained by testing all possible combinations of active MMTMs and training each configuration from scratch. We also evaluated a late fusion approach, in which no MMTMs were active. In addition, two unimodal models were trained, each using a single MRI sequence.

In Paper III, results from a previous study by Leary et al [71] were used for comparison. In the study, they presented a late fusion model based on two ResNet-50 streams, classifying shunt responders versus non-responders using dual-sequence (T2-w and T2 FLAIR) MRI. We implemented a similar model, also consisting of two ResNet-50 streams using the same two sequences, for comparison.

In Paper IV, the qDESH values assessed by two external raters who used the semi-automatic qDESH method were used for comparison with the auto-qDESH pipeline. Rater A was considered to hold the ground truth, since she had the main responsibility of the development of the qDESH method. Rater B represented a typical clinical rater. The intraclass correlation between the two was ICC = 0.99. The images had also been assessed by two neuroradiologists [72], who categorized the images into three groups: No DESH, Mild-Moderate DESH and Severe DESH. In the case the two clinicians did not agree, the more severe label was chosen.

5.9 Statistics and validation

To test whether the CNNs could distinguish differences in brain structure between the participants with HLGD and controls in Paper I, a chi-square test of independence was performed. Statistical significance was defined as $p < 0.05$.

In Papers I-III, model performance was evaluated using B-accuracy, F-score, and AUROC and is presented as the mean \pm standard error of the mean (SEM) across the cross-validation folds. The performance metrics were computed in Python.

In Paper IV, agreement between auto-qDESH and qDESH assessments from Rater A and Rater B was evaluated using the ICC, a pairwise two-

way mixed effects model, for absolute agreement and single measures. For comparisons with the visual DESH categories, the nonparametric test Kruskal-Wallis test with Dunn-Šidák post hoc test was conducted. Statistical analyses were performed using IBM SPSS Statistics (version 25.0.0.2) and MATLAB R2024b.

6 Results

6.1 HLGD classification

In Paper I, all trained single networks could distinguish between participants with HLGD and control according to a chi-square test of independence, $p < 0.001$.

The highest B-accuracy and F-score were obtained by the ensemble K^* , which was a combination of two networks, ResNet34 and DenseNet-169 with a B-accuracy of $76.0 \pm 2.9\%$ and an F-score of 67.7 ± 3.7 . This ensemble also performed better than the combination of all the single networks, C , see Figure 7. The best performing single networks were DenseNet-121 which achieved a B-accuracy of $73.5 \pm 2.3\%$ and ResNet-18 with an F-score of 64.4 ± 2.8 . The results are presented using averaging as the aggregation function, as all aggregation functions provided very similar predictions. The best radiological linear measure to distinguish between these groups is z-EI, presented in a previous study [70]. For this cohort, the best cutoff was determined to $z\text{-EI} = 0.34$, which yielded a B-accuracy of 66.7% and F-score of 52.2%.

6.1.1 Ensemble search

The use of both evaluation and diversity metrics in our search algorithm to find K^* was tested through an ablation study, where all possible combinations of the included evaluation and diversity metrics were tested. Through a brute force search, the ensemble that minimised $F(K)$ on the validation set, for each combination of metrics, was identified. The ensemble obtained using only evaluation metrics (B-accuracy and F-score) was denoted K_E . Many combinations resulted in the same ensemble, which is why only five ensembles are presented in Figure 8. As shown in the figure, K^* achieved the best performance in terms of B-accuracy, F-score and recall.

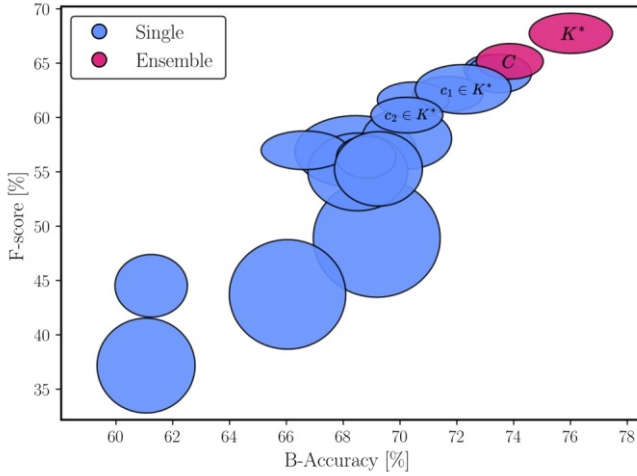


Figure 7, The performance of all the single networks and the ensembles C and K^* in terms of Balanced-Accuracy and F -score. C is the ensemble consisting of all single networks and K^* is the optimal ensemble, determined by the ensemble search algorithm. K^* consists of the two networks c_1 : DenseNet-169 and c_2 : ResNet-34. The lengths of the ellipses correspond to the standard error of the mean of the metric. Reproduced from Paper I [68], © Elsevier.

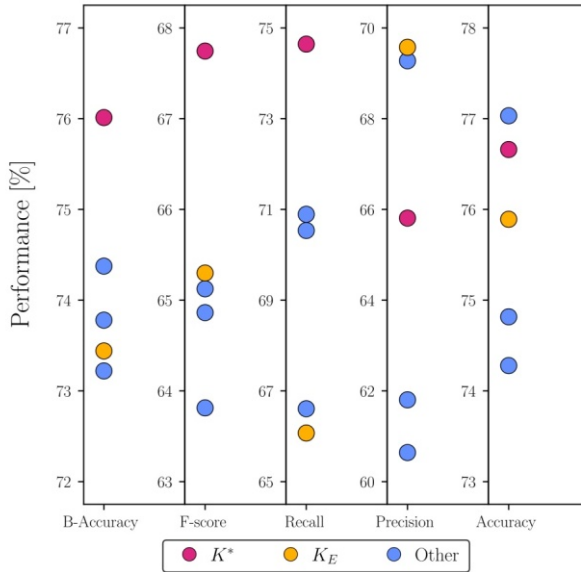


Figure 8, The ablation study. By basing $F(K)$ on all subsets of the included evaluation and diversity metrics, different ensembles were created. K^* is the ensemble based on all five metrics while K_E is based on only the two evaluation metrics. The remaining configurations are denoted Other. Adapted from Paper I [68], © Elsevier.

6.2 Multimodal fusion

When using our newly introduced SFSA to conduct a guided search for the optimal fusion points between two ResNet-18 streams, the resulting model configurations obtained the best results, for both OASIS-3 and Epilepsy datasets. In both cases the SFSA chose a configuration with one MMTM active. For the OASIS-3 dataset the activated MMTM was the second and for the Epilepsy dataset it was the first. The brute force method (testing all possible combinations of active MMTMS) had similar results, while the late fusion and the unimodal models showed inferior performance, see Table 4.

Table 4, Performance metrics on the two cohorts of Paper II. The configuration decided by the sequential fusion search algorithm (SFSA) based on two ResNet-18 streams with 4 multi-modal transfer modules, and its competitors. For the OASIS-3 dataset, the model used T1-weighted and T2-weighted images, whereas for the Epilepsy cohort, the model used T1-weighted and T2 FLAIR images. Results are presented as the mean±standard error of the mean of the test set over the ten cross-validation folds.

Model	OASIS-3			Epilepsy		
	AUROC	B-accuracy	F-score	AUROC	B-accuracy	F-score
SFSA	80.6±0.3	72.3±1.4	80.9±0.9	87.9±3.0	81.5±3.0	83.5±2.6
Brute-force	79.2±1.7	72.0±1.7	79.8±1.3	87.4±2.8	80.5±3.3	82.2±3.2
Late Fusion	75.7±1.4	70.3±1.5	80.6±0.8	84.4±2.5	80.8±2.0	82.7±3.0
Unimodal (T1-w)	79.6±1.8	69.3±1.7	80.0±1.0	81.5±3.0	71.0±3.1	71.5±3.1
Unimodal (T2-w/FLAIR)	70.6±2.2	60.4±2.0	74.9±0.8	74.5±3.5	77.8±3.7	79.6±4.0

6.3 Shunt outcome prediction

For the shunt outcome prediction, none of the networks could distinguish in a clinically relevant way between the shunt responders and the non-responders, with the best performing network in terms of B-accuracy being Linear discriminant analysis based on T1-w radiomic features with $63.8 \pm 7.1\%$. The best model in terms of AUROC was the model mimicking the one presented by Leary et. al [71], with an AUROC of $68.8 \pm 6.7\%$. The best network performance per sequence and modality type is presented in Figure 9. Even though the models did not achieve clinically predictive performance, when testing the group division with a chi-square test of independence, there is still difference between the groups.

To investigate if a clearer group separation would improve the results, in a post hoc analysis, we used all the non-responders ($n = 36$) and only the most improving responders ($n = 36$) and reran the training for all networks. The gait speed improvement in the shunt responder group was then > 0.42 m/s. The results remained modest, see Figure 9. The best-performing model in terms of B-accuracy was ResNet34 on T2-w images with $62.5 \pm 3.7\%$, and in terms of AUROC, the best performing model was the ensemble model found by the search algorithm, based on five CNNs, with $69.2 \pm 5.2\%$.

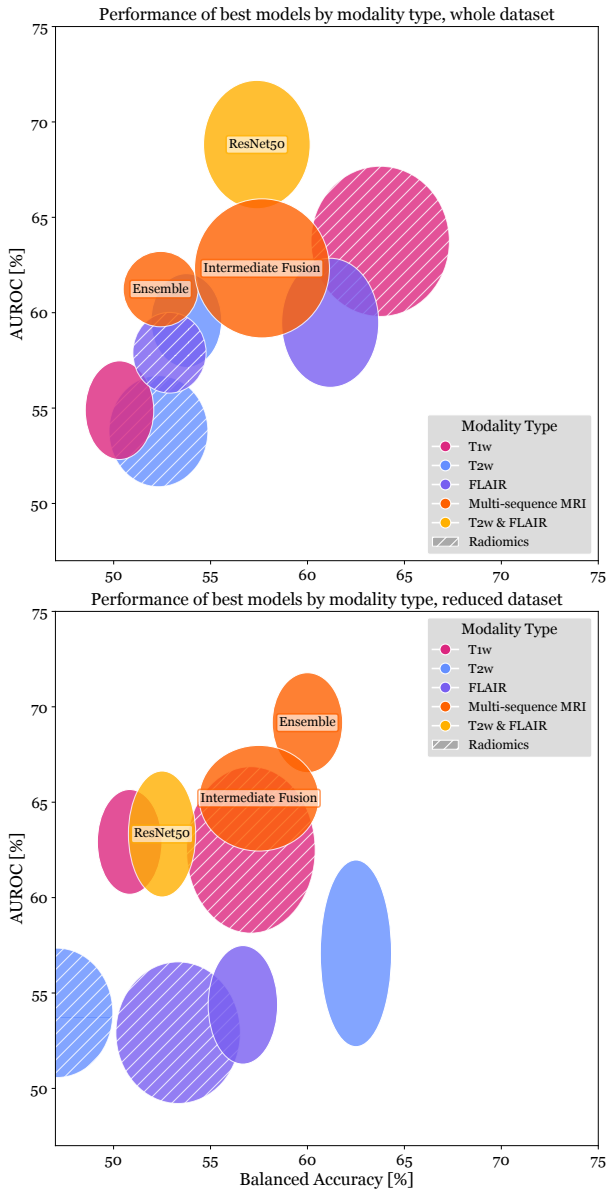


Figure 9, Shunt outcome prediction for the whole dataset (top) and the reduced dataset (bottom). The mean test set performance is presented in terms of balanced accuracy and area under the receiver operating characteristic curve. The lengths of the axes correspond to the standard error of the mean over the ten cross-validation folds. The best performing model for each sequence and modality type is presented.

6.4 Auto-qDESH

During preprocessing, two images did not pass the AC-PC rotation step, hence the evaluation results are based on 64 subjects. After the preprocessing steps were finalized, the algorithm processed each MRI in approximately 2 seconds. When comparing the auto-qDESH with the qDESH values from Rater A and Rater B, an ICC of 0.90 was obtained, see Table 5.

Table 5, The intraclass correlation (ICC) between the auto-qDESH and Rater A and Rater B using the semi-automatic qDESH method. The first section presents the values for the calculated combined measure, and the following sections present the three different volumes that are used in the calculations, High convexities and left and right Sylvian fissure.

Rater Pair	ICC [95% CI]	Mean Diff.	SD of Diff.
<i>qDESH</i>			
<i>Auto-qDESH – qDESH Rater A</i>	0.90 [0.81–0.95]	-0.26	0.55
<i>Auto-qDESH – qDESH Rater B</i>	0.91 [0.83–0.95]	-0.22	0.53
<i>qDESH Rater B – qDESH Rater A</i>	0.99 [0.99–1.00]	0.04	0.18
<i>High convexities volume</i>			
<i>Auto-qDESH –Rater A</i>	0.90 [0.84–0.94]	-0.43	1.62
<i>Auto-qDESH –Rater B</i>	0.89 [0.82–0.93]	-0.46	1.71
<i>Rater B –Rater A</i>	0.99 [0.99–1.00]	-0.03	0.46
<i>Left Sylvian fissure volume</i>			
<i>Auto-qDESH –Rater A</i>	0.90 [0.085–0.97]	-0.72	0.46
<i>Auto-qDESH –Rater B</i>	0.85 [0.26–0.95]	-0.76	0.69
<i>Rater B –Rater A</i>	0.95 [0.92–0.97]	-0.05	0.56
<i>Right Sylvian fissure volume</i>			
<i>Auto-qDESH –Rater A</i>	0.88 [0.60–0.95]	-0.53	0.68
<i>Auto-qDESH –Rater B</i>	0.86 [0.58–0.94]	-0.57	0.76
<i>Rater B –Rater A</i>	0.96 [0.94–0.98]	-0.04	0.47
<i>Abbreviations; CI: confidence interval, SD: standard deviation, Diff: Difference</i>			

The auto-qDESH systematically underestimated all three volumes. However, because the qDESH score is computed as a ratio of the volumes of the Sylvian fissures and the high convexity, this systematic underestimation is partly compensated for. As a result, the ICC for the qDESH score was higher than for the individual volume measurements.

The auto-qDESH values were not normally distributed. Therefore, the ICC was also calculated for the inverse of auto-qDESH and qDESH, as this resulted in values that more closely approximated a normal distribution. The results were ICC = 0.95 (95% CI: 0.90–0.97) between auto-qDESH and Rater A, and ICC = 0.93 (95% CI: 0.86–0.96) between auto-qDESH and Rater B. The difference compared to the original ICC analysis is mainly explained by a reduced influence of the highest qDESH values.

Auto-qDESH was also compared to visual assessment of DESH, see Figure 10. There was a difference in auto-qDESH values between no DESH and mild-moderate DESH as well as no DESH and severe DESH ($p < 0.001$). There was no statistically significant difference between mild-moderate and severe DESH ($p = 0.21$).

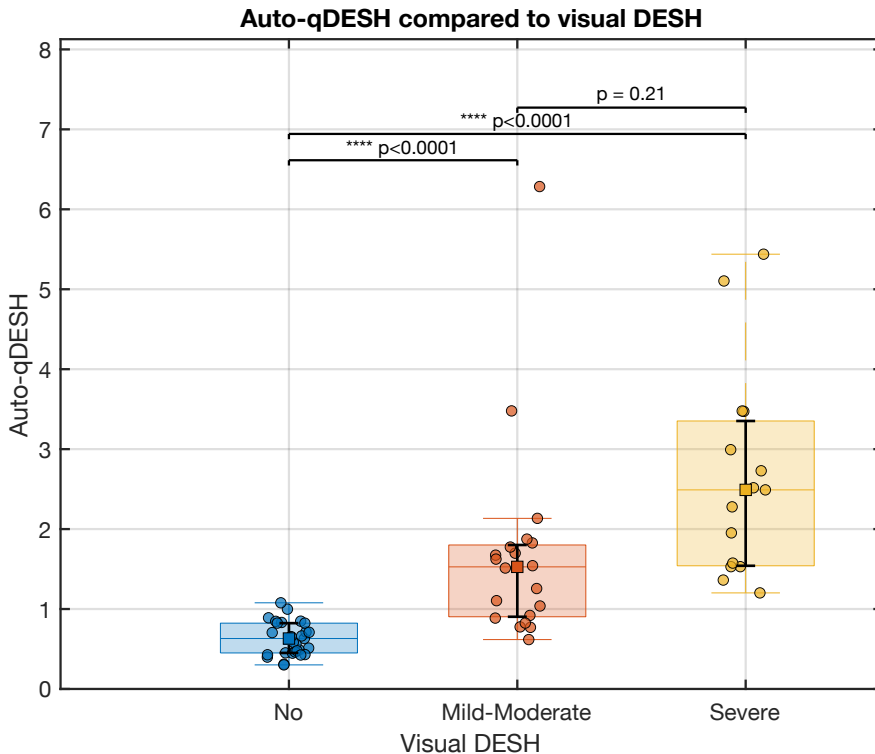


Figure 10, Boxplots of auto-qDESH compared to visual DESH, assessed by two neuroradiologists. If the raters did not agree, the more severe label was chosen.

7 Discussion

In this thesis, ML and DL models were applied to investigate the relationship between brain structure and gait impairment, as well as shunt outcome in INPH. There was a significant relationship in both cases, however, the performance of the models was not sufficient to improve shunt selection in clinical practice. Regarding the developed search algorithms, the ensemble K^* reached the highest performance in classifying HLGD, and the SFSA resulted in the best performing models compared to its competitors, motivating the use of a guided search when working with network ensembles. The proposed auto-qDESH pipeline showed high agreement with the semi-automatic qDESH method, although not as high as the agreement between two raters assessing with qDESH. The results further show that the method can distinguish between the presence and absence of DESH, although it is less accurate in grading the severity of DESH.

7.1 Assessing brain changes with AI

7.1.1 HLGD

For HLGD, all CNNs identified a relationship between brain structure and gait impairment supporting that there is a difference in brain structure between individuals with HLGD and controls (Paper I). Most single models obtained better performance than the best linear radiological measure, z-EI, indicating that DL can capture more patterns linked to HLGD than what is captured by established linear metrics.

The improved results from the ensemble model could indicate that different models complement each other by basing their predictions on distinct features in the brain related to HLGD. This may suggest that there is more than one pattern in the brain that can be related to the gait impairment or that distinct morphological patterns characterize different subgroups of individuals of HLGD, which are detected by different models.

The cohort used in Paper I is unique, as the participants were divided into different gait impairment groups, with focus on the INPH specific gait pattern, independently of the MRI data. Because the cohort is population-based and not selected based on radiological findings, it enables an unbiased investigation of the relationship between brain

structure and gait. This reduces the risk of information leakage or selection bias introduced by imaging-based assessments. The uniqueness of the cohort limits direct benchmarking against other studies. However, a previous study shows associations between gait speed impairment and disproportion between ventricular and sulcal CSF volumes [73] and the INPH Radscale was developed on a population-based cohort, showing a strong correlation between INPH symptoms and Radscale score [38]. In INPH patients, one study has reported associations between gait impairment and white matter hyperintensities in specific white matter tracts [74]. In addition, specific gait patterns in INPH have been associated with the height of the third ventricle [75]. These findings support our conclusion.

The developed model was never intended to be part of a clinical investigation, but rather to assess the relationship between brain structure and symptoms. This is relevant since HLGD is often the earliest symptom in patients with INPH [31], and the gait disturbance is the symptom that most often improves after surgery [2]. Further investigation of this relationship could benefit from the use of explainable AI (XAI) methods [76] which might reveal what brain features the networks rely on when making predictions, i.e. which features are seemingly related to HLGD. In addition, methods that could be used to distinguish brain changes related to gait disturbance from those associated with normal aging could provide further insights [77]. Such approaches may contribute to a better understanding of the underlying mechanisms of INPH.

7.1.2 Shunt outcome prediction in INPH

In Paper III, shunt outcome could not be predicted with sufficient accuracy to be clinically useful by the previously developed models, CNNs, or radiomics-based ML, even though additional MRI sequences were included compared with Paper I. However, a statistical association between brain morphology and shunt outcome was observed in the chi-squared analysis, indicating that structural MRI may contain information related to shunt response, although not enough to support reliable outcome prediction.

The findings highlight that outcome prediction is a complex task that puts high demands on both data quality and model generalizability. This further suggests that structural MRI alone may provide limited prognostic information and that shunt response is influenced by other factors that are not visible on structural MRI. The shunt outcome

prediction performance did not improve with larger group separation in the post-hoc analysis, which implies that the cohort size, rather than the class imbalance, is a limiting factor. This may indicate that the available dataset is too small to capture subtle morphological patterns associated with treatment response, and that additional clinical or physiological information may be necessary to improve prediction performance.

In the previous study by Leary et al [71], who inspired one of our competitor models, an AUROC of 88 % was presented for shunt prediction. This result was indeed promising, and our version of their proposed network architecture performed among the best on the full dataset though not as well as in the previous study or sufficiently well to be clinically useful. The difference in performance can be due to several reasons, such as cohort size and image quality, but the main reason is probably the clinical investigations performed before patients were selected for shunt surgery. While the Leary study selected patients based on clinical assessment and radiological findings, our patients underwent a more rigorous investigation with invasive testing. This means that we investigate whether the images contain additional information to improve shunt selection, when physiological tests have been conducted. It is reasonable to think that this is a more challenging task. The outcome measures also differed between the studies, where Leary et al based their improvement on the modified Rankin scale [78] or the Hellström INPH scale [79], instead of gait speed improvement. This INPH scale was presented after the start of the data collection of our cohort, which is why it was not possible to include in our study. Otherwise, it would have been interesting to train the networks with this outcome measure, to assess how it would affect the results. Our outcome metric, maximum gait speed is a suitable metric for this task since gait and balance are the only symptoms that have been shown to improve in a randomized, double-blinded study [2]. Furthermore, gait speed is an objective and quantitative measure, facilitating comparison between other cohorts.

In terms of other studies that have attempted to predict shunt outcome based on imaging data, one study used conventional ML approaches on clinical and radiological data, obtaining an AUROC of 80% [80]. This indicates that combining imaging with additional clinical information may improve prediction performance. Other studies attempting this task use lumbar drainage response as the outcome measure [81,82], which complicates comparison with our study since this measure is not equivalent to shunt response [83]. The benefit of using lumbar drainage is that the cohorts may be larger, since more of the patients evaluated for

shunt surgery can be included, instead of only the ones undergoing shunt surgery. This can lead to more balanced datasets, since the non-responder group is likely to be larger. The variability in the groups might also increase, since the risk connected to shunt surgery might exclude certain patient groups. However, drainage tests are themselves imperfect predictors of shunt outcome [28,83], meaning that models trained on this outcome are effectively learning from an uncertain ground truth. In contrast, the present study focuses on actual postoperative improvement, which provides a more direct measure of treatment response, although at the cost of smaller cohorts. Future progress in this area will likely require larger multicentre datasets with standardized definitions of shunt outcome.

7.2 Auto-qDESH

Auto-qDESH showed a good to excellent [84] agreement with the semi-automatic qDESH method. However, it did not demonstrate the same level of agreement as the two raters using the qDESH method, with the largest discrepancies observed for large qDESH values. When comparing auto-qDESH with visual DESH ratings, a significant difference in auto-qDESH values was observed between subjects without DESH and those with mild-moderate or severe DESH. This suggests that the method can distinguish between the presence and absence of DESH, even though it compared to the semi-automatic qDESH was less accurate in capturing the exact severity [41]. From a clinical perspective, DESH present or absent is likely the most important distinction regarding DESH, since that is what is considered in the INPH guidelines [10].

Though the auto-qDESH algorithm worked well it tended to underestimate the CSF volumes included in the metric. This limitation mainly arises from the estimation of the dural border. In the semi-automatic qDESH method, the border is manually delineated, whereas in the auto-qDESH method it is derived from the brain segmentation using SynthSeg. Differentiating CSF from bone or the dural border on T1-w MRI is challenging. Therefore, incorporating T2-w MRI into the analysis could improve the estimation of the dural border and thereby increase the accuracy of auto-qDESH. Since acquiring both T1-w and T2-w MRI is common in an INPH investigation, the method would still be feasible for clinical application.

A major advantage of auto-qDESH is the reduction in manual workload. While the semi-automatic qDESH method requires approximately 10–15

minutes of manual work per subject, auto-qDESH performs the analysis automatically, and is user independent. This facilitates the investigation of larger cohorts and supports future clinical implementations.

The search for quantitative and automated radiological markers is ongoing. There are several studies presenting methods that automate the already used metrics, such as EI [85,86], and callosal angle [87], and alternative approaches to qDESH are also presented. The Silver Index is one example, computed through manual delineation of CSF spaces in a 2D-slice [88]. Another example is the DESI index, a volumetric metric computed through a fully automatic U-net based software [89]. Auto-qDESH is both fully automated and volume-based, while maintaining transparency in its computations rather than relying on a black-box model. Another strength is that a semi-automatic version of qDESH exists, which can be used in cases where the automated pipeline fails. In addition, the software used to calculate auto-qDESH, once preprocessing is completed, is lightweight and computationally efficient, which further supports its practical use.

7.3 Development of search algorithms

The classification tasks investigated in this thesis are complex since we do not know which imaging patterns that influence the outcome. Approaches such as ensemble models and multimodal imaging data may therefore be beneficial, as they can capture complementary information. However, determining how to optimally combine networks or imaging modalities is not trivial. Search algorithms were therefore explored as a data-driven approach to identify effective ensemble compositions and multimodal fusion strategies.

7.3.1 Ensemble search

The performance of predicting HLGD versus controls was improved by the ensemble model found by the ensemble search algorithm developed in Paper I. The ensemble, chosen based on both evaluation and diversity metrics, performed better than the ensemble based on only evaluation metrics. These findings highlight the importance of diversity in the search algorithm, suggesting that this contributes to improved performance.

The ensemble search showed good performance in Paper I but did not reach highest performance on the full dataset in Paper III. This could be an indication that the cohort size was too small, creating large

differences between the cross-validation folds. The search algorithm is optimized on the validation set, but if this set does not accurately reflect the data distribution in the test set, the chosen ensemble may perform poorly. This can occur particularly when the dataset is small, as the composition of individual folds may vary substantially. In such cases, the search algorithm may favour ensembles that perform well on specific characteristics present in the validation set but that are not representative of the overall dataset. Consequently, the selected ensemble may capture patterns that do not generalize well, leading to reduced performance on the independent test set.

The importance of diversity between models is not something new [90], but it can be approached in different ways. Bagging (or bootstrapping) is a common way of creating diverse models, but is infeasible when the data set is small [91]. Another common approach is to pick one architecture and train it multiple times, initialized with different random seeds or other hyperparameters [92]. However, this raises the question of *which* architecture to choose, that may not always be straightforward. The ensemble algorithm presented here is one approach to facilitate this. The different aggregation functions did not alter the ensemble performance in Paper I, indicating that the selection of which networks to combine was more important than how the predictions were aggregated. This could suggest that the individual CNNs produced reasonably calibrated probabilities, assigning lower confidence to predictions when the models were less certain. As a result, the ensemble predictions were relatively insensitive to the specific aggregation function among those evaluated here.

Ensemble models are used in medical imaging research, but search algorithms for ensemble selection do not appear to be widely used. The approach proposed in this thesis therefore contributes with a simple method for identifying optimal ensemble combinations. Ensembles are more often manually designed than identified through systematic search [93]. In the broader AI field, studies have explored searching for or pruning ensembles based on diversity [94]. One advantage of the ensemble search method presented in this thesis is its simplicity and ease of implementation which means it can be used in other similar tasks in the future.

7.3.2 Sequential Fusion Search Algorithm

The model configuration decided by the SFSA, developed in Paper II, provided the best results on the two classification tasks, while

substantially reducing computational cost, compared to the brute force method. This indicates that guided fusion can capture complementary multimodal information without increasing the computational burden.

Apart from the computational burden, one difference between the brute-force approach and the model identified by the SFSA is that, in the SFSA, the weights are retained from previous training when a new fusion module is activated. This type of progressive training may facilitate optimization, as the weights can adapt to each sequence before more complex multimodal representations are introduced.

In medical brain imaging, relatively few studies have investigated multimodal network architecture search for classification tasks [95]. Among the limited studies, one study [96] searched the modality-specific feature extraction networks, but it did not explicitly search the fusion timing or fusion strategy. Existing approaches surrounding this topic involve fusing complementary information from multiple medical imaging modalities into a single image [97,98]. Another example [99] targeted searching for the optimal architecture but it is not a multimodal framework. Outside the medical field, there are approaches which investigate more complex ways of combining modalities [100,101]. Our method focuses specifically on identifying the optimal fusion timing, i.e., where the fusion between sequence types should occur. By limiting the search to four fusion modules, the proposed SFSA substantially reduces the complexity of the search while still providing good performance. Overall, these works indicate that multimodal network architecture search for brain-image classification remains an interesting research area.

The work in Paper II was conducted to compare fusion strategies, rather than maximizing performance on the task. Nonetheless, the performance is similar to another study on the OASIS-3 dataset when using only MRI data [102]. Interestingly, a study using only T2-w images achieved an accuracy of 91 %, but they used substantially more images since they did not need to consider the multimodal aspect. The epilepsy cohort also includes segmentations of lesions, which has made it suitable for segmentation tasks [103,104], but one study, using a subset of patients and controls, achieved an accuracy of 97.5 % when trained on 2D image slices from the 3D MRI volumes [105]. The downside of this methodology is that not all slices can be used, and it requires manual selection of anatomically appropriate slices. Our results should however primarily be interpreted in relation to the methodological aim of the study.

7.4 Methodological considerations

Even though the ensemble search algorithm was implemented for specific medical brain imaging tasks in this thesis, the algorithm is generalizable to a wide range of classification tasks and model architectures. Depending on the task, the cost function can be defined using different evaluation and diversity metrics. For example, in some applications specificity may be particularly important, or certain diversity metrics may be more suitable for multilabel classification. Similarly, the SFSA is not restricted to specific input types or network architectures. Rather, it is a methodology that can be integrated into multimodal networks with predefined fusion points and arbitrary fusion modules. The approaches developed in this work may therefore also be relevant for other neurological disorders in which MRI is central to diagnosis, suggesting a broader applicability of the proposed methods within neuroimaging research.

A limitation with the SFSA is that the fusion points must be predefined. There might be better places to fuse the data, that are not explored with this search algorithm. However, in Paper II all dual-sequence models showed good performance, highlighting the advantage of using multiple MRI sequences when available.

In this thesis, the focus was not on extensive hyperparameter tuning of a single model. Instead, many different network architectures were trained. The intention was to introduce diversity through different network designs and allow different CNNs to capture complementary information. If the results in Paper III had been substantially stronger, further exploration of hyperparameter tuning and architecture optimization would have been motivated. With the current results, and given the large number of models already evaluated, such optimization would likely not have substantially changed the results.

Expected anatomical inter-individual variations unrelated to shunt outcome may have interfered with relevant information during training, particularly as the dataset in Paper III was limited. As the cohort is retrospective, different scanners and imaging protocols were used. These differences may also have affected the networks, even though the images underwent the same preprocessing. While variability in the dataset reflects real clinical conditions, the relatively small number of subjects in each cross-validation fold means that an unfavourable distribution of outliers, either in terms of brain morphology or imaging-related factors, may still have affected the results. In addition, the datasets in both

Papers I and III were imbalanced. This was handled through oversampling during training and a weighted loss function, but the risk of not capturing the full variability of the data, particularly in the smaller group, remains. A larger cohort could have been obtained in Paper I by including individuals with “other neurological reasons for gait impairment”, either as part of a general gait disorder group or as a separate third group. However, since the aim was not general gait classification, but understanding the relationship between brain structure and HLGD, the first option would have altered the research question. In the second case, the HLGD group would have become proportionally even smaller, and therefore would probably not have improved performance.

For the auto-qDESH pipeline, openly available tools such as *acpcdetect* and SynthSeg were used. These tools have been validated across many studies and using established softwares facilitates reproducibility and enables other researchers to apply the same pipeline. After further development of the dural border estimation, the aim is to make the pipeline openly available to the research community. The intention is to implement the finalized pipeline in a user-friendly manner, where the user only needs to select a file or folder containing MRI volumes and the analysis (from pre-processing to computation) is then performed automatically. This would lower the threshold for clinically oriented users to apply the method in research settings.

7.5 Future outlook

In this thesis, several approaches have been presented to analyse and assess brain MRI. With increasing computational power and improved image analysis algorithms and models, such methods are likely to become more accessible for both research centres and clinics.

A major limitation in AI research, particularly in medical imaging, is the lack of large datasets. Many models are benchmarked on relatively simple datasets, whereas complex diseases such as INPH likely require substantially larger cohorts to capture the full variability of the disorder. To achieve this, datasets collected across multiple clinical centres will likely be necessary. Future AI-approaches for INPH should also include multimodal data. Imaging could be combined with both clinical and physiological measures, such as duration of symptoms, cognitive level and CSF outflow resistance, as well as CSF biomarkers [106,107]. This

would likely better capture the complexity of the condition and thereby improve the potential for predicting shunt response.

Standardized data collection protocols would facilitate such efforts. Ideally INPH investigations should include consistent MRI sequences, standardized clinical assessments, and comparable time intervals for evaluation before and after shunting. A wider range of quantitative outcome measures should also be consistently recorded. Several initiatives already aim to standardize INPH assessment, including the INPH scale and INPH Radscale. The development of automatically computed metrics such as auto-qDESH also represents a step towards this standardized analysis.

The auto-qDESH pipeline already shows sufficient reliability for use in research cohorts where analyses are performed at the group level rather than for individual clinical decisions. Even though DESH is a commonly used metric to assess INPH, other measures should be adopted into the pipeline for best clinical use. Volumetric measures of both cortical regions and CSF spaces are easily extracted from the segmentation process, and have shown potential for shunt prediction [82], and could therefore be included. Other metrics from the INPH Radscale could also be added, such as callosal angle or width of the temporal horns. Further improvements of the algorithm could involve evaluating alternative segmentation approaches in addition to SynthSeg, for example methods using more age-specific atlases [108], or alternative techniques for identifying the AC-PC landmarks [109], which could improve stability and segmentation accuracy. In particular, improving the delineation of the dural border should be a priority, as this remains a key source of uncertainty in the current pipeline. With such modifications, and validation on additional datasets, the pipeline can hopefully be incorporated into future clinical workflows.

Finally, even though none of the models presented in this thesis were able to predict shunt outcome based only on MRI, this type of analysis remains valuable. Applying neural networks to investigate whether patterns in the data are associated with clinical outcomes can serve as an exploratory step. If such relationships are detected, they can guide further research into the underlying mechanisms and inform the development of more advanced predictive models.

8 Conclusions

This thesis demonstrates that MRI contains meaningful information related to HLGD that can be assessed by CNN classifiers, applied to brain MRI. This motivates further investigations on INPH cohorts since HLGD is the gait impairment typically seen in patients with INPH.

The fact that both the DL ensemble model and the multimodal model, identified by the search algorithms developed within this thesis, showed improved performance highlights the benefit of a guided search for the optimal network configuration. The efficient search strategies enabled these improvements with a relatively low computational cost, which supports their practical applicability in clinical imaging studies.

Despite these advances, MRI alone was insufficient for reliable prediction of shunt outcome for INPH patients. Neither DL-based nor radiomics-based ML models were able to achieve clinically useful performance, suggesting that treatment response is influenced by additional clinical markers that are not captured by structural brain MRI. Future studies should be conducted with clinical and physiological data, in addition to MRI.

Complementing the AI-based analyses, automated calculation of the quantitative measure qDESH was shown to be feasible for research cohorts, which supports future objective and scalable investigations of INPH.

Overall, the methodological developments and findings of this work demonstrate both the potential and the current limitations of using AI-based and automated MRI analysis in INPH for shunt decision support.

9 Acknowledgements

When I was admitted to this PhD programme, the four studies originally planned were not the same as the ones that ended up in this thesis. It has been an eventful time, and by my side I have always had my main supervisor, Sara Qvarlander. The door to your office has always been open for my thoughts and questions, and you have been very helpful and flexible regarding both my wishes to work abroad, and my occasionally odd working hours. Thank you for your time, dedication and support. I've also had big support from Anders Eklund, improving my critical thinking and writing skills. It is always inspiring talking with you, and I've learnt a lot. Jan Malm has been the medical expert during these years, thank you for your time, knowledge, and honest answers. Imaging questions have always gone straight to Anders Wåhlin, your way of tackling the problems that arise is inspiring. After a year I got an additional co-supervisor, Paolo Soda. Thank you for your valuable input in the AI field and for welcoming me to your lab in Rome.

I feel very fortunate to be part of such a large, supportive, and friendly research group, and to have such helpful co-authors. Thank you all so much. I especially would like to thank: Jenny and William, for the work during the first years with the VESPR cohort. Cicci and Tomas, for handing me scripts, computing resources, and all the laughter together. Axel, for sharing good moments and thesis anxiety together. Sofia, for being an encyclopaedia regarding the different cohorts I've used. Pontus and Arvid, for the time shared in the PhD room. Andrea, for making my home office a shared office space sometimes. Johan, Afroditi, and Lars-Owe, for being supportive co-authors. Emil, Nina, and Tommy, for reading this thesis and helping me improve it. Grazie mille, Mirri, Saris, Lorre, Filippo and Giulia for coming and sharing your energy.

Petter, having a reference person may be mandatory, but our lunches and phone calls have really made a difference, thank you. I'd also like to thank the staff at the department. You have taken good care of me and have been quick with answers when time has been limited. A big thank you to all my colleagues at MT-FoU, you are all very clever and friendly.

During two of my years I've been part of the Human Brain Clearance Imaging consortium and would like to thank all its members, for the interesting discussions and all well-planned visits. I have also spent several months at the unit of Artificial Intelligence and Computer

Systems at Campus Bio-Medico University of Rome. Thank you, for being so welcoming during my time with you. And, Vito, for showing me Rome, and Daniele, for giving me a home there. Many lunches have been spent at Britta's lunchroom over the years. Thank you to everyone there for the happy moments.

There are several pairs of parents I would like to thank. First, my biological ones, Pi & Karin, for always being there when I need you. Thank you to Emanuela & Camillo for being my extra parents during my stay in Rome, and to Lisen & Leon, and Bettan & Lillen, for your support and kindness. To my mormor, siblings, nieces, nephews, and extended family, you are all amazing people, and I am grateful to be part of such a creative and generous pack. Special thanks to Tuva, Lovis, Jonas, Stina, and Johanna for the help with the cover of this book.

This period would not have been as easy without the help from the people around me. I would like to thank: My neighbours, Nikanor & Ebba, for reminding me when to stop working on Fridays. Andre, for cooking dinners when my energy's been out. Alina, for stepping in when my dog-walking possibilities been limited. Tuve, Tove, and Linde, for keeping the Sjövik spirit close. The Umeå Student Choir and Sworn by the Horn, especially Alex, Rikard, Kajsa, Alva, and Klarisen, for good moments and lots of good music. Sandra and Kicki, for always being just a phone call away.

Valle, I could not have asked for a better companion along the way. This thesis would certainly not have been the same without you. I am so grateful that we got to know each other through it.

Finally, thank you to my little family. Britta, for being with me from the admission seminar until today, illustrating my thesis and presentations, and always supporting me. Markus, for being so thoughtful and generous, for taking care of the whole household at times, and for reminding me that research is not all there is. Shanti, Grayson & Sonic, for making our household better. Sören has definitely not made writing this thesis easier, but you deserve a mention too.

10 Funding

This work was supported by the Swedish Foundation for Strategic Research, grant number RMX18-0152, the Swedish Research Council, grant number 2021-00711_VR/JPND, and a regional agreement between Umeå University and Region Västerbotten (ALF).

Computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreements number 2022-06725 and 2018-05973.

Travel was supported by the C.M. Leric Foundation, the Kempe foundation and travel funds from the Faculty of Medicine at Umeå University.

11 References

1. Toma AK, Papadopoulos MC, Stapleton S, Kitchen ND, Watkins LD. Systematic review of the outcome of shunt surgery in idiopathic normal-pressure hydrocephalus. *Acta Neurochir.* 2013;155: 1977–1980. doi:10.1007/s00701-013-1835-5
2. Luciano MG, Williams MA, Hamilton MG, Katzen HL, Dasher NA, Moghekar A, et al. A Randomized Trial of Shunting for Idiopathic Normal-Pressure Hydrocephalus. *N Engl J Med.* 2025;393: 2198–2209. doi:10.1056/NEJMoa2503109
3. Andersson J, Rosell M, Kockum K, Lilja-Lund O, Söderström L, Laurell K. Prevalence of idiopathic normal pressure hydrocephalus: A prospective, population-based study. Burger MC, editor. *PLoS ONE.* 2019;14: e0217705. doi:10.1371/journal.pone.0217705
4. Constantinescu C, Wikkelsø C, Westman E, Ziegelitz D, Jaraj D, Rydén L, et al. Prevalence of Possible Idiopathic Normal Pressure Hydrocephalus in Sweden: A Population-Based MRI Study in 791 70-Year-Old Participants. *Neurology.* 2024;102: e208037. doi:10.1212/WNL.0000000000208037
5. Iseki C, Takahashi Y, Adachi M, Igari R, Sato H, Koyama S, et al. Prevalence and development of idiopathic normal pressure hydrocephalus: A 16-year longitudinal study in Japan. *Acta Neurol Scand.* 2022;146: 680–689. doi:10.1111/ane.13710
6. Sundström N, Malm J, Laurell K, Lundin F, Kahlon B, Cesarini KG, et al. Incidence and outcome of surgery for adult hydrocephalus patients in Sweden. *British Journal of Neurosurgery.* 2017;31: 21–27. doi:10.1080/02688697.2016.1229749
7. Mercado-Diaz LR, Prakash N, Gong GX, Posada-Quintero HF. Artificial Intelligence Approaches for the Detection of Normal Pressure Hydrocephalus: A Systematic Review. *Applied Sciences.* 2025;15: 3653. doi:10.3390/app15073653
8. Malm J, Eklund A. Idiopathic normal pressure hydrocephalus. *Pract Neurol.* 2006;6: 14–27. doi:10.1136/jnnp.2006.088351
9. Jacobsson J, Qvarlander S, Eklund A, Malm J. Comparison of the CSF dynamics between patients with idiopathic normal pressure

- hydrocephalus and healthy volunteers. *Journal of Neurosurgery*. 2019;131: 1018–1023. doi:10.3171/2018.5.JNS173170
10. Nakajima M, Yamada S, Miyajima M, Ishii K, Kuriyama N, Kazui H, et al. Guidelines for Management of Idiopathic Normal Pressure Hydrocephalus (Third Edition): Endorsed by the Japanese Society of Normal Pressure Hydrocephalus. *Neurol Med Chir (Tokyo)*. 2021;61: 63–97. doi:10.2176/nmc.st.2020-0292
 11. Relkin N, Marmarou A, Klinge P, Bergsneider M, Black PMcL. Diagnosing Idiopathic Normal-pressure Hydrocephalus. *Neurosurgery*. 2005;57: S2-4-S2-16. doi:10.1227/01.NEU.0000168185.29659.C5
 12. Virhammar J, Laurell K, Cesarini KG, Larsson E-M. Preoperative Prognostic Value of MRI Findings in 108 Patients with Idiopathic Normal Pressure Hydrocephalus. *AJNR Am J Neuroradiol*. 2014;35: 2311–2318. doi:10.3174/ajnr.A4046
 13. Hashimoto M, Mori E, Kuwana N, The study of INPH on neurological improvement (SINPHONI). Diagnosis of idiopathic normal pressure hydrocephalus is supported by MRI-based scheme: a prospective cohort study. *Fluids Barriers CNS*. 2010;7: 18. doi:10.1186/1743-8454-7-18
 14. Mihalj M, Dolić K, Kolić K, Ledenko V. CSF tap test — Obsolete or appropriate test for predicting shunt responsiveness? A systemic review. *Journal of the Neurological Sciences*. 2016;362: 78–84. doi:10.1016/j.jns.2016.01.028
 15. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*. 2017;42: 60–88. doi:10.1016/j.media.2017.07.005
 16. Sakka L, Coll G, Chazal J. Anatomy and physiology of cerebrospinal fluid. *European Annals of Otorhinolaryngology, Head and Neck Diseases*. 2011;128: 309–316. doi:10.1016/j.anorl.2011.03.002
 17. Iliff JJ, Wang M, Liao Y, Plogg BA, Peng W, Gundersen GA, et al. A Paravascular Pathway Facilitates CSF Flow Through the Brain Parenchyma and the Clearance of Interstitial Solutes, Including Amyloid β . *Sci Transl Med*. 2012;4. doi:10.1126/scitranslmed.3003748

18. Yamada S, Otani T, Ii S, Kawano H, Nozaki K, Wada S, et al. Aging-related volume changes in the brain and cerebrospinal fluid using artificial intelligence-automated segmentation. *Eur Radiol.* 2023;33: 7099–7112. doi:10.1007/s00330-023-09632-x
19. Qvarlander S, Sundström N, Malm J, Eklund A. CSF formation rate—a potential glymphatic flow parameter in hydrocephalus? *Fluids Barriers CNS.* 2024;21: 55. doi:10.1186/s12987-024-00560-6
20. Ekstedt J. CSF hydrodynamic studies in man. 2. Normal hydrodynamic variables related to CSF pressure and flow. *Journal of Neurology, Neurosurgery & Psychiatry.* 1978;41: 345–353. doi:10.1136/jnnp.41.4.345
21. Vinje V, Ringstad G, Lindstrøm EK, Valnes LM, Rognes ME, Eide PK, et al. Respiratory influence on cerebrospinal fluid flow – a computational study based on long-term intracranial pressure measurements. *Sci Rep.* 2019;9: 9732. doi:10.1038/s41598-019-46055-5
22. Nedergaard M, Goldman SA. Glymphatic failure as a final common pathway to dementia. *Science.* 2020;370: 50–56. doi:10.1126/science.abb8739
23. Malm J, Jacobsson J, Birgander R, Eklund A. Reference values for CSF outflow resistance and intracranial pressure in healthy elderly. *Neurology.* 2011;76: 903–909. doi:10.1212/WNL.0b013e31820f2ddo
24. Giordan E, Palandri G, Lanzino G, Murad MH, Elder BD. Outcomes and complications of different surgical treatments for idiopathic normal pressure hydrocephalus: a systematic review and meta-analysis. *Journal of Neurosurgery.* 2019;131: 1024–1036. doi:10.3171/2018.5.JNS1875
25. Malm J, Graff-Radford NR, Ishikawa M, Kristensen B, Leinonen V, Mori E, et al. Influence of comorbidities in idiopathic normal pressure hydrocephalus — research and clinical care. A report of the ISHCSF task force on comorbidities in INPH. *Fluids Barriers CNS.* 2013;10: 22. doi:10.1186/2045-8118-10-22
26. Klinge P, Hellström P, Tans J, Wikkelsø C, On behalf of the European iNPH Multicentre Study Group. One-year outcome in the

- European multicentre study on iNPH. *Acta Neurol Scand.* 2012;126: 145–153. doi:10.1111/j.1600-0404.2012.01676.x
27. Gasslander J, Sundström N, Eklund A, Koskinen L-OD, Malm J. Risk factors for developing subdural hematoma: a registry-based study in 1457 patients with shunted idiopathic normal pressure hydrocephalus. *Journal of Neurosurgery.* 2021;134: 668–677. doi:10.3171/2019.10.JNS191223
 28. Marmarou A, Bergsneider M, Klinge P, Relkin N, Black PMcL. The Value of Supplemental Prognostic Tests for the Preoperative Assessment of Idiopathic Normal-pressure Hydrocephalus. *Neurosurgery.* 2005;57: S2-17-S2-28. doi:10.1227/01.NEU.0000168184.01002.60
 29. Thavarajasingam SG, El-Khatib M, Rea M, Russo S, Lemcke J, Al-Nusair L, et al. Clinical predictors of shunt response in the diagnosis and treatment of idiopathic normal pressure hydrocephalus: a systematic review and meta-analysis. *Acta Neurochir.* 2021;163: 2641–2672. doi:10.1007/s00701-021-04922-z
 30. Thavarajasingam SG, El-Khatib M, Vemulapalli K, Iradukunda HAS, K. SV, Borchert R, et al. Radiological predictors of shunt response in the diagnosis and treatment of idiopathic normal pressure hydrocephalus: a systematic review and meta-analysis. *Acta Neurochir.* 2022;165: 369–419. doi:10.1007/s00701-022-05402-8
 31. Andrén K, Wikkelsø C, Laurell K, Kollén L, Hellström P, Tullberg M. Symptoms and signs did not predict outcome after surgery: a prospective study of 143 patients with idiopathic normal pressure hydrocephalus. *J Neurol.* 2024;271: 3215–3226. doi:10.1007/s00415-024-12248-w
 32. Nutt JG, Marsden CD, Thompson PD. Human walking and higher-level gait disorders, particularly in the elderly. *Neurology.* 1993;43: 268–268. doi:10.1212/WNL.43.2.268
 33. Nutt JG. Higher-level gait disorders: An open frontier. *Movement Disorders.* 2013;28: 1560–1565. doi:10.1002/mds.25673
 34. Evans WA. An encephalographic ratio for estimating ventricular enlargement and cerebral atrophy. *Archives of Neurology &*

- Psychiatry. 1942;47: 931–937.
doi:10.1001/archneurpsyc.1942.02290060069004
35. Yamada S, Ishikawa M, Yamamoto K. Optimal Diagnostic Indices for Idiopathic Normal Pressure Hydrocephalus Based on the 3D Quantitative Volumetric Analysis for the Cerebral Ventricle and Subarachnoid Space. *AJNR Am J Neuroradiol.* 2015;36: 2262–2269. doi:10.3174/ajnr.A4440
 36. Ambarki K, Israelsson H, Wåhlin A, Birgander R, Eklund A, Malm J. Brain Ventricular Size in Healthy Elderly: Comparison Between Evans Index and Volume Measurement. *Neurosurgery.* 2010;67: 94–99. doi:10.1227/01.NEU.0000370939.30003.D1
 37. Ishii K, Kanda T, Harada A, Miyamoto N, Kawaguchi T, Shimada K, et al. Clinical impact of the callosal angle in the diagnosis of idiopathic normal pressure hydrocephalus. *Eur Radiol.* 2008;18: 2678–2683. doi:10.1007/s00330-008-1044-4
 38. Kockum K, Lilja-Lund O, Larsson E -M., Rosell M, Söderström L, Virhammar J, et al. The idiopathic normal-pressure hydrocephalus Radscale: a radiological scale for structured evaluation. *Eur J Neurol.* 2018;25: 569–576. doi:10.1111/ene.13555
 39. Whitley H, Skalický P, Zazay A, Bubeníková A, Bradáč O. Volumes and velocities: Meta-analysis of PC-MRI studies in normal pressure hydrocephalus. *Acta Neurochir.* 2024;166: 463. doi:10.1007/s00701-024-06333-2
 40. Lalou AD, Bryngelsson AV, Wåhlin A, Asplund P, Larsson JM, Qvarlander S. No Predictive Value of Aqueduct CSF Flow Dynamics for Shunt Response in Idiopathic Normal Pressure Hydrocephalus. *In Review;* 2025. doi:10.21203/rs.3.rs-7400171/v1
 41. Behndig S, Lalou A, Axelsson J, Larsson J, Wåhlin A, Ryska P, et al. qDESH: a method to quantify disproportionately enlarged subarachnoid space hydrocephalus. *Fluids Barriers CNS.* 2025;22: 67. doi:10.1186/s12987-025-00677-2
 42. Krauss JK, Droste DW, Vach W, Regel JP, Orszagh M, Borremans JJ, et al. Cerebrospinal Fluid Shunting in Idiopathic Normal-Pressure Hydrocephalus of the Elderly: Effect of Periventricular and Deep White Matter Lesions. *Neurosurgery.* 1996;39: 292–300.

43. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Kötter R, editor. *Phil Trans R Soc Lond B*. 2001;356: 1293–1322. doi:10.1098/rstb.2001.0915
44. Fischl B. FreeSurfer. *NeuroImage*. 2012;62: 774–781. doi:10.1016/j.neuroimage.2012.01.021
45. Ashburner J. SPM12 Manual. Wellcome Trust Centre for Neuroimaging; 2021.
46. Ashburner J, Friston KJ. Unified segmentation. *NeuroImage*. 2005;26: 839–851. doi:10.1016/j.neuroimage.2005.02.018
47. Billot B, Greve DN, Puonti O, Thielscher A, Van Leemput K, Fischl B, et al. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*. 2023;86: 102789. doi:10.1016/j.media.2023.102789
48. Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*. 2020;219: 117012. doi:10.1016/j.neuroimage.2020.117012
49. Talaei Khoei T, Kaabouch N. Machine Learning: Models, Challenges, and Research Directions. *Future Internet*. 2023;15: 332. doi:10.3390/fi15100332
50. Chen X, Wang X, Zhang K, Fung K-M, Thai TC, Moore K, et al. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*. 2022;79: 102444. doi:10.1016/j.media.2022.102444
51. Ahmed SF, Alam MdSB, Hassan M, Rozbu MR, Ishtiaq T, Rafa N, et al. Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artif Intell Rev*. 2023;56: 13521–13617. doi:10.1007/s10462-023-10466-8
52. Taye MM. Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. *Computation*. 2023;11: 52. doi:10.3390/computation11030052

53. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*. 2022;115: 105151. doi:10.1016/j.engappai.2022.105151
54. Hansen LK, Salamon P. Neural network ensembles. *IEEE Trans Pattern Anal Machine Intell*. 1990;12: 993–1001. doi:10.1109/34.58871
55. Ju C, Bibaut A, Van Der Laan M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*. 2018;45: 2800–2818. doi:10.1080/02664763.2018.1441383
56. Larsson J, Hansson W, Israelsson Larsen H, Koskinen L-OD, Eklund A, Malm J. Higher-level gait disorders: a population-based study on prevalence, quality of life, depression and confidence in gait and balance. *BMJ Neurol Open*. 2025;7: e000992. doi:10.1136/bmjno-2024-000992
57. Tilson JK, Sullivan KJ, Cen SY, Rose DK, Koradia CH, Azen SP, et al. Meaningful Gait Speed Improvement During the First 60 Days Poststroke: Minimal Clinically Important Difference. *Physical Therapy*. 2010;90: 196–208. doi:10.2522/ptj.20090079
58. Ardekani BA, Bachman AH. Model-based automatic detection of the anterior and posterior commissures on MRI scans. *NeuroImage*. 2009;46: 677–682. doi:10.1016/j.neuroimage.2009.02.030
59. Itzcovich I. DeepBrain. Available: <https://github.com/iitco/deepbrain>
60. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE; 2017. pp. 2261–2269. doi:10.1109/CVPR.2017.243
61. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016. pp. 770–778. doi:10.1109/CVPR.2016.90
62. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern

- Recognition. Salt Lake City, UT: IEEE; 2018. pp. 7132–7141. doi:10.1109/CVPR.2018.00745
63. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv; 2020. doi:10.48550/arXiv.1905.11946
 64. Peng H, Gong W, Beckmann CF, Vedaldi A, Smith SM. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*. 2021;68: 101871. doi:10.1016/j.media.2020.101871
 65. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE; 2009. pp. 248–255. doi:10.1109/CVPR.2009.5206848
 66. Chen S, Ma K, Zheng Y. Med3D: Transfer Learning for 3D Medical Image Analysis. arXiv; 2019. doi:10.48550/arXiv.1904.00625
 67. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*. 2016;19: 1523–1536. doi:10.1038/nn.4393
 68. Mogensen K, Guarrasi V, Larsson J, Hansson W, Wåhlin A, Koskinen L-O, et al. An optimized ensemble search approach for classification of higher-level gait disorder using brain magnetic resonance images. *Computers in Biology and Medicine*. 2025;184: 109457. doi:10.1016/j.combiomed.2024.109457
 69. Vaezi Joze HR, Shaban A, Iuzzolino ML, Koishida K. MMTM: Multimodal Transfer Module for CNN Fusion. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE; 2020. pp. 13286–13296. doi:10.1109/CVPR42600.2020.01330
 70. Larsson J. Population-based studies of higher-level gait disorders and hydrocephalus: focused on brain ventricular morphometry and patient outcomes following shunt surgery. Umeå: Department of Clinical Sciences, Neurosciences and Department of Radiation Sciences, Umeå University; 2022.
 71. Leary OP, Zhong Z, Bi L, Jiao Z, Dai Y-W, Ma K, et al. MRI-Based Prediction of Clinical Improvement after Ventricular Shunt

- Placement for Normal Pressure Hydrocephalus: Development and Evaluation of an Integrated Multisequence Machine Learning Algorithm. *AJNR Am J Neuroradiol*. 2024;45: 1536–1544. doi:10.3174/ajnr.A8372
72. Ryska P, Slezak O, Eklund A, Malm J, Salzer J, Zizka J. Radiological markers of idiopathic normal pressure hydrocephalus: Relative comparison of their diagnostic performance. *Journal of the Neurological Sciences*. 2020;408: 116581. doi:10.1016/j.jns.2019.116581
73. Palm WM, Saczynski JS, Van Der Grond J, Sigurdsson S, Kjartansson O, Jonsson PV, et al. Ventricular dilation: Association with gait and cognition. *Annals of Neurology*. 2009;66: 485–493. doi:10.1002/ana.21739
74. Tang Y, Yao Y, Xu S, Li X, Hu F, Wang H, et al. White Matter Microstructural Damage Associated With Gait Abnormalities in Idiopathic Normal Pressure Hydrocephalus. *Front Aging Neurosci*. 2021;13: 660621. doi:10.3389/fnagi.2021.660621
75. Nicolosi S, Todisco M, Paoletti M, Caverzasi E, Tarantino F, Ballante E, et al. Radiological features of gait phenotypes in patients with idiopathic normal pressure hydrocephalus. *Front Aging Neurosci*. 2025;17: 1554642. doi:10.3389/fnagi.2025.1554642
76. Zhang K, Wang D, Lin F, Xie J, Zhou W. A comprehensive review of explainable artificial intelligence in healthcare methods, evaluation, and clinical integration. *iScience*. 2026;29: 115026. doi:10.1016/j.isci.2026.115026
77. Fu J, Ferreira D, Smedby Ö, Moreno R. Decomposing the effect of normal aging and Alzheimer’s disease in brain morphological changes via learned aging templates. *Sci Rep*. 2025;15: 11813. doi:10.1038/s41598-025-96234-w
78. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJA, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke*. 1988;19: 604–607.
79. Hellström P, Klinge P, Tans J, Wikkelso C. A new scale for assessment of severity and outcome in iNPH. *Acta Neurol Scand*. 2012;126: 229–237. doi:10.1111/j.1600-0404.2012.01677.x

80. Sotoudeh H, Sadaatpour Z, Rezaei A, Shafaat O, Sotoudeh E, Tabatabaie M, et al. The Role of Machine Learning and Radiomics for Treatment Response Prediction in Idiopathic Normal Pressure Hydrocephalus. *Cureus*. 2021;13. doi:10.7759/cureus.18497
81. Lang S, Dimond D, Isaacs AM, Dronyk J, Vetkas A, Conner CR, et al. Use of cortical volume to predict response to temporary CSF drainage in patients with idiopathic normal pressure hydrocephalus. *Journal of Neurosurgery*. 2023;139: 1776–1783. doi:10.3171/2023.3.JNS222787
82. Wu D, Moghekar A, Shi W, Blitz AM, Mori S. Systematic volumetric analysis predicts response to CSF drainage and outcome to shunt surgery in idiopathic normal pressure hydrocephalus. *Eur Radiol*. 2021;31: 4972–4980. doi:10.1007/s00330-020-07531-z
83. Mahr CV, Dengl M, Nestler U, Reiss-Zimmermann M, Eichner G, Preuß M, et al. Idiopathic normal pressure hydrocephalus: diagnostic and predictive value of clinical testing, lumbar drainage, and CSF dynamics. *Journal of Neurosurgery*. 2016;125: 591–597. doi:10.3171/2015.8.JNS151112
84. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*. 2016;15: 155–163. doi:10.1016/j.jcm.2016.02.012
85. Barough SS, Bilgel M, Moghekar A, Ventura C, Luciano G, Moghekar A. Fully Automated Deep Learning-Based Pipeline for Evans Index Measurement from Raw 3D MRI. *MedRxiv*. 2025. doi:10.64898/2025.11.30.25341302
86. Wang Y, Feng A, Xue Y, Zuo L, Liu Y, Blitz AM, et al. Automated Ventricle Parcellation and Evan’s Ratio Computation in Pre- and Post-Surgical Ventriculomegaly. 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). Cartagena, Colombia: IEEE; 2023. pp. 1–5. doi:10.1109/ISBI53787.2023.10230729
87. Shirzadeh Barough S, Bilgel M, Ventura C, An L, Moghekar A, Albert MS, et al. Automated deep learning pipeline for callosal angle quantification. *Fluids Barriers CNS*. 2025;23: 17. doi:10.1186/s12987-025-00750-w
88. Benedetto N, Gambacciani C, Aquila F, Di Carlo DT, Morganti R, Perrini P. A new quantitative method to assess disproportionately

- enlarged subarachnoid space (DESH) in patients with possible idiopathic normal pressure hydrocephalus: The SILVER index. *Clinical Neurology and Neurosurgery*. 2017;158: 27–32. doi:10.1016/j.clineuro.2017.04.015
89. Barough SS, Ohno S, Bilgel M, Sair HI, Moghekar A. Disproportionately Elevated Sulcal Index (DESI): An automatically driven index representing disproportionate subarachnoid space enlargement in brain MRI scans. *MedRxiv*. 2025. doi:10.64898/2025.12.01.25341388
 90. Kuncheva LI, Whitaker CJ. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*. 2003;51: 181–207.
 91. Vu T, Braga-Neto U. Is Bagging Effective in the Classification of Small-Sample Genomic and Proteomic Data? *EURASIP J Bioinform Syst Biol*. 2009;2009: 158368. doi:10.1155/2009/158368
 92. Tanveer M, Rashid AH, Ganaie MA, Reza M, Razzak I, Hua K-L. Classification of Alzheimer’s Disease Using Ensemble of Deep Neural Networks Trained Through Transfer Learning. *IEEE J Biomed Health Inform*. 2022;26: 1453–1463. doi:10.1109/JBHI.2021.3083274
 93. Supriyadi MR, Samah ABA, Muliadi J, Awang RAR, Ismail NH, Majid HA, et al. A systematic literature review: exploring the challenges of ensemble model for medical imaging. *BMC Med Imaging*. 2025;25: 128. doi:10.1186/s12880-025-01667-4
 94. Chen M, Peng H, Fu J, Ling H. One-Shot Neural Ensemble Architecture Search by Diversity-Guided Search Space Shrinking. *arXiv*; 2021. doi:10.48550/arXiv.2104.00597
 95. Chaiyarin S, Rojbundit N, Piyabenjarad P, Limpitigranon P, Wisitthipakdeekul S, Nonthasaen P, et al. Neural architecture search for medicine: A survey. *Informatics in Medicine Unlocked*. 2024;50: 101565. doi:10.1016/j.imu.2024.101565
 96. Li T, Hou N, Yu J, Zhao Z, Sun Q, Chen M, et al. Evolutionary neural architecture search for automated MDD diagnosis using multimodal MRI imaging. *iScience*. 2024;27: 111020. doi:10.1016/j.isci.2024.111020

97. Ye S, Wang T, Ding M, Zhang X. F-DARTS: Foveated Differentiable Architecture Search Based Multimodal Medical Image Fusion. *IEEE Trans Med Imaging*. 2023;42: 3348–3361. doi:10.1109/TMI.2023.3283517
98. Mu P, Wu G, Liu J, Zhang Y, Fan X, Liu R. Learning to Search a Lightweight Generalized Network for Medical Image Fusion. *IEEE Trans Circuits Syst Video Technol*. 2024;34: 5921–5934. doi:10.1109/TCSVT.2023.3342808
99. Wang Y, Zhen L, Zhang J, Li M, Zhang L, Wang Z, et al. MedNAS: Multiscale Training-Free Neural Architecture Search for Medical Image Analysis. *IEEE Trans Evol Computat*. 2024;28: 668–681. doi:10.1109/TEVC.2024.3352641
100. Pérez-Rúa J-M, Vielzeuf V, Pateux S, Baccouche M, Jurie F. MFAS: Multimodal Fusion Architecture Search. *arXiv*; 2019. doi:10.48550/arXiv.1903.06496
101. Cui Z, Sun S, Guo Q, Liang X, Qian Y, Zhang Z. A Fast Neural Architecture Search Method for Multi-Modal Classification via Knowledge Sharing. *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. 2025; 5003–5011.
102. Salami F, Bozorgi-Amiri A, Hassan GM, Tavakkoli-Moghaddam R, Datta A. Designing a clinical decision support system for Alzheimer’s diagnosis on OASIS-3 data set. *Biomedical Signal Processing and Control*. 2022;74: 103527. doi:10.1016/j.bspc.2022.103527
103. Hom KL, Illapani VSP, Xie H, Oluigbo C, Vezina LG, Gaillard WD, et al. Application of preoperative MRI lesion identification algorithm in pediatric and young adult focal cortical dysplasia-related epilepsy. *Seizure: European Journal of Epilepsy*. 2024;122: 64–70. doi:10.1016/j.seizure.2024.09.024
104. Kersting LN, Walger L, Bauer T, Gnatkovsky V, Schuch F, David B, et al. Detection of focal cortical dysplasia: Development and multicentric evaluation of artificial intelligence models. *Epilepsia*. 2025;66: 1165–1176. doi:10.1111/epi.18240
105. Shankar A, Saikia MJ, Dandapat S, Barma S. Focal cortical dysplasia (type II) detection with multi-modal MRI and a deep-learning framework. *npj Imaging*. 2024;2: 31. doi:10.1038/s44303-024-00031-5

106. Lukkarinen H, Jeppsson A, Wikkelsö C, Blennow K, Zetterberg H, Constantinescu R, et al. Cerebrospinal fluid biomarkers that reflect clinical symptoms in idiopathic normal pressure hydrocephalus patients. *Fluids Barriers CNS*. 2022;19: 11. doi:10.1186/s12987-022-00309-z
107. Ying Y, Lin J, Gao W, Yue L, Zeng Q, Bartas K, et al. Proteomic profiling in cerebrospinal fluid reveal biomarkers for shunt outcome in idiopathic normal-pressure hydrocephalus. *Journal of Advanced Research*. 2026;80: 671–682. doi:10.1016/j.jare.2025.04.043
108. Wu D, Ma T, Ceritoglu C, Li Y, Chotiyanonta J, Hou Z, et al. Resource atlases for multi-atlas brain segmentations with multiple ontology levels based on T1-weighted MRI. *NeuroImage*. 2016;125: 120–130. doi:10.1016/j.neuroimage.2015.10.042
109. Barough SS, Ventura C, Bilgel M, Albert MS, Miller MI, Moghekar A. BrainSignsNET: Deep Learning-Based 3D Anatomical Landmark Detection in Human Brain Imaging. *MedRxiv*. 2025. doi:10.1101/2025.07.31.25332457