Artificial Intelligence for Prediction of Shunt Response in Idiopathic Normal Pressure Hydrocephalus: A Systematic Review

Rafael Tiza Fernandes[1,2], Filipe Wolff Fernandes[1,3], Mrinmoy Kundu[1,4], Daniele S.C. Ramsay[1,5], Ahmed Salih[1,5], Srikar N. Namireddy[1,5], Dragan Jankovic[6], Darius Kalasauskas[6], Malte Ottenhausen[6], Andreas Kramer[6], Florian Ringel[6], Santhosh G. Thavarajasingam[1,6]

■ BACKGROUND: **Idiopathic normal pressure hydrocephalus (iNPH) is a reversible cause of dementia, typically treated with shunt surgery, although outcomes vary. Artificial intelligence (AI) advancements could improve predictions of shunt response (SR) by analyzing extensive datasets.**

■ METHODS: **We conducted a systematic review to assess AI's effectiveness in predicting SR in iNPH. Studies using AI or machine learning algorithms for SR prediction were identified through searches in MEDLINE, Embase, and Web of Science up to September 2023, adhering to Synthesis Without Meta-Analysis reporting guidelines.**

■ RESULTS: **Of 3541 studies identified, 33 were assessed for eligibility, and 8 involving 479 patients were included. Study sample sizes varied from 28 to 132 patients. Common data inputs included imaging/radiomics (62.5%) and demographics (37.5%), with Support Vector Machine being the most frequently used machine learning algorithm (87.5%). Two studies compared multiple algorithms. Only 4 studies reported the Area Under the Curve values, which ranged between 0.80 and 0.94. The results highlighted inconsistency in outcome measures, data heterogeneity, and potential biases in the models used.**

■ CONCLUSIONS: **While AI shows promise for improving iNPH management, there is a need for standardized data and extensive validation of AI models to enhance their clinical utility. Future research should aim to develop robust and generalizable AI models for more effective diagnosis and management of iNPH.**

INTRODUCTION

Idiopathic normal pressure hydrocephalus (iNPH) is a dementia subtype that can be completely or partially reversed with cerebrospinal fluid (CSF) diversion surgery. According to a study conducted in Sweden, iNPH has a prevalence of 3.7% among individuals aged 65 years and more.[1] With an increasing elderly population, the prevalence of iNPH is set to rise.

The pathophysiology of iNPH is, therefore, complex and multifactorial, involving both mechanical and biochemical

Key words
- Artificial intelligence
- Idiopathic normal pressure hydrocephalus
- iNPH
- Normal pressure hydrocephalus
- Prediction
- Shunt response

Abbreviations and Acronyms
AI: Artificial intelligence
AUC: Area under the curve
CSF: Cerebrospinal fluid
ELD: External lumbar drainage
iNPH: Idiopathic normal pressure hydrocephalus
LVLP: Large volume lumbar puncture
ML: Machine learning
MRI: Magnetic resonance imaging
PROBAST: Prediction model Risk Of Bias Assessment Tool
ROB: Risk of bias
SR: Shunt response
SVM: Support Vector Machine
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
TT: Tap test

From the [1]Imperial Brain & Spine Initiative, Imperial College London, London, United Kingdom; [2]Department of Neurosurgery, ULS São José, Lisbon, Portugal; [3]Department of Neurosurgery, Hannover Medical School, Hannover, Germany; [4]Institute of Medical Sciences and SUM Hospital, Bhubaneswar, India; [5]Faculty of Medicine, Imperial College London, London, United Kingdom; and [6]Department of Neurosurgery, University Medical Center Mainz, Mainz, Germany

To whom correspondence should be addressed: Santhosh G. Thavarajasingam, M.B.B.S., B.Sc. [E-mail: santhosh.thavarajasingam16@imperial.ac.uk]

Citation: World Neurosurg. (2024) 192:e281-e291.
https://doi.org/10.1016/j.wneu.2024.09.087

*Journal homepage: **www.journals.elsevier.com/world-neurosurgery***

*Available online: **www.sciencedirect.com***

factors.[2] There is still a significant gap of time between diagnosis and the start of symptoms. A shorter duration of the disorder is reported to have a better prognosis compared to a longer duration.[3] Thus, early recognition and timely treatment are crucial for a better outcome and quality of life for these patients.[4] However, accurate diagnosis and prediction of response remains a challenge, especially in the setting of concomitant neurodegenerative disorders, including Parkinson and Alzheimer disease. Currently, it requires a combination of clinical, neuroimaging, and CSF pressure measurements, as well as the response to CSF removal or temporary diversion procedures.[5] Recent studies underscore the limitations of conventional neuroimaging and basic CSF removal procedures such as tap test (TT)—which are often the only tests that make up the diagnostic algorithm of neurosurgical units preshunt insertion—as they are not able to reliably predict shunt response (SR) and shunt nonresponse. Novel noninvasive diagnostic tools such as CSF and venous biomarkers have shown promising results; however, they are not effective in isolation.

Ultimately, the question arises of how we can combine the results from different diagnostic modalities and maximize the use of single diagnostic modality data to optimize diagnostic performance. The current diagnostic dilemma in iNPH underscores the potential role of artificial intelligence (AI). In recent years, the integration of AI into medical research and clinical practice has opened new avenues for predicting treatment outcomes. Understanding the role of AI in this context is imperative for optimizing diagnostic accuracy, treatment planning, and overall patient care. AI algorithms emerge as a promising tool for decision-making, being able to analyze complex datasets and identify subtle patterns, and correlations that could otherwise not be assessed by traditional methods. AI could potentially help to identify novel biomarkers, patient selection, prediction of outcome, and optimize shunt adjustments in patients with iNPH.

By conducting a systematic review, this study aims to critically evaluate the existing body of literature on AI's predictive capabilities for SR in iNPH. The findings of this review hold the potential to shed light on the efficacy and accuracy of AI models in guiding clinical decision-making and improving patient outcomes in the management of iNPH.

METHODS

Literature Search

This systematic review was conducted following the Cochrane Collaboration guidelines and Preferred Reporting Items for Systematic Reviews and Meta-Analyses. It was previously registered on the International Prospective Register of Systematic Reviews. A comprehensive search of Embase, Scopus, PubMed, and Web of Science was performed until September 2023. Search strings were created for the following research question: "Is AI an effective and accurate tool for predicting SR in patients with iNPH?" Before undertaking the comprehensive literature search as described above, a preliminary literature search was conducted to identify any existing studies on the topic. However, the initial search yielded a limited number of articles. Recognizing the potential significance of the research question and the necessity to encompass a broader scope to ensure the inclusion of the most pertinent and comprehensive studies in this domain, an expanded and more extensive search strategy was performed. This methodological approach was employed to maximize the retrieval of relevant scientific evidence and ensure the robustness of the systematic review.

Study Selection

Two pairs of researchers screened all identified abstracts for inclusion. Each evaluator worked independently to assess the eligibility of the studies based on predefined inclusion and exclusion criteria. Two researchers then performed full-text screening. Any discrepancies during abstract or full-text screening were resolved by consulting a third researcher.

Inclusion and Exclusion Criteria

Inclusion Criteria

- Studies employing AI models, machine learning (ML), or algorithms to forecast shunt treatment response in iNPH patients.
- Participants with a possible/probable/definite diagnosis of iNPH.
- Studies reporting outcome measures for AI-based SR prediction in iNPH.

Exclusion Criteria

- Studies including solely participants diagnosed with secondary normal pressure hydrocephalus or alternative forms of hydrocephalus, or comorbidities potentially confounding result interpretation.
- Patients subjected to endoscopic third ventriculostomy.

In adherence to a predefined framework, this study implemented inclusion and exclusion criteria. Included were studies reporting the development and/or validation of ML models, as well as incremental value studies involving adult patients (aged >18 years) diagnosed with iNPH, with or without healthy controls. Conversely, exclusion criteria were applied to studies involving patients with disorders other than iNPH and those falling into categories such as narrative reviews, scoping reviews, systematic reviews, commentaries, conference abstracts, animal studies, or letters to editors. Additionally, studies that were not retrievable were excluded.

Data Extraction for Systematic Review Without Meta-Analysis

In the extraction process, relevant data encompassing various domains were systematically obtained. Study characteristics, including first author, publication year, country/institution, study type, setting, population, follow-up duration, age, and gender, were documented. Detailed information on the models/algorithms employed for SR prediction was recorded, emphasizing the input features associated with each ML method. These inputs encompassed demographic parameters, clinical invasive or noninvasive data, CSF biomarkers, and imaging features. The analysis focused on evaluating the sensitivity, specificity, and area under the curve (AUC) of the models/algorithms in predicting outcomes. The datasets used for training and testing, along with details on statistical analysis, including the handling of missing data and predictor selection, were documented. The Excel template

provided by Fernandez-Felix et al.[6] was used for data extraction and risk of bias (ROB) assessment of each article.

Critical Appraisal

Two evaluators independently used the Prediction model Risk Of Bias Assessment Tool (PROBAST) to gauge potential biases in the studies analyzed. PROBAST examines 4 key aspects: participants, predictors, outcomes, and analysis. Within these areas, biases related to participant selection, prediction methods, outcome determination, and data analysis were scrutinized using specific guiding questions. Discrepancies in study quality were resolved by a third reviewer.[7]

In our review, adherence to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines was rigorously evaluated by 2 independent researchers for each included study. TRIPOD provides a comprehensive checklist of 22 essential items aimed at enhancing the transparency and completeness of reporting in studies developing, validating, or updating prediction models for diagnostic or prognostic purposes.[8,9]

Data Analysis and Reporting

Due to methodological heterogeneity observed across the studies, it was deemed unfeasible to conduct a meta-analysis. Consequently, a

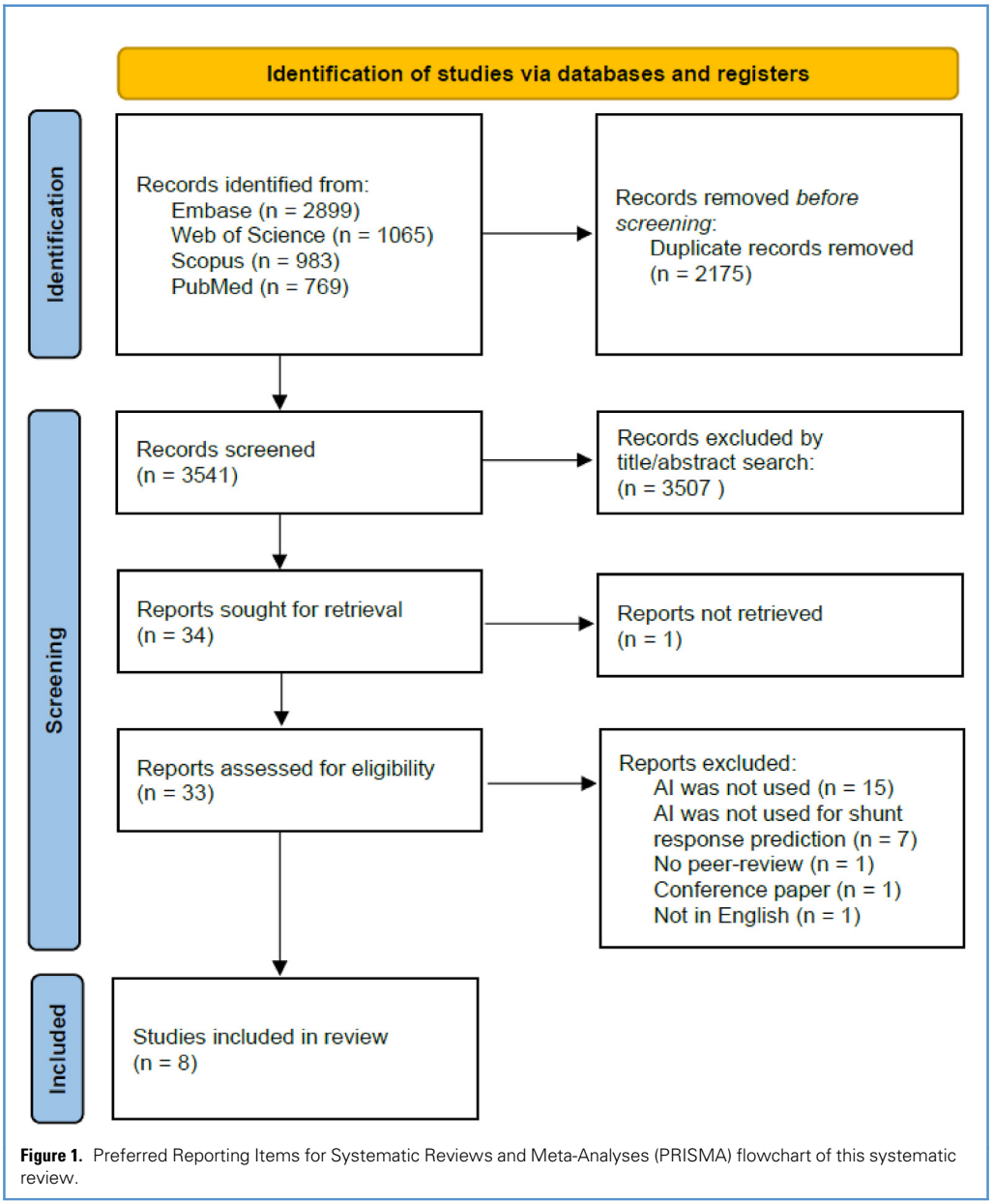

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart of this systematic review.

qualitative synthesis was undertaken without the integration of meta-analytic techniques, adhering to the Synthesis Without Meta-analysis guidelines.[10] Data interpretation was facilitated through graphical representations, which were generated using the Google Sheets platform. Additionally, Sankey diagrams were crafted using the Python programming language within the Google Colab framework, employing the Plotly library.

RESULTS

The literature search retrieved a total of 3541 articles for abstract screening (following the removal of duplicates), of which 33 articles underwent full-text review and 8 studies were included. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses chart is shown in Figure 1.

Study Characteristics and Design

Among the articles that included details about their setting and region, all were single-centric. Four studies were retrospective cohorts and the other 4 were prospective cohorts. Follow-up period varied from 1 year and 5 months (Levin et al.)[11] to 12 years (Sotoudeh et al.)[12] (Table 1, Figure 2).

Patient Characteristics

The sample size (with follow-up) ranged from 28 to 132 patients. The total number of patients (with follow-up) was 479 patients and the total number of responders was 262 patients (55.85%). Every article provided data regarding shunt/drainage responders versus nonresponders. SR ranged from 34% (Wu et al.)[15] to 80.5% (Mladek et al.).[17] Only 50% (4/8) of the studies included in the review disclosed demographic information (age and sex/gender) about participants in each group (Table 1).

Data Input for AI Models

Demographics. Three articles incorporated demographic data as inputs for their ML models. (Table 2, Figure 3) In each case, age was used, with additional demographic inputs consisting of either sex/gender or education level. Notably, Lang et al.[18] explored the inclusion of sex as an input in some algorithms but determined that their optimal model exclusively included age.

Noninvasive Clinical Tests. Two articles relied on clinical noninvasive data, such as grading scales (Sotoudeh et al.)[12] or parameters like disease duration, walking speed, step length, step width, Timed Up and Go test, categorical verbal fluency, Free and Cued Selective Reminding Test, Wechsler Adult Intelligence Scale III, and Starkstein scale for apathy (Griffa et al.)[16] as inputs for their models.

Invasive Clinical Tests. Two studies included clinical invasive data for their ML models (Mazzone et al.[13]; Mladek et al.[17]). Mazzone et al.[13] used CSF pressure patterns obtained through lumbar puncture, classified as "typical" by a trained neurosurgeon. Mladek et al.[17] incorporated intracranial pressure features

Table 1. A Summary of the Patient Characteristics of the Included Studies

					Participant Characteristics				
Authors & Year	Study Design	Enrolment Period	Study Region	Sample Size	Proportion of R	R: Age	R: Male	Non-R: Age	Non-R: Male
Mazzone et al. 1996[13]	Prospective cohort	NR	Italy	28	15/28 (54%)	NR	NR	NR	NR
Rau et al. 2021[14]	Retrospective cohort	2010–2017	Germany	28	20/28 (71%)	NR	NR	NR	NR
Sotoudeh et al. 2021[12]	Retrospective cohort	2009–2021	United States	78	34/78 (44%)	71.21 (9.87)	22/34 (65%)	68.14 (10.25)	19/44 (43%)
Wu et al. 2021[15]	Retrospective cohort	2009–2016	NR	104	35/104 (34%)	74.0 (8.4)	20/35 (57%)	73.5 (10.5)	39/69 (57%)
Griffa et al. 2022[16]	Prospective cohort	2017–2021	Switzerland	30	16/30 (53%)	79.3 (6.6)	10/16 (62.5%)	79.4 (5.2)	8/14 (57%)
Mladek et al. 2022[17]	Prospective cohort	2016–2021	Czech Republic	41	33/41 (80.5%)	NR	NR	NR	NR
Lang et al. 2023[18]	Retrospective cohort	2018–2021	Canada	132	87/132 (65.9%)	77.5 (5.9)	53/87 (60.9%)	77.6 (4.9)	29/45 (64.4%)
Levin et al. 2023[11]	Prospective cohort	2020–2021	United States	38	22/38 (57.9%) gait only	NR	NR	NR	NR

Table 1 shows a comprehensive summary of patient characteristics from the included studies, detailing aspects such as authors and publication year, study design, enrolment period, study region, and sample size. It further dissects the participant characteristics into 2 main groups: those who responded and those who did not respond to the treatments or interventions assessed in these studies.

Non-R, shunt nonresponders; NR, not reported; R, shunt responders.

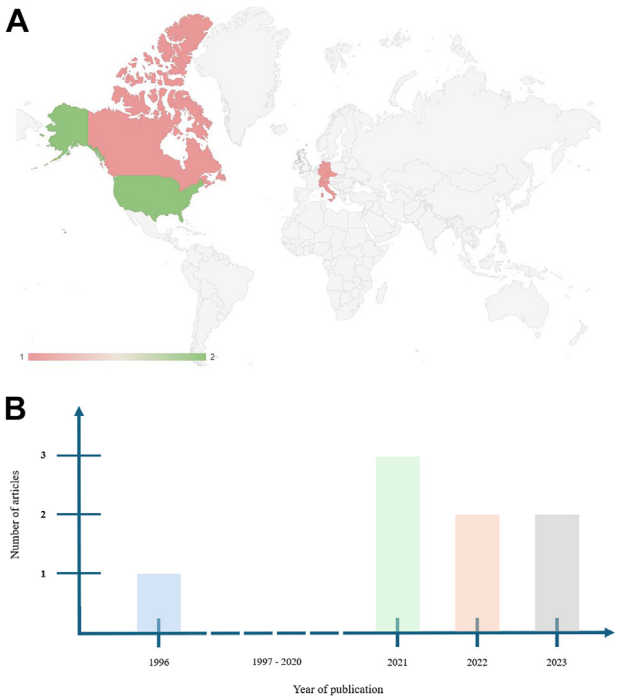

Figure 2. World map and publication date of the included studies. (**A**) The geographical origins of the included studies, with Germany, the United States, Switzerland, the Czech Republic, and Canada distinctly marked. Each country highlighted in red had 1 study included. The United States is marked with green denoting the 2 studies included from there. (**B**) The number of articles published per year across a specified time frame. In 1996, there was a notable publication of a single article. From 1997 to 2020, there were no articles published. In the subsequent years, there is a clear resurgence of publishing activity with 3 articles in 2021, 2 articles in 2022, and other 2 articles in 2021, indicating a rising trend in research output. This visualization may imply a renewed interest or advancements in the field, as evidenced by the increasing number of articles over the most recent years.

derived from a lumbar infusion test, consisting of 48 features grouped into 7 classes: temporal dynamics, integral, nonlinear, continuous wavelet transform, recurrence quantification analysis, heart rate, and electrocardiogram locking.

CSF Biomarkers. Levin et al.[11] incorporated CSF transcriptome data as inputs for their ML model. This encompasses the collective set of RNA molecules expressed in the CSF, which was collected during shunt surgery.

Neuroradiological Imaging and Radiomics

Among the selected articles, 5 employed imaging data, with 1 study (Rau et al.)[14] exclusively using imaging inputs. They used T1-weighted MP-RAGE scans, extracting information related to the volume of gray matter and CSF from 272 regions. Their primary objective was to diagnose iNPH rather than discerning responders from nonresponders. Nevertheless, they explored the adaptability of their algorithm for the latter classification.

Sotoudeh et al.[12] also employed magnetic resonance imaging (MRI) data, specifically T2-weighted scans, to develop their model. They extracted 120 features from the ventricular system and identified 12 features deemed most predictive for input into their model. Wu et al.[15] used T1-weighted MP-RAGE scans for their ML model, incorporating information from 283 regions of interest at various levels of granularity. The optimal number of features selected for input was 79, with a granularity level set at 5. Griffa et al.[16] adopted a comprehensive approach by integrating multiple MRI modalities, including T1-weighted, T2-weighted, diffusion-weighted imaging, arterial spin labeling, and resting-state functional imaging. They extracted 13 imaging features related to ventricle volume, posterior callosal marginal fissure, calcarine fissure relative volumes, orientation dispersion index of the posterior limb of the internal capsule, intra-axonal volume fraction (Vic) of the posterior limb of the internal capsule, orientation dispersion index of the cingulum, intra-axonal volume fraction (Vic) of the cingulum, functional connectivity within the default mode network, functional connectivity between the default mode network and somatomotor and visual regions, functional connectivity between the default mode network and executive-control regions, thalamus-perfusion, posterior cingulate cortex-perfusion, and Fazekas score. Lang et al.[18] focused on MRI data as well, using T1-weighted scans for gray matter volume, T2-weighted scans for the relative volume of CSF, and white matter hyperintensities. The top-performing model included average gray matter volume from specific brain regions, including the right supplementary motor area and right posterior parietal cortex, along with callosal angle. Other models included disproportionately enlarged subarachnoid space hydrocephalus and Evans index but did not achieve the same results.

Algorithms and Validation of AI Models

Machine Learning Algorithms and Performance. The predominant ML algorithm employed across the selected articles was the Support Vector Machine (SVM), used in every article except for the one by Mazzone et al.,[13] who built a neural network (perceptron). Additionally, 2 studies (Sotoudeh et al. and[12] Mladek et al.[17]) implemented multiple ML algorithms to compare their relative effectiveness (**Figure 4**). This approach involved a comparative analysis of different algorithms to determine which one yielded the most favorable results in the specific context of each study. Sotoudeh et al.,[12] when using only clinical data (4 clinical features: cognitive impairment, gait disturbance, urinary incontinence, and Modified Rankin Scale), the best performing model was the random forest (AUC = 0.71). When integrating imaging data, the predictive performance increased, and the best-performing model was the SVM (AUC = 0.80). In Mladek et al.,[17] the authors obtained performance measures for each model using all 48 features. The extreme gradient boosting model (AUC = 0.887), despite a slightly lower AUC than GradientBoost (AUC = 0.895), was selected for calibration after achieving the best discrimination capabilities (higher accuracy and specificity). The highest AUC (0.891), accuracy (0.823), and sensitivity (0.861) were achieved by incorporating 8 features, and the highest specificity (0.783) with 7 features (**Table 3**, **Figure 5**).

Table 2. A Summary of the Artificial Intelligence Models of the Included Studies

			Data Input for AI Models				
Authors & Year	**ML Algorithm**	**Validation Method**	**Demographic**	**Clinical noninvasive**	**Imaging**	**Clinical invasive**	**CSF biomarkers**
Mazzone et al. 1996[13]	NN (Perceptron)	Cross-validation				CSF pressure pattern	
Rau et al. 2021[14]	SVM	k-fold cross-validation			MRI (T1-W)		
Sotoudeh et al. 2021[12]	knn; DT; RF; SVM; LR; AdaBoost; NN	Cross-validation		iNPH grading Scale + Modified Rankin Scale	MRI (T2-W)		
Wu et al. 2021[15]	SVM	Cross-validation	Age + Sex		MRI (T1-W)		
Griffa et al. 2022[16]	SVM	Cross-validation	Age + Education level	9 different clinical parameters	MRI (T1 + T2-W + DWI + ASL + Rs-fMRI)		
Mladek et al. 2022[17]	RF; LR; GNB; SVM; Gradient boosting; XGBoost	Cross-validation				ICP features	
Lang et al.,2023[18]	SVM	Cross-validation	Age-only (top-performing model) + Sex (others)		MRI (T1 + T2-W)		
Levin et al. 2023[11]	SVM	Cross-validation					CSF transcriptome

Table 2 shows a systematic overview of the AI models employed across the included studies, detailing the ML algorithms used, the validation methods, and the data inputs leveraged for AI model development. The ML algorithms range from classical approaches like SVM and NN to ensemble methods including RF and Gradient boosting. The data inputs for these models encompass a diverse array of categories, including demographic information, noninvasive and invasive clinical parameters, imaging data (e.g., MRI with various weightings and sequences), and CSF biomarkers.

ML, machine learning; AI, artificial intelligence; CSF, cerebrospinal fluid; NN, neural network; SVM, Support Vector Machine; MRI, magnetic resonance imaging; T1-W, T1 weighted image; knn, k-nearest neighbors; DT, decision tree; RF, random forest; LR, logistic regression; iNPH, idiopathic normal pressure hydrocephalus; T2-W, T2 weighted image; DWI, diffusion-weighted imaging; ASL, arterial spin labeling; Rs-fMRI, resting-state functional imaging; GNB, Gaussian naive bayes; XGBoost, extreme gradient boosting; ICP, intracranial pressure.

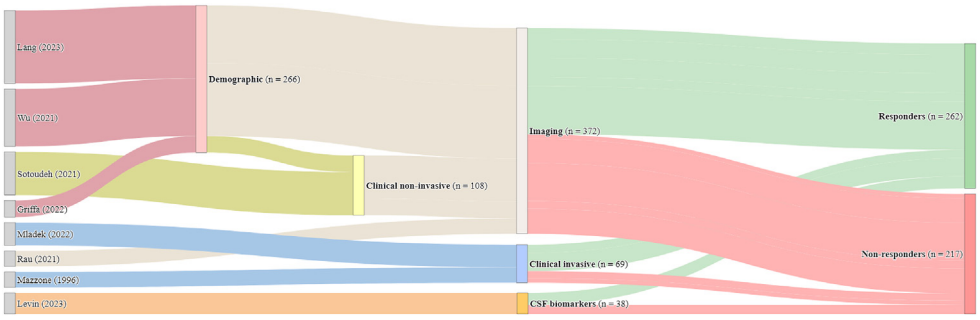

Figure 3. Data provided to each study and responders distribution. Figure 3 shows a Sankey diagram depicting the inputs each article's algorithm received to predict the response to shunt treatment, along with the subsequent distribution of responders and nonresponders among the full patient sample. This visualization categorizes the inputs into 5 primary groups: demographic data, clinical noninvasive data, imaging data, clinical invasive data, and CSF biomarkers. These inputs stem from a range of studies published between 1996 and 2023 included in this review. Each node within the diagram is distinctly color-coded to differentiate between the types of input data, employing lighter shades for the inputs and darker shades for the outcomes, namely responders and nonresponders. The links connecting input types to outcomes are likewise color-coded, representing the flow and volume of data contributing to each outcome. The link widths are proportionate to the volume of data from each input category used in predicting patient responses, thus quantitatively illustrating their distribution across the reviewed studies.

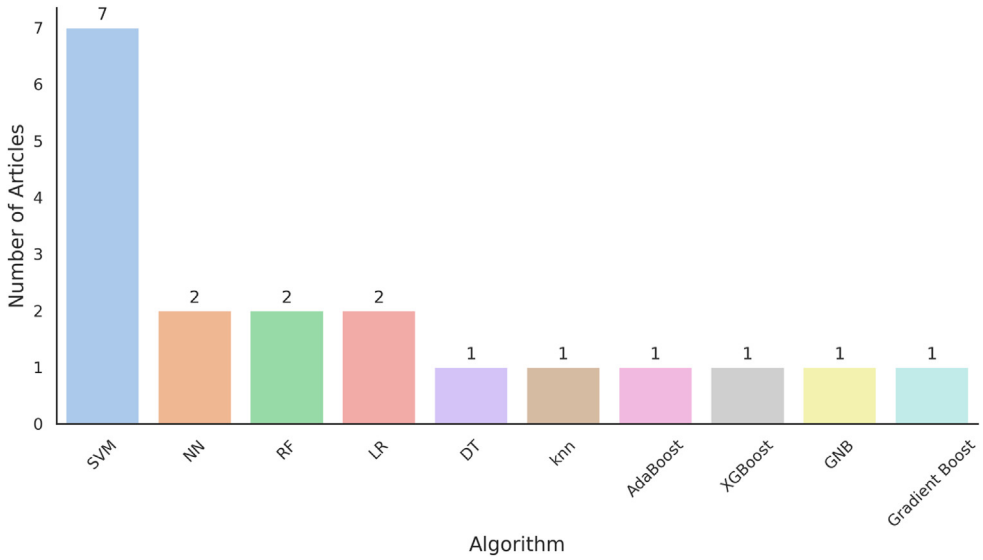

Figure 4. Algorithms employed. Figure 4 illustrates the distribution of algorithms used in the reviewed literature for predicting shunt response in patients. The bar graph delineates the frequency with which each algorithm appears across the collected articles. It is immediately apparent that support vector machine (SVM) is the most frequently employed algorithm, cited in 7 articles. This is followed by neural network (NN) and random forest (RF), each referenced in 2 articles, as well as logistic regression (LR) also appearing in 2 articles. Other algorithms are each reported in a single article. DT, decision tree; GNB, Gaussian naive bayes; knn, k-nearest neighbors; LR, logistic regression; NN, neural network; RF, random forest; SVM, Support Vector Machine; XGBoost, extreme gradient boosting.

Table 3. Outcome Measurements and Artificial Intelligence Performance Metrics

		AI Performance Metrics				
Authors & Year	**Outcome Measurement**	**AUC**	**Accuracy**	**Precision**	**Specificity**	**Sensitivity**
Mazzone et al. 1996[13]	Shunt surgery - Stein and Langfitt scale	NR	0.54	NR	NR	NR
Rau et al. 2021[14]	Shunt surgery - Nonreported	0.53*	NR	NR	NR	NR
Sotoudeh et al. 2021[12]	Shunt surgery - iNPH scale	0.8	0.76	NR	0.84	0.67
Wu et al.,2021[15]	CSF drainage - LVLP - TUG OR Tinetti gait assessment	0.94	0.9	0.89	0.91	0.84
Griffa et al. 2022[16]	CSF drainage - TT - TUG test OR Walking speed	0.83	0.7	NR	0.64	0.75
Mladek et al.,2022[17]	CSF drainage - ELD - Nonspecified	0.891	0.823	NR	0.783	0.861
Lang et al.,2023[18]	CSF drainage - LVLP or ELD-Gait velocity AND/OR MoCA score changes	NR	0.758	0.757	0.422	0.931
Levin et al. 2023[11]	Shunt surgery - Gait OR Urinary OR Cognitive improvement	~0.80	NR	NR	NR	NR

Table 3 encapsulates the outcome measures and AI model performance metrics across the reviewed studies. The AI performance is quantitatively evaluated through metrics such as AUC, Accuracy, Precision, Specificity, and Sensitivity. AUC values span from marginal (0.53) to excellent (0.94) predictive performance. Entries corresponding to shunt surgery are highlighted in green, while those concerning other CSF drainage techniques are shown in red. This summary underscores the heterogeneity in AI model efficacy and the complexity of evaluating their performance in clinical applications, highlighting the necessity of employing a multimetric approach for a comprehensive assessment.

AI, artificial intelligence; AUC, area under the curve; NR, not reported; iNPH, idiopathic normal pressure hydrocephalus; CSF, cerebrospinal fluid; LVLP, large volume lumbar puncture; TUG, Timed Up and Go; TT, tap test; ELD, external lumbar drainage; MoCA, Montreal-cognitive-assessment.

*value obtained upon contact with the authors.

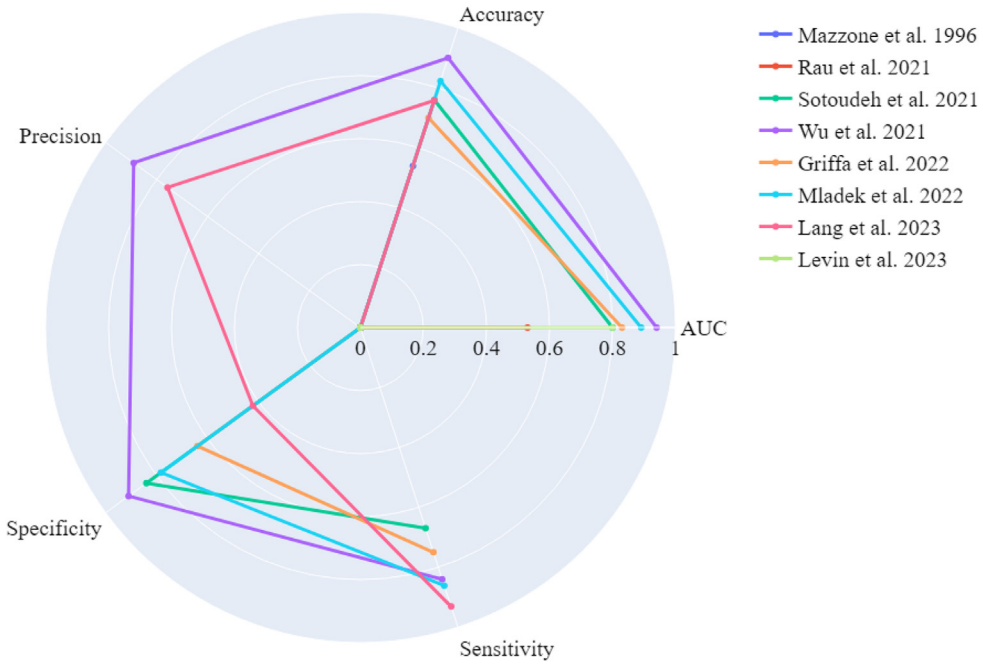

Figure 5. AI algorithms performance metrics. Figure 5 illustrates the performance metrics of various AI algorithms, as reported by each article's authors. The radar chart specifically showcases 5 key metrics: Area Under the Curve (AUC), Accuracy, Precision, Specificity, and Sensitivity. Each vertex of the radar chart corresponds to one of these metrics, scaled from 0 to 1, where 1 denotes optimal performance. For the sake of graphical illustration, a metric value of 0 is equivalent to "NR" (Not Reported), indicating that the specific performance metric was not disclosed in the respective study. The algorithms reported by "Wu et al. 2021" and "Mladek et al. 2022" exhibit comprehensive proficiency across most metrics, with particularly high AUC and Accuracy. Conversely, other studies present more variable results, with several metrics marked as NR, denoting a lack of reported data. AI, artificial intelligence.

Validation Methods. Each study employed cross-validation techniques, with a particular emphasis on leave-one-out validation. Only Rau et al.[14] used k-fold cross-validation as their chosen approach.

Output and Outcome Measures. Four articles determined their outcomes after shunt surgery and the other half after CSF drainage procedures. Wu et al.[15] assessed large volume lumbar puncture (LVLP) response, Griffa et al.[16] employed the TT, and Mladek et al.[17] used external lumbar drainage (ELD). Lang et al.[18] used either LVLP or ELD. To assess drainage response, score changes in clinical scales or tests like the Timed Up and Go were employed (**Table 3**).

The mean AUC of the studies comparing the clinical outcome (7 of 8) was 0.8522 (range: 0.53 to 0.891). (**Figure 5**).

Appraisal of AI Models and Studies

TRIPOD Assessment. TRIPOD adherence, excluding the items considered "not applicable," ranged from 42.86% to 85.71%. Mean adherence was 72.41%. 7 items had total (100%) adherence: 2, 3a, 3b, 4a, 6a, 10b, 13b, 14b, and 19b. Items 10c, 10e, 11, 12, 13c, 17, and 19a were considered "not applicable" to the majority of the articles. This derives from the model development nature of each article and the lack of validation on other datasets.

PROBAST Assessment. After a thorough evaluation, 50% of the studies included in this review were considered to have a high ROB, predominantly due to inadequate outcome determination, for example, by including nonobjective measurements, such as "clinical expertise" to evaluate drainage response. 3 of 8 (37.5%) have a low ROB followed by 1 with unclear ROB. Outcome applicability was considered unclear in the majority of the articles. This was mostly due to the usage of CSF temporary drainage methods, such as TT, LVLP, and ELD, and the lack of shunt surgery results. This led to an overall unclear and/or high concern for applicability (**Table 4**).

DISCUSSION

This systematic review synthesizes the findings from 8 studies employing AI and ML algorithms in predicting SR or CSF drainage in patients with iNPH. The review highlights the evolving landscape of AI/ML applications in this field, with a notable sharp increase in publications from 2021 onwards.

The use of diverse inputs, including demographic data, clinical noninvasive and invasive data, imaging/radiomics, and CSF biomarkers, demonstrates the multidimensional nature of iNPH and the complexity of predicting treatment outcomes. The

Table 4. Critical Appraisal of Included Studies

	Risk of Bias				Applicability			Overall	
Author, Year	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	Risk of Bias	Applicability
Mazzone et al., 1996[13]	-	+	-	-	-	?	-	-	-
Rau et al., 2021[14]	?	?	?	+	?	-	-	?	-
Sotoudeh et al., 2021[12]	+	+	+	+	+	+	+	+	+
Wu et al., 2021[15]	+	-	+	+	+	?	?	-	?
Griffa et al., 2022[16]	+	+	+	+	+	+	?	+	+
Mladek et al., 2022[17]	+	+	+	+	+	+	?	+	+
Lang et al., 2023[18]	+	+	-	?	+	?	?	-	?
Levin et al., 2023[11]	?	+	-	+	-	+	?	-	-

Table 4 shows each article's critical appraisal according to the Prediction model Risk Of Bias Assessment Tool (PROBAST).

predominant use of SVMs, with varied performance across different studies, indicates a preference for this algorithm in the field, although the comparative effectiveness of multiple algorithms in some studies suggests that no single model is universally superior. SVM's popularity for healthcare applications has been rising since the beginning of this century.[19,20] It has been used both for diagnostic and prognostic purposes and as of today still seems to be the most commonly employed ML algorithm for disease prediction.[21] This might be due to a variety of factors: ability to efficiently handle high-dimensional data, robustness against overfitting, versatility through different kernel functions, and effectiveness in small sample sizes.[22] The top-performing model in Lang et al.[18] intriguingly employed only age as a demographic parameter, complemented by extracted average grey matter volume and callosal angle. This combination of parameters demonstrates a key insight: sophisticated and accurate predictive models do not necessarily require an extensive array of variables. Instead, strategic selection of highly relevant features can lead to robust predictions. Dimensionality reduction techniques like Principal Component Analysis are often employed to identify the most relevant features from large datasets, reducing the complexity of models without significantly compromising their predictive accuracy. However, in regards to brain imaging, other techniques have been employed as Principal Component Analysis can limit each component's interpretation. For instance, Varikuti et al.[23] evaluated non-negative matrix factorization approaches for representing grey matter volume data in age prediction frameworks, suggesting that this technique could yield more interpretable and clinically relevant results from complex brain imaging data.

The TRIPOD adherence ranging from 42.86% to 85.71% indicates variable reporting quality, with certain key items universally adhered to and others deemed not applicable. This suggests a need for tailored reporting guidelines for AI/ML models in healthcare. The PROBAST assessment, revealing a high ROB in half of the studies, mainly due to poor outcome determination methods, can limit the reliability of the findings. The landscape of ML-based prediction model studies is rapidly expanding, creating, validating, and updating models. Accurate reporting of essential details is crucial for readers to assess study quality and grasp findings, particularly those related to these predictive models. Compliance with TRIPOD guidelines enhances the transparency of ML models in neurosurgery. While ML-based neurosurgical prediction models demonstrated reasonable adherence to these guidelines, there is room for improvement. As ML tools are increasingly employed for more complex neurosurgical procedures, thorough and accurate reporting of these algorithms is crucial to maintaining the trust of clinicians.[24]

To put the AI studies and this review into the context of the current literature, a recent triad of meta-analyses conducted by Thavarajasingam et al. from 2021 to 2023[25-27] elucidated the most relevant predictors for SR in different domains. They identified intracranial pressure monitoring as the most effective clinical predictor of shunt responsiveness in iNPH. The study found that a patient with shunt-responsive iNPH is significantly more likely to have positive intracranial pressure monitoring results compared to shunt-unresponsive iNPH patients. The mean pulse amplitude cut-off of $\geq$4 mmHg in 70% of recording times was particularly predictive of SR.[25] This somewhat aligns with the inclusion of clinical invasive data in some of the AI models in our review, particularly those by Mazzone[13] and Mladek,[17] which used CSF pressure patterns and intracranial pressure features. Levin et al.[11] found no significant difference in Aβ42 and p-Tau181 levels, as well as the p-Tau181/total-Tau ratio, between patients who showed symptom improvement and those who did not after shunt surgery. This suggests that, in their study, these biomarkers were not predictive of shunt responsiveness. In contrast, Thavarajasingam et al. (2022)[26] reported that elevated CSF levels of Tau proteins, including P-Tau, were indicative of nonresponsiveness to shunt surgery, which aligns with the understanding that Tau proteins are linked to neurodegenerative processes, suggesting their potential role in influencing treatment outcomes. Regarding imaging data, in 2023 they reported that only the callosal angle and periventricular white matter changes were significant, albeit weak, radiological predictors for differentiating iNPH shunt responders from

nonresponders. Other radiological markers, such as Evan's index, disproportionately enlarged subarachnoid space hydrocephalus, cerebral blood flow, and computed tomography cisternography, were not significantly effective in this differentiation.[27] The majority of the studies in our review used various MRI data, with the model by Lang et al.[18] including the callosal angle as a predictor. While the AI studies included seem to have recognized the value of all markers to predict SR, which are included in the Japanese Guidelines for iNPH management too, the absence of any mention of the potential role of AI or ML by the guidelines in this condition is unfortunate and must be changed.

Collectively, these findings underscore the need for methodological advancements, including multicenter collaborations and the incorporation of comprehensive, multifaceted data inputs, as well as consensus on outcome measurements, to enhance the accuracy and applicability of AI/ML models in predicting shunt responsiveness in iNPH patients. The systematic implementation of comparative evaluations of ML methodologies in the studies reviewed is critical for enriching our understanding of the appropriateness and efficacy of various algorithms in this specific field. Through methodical examination and juxtaposition of different ML strategies, we can acquire vital insights into the subtle complexities of model performance and adaptability to the specificities of the datasets and research questions.[28,29]

Limitations

Our review has several limitations that should be considered when interpreting the results. First, the number and quality of the included studies were limited, as most of them were retrospective, single-center, and small-scale studies, which may introduce bias and confounding factors, and limit the generalizability and reproducibility of the findings. Moreover, the lack of international diversity (**Figure 2A**), with all studies conducted in a single country, reduces the representativeness and applicability of the results to different populations and settings. Second, the heterogeneity of the data and methods used in the studies, such as the definition and measurement of SR or CSF drainage, the selection and extraction of features, the choice and evaluation of algorithms, and the reporting and interpretation of results, made it difficult to compare and synthesize the findings across studies, and to enhance the reliability and robustness of the models. Third, the variation in sample sizes and the range of SR rates suggest diverse patient populations and differences in treatment efficacy, which may affect the accuracy and validity of the models. Fourth, most of the articles included in our review did not provide sufficient information regarding specific symptom improvements such as gait, cognition, or urinary function. The outcome measures were generally reported as overall SR or CSF drainage response, often without detailed breakdowns of individual symptom trajectories. Therefore, while we acknowledge that symptom-specific data could provide valuable insights, the current body of literature does not provide enough consistent or detailed information to facilitate a comprehensive discussion on this aspect. Fifth, the lack of external validation or prospective evaluation of the AI/ML models in the studies, as well as the absence of comparison with conventional methods or clinical judgment, raised questions about the validity and applicability of the models in real-world settings. Sixth, the ethical, legal, and social implications of using AI/ML models for SR prediction in iNPH, such as the issues of data privacy, security, ownership, consent, accountability, transparency, and trust, were not addressed or discussed in the studies, which may pose challenges and risks for the adoption and implementation of the models in practice.

CONCLUSION

In conclusion, this study summarizes the current applications of AI in predicting treatment response in iNPH. The highest AUC reported was 0.94, indicating an AI model with strong overall predictive power, with the other models included also all demonstrating moderate to high accuracy, specificity, and sensitivity. These performance metrics highlight the potential of AI to support clinical decision-making in iNPH treatment, yet our findings also reveal the necessity for standardized outcome measures, as well as further validation in prospective studies to enhance the robustness and clinical utility of AI applications. The integration of such advanced methodologies holds promise for a more nuanced and precise management of iNPH, provided future research adheres to rigorous standards and ethical considerations within a scientifically sound and clinically relevant framework.

AVAILABILITY OF DATA AND MATERIAL

All data used for the study have been included in the manuscript.

CRediT AUTHORSHIP CONTRIBUTION STATEMENT

Rafael Tiza Fernandes: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Filipe Wolff Fernandes:** Writing – original draft, Validation, Investigation, Formal analysis, Data curation, Conceptualization. **Mrinmoy Kundu:** Writing – original draft, Validation, Investigation, Formal analysis, Data curation. **Daniele S.C. Ramsay:** Writing – review & editing, Writing – original draft, Conceptualization. **Ahmed Salih:** Writing – review & editing, Writing – original draft, Conceptualization. **Srikar N. Namireddy:** Writing – review & editing, Writing – original draft, Conceptualization. **Dragan Jankovic:** Writing – review & editing, Writing – original draft, Conceptualization. **Darius Kalasauskas:** Writing – review & editing, Writing – original draft, Conceptualization. **Malte Ottenhausen:** Writing – review & editing, Writing – original draft, Conceptualization. **Andreas Kramer:** Writing – review & editing, Writing – original draft, Conceptualization. **Florian Ringel:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Investigation, Formal analysis, Conceptualization. **Santhosh G. Thavarajasingam:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

REFERENCES

1. 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*. 2019;14:e0217705.
2. Wang Z, Zhang Y, Hu F, Ding J, Wang X. Pathogenesis and pathophysiology of idiopathic normal pressure hydrocephalus. *CNS Neurosci Ther*. 2020;26:1230-1240.
3. Sundström N, Lundin F, Arvidsson L, Tullberg M, Wikkelsø C. The demography of idiopathic normal pressure hydrocephalus: data on 3000 consecutive, surgically treated patients and a systematic review of the literature. *J Neurosurg*. 2022;137:1310-1320.
4. Isaacs AM, Hamilton M. Natural history, treatment outcomes and quality of life in idiopathic normal pressure hydrocephalus (iNPH). *Neurol India*. 2021;69(Supplement):S561-S568.
5. Research Committee of Idiopathic Normal Pressure Hydrocephalus, Nakajima M, Yamada S, Miyajima M, 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.
6. Fernandez-Felix BM, López-Alcalde J, Roqué M, Muriel A, Zamora J. CHARMS and PROBAST at your fingertips: a template for data extraction and risk of bias assessment in systematic reviews of predictive models. *BMC Med Res Methodol*. 2023;23:44.
7. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51-58.
8. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br Med J*. 2015;350:g7594.
9. Heus P, Damen JAAG, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open*. 2019;9:e025611.
10. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *Br Med J*. 2020;368:l6890.
11. Levin Z, Leary OP, Mora V, et al. Cerebrospinal fluid transcripts may predict shunt surgery responses in normal pressure hydrocephalus. *Brain*. 2023;146:3747-3759.
12. Sotoudeh H, Sadaatpour Z, Rezaei A, et al. The role of machine learning and radiomics for treatment response prediction in idiopathic normal pressure hydrocephalus. *Cureus*. 2021;13:e18497.
13. Mazzone P, Fortuna L, Arena P, Pisani R. Multilayer neural network analysis of cerebrospinal fluid pressure patterns in idiopathic normal-pressure hydrocephalus. *Technol Health Care*. 1996;4:393-401.
14. Rau A, Kim S, Yang S, et al. SVM-based normal pressure hydrocephalus detection. *Clin Neuroradiol*. 2021;31:1029-1035.
15. 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.
16. Griffa A, Bommarito G, Assal F, et al. CSF tap test in idiopathic normal pressure hydrocephalus: still a necessary prognostic test? *J Neurol*. 2022;269:5114-5126.
17. Mládek A, Gerla V, Skalický P, et al. Prediction of shunt responsiveness in suspected patients with normal pressure hydrocephalus using the lumbar infusion test: a machine learning approach. *Neurosurgery*. 2022;90:407-418.
18. Lang S, Dimond D, Isaacs AM, et al. Use of cortical volume to predict response to temporary CSF drainage in patients with idiopathic normal pressure hydrocephalus. *J Neurosurg*. 2023;139:1776-1783.
19. Noble WS. What is a support vector machine? *Nat Biotechnol*. 2006;24:1565-1567.
20. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database (Oxford)*. 2020;2020:baaa010.
21. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inf Decis Making*. 2019;19:281.
22. Burbidge R. Adaptive kernels for support vector classification. In: Denison DD, Hansen MH, Holmes CC, Mallick B, Yu B, eds. *Nonlinear Estimation and Classification. Lecture Notes in Statistics*. 171. New York, NY: Springer; 2003.
23. Varikuti DP, Genon S, Sotiras A, et al. Evaluation of non-negative matrix factorization of grey matter in age prediction. *Neuroimage*. 2018;173:394-410.
24. Warman A, Kalluri AL, Azad TD. Machine learning predictive models in neurosurgery: an appraisal based on the TRIPOD guidelines. Systematic review. *Neurosurg Focus*. 2023;54:E8.
25. Thavarajasingam SG, El-Khatib M, Rea M, 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 (Wien)*. 2021;163:2641-2672.
26. Thavarajasingam SG, El-Khatib M, Vemulapalli KV, et al. Cerebrospinal fluid and venous biomarkers of shunt-responsive idiopathic normal pressure hydrocephalus: a systematic review and meta-analysis. *Acta Neurochir (Wien)*. 2022;164:1719-1746.
27. Thavarajasingam SG, El-Khatib M, Vemulapalli K, 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 (Wien)*. 2023;165:369-419.
28. Senders JT, Staples PC, Karhade AV, et al. Machine learning and neurosurgical outcome prediction: a systematic review. *World Neurosurg*. 2018;109:476-486.e1.
29. Senders JT, Zaki MM, Karhade AV, et al. An introduction and overview of machine learning in neurosurgical care. *Acta Neurochir (Wien)*. 2018;160:29-38.

Conflict of interest statement: The authors declare that the article content was composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received 14 June 2024; accepted 17 September 2024

Citation: World Neurosurg. (2024) 192:e281-e291. https://doi.org/10.1016/j.wneu.2024.09.087

*Journal homepage: **www.journals.elsevier.com/world-neurosurgery***

*Available online: **www.sciencedirect.com***