Full title page:

Deep Learning Achieves Neuroradiologist-Level Performance in Detecting Hydrocephalus

Yu Huang, PhD[1]*, Raquel Moreno, MD[1]*, Rachna Malani, MD[2,3], Alicia Meng, MD[1,3], Nathaniel Swinburne, MD[1,3], Andrei I Holodny, MD[1], Ye Choi, MD[1], Lucas C Parra, PhD[4], Robert J Young, MD[1,3]

[1] Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY 10065
[2] Department of Neurology, Memorial Sloan Kettering Cancer Center, New York, NY 10065
[3] Brain Tumor Center, Memorial Sloan Kettering Cancer Center, New York, NY 10065
[4] Department of Biomedical Engineering, City College of New York, New York, NY 10031

* contributed equally as co-first authors

Manuscript Type: Original Research
Word Count of Body Text: 2786

Funding Information: Support for this work was provided in part through the following grants from the National Institutes of Health: P30 CA008748, R01MH111896, R01NS095123, R01 CA247910, R01CA247910, R01MH111896, R21NS115018. Support was also provided by the Memorial Sloan Kettering Cancer Center Department of Radiology.

Originating Institution: Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065

Corresponding Author: Lucas Parra, PhD, City College of New York, 160 Convent Ave, Steinman Hall Room 401, New York, NY 10031. Phone: 212-650-7211. Email: parra@ccny.cuny.edu

Senior Author: Robert J Young, MD, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065

Acknowledgements

The authors are grateful for the expert editorial advice of Alyssa Duck, PhD, Editor, Department of Radiology, Memorial Sloan Kettering Cancer Center.

Key: R1 = Raquel Moreno, R2 = Nathaniel Swinburne, R3 = Rob Young, R4 = Alicia Meng, R5 = Rachna Malani. In the final manuscript the initials will be inserted again.

Blinded title page:
Title: Deep Learning Achieves Neuroradiologist-Level Performance in Detecting Hydrocephalus

Article Type: Original Research

Abbreviations (no more than 10):

MRI = magnetic resonance imaging, 2D/3D = two-dimensional/three-dimensional, CNN = convolutional neural network, TPM = tissue probability map, CSF = cerebrospinal fluid, NPH = normal pressure hydrocephalus, ROC = receiver operating characteristic, AUC = area under the curve, SPM = statistical parametric mapping, FSL = FMRIB software library

Summary statement: A two-stage automated pipeline was developed to segment head MRI and extract volumetric features to accurately and efficiently detect hydrocephalus that required shunting and achieved performance comparable to that of trained neuroradiologists.

Key Points:

- We developed a state-of-the-art 3D deep convolutional network to perform fully automated segmentation of the ventricles, extraventricular cerebrospinal fluid, and brain tissues in anisotropic MRI brain scans in a heterogeneous patient population.
- Volumetric features extracted from anatomical segmentations can be used to classify hydrocephalus (which may require neurosurgical intervention) vs. non-hydrocephalus.
- When tested in an independent dataset, the network achieved performance comparable to that of expert neuroradiologists.

Abstract

Purpose: We aim to develop automated detection of hydrocephalus requiring treatment in a heterogeneous patient population referred for MRI brain scans, and compare performance to that of neuroradiologists.

Materials and Methods: We leveraged 496 clinical MRI brain scans (259 hydrocephalus) collected retrospectively at a single clinical site from patients aged 2–90 years (mean 54) referred for any reason. Sixteen MRI scans (ten hydrocephalus) were segmented semi-automatically in 3D to delineate ventricles, extraventricular CSF, and brain tissues. A 3D CNN was trained on these segmentations and subsequently used to automatically segment the remaining 480 scans. To detect hydrocephalus, volumetric features such as volumes of ventricles and temporal horns were computed from the segmentation and were used to train a linear classifier. Machine performance was evaluated in a diagnosis dataset where hydrocephalus was confirmed as requiring surgical intervention, and compared to four neuroradiologists on a random subset of 240 scans. The pipeline was tested on a separate screening dataset of 205 scans collected from a routine clinical population aged 1–95 years (mean 56) to predict the majority reading from four neuroradiologists using images alone.

Results: When compared to the neuroradiologists at a matched sensitivity, the machine did not show a significant difference in specificity (proportions test, $p > 0.05$). The machine demonstrated comparable performance in independent diagnosis and screening datasets. Overall ROC performance compared favorably with the state-of-the-art (AUC 0.82–0.93).

Conclusion: Hydrocephalus can be detected automatically from MRI in a heterogeneous patient population with performance equivalent to that of neuroradiologists.

Introduction

Hydrocephalus is a common neurological disorder resulting from the abnormal accumulation of CSF that affects about one million people in the United States (1). Hydrocephalus commonly manifests with abnormal ventricular enlargement on brain imaging, either resulting from an obstructing mass lesion in the ventricles blocking CSF outflow (obstructive hydrocephalus) or from impaired CSF resorption (communicating hydrocephalus). This paper focuses on the diagnostically more difficult communicating hydrocephalus, including NPH in which ventricles slowly enlarge but intraventricular pressure does not increase. Clinical diagnosis of hydrocephalus remains difficult due to the wide spectrum of imaging results, overlap between normal and pathologically dilated ventricles, and highly variable signs and symptoms. The diagnosis often requires a combination of imaging and clinical abnormalities along with a high degree of suspicion by a well-trained radiologist.

While there have been attempts to standardize the assessment of ventricular size with measurements such as the callosal angle and Evans' index (2–5), these 2D measurements are obtained manually and thus can be time-consuming and less accurate than volumetric features (3,6,7). To improve the imaging diagnosis of hydrocephalus, rapid and reliable automated 3D segmentation and measurement of ventricle size are necessary. Modern neuroimaging software such as SPM (8) and FSL (9) are not specifically designed to generate accurate segmentation in patients with substantial intracranial pathology such as brain tumors. FreeSurfer (10,11) provides adequate segmentations in the presence of abnormal ventricles, but typically takes hours to generate one segmentation (3). Recently, deep learning methods have achieved great success in medical image segmentation, especially in applications where conventional neuroimaging packages fail due to atypical anatomy (12–14).

Previous machine learning efforts to diagnose hydrocephalus using MRIs have compared NPH with healthy volunteers or specific patient populations (*e.g.*, Alzheimer's Disease) with over 90% accuracy but on small sample sizes (typically less than 40 patients) (3,6,7,15). However, these methods have not been tested in a broader, unselected clinical population with the goal of detecting hydrocephalus among the heterogeneous conditions observed in neuroradiology practice. We hypothesize that a properly trained 3D deep CNN will generate accurate segmentation of the ventricles and other brain tissues, generating volumetric features that will enable accurate detection of hydrocephalus across a patient population referred for MRI brain scans. The purpose of this research is to design a machine to detect hydrocephalus requiring treatment with performance equivalent to trained neuroradiologists in this heterogeneous population.

Materials and Methods

Study Design

This retrospective study leverages two datasets (the Diagnosis Dataset and the Screening Dataset). The Diagnosis Dataset consists of 496 clinical MRI head scans performed between 2004 and 2019 (Fig. 1, see Human Subjects). Sixteen of these scans (10 hydrocephalus) were used to train a 3D deep CNN for segmentation based on the MultiPrior (Fig. 2 and Fig. S1) (14). The trained network was applied to the remaining 480 scans. Hand-selected volumetric features were then extracted from the segmentation to train a logistic regression classifier to detect hydrocephalus (Fig. 2, see Volumetric Features). For a random subset of 240 scans in the Diagnosis Dataset (120 hydrocephalus, Fig. 1), we obtained independent readings based on multiple views from four neuroradiologists blinded to clinical information (see Reader Study). The goal of this dataset was to determine if the machine could diagnose clinically significant hydrocephalus requiring shunting using surgical intervention as the truth data.

We next tested the trained segmentation network on the independent Screening Dataset of 205 scans randomly selected from brain MRI performed between 2011 and 2017 (Fig. 1, see Human Subjects). Since none of these patients underwent shunting, the surgical diagnosis of hydrocephalus used for the Diagnosis Dataset was not available and hydrocephalus was instead diagnosed by majority consensus of three neuroradiologists (see Reader Study). The hydrocephalus classifier was tested on this dataset and predictions were compared to the majority of readings. The purpose of this dataset was to determine if the machine could replicate the readings of neuroradiologists in a cohort with similar incidence of hydrocephalus as the general clinical population (1–6%) (16,17).

Human Subjects

This study was approved by the local Institutional Review Board and Privacy Board and written informed consent was waived. All data handling complied with HIPAA regulations. The Diagnosis Dataset consists of scans of 496 patients including an enriched group with 259 hydrocephalus and 237 non-hydrocephalus subjects. The age range of this group was 2–90 (mean 54) for 225 males and 2–89 (mean 55) for 271 females. The Screening Dataset consists of scans of 205 patients who did not have prior clinical or imaging diagnosis of hydrocephalus or prior surgical shunting. The age range was 1–95 years (mean 54) for 85 males and 4–88 years (mean 58) for 120 females. See Human Subjects in the Supplement for details.

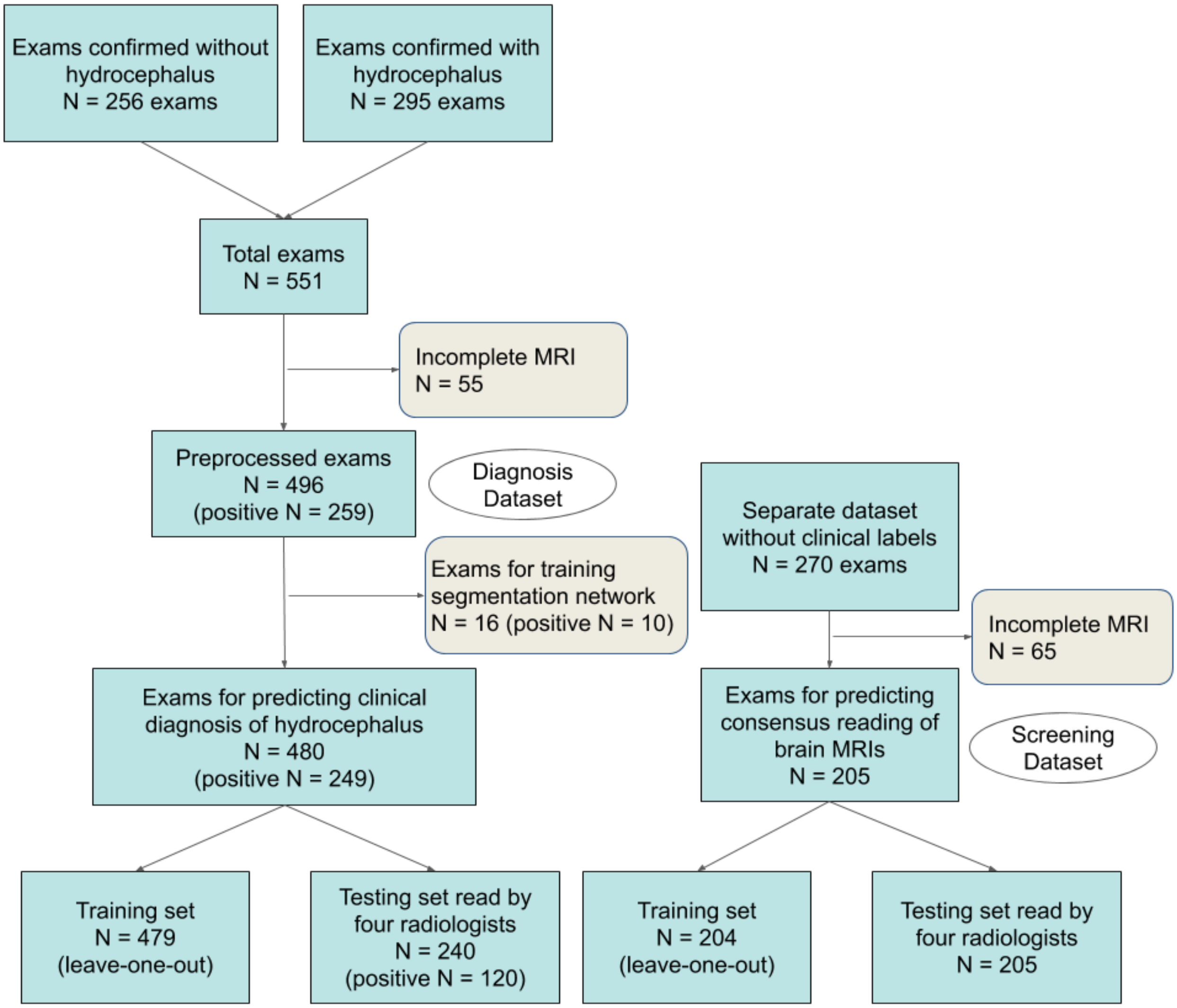
Exams confirmed without hydrocephalus
N = 256 exams
Exams confirmed with hydrocephalus
N = 295 exams
Total exams
N = 551
Incomplete MRI
N = 55
Preprocessed exams
N = 496
(positive N = 259)
Diagnosis Dataset
Exams for training segmentation network
N = 16 (positive N = 10)
Separate dataset without clinical labels
N = 270 exams
Incomplete MRI
N = 65
Exams for predicting clinical diagnosis of hydrocephalus
N = 480
(positive N = 249)
Exams for predicting consensus reading of brain MRIs
N = 205
Screening Dataset
Training set
N = 479
(leave-one-out)
Testing set read by four radiologists
N = 240
(positive N = 120)
Training set
N = 204
(leave-one-out)
Testing set read by four radiologists
N = 205

Figure 1: Number of scans used in training and testing. Note that there is overlap between the training set and testing set, as the training was performed using leave-one-out cross-validation, *i.e.*, for each test scan a different classifier was trained on the training data leaving out the one test scan.

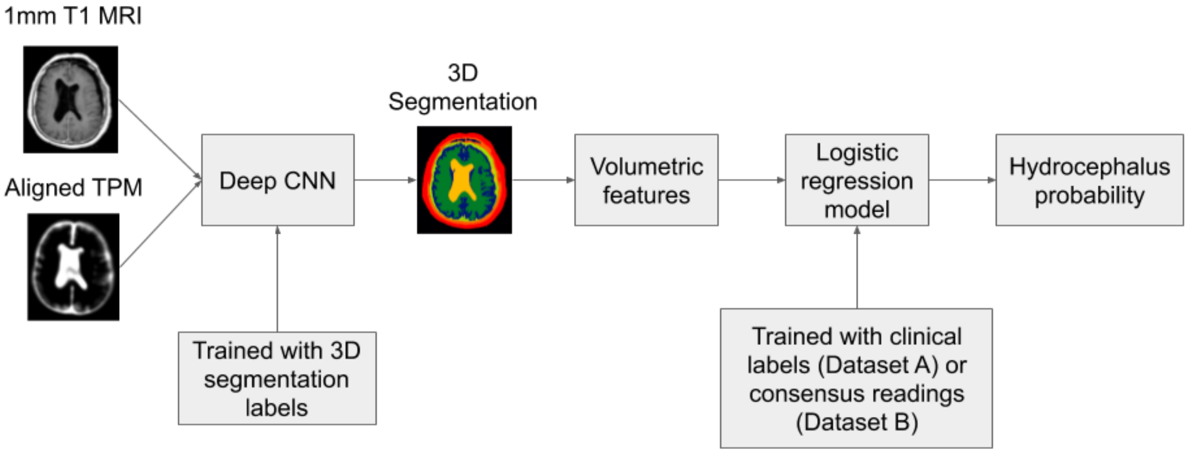
1mm T1 MRI
Aligned TPM
Deep CNN
Trained with 3D segmentation labels
3D Segmentation
Volumetric features
Logistic regression model
Trained with clinical labels (Dataset A) or consensus readings (Dataset B)
Hydrocephalus probability

Figure 2: Flowchart of the automated pipeline for machine detection of hydrocephalus.

Automated Pipeline for Machine Detection of Hydrocephalus

The major steps in the pipeline are: 1) tissue segmentation by a deep CNN, 2) extraction of volumetric features, and 3) logistic regression to detect hydrocephalus (Fig. 2). Preprocessing to harmonize input MRIs and align TPM to individual MRIs is shown in Fig. S1. The deep CNN was trained with 3D segmentation labels for the ventricles, extraventricular CSF, gray and white matter, air cavities, skull, and other soft tissue (see Training of Segmentation Network in Supplement). For the Diagnosis Dataset, logistic regression was trained with clinical labels as the ground truth. For the Screening Dataset, logistic regression was trained with majority readings from three neuroradiologists.

Volumetric Features

To train the hydrocephalus classifier using logistic regression, nine features were computed from the 3D segmentation masks, including previously proposed measures such as the Evans' index (3) and various volumetric features. Feature selection was performed to identify the subset of the ten features (nine from segmentation plus age) providing the best training-set performance on a subset of the Diagnosis Dataset (see Feature Extraction from Segmentation Data in Supplement). This feature selection was performed on a subset of the data that did not include the test set used in Fig. 4A. Three representative features are shown in Fig. 3: ratio of total ventricle volume to extraventricular CSF volume (R_{VC}), volume of the temporal horns (V_H), and average 2D extent of the lateral ventricles in the coronal slices (E_{2c}). A complete plot of all features is shown in Fig. S2.

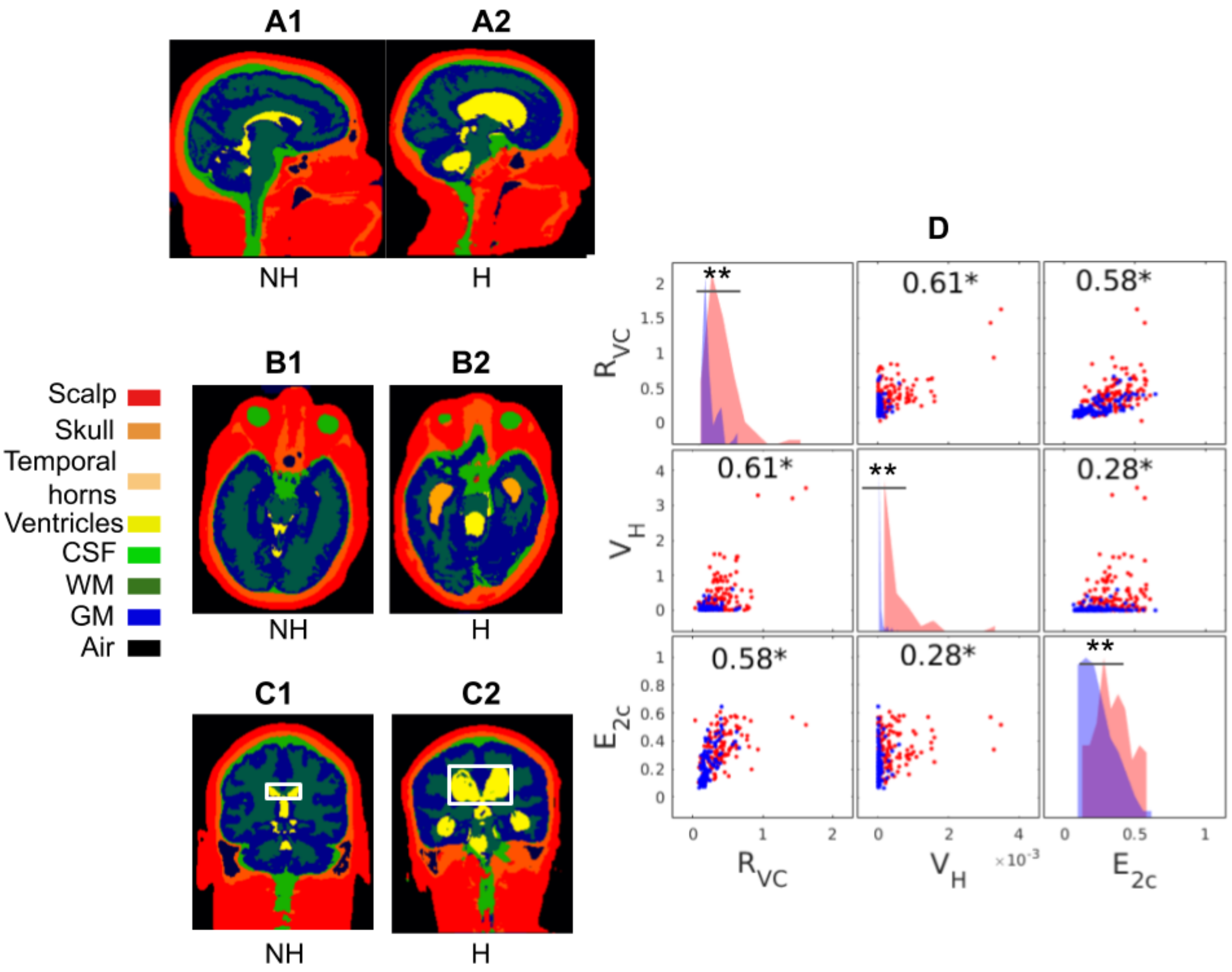

Figure 3: Example head segmentation for non-hydrocephalus patients (NH) and hydrocephalus patients (H) from the Diagnosis Dataset. Scatter plots of three representative features extracted from these segmentations are shown in panel (D), with red and blue dots representing hydrocephalus and non-hydrocephalus, respectively. R_{VC}: Ratio between ventricle volume and CSF volume; V_H: Volume of the temporal horns; E_{2c}: Average 2D extent of the lateral ventricles in the coronal slices. Extent is defined as the ratio of areas between the ventricles and the bounding box (white rectangles in panels C1 and C2), and is averaged across 20 coronal slices around the posterior commissure. Correlation coefficients (r) of each pair of features are noted in panel D (*: $p < 0.05$). Histograms of each feature are shown on the diagonal, with red and blue indicating hydrocephalus and non-hydrocephalus, respectively (**: $p < 0.001$).

Reader Study

To evaluate the performance of the automated hydrocephalus classification, four neuroradiologists (R1–R4; years of experience: 7, 7, 20, and 6, respectively) reviewed the scans in the Diagnosis Dataset and Screening Dataset. See Reader Study in Supplement for details.

Primary Outcome Measure and Statistical Methods

Segmentation performance was measured in Dice score (18). Significant differences of individual volumetric features were tested using the Wilcoxon rank sum test (Fig. 3D, diagonal panels). Detection accuracy was measured using ROC curves (Fig. 4). Comparison between ROC curves and neuroradiologist readings was performed by selecting a point on the ROC curve that matched neuroradiologist sensitivity. At that point, differences in specificity between the machine and neuroradiologists were tested for proportions (19). To establish the strength of the bully hypothesis, Bayes Factor was computed for this test of proportions using proportionBF() in the BayesFactor package in R (20). Inter-rater agreements between neuroradiologists and the clinical ground truth were quantified by Cohen's Kappa (21) (Fig. 5).

Results

We used 16 MRI scans (10 hydrocephalus) to train a segmentation network, separately labeling ventricles from extraventricular CSF and tissues usually included in head segmentation, *e.g.,* gray and white matter, bone, and other soft tissue (see Training of Segmentation Network in Supplement). This automated segmentation achieved a Dice score of 0.93 on a validation set of four scans. The trained network was applied to the other 480 scans in the Diagnosis Dataset and to the 205 scans in the Screening Dataset. Results of segmentation (Fig. 3) and hydrocephalus classification (Fig. 4) are described in the following.

Representative head segmentations on two subjects from the Diagnosis Dataset are shown in Fig. 3. In hydrocephalus patients, the deep network correctly identified the enlarged ventricles (Figs. 3A2 and C2, yellow) and temporal horns (Fig. 3B2, light orange). The segmentations obtained with the CNN better captured atypical anatomy as compared to conventional head segmentation methods (see Fig. S4; training set average Dice score: CNN = 0.92, SPM = 0.64, N=16). Three representative features extracted from volumetric segmentations of the Diagnostic Dataset are shown in Fig. 3D (volume ratio of ventricle vs. CSF, temporal horn volume, and ventricle extent). These features are significantly correlated with one another ($p < 0.05$, N = 240), and all differ significantly between hydrocephalus and non-hydrocephalus ($p < 0.001$, Wilcoxon rank sum test, N = 240). Statistics for the complete set of features are shown in Fig. S2.

On a subset of the Diagnosis Dataset that did not include any testing data, we performed feature selection and found that the best training-set performance was achieved when all of the nine features from segmentation were used (see Fig. S3, Feature Extraction from Segmentation Data in Supplement). These features do not include age, which did not provide additional discriminant information despite known age effects (17). Logistic regression was used to detect hydrocephalus. The ROC curves of the machine were computed for the Diagnosis Dataset using surgical intervention as ground truth (Fig. 4A) and for the Screening Dataset using a majority vote based on independent readings (Fig. 4B). The specificity and sensitivity of the four neuroradiologists are also indicated. For the Diagnosis Dataset, the machine was trained on 479 cases and test set performance was evaluated with a leave-one-out procedure. Evaluation was limited to a subset of 240 scans, which were reviewed by the four neuroradiologists (randomly chosen from the 480 scans, half with and half without hydrocephalus; this test set was also excluded in the feature selection procedure). The leave-one-out test-set ROC had an AUC of 0.93. The four neuroradiologists achieved accuracies of 85.4%, 86.3%, 82.9%, and 85.8% (sensitivity of 0.83, 0.79, 0.78, and 0.78, and specificity of 0.88, 0.93, 0.88, and 0.94, respectively). When selecting a classification threshold with the same sensitivity as each of the neuroradiologists, the machine achieved a specificity of 0.89, 0.91, 0.91, and 0.91, respectively (Fig. 4A). These values are not significantly different from those of the neuroradiologists (proportion test, p = 0.84, 0.47, 0.53, and 0.33, respectively; N = 240), although Bayes Factors in favor of the null hypothesis were weak (BF = 2.83, 1.72, 1.99, and 1.13, respectively).

For the Screening Dataset, the machine was trained on 205 scans and test-set performance was computed using leave-one-out cross-validation. As truth labels were obtained from a majority expert opinion, the training and testing were performed separately for each neuroradiologist while excluding that neuroradiologist from the majority to avoid bias (Fig. 4B). The machine achieved an AUC of 0.83, 0.85, and 0.82 for the three majority expert opinions (Fig. 4B). When comparing the neuroradiologists to the machine trained on the Screening Dataset (Fig. 4B), R1 fell exactly on the ROC curve, and there was no significant difference in specificity between the machine and the neuroradiologists at the same sensitivity for the other two neuroradiologists (proportion test, p = 0.08, BF = 0 for R2; p = 0.31, BF = 0.76 for R3; N = 205). The discrete steps of the ROC curves in Fig. 4B reflect the small number of hydrocephalus cases in the Screening Dataset, with 7, 14, and 13 cases for R1, R2, and R3, respectively, yielding 3.4–6.8% prevalence, which is typical for an unfiltered population in which most patients do not have hydrocephalus.

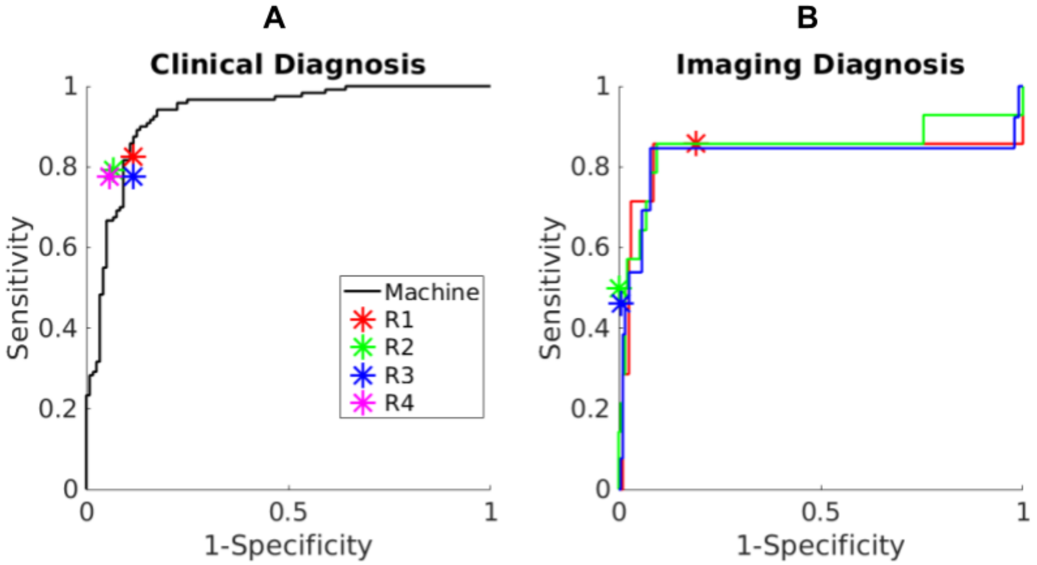

Figure 4: Performances of the machine and neuroradiologists in detecting hydrocephalus. (A) Machine vs. four neuroradiologists in predicting the clinical diagnosis of hydrocephalus in 240 scans (120 positive) in the Diagnosis Dataset; (B) Performance in predicting the majority readings of three neuroradiologists using images only in the 205 scans in the Screening Dataset. For each radiologist (R1–R3), a slightly different majority diagnosis serves as "ground truth," hence different ROC curves.

To test the neuroradiologist majority readings vs. the clinical truth labels, we computed the inter-rater agreement on the 240 scans in the Diagnosis Dataset reviewed by the neuroradiologists. The four neuroradiologists agreed with one another more than they agreed with the clinical truth labels (Fig. 5; we used Cohen's Kappa in all cases to allow direct comparison between pairs).

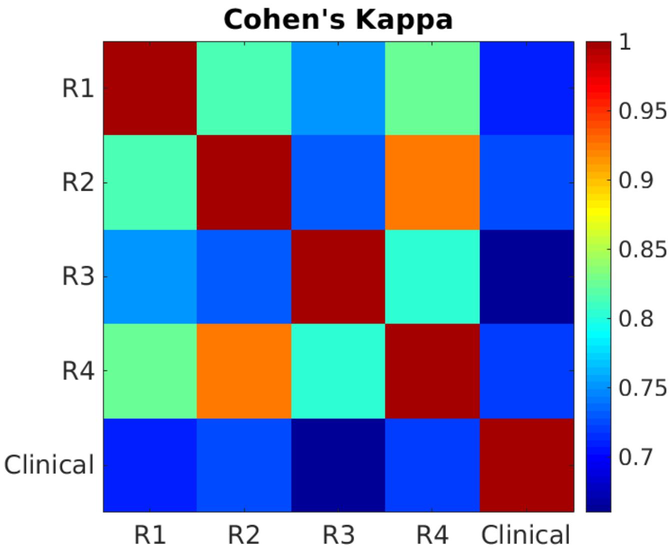

Figure 5: Inter-rater agreement on 240 scans in the Diagnosis Dataset between the four neuroradiologists (R1–R4) and the surgical intervention truth labels (Clinical).

Discussion

Hydrocephalus is a common neurological disorder that can mimic common dementias. Diagnosis is often difficult due to inconsistent imaging abnormalities and gradual onset of clinical symptoms. Current methods to quantify hydrocephalus on MRI scans are not standard of care because they are tedious and difficult to perform accurately and reproducibly (2,3). We trained a network to automatically provide volumetric segmentations of the brain and ventricles even in the presence of atypical brain anatomy. We automatically extracted various volumetric features and achieved detection performance comparable to that of neuroradiologists.

Several automated (6,15) and semi-automated methods (3,4,7) have been proposed for detection of hydrocephalus. These studies focused on distinguishing between NPH and healthy controls, or distinguishing hydrocephalus from specific disorders such as Alzheimer's disease. Our clinical dataset includes a much broader, unselected population of patients referred for MRI brain scans, with variable pathologies including brain tumors, surgical cavities, and infarcts. We found that discrimination was more challenging in this heterogeneous dataset compared to earlier studies with relatively small datasets of <30 cases in each group (3,6,7,15). Here we have leveraged a significantly larger dataset with a total >700 patients, including >200 cases of hydrocephalus requiring shunting and >400 cases with imaging evaluation.

The four neuroradiologists achieved mean performance accuracy of 85.2%, which is in line with previous studies reporting 75–95% (3,6,15). The wide range suggests that it is difficult to compare performance across studies differing in discrimination tasks, patient populations, and data quality. Here we compared the performance of the machine and neuroradiologist using identical tasks and datasets. The machine achieved comparable performance to the four neuroradiologists when using surgical intervention as truth data. Notably, the neuroradiologists agreed with each other more frequently than they agreed with the surgical intervention label, justifying predicting surgical intervention labels (Diagnosis Dataset) as a different task than predicting majority readings (Screening Dataset). Although neuroradiologists routinely scrutinize the ventricles as part of their clinical interpretation process, explicit measurements of size are not performed and cases of clinically important hydrocephalus requiring treatment may be missed. Completely automated segmentation and quantification of ventricle size and prediction of hydrocephalus may provide a useful screening tool to prioritize cases that require emergent reads, and also provide a useful adjunct to reading radiologists by increasing confidence in diagnosing unsuspected hydrocephalus.

Clinical brain MRIs usually have anisotropic resolution with higher in-plane resolution. To increase robustness, we resampled all images to isotropic 1 mm resolution, which allowed us to analyze the anatomy regardless of the original scan orientation. We leveraged previous work on segmentation of atypical head anatomy (14) and trained a deep CNN to segment enlarged ventricles that are often misclassified by conventional neuroimaging software (6,22). Compared to other recent studies using deep learning for segmenting ventricles from hydrocephalus (22–25), our 3D deep network achieved higher Dice scores. Nevertheless, we concur with the recommendation to use higher-resolution isotropic MRIs whenever possible (3).

In contrast to previous studies that use semi-automated methods (3,4,7), our goal was fully-automated processing to yield a reproducible and scalable approach. Our deep CNN provided volumetric segmentations within two minutes (14). This is significantly faster than alternatives such as FreeSurfer, which can take up to eight hours to segment the ventricles in MRI for detecting hydrocephalus (3,7). The speedup potentially provides an efficient, fully-automated tool for hydrocephalus detection in future clinical “big data” studies and population-level studies (26).

We encountered potential limitations. First, the Diagnosis Dataset defined hydrocephalus as clinical and imaging evidence of hydrocephalus requiring surgical intervention. The decision for surgical intervention, however, is complex and multifactorial including data such as patient symptoms, comorbidities, performance status, predicted improvement after shunting, and life expectancy. While shunt risks (*e.g.*, catheter malfunction, hemorrhage, infection) are beyond the scope of this project, we believe that these clinical, imaging, and surgical truth labels provide maximal confidence of hydrocephalus. To better simulate real-life conditions, we also tested these truth labels against majority readings of neuroradiologists for patients in both datasets. Second, we did not explicitly segment the temporal horns, as their posterior margins are arbitrarily defined, instead adopting a pragmatic estimation of their volumes (details in Supplement). We found that this estimate correlated with the presence of hydrocephalus. Future work could train the network to explicitly segment the temporal horns to calculate their volumes more accurately.

In conclusion, we trained a deep CNN to rapidly diagnose hydrocephalus with performance comparable to that of expert neuroradiologists. This network has the potential to assist the diagnosis of unsuspected hydrocephalus, enable automated quantification of hydrocephalus, expedite and augment neuroradiology reads, and ultimately improve patient care. To facilitate future studies of hydrocephalus and ventricle segmentation, we have made the pre-trained network and hydrocephalus classifier publicly available at https://github.com/andypotatohy/hydroDetector.

References

1. Dewan MC, Rattani A, Mekary R, Glancz LJ, Yunusa I, Baticulon RE, et al. Global hydrocephalus epidemiology and incidence: systematic review and meta-analysis. Journal of Neurosurgery. 2018 Apr 27;130(4):1065–79.
2. 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 Jul;67(1):94–9; discussion 99.
3. Miskin N, Patel H, Franceschi AM, Ades-Aron B, Le A, Damadian BE, et al. Diagnosis of Normal-Pressure Hydrocephalus: Use of Traditional Measures in the Era of Volumetric MR Imaging. Radiology. 2017;285(1):197–205.
4. 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. American Journal of Neuroradiology [Internet]. 2015 Dec 1 [cited 2020 Nov 27];36(12):2262–9. Available from: http://www.ajnr.org/content/36/12/2262
5. 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. European Journal of Neurology [Internet]. 2018 [cited 2020 Nov 27];25(3):569–76. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/ene.13555
6. Serulle Y, Rusinek H, Kirov II, Milch H, Fieremans E, Baxter AB, et al. Differentiating shunt-responsive normal pressure hydrocephalus from Alzheimer disease and normal aging: pilot study using automated MRI brain tissue segmentation. J Neurol [Internet]. 2014 Oct 1 [cited 2020 Nov 27];261(10):1994–2002. Available from: https://doi.org/10.1007/s00415-014-7454-0
7. Quattrone A, Sarica A, Torre DL, Morelli M, Vescio B, Nigro S, et al. Magnetic Resonance Imaging Biomarkers Distinguish Normal Pressure Hydrocephalus From Progressive Supranuclear Palsy. Movement Disorders. 2020;35(8):1406–15.
8. Ashburner J, Friston KJ. Unified segmentation. NeuroImage. 2005 Jul 1;26(3):839–51.
9. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging. 2001 Jan;20(1):45–57.
10. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage. 1999 Feb;9(2):179–94.
11. Fischl B, Sereno MI, Dale AM. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage. 1999 Feb;9(2):195–207.
12. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis. 2017 Feb 1;36:61–78.
13. Guha Roy A, Conjeti S, Navab N, Wachinger C. QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage. 2019 Feb 1;186:713–27.
14. Hirsch L, Huang Y, Parra LC. Segmentation of lesioned brain anatomy with deep volumetric neural networks and multiple spatial priors achieves human-level performance. arXiv:190510010 [cs, eess, q-bio, stat] [Internet]. 2020 May 11 [cited 2020 Jul 15]; Available from: http://arxiv.org/abs/1905.10010
15. Irie R, Otsuka Y, Hagiwara A, Kamagata K, Kamiya K, Suzuki M, et al. A Novel Deep Learning Approach with a 3D Convolutional Ladder Network for Differential Diagnosis of

Idiopathic Normal Pressure Hydrocephalus and Alzheimer's Disease. Magnetic Resonance in Medical Sciences. 2020;advpub.

16. Tanaka N, Yamaguchi S, Ishikawa H, Ishii H, Meguro K. Prevalence of possible idiopathic normal-pressure hydrocephalus in Japan: the Osaki-Tajiri project. Neuroepidemiology. 2009;32(3):171–5.
17. Jaraj D, Rabiei K, Marlow T, Jensen C, Skoog I, Wikkelsø C. Prevalence of idiopathic normal-pressure hydrocephalus. Neurology. 2014 Apr 22;82(16):1449–54.
18. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945 Jul 1;26(3):297–302.
19. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med. 1998 Apr 30;17(8):873–90.
20. proportionBF: Function for Bayesian analysis of proportions in BayesFactor: Computation of Bayes Factors for Common Designs [Internet]. [cited 2020 Dec 9]. Available from: https://rdrr.io/cran/BayesFactor/man/proportionBF.html
21. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012 Oct 15;22(3):276–82.
22. Ishii K, Kawaguchi T, Shimada K, Ohkawa S, Miyamoto N, Kanda T, et al. Voxel-Based Analysis of Gray Matter and CSF Space in Idiopathic Normal Pressure Hydrocephalus. DEM [Internet]. 2008 [cited 2020 Nov 28];25(4):329–35. Available from: https://www.karger.com/Article/FullText/119521
23. Ono K, Iwamoto Y, Chen Y-W, Nonaka M. Automatic Segmentation of Infant Brain Ventricles with Hydrocephalus in MRI Based on 2.5D U-Net and Transfer Learning. JOIG [Internet]. 2020 [cited 2020 Nov 27];42–6. Available from: http://www.joig.org/index.php?m=content&c=index&a=show&catid=63&id=236
24. Grimm F, Edl F, Kerscher SR, Nieselt K, Gugel I, Schuhmann MU. Semantic segmentation of cerebrospinal fluid and brain volume with a convolutional neural network in pediatric hydrocephalus—transfer learning from existing algorithms. Acta Neurochir. 2020 Oct 1;162(10):2463–74.
25. Ren X, Huo J, Xuan K, Wei D, Zhang L, Wang Q. Robust Brain Magnetic Resonance Image Segmentation for Hydrocephalus Patients: Hard and Soft Attention. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). 2020. p. 385–9.
26. Andersson J, Rosell M, Kockum K, Lilja-Lund O, Söderström L, Laurell K. Prevalence of idiopathic normal pressure hydrocephalus: A prospective, population-based study. PLOS ONE [Internet]. 2019 May 29 [cited 2020 Nov 27];14(5):e0217705. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0217705
27. Relkin N, Marmarou A, Klinge P, Bergsneider M, Black PM. Diagnosing Idiopathic Normal-pressure Hydrocephalus. Neurosurgery [Internet]. 2005 Sep 1 [cited 2020 Nov 27];57(suppl_3):S2-4-S2-16. Available from: https://academic.oup.com/neurosurgery/article/57/suppl_3/S2-4/2744115
28. Huang Y, Dmochowski JP, Su Y, Datta A, Rorden C, Parra LC. Automated MRI segmentation for individualized modeling of current flow in the human head. J Neural Eng. 2013 Dec 1;10(6):066004.
29. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv:14126980 [cs] [Internet]. 2017 Jan 29 [cited 2020 May 4]; Available from: http://arxiv.org/abs/1412.6980
30. Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: is there a better way? Radiology. 2010 Oct;257(1):14–7.

Table 1: Symptomatology of the patients in both datasets. Note that each patient may have demonstrated more than one symptom. *Hydrocephalus was defined as clinical and imaging evidence of hydrocephalus that required ventricular shunting. **Hydrocephalus was defined as already ventricular shunted.

Diagnostic Dataset	Hydrocephalus* (N=259)	Non-hydrocephalus (N=237)
"Classic Triad"	52	0
Gait disturbance	163	N/A
Urinary urgency / incontinence	62	N/A
Cognitive impairment	150	N/A
Headaches	128	N/A
Nausea / vomiting	72	N/A
Screening Dataset	Hydrocephalus** (N=5)	Non-hydrocephalus (N=200)
Aseptic meningitis	0	1
Brain metastases	0	66
CNS infection	0	1
CNS lymphoma	0	7
Cognitive impairment	1	4
CNS vascular abnormality	0	9
Encephalopathy	0	1
Epilepsy	0	45
CNS tumors	2	73
Headaches	2	16
CNS hemorrhage	0	5
Leptomeningeal disease	0	10
Multiple sclerosis	0	1
Radiation necrosis	0	8
Screening	0	25

Supplementary Material

Human Subjects

All scans for the two datasets were extracted from 25,595 consecutive MRI brain scans performed over a five-year period in patients referred to a National Cancer Institute Comprehensive Cancer Center. The patient characteristics are summarized in Table 1.

The Diagnosis Dataset consists of scans of 496 patients including an enriched group with hydrocephalus, using surgical intervention as the truth data. We first identified all patients who underwent ventricular draining or shunting < 100 days after the MRI. We then excluded patients who did not have a clinical diagnosis of hydrocephalus based on chart review by a neuro-oncologist (R5, 8 years of experience) and imaging diagnosis of hydrocephalus based on imaging review by a neuroradiologist (R1, 7 years of experience) using established criteria (27). Given these expanded criteria, we expected a systematic difference between imaging findings and final clinical diagnosis. We found 259 patients who had clinical and imaging diagnoses of hydrocephalus and required surgical intervention. The age range of this group was 4–90 (mean 54) for 120 males and 2–89 (mean 56) for 139 females. To achieve an approximate 1:1 class balance, we next randomly selected 237 age- and sex-matched non-hydrocephalus patients who had no hydrocephalus or focal abnormalities on their MRI scans, had no clinical signs or symptoms consistent with hydrocephalus, and did not require surgical intervention. The age range of this group was 2–85 (mean 54) for 105 males and 2–87 (mean 54) for 132 females.

The Screening Dataset consists of 205 randomly selected scans from the remaining 25,099 scans (Fig. 1). After chart review, this included N=5 patients who had prior surgical shunting for hydrocephalus and N=200 patients who had no prior clinical or imaging diagnosis of hydrocephalus. The age range was 1–95 years (mean 54) for 85 males and 4–88 years (mean 58) for 120 females. The purpose of the Screening Dataset was to evaluate the performance of the hydrocephalus classification pipeline in a more generalizable patient cohort in which a minority of patients would be expected to have hydrocephalus.

Data Description and Harmonization

All brain MRI examinations were acquired on either a 1.5 or 3.0 Tesla GE scanner (GE Medical Systems, Waukesha, WI). MRI scans that met one of the following two conditions were treated as complete scans: (1) inclusion of a T1 post-contrast scan with isotropic resolution of 1 mm; (2) inclusion of T1 post-contrast scans with anisotropic resolutions that were acquired in three orthogonal planes: sagittal, coronal, axial. Scans that did not meet either of these conditions were declared incomplete and discarded. If both conditions were met, only the isotropic scans were used. Anisotropic scans (sagittal, coronal, and axial) have higher in-plane resolutions (0.39–1.02 mm) and lower out-of-plane (lateral) resolutions (3.00–7.50 mm). To harmonize images, all scans were resampled into 1 mm isotropic resolution and normalized in intensity by dividing with the 95-percentile of pixel intensity before entering the deep CNN for segmentation (Fig. S1A). Despite originating from different orientations (sagittal, axial, or isotropic), the final images entering the network all have the same orientation.

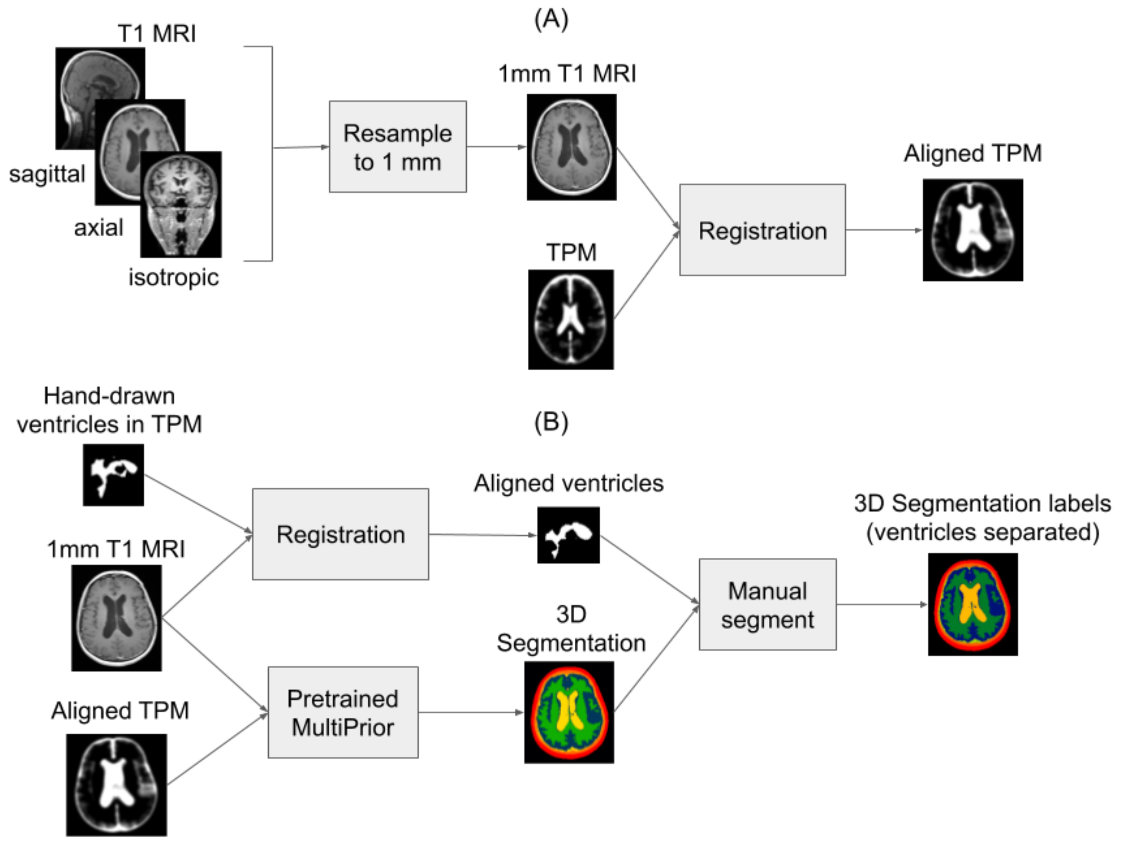

Figure S1: Flowcharts of preprocessing MRI scans (A) and generating segmentation labels for training the deep CNN (B) used in the hydrocephalus detection pipeline depicted in Fig. 2.

Training of Segmentation Network

Previously developed deep CNN (MultiPrior) (14) was used for segmentation of the head tissues. The MultiPrior architecture requires a spatial prior known as the TPM. The TPM we used covers the entire head down to the neck (28) and was aligned with individual MRIs by the non-linear registration algorithm implemented in SPM8 (8) before entering the CNN (Fig. S1).

We used sagittal scans from 16 scans in the Diagnosis Dataset to train the CNN (Fig. 1). To better extract the features from the temporal horns, which are easier seen in the axial slices (Fig. 3), in six of these 16 scans we also used the axial scans for training. To further boost the robustness of the network we used another four scans with 1 mm isotropic resolution from our previous study (28). In total, we used 26 scans (16 sagittal, six axial, four isotropic) to train this 3D CNN. In order to maintain input uniformity, all input scans were first resampled to 1 mm isotropic resolution before entering the CNN for training (Fig. S1A). Segmentation labels for training were first generated by pretrained MultiPrior (14), and then ventricles were manually separated out (see Generating 3D Segmentation Labels for Training).

During training, performance was monitored in a validation set of four scans and training was terminated when the loss function on the validation set did not change by more than 0.01 during four consecutive epochs, or when 100 epochs were reached, whichever occurred first. No strong overfitting was observed, so regularization was kept low, with no dropout and L2 penalty on model weights of 10^{-5}. Learning rate was set to 5×10^{-5} and adapted automatically with the Adam optimizer (29). For details on the training, see Hirsch et al. (14). The network that performed best on the validation set was applied on the remaining 480 scans in the Diagnosis Dataset and on the separate 205 scans in the Screening Dataset. While the network was trained on resampled images originating from axial, sagittal, or isotropic scans, when applied to the full dataset, we were able to use either axial, coronal, or isotropic scans, all resampled to 1 mm (Fig. S1A, Fig. 2).

Generating 3D Segmentation Labels for Training

The pretrained MultiPrior CNN from Hirsch et al. (14) was first applied to the 26 scans for training (Fig. S1B), which gives segmentation of seven classes (background, air cavities, gray matter, white matter, CSF, skull, and scalp). In this work, we defined features that were automatically extracted from the ventricle segmentation as the inputs for the hydrocephalus classifier. To train the network to segment ventricles, we manually separated out ventricles from the extraventricular CSF. To this end, ventricles were first drawn in the TPM (by the first author) and saved as a mask (Fig. S1B). This mask was warped to each head by SPM8 (8), giving an initial ventricle segmentation which was then manually improved based on the MRI intensity (Fig. S1B). Manual segmentation was performed by the first author using ScanIP (Synopsys, Mountain View, CA) and confirmed by neuroradiologists (R2 and R3). Finally the manual segmentation, along with the 26 MRIs and the aligned TPM, was used to train the CNN.

Feature Extraction from Segmentation Data

The network produced segmentation masks for gray matter, white matter, extraventricular CSF, ventricles, skull, scalp, and air cavities (Fig. 3). From these masks, we calculated the following features:

(i) Total volume of the ventricles (V_V), normalized by the intracranial volume (volumes of brain and CSF). (ii) Ratio of total ventricle volume to extraventricular CSF volume (R_{VC}). (iii) Ratio of total ventricle volume to brain volume (R_{VB}). (iv) Volume of the temporal horns (V_H), normalized by the intracranial volume. As there is no mask for temporal horns, we took an approximate approach: we identified the two horns in the TPM and the coordinates were mapped to each individual head using the mapping produced when the TPM is first coregistered to the individual MRI during preprocessing (Fig. S1A). A sphere of 1 cm radius was generated at each mapped location and intersected with the ventricle segmentation, giving us an estimate of the volume of the horns (Fig. 3B). (v) Evans' index as defined in Miskin et al. (3). To calculate the Evans' index, frontal horns were also identified in the TPM and mapped to individual axial MRIs as before. The largest left-to-right width of the frontal horns was determined by searching through 20 axial slices in the ventricle segmentation around the registered axial location. (vi) Magnetic Resonance Hydrocephalic Index (MRHI) as defined in Quattrone et al. (7). Likewise, collateral trigones of the lateral ventricles were mapped from the TPM to axial MRIs and the largest left-to-right width was found by reviewing 20 axial slices around the registered axial location. (vii) 3D extent of lateral ventricles in the axial scans. Similarly, posterior commissure was mapped to individual axial MRIs. Lateral ventricles above the posterior commissure were fed into function regionprops3() in Matlab (R2017b, MathWorks, Natick, MA) to calculate its extent (defined as the ratio of volumes between the ventricles and the bounding box). (viii) 3D extent of lateral ventricles in the coronal scans. This is similar to (vii) but was performed in the coronal scans. (ix) Averaged 2D extent of lateral ventricles in the coronal scans. This is similar to (vii) but the extent was calculated in 2D slice by slice across 20 coronal slices around the posterior commissure and averaged (Fig. 3C). We calculated the extent of the lateral ventricles to approximate the callosal angle (3), as it is not straightforward to directly calculate this angle from the ventricle segmentation.

Feature selection was performed using a subset of the Diagnosis Dataset (the 240 scans that were not in the test set), with the clinical labels as the ground truth. Features considered are the nine features mentioned above and the age of the subjects. These ten features are plotted in Fig. S2. All features are significantly correlated with one another ($p < 0.05$, $N = 240$) except age. The AUC as a function of different feature combinations is shown in Fig. S3. We found that the best AUC of 0.91 was achieved when all the nine features from segmentation were used, while age did not improve the classification. We also used these nine features for the Screening Dataset.

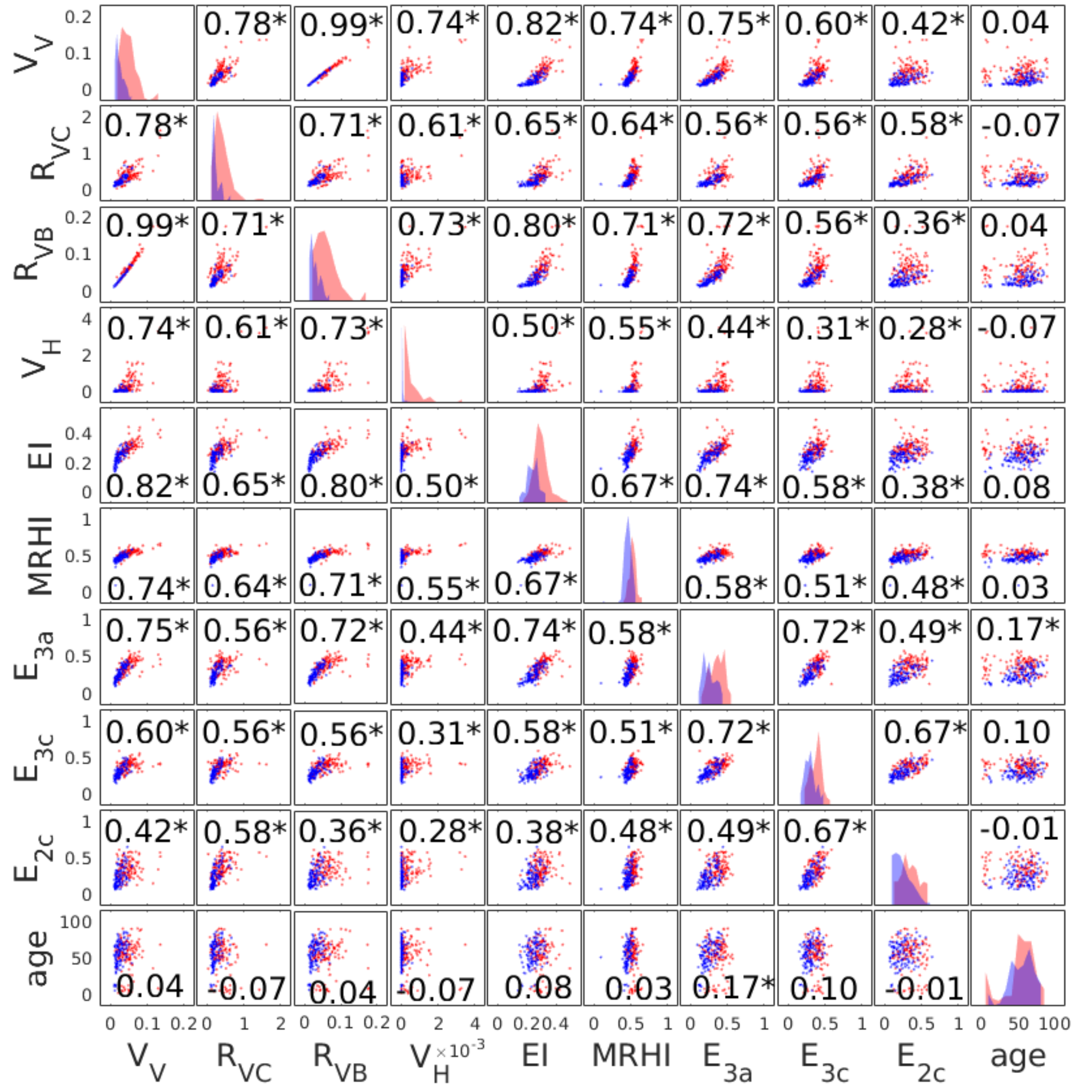

Figure S2: Scatter plots of all pairs of the ten features considered. Features are from a subset of the Diagnosis Dataset that were not included in the test set, with red and blue dots representing hydrocephalus and non-hydrocephalus subjects, respectively. The correlation coefficients (r) of each pair of features are noted in each panel (*: p < 0.05). Histograms of each feature are shown on the diagonal, with red and blue indicating hydrocephalus and non-hydrocephalus, respectively.

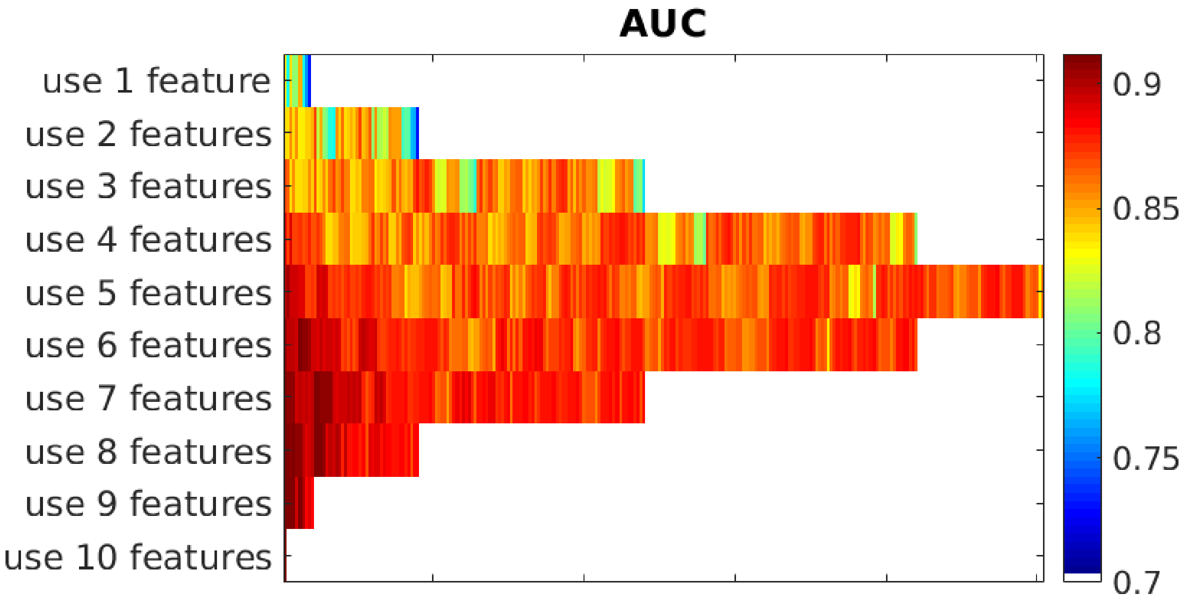

Figure S3: Selection of ten features for hydrocephalus classification. The best AUC of 0.91 was achieved when all nine features from segmentation were used.

Comparing Segmentation between Deep CNN and SPM

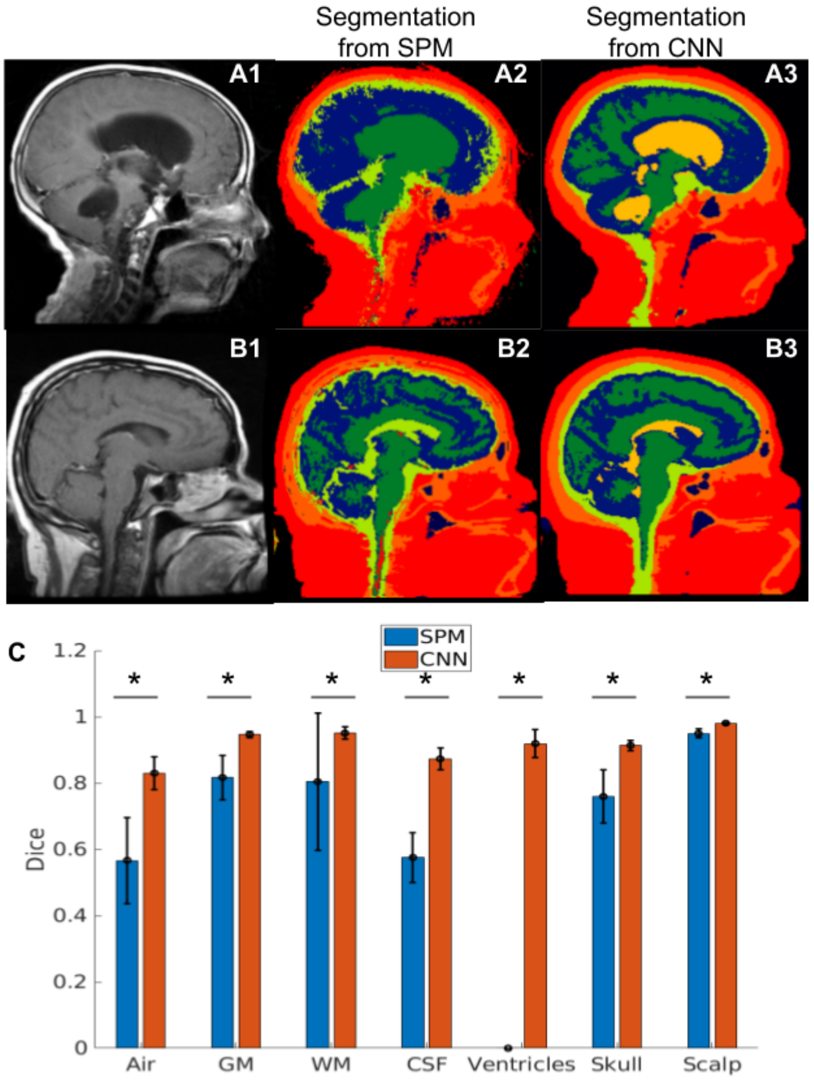

Figure S4: Head segmentation on a hydrocephalus patient (panel A) and a non-hydrocephalus patient (panel B) using SPM (A2 and B2) and deep CNN (A3 and B3). Dice scores for the segmentation of each tissue are shown across the 16 scans in the training set (panel C). *: p<0.01.

Representative head segmentations on two subjects from the Diagnosis Dataset using SPM and deep CNN are shown in Fig. S4. For the non-hydrocephalus subject (Fig. S4 panel B), SPM classified ventricles as part of the CSF (light green). However, SPM incorrectly labelled ventricles as white matter (dark green) for the hydrocephalus patient (Fig. S4 panel A2). The trained deep CNN correctly identified ventricles for both non-hydrocephalus and hydrocephalus patients (light orange, panels A3 and B3). The Dice scores for each tissue segmentation are shown in panel C for the 16 scans used to train the deep CNN. A formal analysis between SPM and deep CNNs used here is presented in Hirsch et al. (14).

Reader Study

For the Diagnosis Dataset, R1–R4 independently reviewed a subset of 240 randomly selected scans (half with hydrocephalus, Fig. 1) while blinded to the ground truth surgical intervention and clinical and demographic information. The neuroradiologists reviewed six different slices in the brain (Fig. S5): one sagittal midline slice; two coronal slices at the level of the third ventricle and of the posterior commissure; and three axial slices at the level of the body, left, and right temporal horns of the lateral ventricles. The reader diagnoses were based on established criteria examining the size and shape of the lateral ventricles, including temporal horns, third ventricle, Evans' index, and the callosal angle (27).

For the Screening Dataset, majority readings were used to train the machine and test performance of both machine and neuroradiologists. R1–R3 read all 205 scans in the Screening Dataset independently. R4 provided an independent rating only when the R1–R3 ratings were not unanimous. To prevent bias, these majority readings excluded the neuroradiologist being evaluated (*e.g.*, to evaluate R1, the diagnosis is the majority vote from R2–R4). With this construct, we were able to evaluate performance by majority for each of R1–R3 while avoiding some of the biases of consensus reads (30).

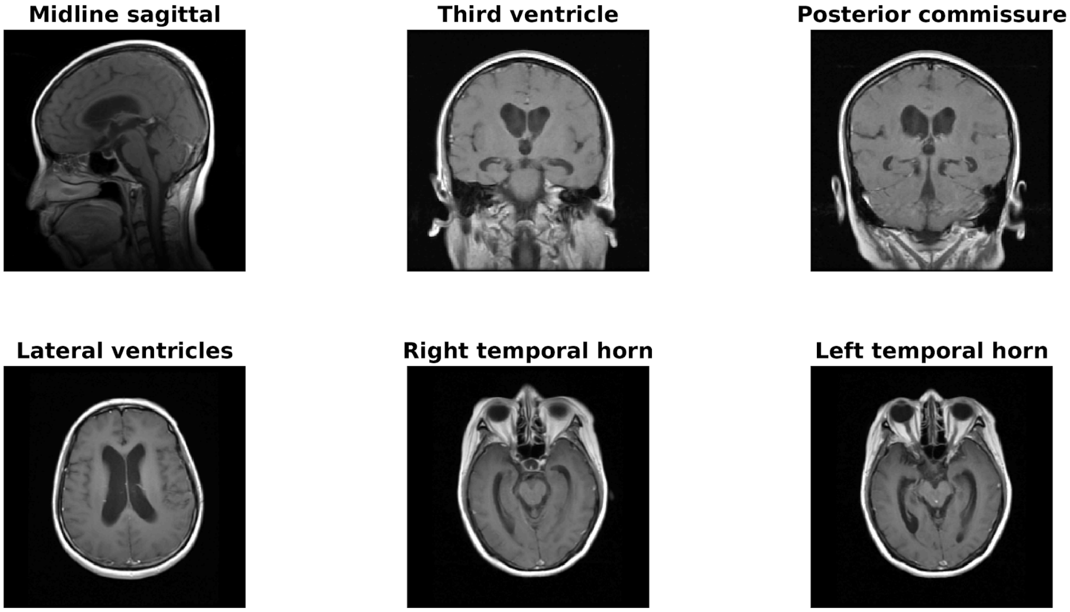

Figure S5: Representative images provided to neuroradiologists to make an imaging diagnosis of hydrocephalus. This is an example for a single scan.