Automated Identification of Multiple Findings on Brain MRI for Improving Scan Acquisition and Interpretation Workflows
A Systematic Review
Sheng, Kaining; Offersen, Cecilie Morck; Middleton, Jon; Carlsen, Jonathan Frederik; Truelsen, Thomas Clement; Pai, Akshay; Johansen, Jacob; Nielsen, Michael Bachmann

Published in:
Diagnostics

DOI:
10.3390/diagnostics12081878

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Automated Identification of Multiple Findings on Brain MRI for Improving Scan Acquisition and Interpretation Workflows: A Systematic Review

Kaining Sheng 1,2,*, Cecilie Mørck Offersen 1,2, Jon Middleton 3,4, Jonathan Frederik Carlsen 1,2, Thomas Clement Truelsen 2,5, Akshay Pai 1,4, Jacob Johansen 3,4 and Michael Bachmann Nielsen 1,2

Abstract: We conducted a systematic review of the current status of machine learning (ML) algorithms’ ability to identify multiple brain diseases, and we evaluated their applicability for improving existing scan acquisition and interpretation workflows. PubMed Medline, Ovid Embase, Scopus, Web of Science, and IEEE Xplore literature databases were searched for relevant studies published between January 2017 and February 2022. The quality of the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 tool. The applicability of ML algorithms for successful workflow improvement was qualitatively assessed based on the satisfaction of three clinical requirements. A total of 19 studies were included for qualitative synthesis. The included studies performed classification tasks (n = 12) and segmentation tasks (n = 7). For classification algorithms, the area under the receiver operating characteristic curve (AUC) ranged from 0.765 to 0.997, while accuracy, sensitivity, and specificity ranged from 80% to 100%, 72% to 100%, and 65% to 100%, respectively. For segmentation algorithms, the Dice coefficient ranged from 0.300 to 0.912. No studies satisfied all clinical requirements for successful workflow improvements due to key limitations pertaining to the study’s design, study data, reference standards, and performance reporting. Standardized reporting guidelines tailored for ML in radiology, prospective study designs, and multi-site testing could help alleviate this.

Keywords: artificial intelligence; machine learning; brain MRI; brain diseases; workflow

1. Introduction

Brain magnetic resonance imaging (MRI) is recognized as the imaging modality that produces the best images of brain tissues, body fluids, and fat [1]. It remains the most appropriate modality for diagnosing patients with symptoms of multiple brain diseases including inflammatory diseases, dementia, neurodegenerative disease, cerebrovascular disease, and brain tumors [2–5]; hence, it plays an important role in multiple clinical scenarios ranging from acute diagnostics to routine follow-ups.

A brain MRI scan typically consists of several scan sequences, the most commonly included being T1-weighted (T1) and T2-weighted (T2) sequences, a diffusion-weighted imaging (DWI) sequence, a fluid attenuated inversion-recovery (FLAIR) sequence, and a bleeding sensitive sequence, e.g., T2* gradient-recall-echo (T2*-GRE) [6]. Selecting the appropriate sequences a priori can be challenging, because many brain diseases often
present with the same symptoms [7] while requiring different combinations of sequences for correct diagnosis. Inefficient MR sequence selection can increase the risk of inconclusive scans, scan recalls, and inappropriate usage of gadolinium contrast agents [8]. It can also cause the redundant scanning of patients and result in increased patient inconvenience [8], higher risks of radiologist burnout due to increased workload [9], and prolonged reporting time of potentially time-sensitive diseases [10].

In recent years, machine learning (ML) has been increasingly applied in neuroimaging to alleviate some of these challenges using automated workflow improvements [11]. A potential application of ML algorithms could be to automate scan-sequence acquisition alterations based on real-time image analysis while the patient is still in the scanner [12]. Another application could be to improve scan interpretation efficiencies by prioritizing the reading list of essential and acute/critical findings [11]. However, to improve any existing workflows, ML algorithms require the satisfaction of at least three essential requirements. First, the ML algorithms must be developed and tested in a scenario that reflects clinical practice [13,14]. For improving scan acquisition and interpretation workflows, this means ML algorithms capable of automatically identifying or differentiating between multiple brain diseases with the identification of brain infarcts, hemorrhages, and tumors, being a must due to their frequent and time-critical nature [15,16]. This also requires consecutive datasets not prone to spectrum biases and ground-truth labels unaffected by selection bias [13]. Secondly, tests of the ML algorithm should be performed on an out-of-distribution dataset to account for potential overfitting [13,17]. This could for instance be achieved by testing the algorithms on an external dataset sourced from a different point in time or geographical location compared to the training dataset. Finally, if ML algorithms are to gain widespread trust and usage, their technical performance results should be acceptable with respect to balancing increased workloads from false-positive findings and risk of missing important findings from false-negative results [18]. One method of assessing whether an ML algorithm performs a certain task acceptably could be to compare its performance to that of domain experts performing similar tasks.

Many studies about ML algorithms used within neuroradiology exist [19]. However, few of them address the important question of whether these algorithms could bring actual benefits to clinicians and patients if they were deployed today. To address this knowledge gap, we conducted a systematic review on how well the most recent ML algorithms could identify multiple brain diseases with the aim of evaluating their applicability for improving existing scan acquisition and interpretation workflows based on the satisfaction of the aforementioned requirements.

2. Materials and Methods

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [20]. The study protocol was registered in the Prospective Register of Systematic Reviews (PROSPERO) under number CRD42022329801 during the research process.

2.1. Literature Search

The literature was searched in MEDLINE (accessed through PubMed), Ovid Embase, Elsevier Scopus, Clarivate Web of Science, and IEEE Explore in order to find studies covering both clinical and technical aspects of the review question. The search period ranged between 1 January 2017 and 10 February 2022. This relatively short period was selected due to the nature of deep learning research, with rapid development cycles rendering older studies less relevant for the review question. Structured search terms (MeSH, Emtree), such as “magnetic resonance imaging”, “machine learning”, and “brain diseases”, were combined using Boolean operators and supplemented with free keyword search terms, such as “detection”, “classification”, “triaging”, or “workflow”. Multiple additional keywords describing pathologies of interest, such as “neoplasm”, “stroke”, “intraparenchymal hem-
orrhage”, “subarachnoid hemorrhages”, and “subdural hemorrhages”, were also included in the search string. The full search string can be found in Appendix A.

2.2. Study Selection

Records that developed ML algorithms for the automated identification of normal and abnormal brain diseases were screened. Main inclusion and exclusion criteria are listed in Table 1.

Table 1. Inclusion and exclusion criteria.

<table>
<thead>
<tr>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studies focusing on abnormal brain diseases that included either brain infarct, hemorrhage, or tumor on brain MRI</td>
<td>Studies focusing on tasks not relevant for identification of brain diseases</td>
</tr>
<tr>
<td>Studies developing algorithms tested on a dataset that was separate from the training dataset</td>
<td>Studies focusing on identification of a single brain disease only</td>
</tr>
<tr>
<td>Peer-reviewed studies in English</td>
<td>Studies focusing on development of ML for specialized MR sequences (e.g., MR elastography, functional MRI) or other imaging modalities (e.g., SPECT, PET, CT, US)</td>
</tr>
<tr>
<td></td>
<td>Studies with primarily non-adult populations</td>
</tr>
<tr>
<td></td>
<td>Editorials, case series, letters, conference proceedings, reviews, and inaccessible papers</td>
</tr>
</tbody>
</table>

Two medical doctors (K.S. and C.M.O.) served as reviewers. They independently screened all records based on title and abstract. This was followed by the extraction of relevant reports for full-text screening and final study inclusion. For the process of record and report screening, Covidence (Melbourne, Australia) was used. Discussions between both reviewers were held to resolve any conflicts, but if a consensus was not reached, a third reviewer (J.F.C.) was consulted.

2.3. Data Extraction and Analysis

The reviewers independently extracted data from the included studies according to a pre-defined datasheet. Study and algorithm characteristics were extracted and comprised of the following: (a) study information, (b) population/dataset characteristics including number of patients or images, pathology in the dataset, and MR sequences available, (c) aim of algorithms, (d) type of algorithm, and (e) training and testing strategies including how data-splits were performed. Reported performance metrics together with confidence intervals were also extracted, including accuracy, sensitivity, specificity, F1-score, negative predictive value (NPV), positive predictive value (PPV), and area under the receiver operating characteristic curve (AUC). Furthermore, the Dice score coefficient (DSC), which is one of the most common evaluation metrics used in medical segmentation tasks [21], was extracted where applicable in brain segmentation studies. Performance numbers were summarized using descriptive statistics. If multiple results were reported for different variations of the same algorithm, only the best performance result was extracted unless otherwise stated. When available, performance results were extracted from external test datasets. Included studies were divided by tasks of the included algorithms. The analysis of data was primarily conducted using pivot tables and the in-built analysis tools of Microsoft Excel.

2.4. Quality Assessment of Included Studies

The two reviewers independently assessed the quality of the included studies by using the tailored questionnaire Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) [22] with signaling questions covering risk of bias and concern for applicability in the domains of patient selection, index test, reference test, and flow and
For each study, the respective domains were graded as high-, unclear-, or low-risk of bias/concern for applicability. Discordance between the reviewers was resolved through discussion.

2.5. Evaluation of Applicability for Workflow Improvements

The applicability of each included ML algorithm for improving scan acquisition and interpretation workflows was qualitatively assessed based on three essential requirements previously mentioned in the Introduction (Section 1): (A) reflection of clinical practice, (B) testing on an external out-of-distribution dataset, and (C) acceptable performance results. Each requirement was indicated as being ‘Satisfied’ (S) or ‘Not Satisfied’ (NS). The first requirement (A) was satisfied if the patient population was consecutively sampled, if the disease distribution was well reported, and if the study was assessed as having low risk of bias/concern for applicability for the review question. The second requirement (B) was satisfied if external test datasets with data from a different time-period and geographical location were used to produce performance results. Lastly, the third requirement (C) was graded satisfied if a majority of the abovementioned result metrics exceeded a pre-defined threshold of 85% of the maximum attainable value. This threshold for acceptable performance was selected because it reflected the performance levels of a neuroradiologist when performing similar disease identification tasks [23].

3. Results

3.1. Study Selection and Data Extraction

The search of electronic databases returned 5688 records. The removal of duplicates resulted in 3542 records. Screening record titles and abstracts resulted in 81 reports selected for full-text eligibility assessments, of which 19 studies were included for qualitative review. The study inclusion process is illustrated in Figure 1.

Details about the study’s characteristics are summarized in Table 2. All 19 studies included were of retrospective design. Study populations varied with regard to source and size. Twelve out of nineteen studies (63%) used public datasets for development and testing of their algorithms. These datasets included The Whole Brain Atlas from Harvard Medical School (HMS) [24], The Cancer Imaging Archive (TCIA) [25], the Brain Tumor Segmentation (BRATS) challenge dataset [26], and the Ischemic Stroke Lesion Segmentation (ISLES) challenge set [27]. Study populations varied between 100 and 500 patients with one study (5%) having a population of less than 100 patients and five studies (26%) having a population larger than 1000 patients. All large study populations were private, i.e., part of a local in-house dataset not publicly available to researchers outside of the research institution in question. Six studies (31%) only reported on data size as the number of 2D images ranging from 200 images to 4600 images. Training, validation, and testing of algorithms were on average performed using 69%, 3%, and 28% of all available data, respectively. Validation was performed only in six (31%) studies. Testing was mostly performed on data split out from the same data source; however, an external dataset with data from a different time-period and geographical location was used in three (15%) studies.
Figure 1. PRISMA workflow: 5688 records screened, and 19 studies were included for qualitative review.
Table 2. Study population and data characteristics.

<table>
<thead>
<tr>
<th>Author</th>
<th>Data Source</th>
<th>No. Patients/Images</th>
<th>Training Data</th>
<th>Validation Data</th>
<th>Testing Data</th>
<th>Disease Distribution in Data</th>
<th>MR Sequences Utilized</th>
<th>MR Field Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahmadi et al., 2021 [28]</td>
<td>Private + Harvard Medical School Whole Brain Atlas</td>
<td>1200 images</td>
<td>1120</td>
<td>N/A</td>
<td>80</td>
<td>12.5% normals 87.5% abnormal incl. glioma, Huntington’s disease, Meningioma, and Alzheimer</td>
<td>2D single slice of: Ax T2</td>
<td>1.5 T</td>
</tr>
<tr>
<td>Baur et al., 2021 [29]</td>
<td>Private WMH TCIA</td>
<td>259 patients</td>
<td>100</td>
<td>18</td>
<td>141</td>
<td>42% normal used for unsupervised training 19% multiple sclerosis 15% glioma &amp; glioblastoma 4% microangiopathy 20% WMH</td>
<td>Ax T2-FLAIR</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Duong et al., 2019 [30]</td>
<td>Private</td>
<td>387 patients</td>
<td>295</td>
<td>N/A</td>
<td>92</td>
<td>Normal and 19 different abnormalities incl. MS, high grade glioma, and vascular (acute or subacute ischemia)</td>
<td>Ax T2-FLAIR</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Fayaz et al., 2021 [31]</td>
<td>Harvard Medical School Whole Brain Atlas</td>
<td>4100 images</td>
<td>2870</td>
<td>N/A</td>
<td>1230</td>
<td>50% normal 50% abnormal incl. glioma, meningioma, and Alzheimer</td>
<td>2D single slice of: Ax T2</td>
<td>1.5 T</td>
</tr>
<tr>
<td>Felipe Fattori Alves et al., 2020 [32]</td>
<td>Private</td>
<td>67 patients</td>
<td>50</td>
<td>N/A</td>
<td>17</td>
<td>45% inflammatory lesion (incl. MS, vasculitis, toxoplasmosis, pyogenic and septic-embolic brain abscess, etc.) 55% brain tumors (incl. Glioblastoma, anaplastic astrocytoma, anaplastic ependymoma)</td>
<td>Ax T1 &amp; T1 + C  Ax T2  Ax T2-FLAIR  Ax DWI</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Gauriau et al., 2021 [33]</td>
<td>Private</td>
<td>10,770 patients</td>
<td>7795</td>
<td>473</td>
<td>2502</td>
<td>Normal and 8+ different abnormalities including infarct, hemorrhage, neoplasm, demyelination, and infections</td>
<td>Ax T2-FLAIR</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Gilanie et al., 2018 [34]</td>
<td>Harvard Medical School Whole Brain Atlas</td>
<td>4589 images</td>
<td>3029</td>
<td>N/A</td>
<td>1560</td>
<td>11% normal 89% abnormal incl. cerebrovascular, neoplasm, neurodegenerative, and inflammatory disease</td>
<td>2D single slices of: Ax T1 &amp; T1 + C  Ax T2  Ax T2-FLAIR  Ax DWI</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Han et al., 2020 [35]</td>
<td>OASIS-3 Private</td>
<td>1162 patients</td>
<td>543</td>
<td>N/A</td>
<td>619</td>
<td>47% normals used for unsupervised training 19% normals used for testing 21% dementia of varying degree 7% brain metastasis 6% various disease incl. small infarct, hemorrhage, and white matter lesions</td>
<td>Ax T1 &amp; Ax T1 + c</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Hu et al., 2020 [36]</td>
<td>BRATS 2019 ISLES 2017</td>
<td>459 patients</td>
<td>317</td>
<td>N/A</td>
<td>142</td>
<td>84% glioma (HGG, LGG) 16% acute &amp; subacute infarct</td>
<td>Ax T1 &amp; T1 + C  Ax T2  Ax T2-FLAIR  Ax DWI, ADC, perfusion</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Kamnitas et al., 2017 [37]</td>
<td>Private BRATS 2015 ISLES 2015</td>
<td>509 patients</td>
<td>348</td>
<td>N/A</td>
<td>161</td>
<td>75% tumor (high grade glioma, low grade glioma) 13% acute &amp; subacute infarct 12% traumatic brain injury</td>
<td>Ax or Sag T1 &amp; T1 + C  Ax T2 &amp; Ax PD  Ax T2-FLAIR  Ax T2 * GRE  Ax DWI &amp; ADC</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Kim et al., 2021 [38]</td>
<td>BRATS 2019 ISLES 2015</td>
<td>259 patients</td>
<td>239</td>
<td>N/A</td>
<td>26</td>
<td>36% normal 60% glioma 4% acute &amp; subacute infarct</td>
<td>2D slices of: Ax T1 &amp; T1 + C  Ax T2  Ax T2-FLAIR  Ax DWI, ADC, perfusion</td>
<td>1.5 T 3 T</td>
</tr>
</tbody>
</table>
Table 2. Cont.

<table>
<thead>
<tr>
<th>Author</th>
<th>Data Source</th>
<th>No. Patients /Images</th>
<th>Training Data</th>
<th>Validation Data</th>
<th>Testing Data</th>
<th>Disease Distribution in Data</th>
<th>MR Sequences Utilized</th>
<th>MR Field Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lu et al., 2021 [39]</td>
<td>Private</td>
<td>7134 patients</td>
<td>* 5002</td>
<td>1061</td>
<td>1071</td>
<td>13% acute/subacute stroke &lt;br&gt;87% non-stroke abnormalities incl. tumor, hemorrhage and normals</td>
<td>Axial T2-FLAIR &lt;br&gt;Axial DWI + ADC</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Lu, Lu et Zhang., 2019 [40]</td>
<td>Harvard Medical School Whole Brain Atlas</td>
<td>291 images</td>
<td>204</td>
<td>N/A</td>
<td>87</td>
<td>39% normal &lt;br&gt;61% abnormal incl. neoplasm, neurodegenerative, and inflammatory disease</td>
<td>2D single slice of: &lt;br&gt;Ax T2</td>
<td>1.5 T</td>
</tr>
<tr>
<td>Nael et al., 2021 [41]</td>
<td>Private</td>
<td>13,215 patients</td>
<td>9845</td>
<td>1248</td>
<td>2122</td>
<td>17% normal &lt;br&gt;11% acute infarction &lt;br&gt;5% acute hemorrhage &lt;br&gt;4% intracranial mass effect &lt;br&gt;63% other abnormalities including white matter lesions</td>
<td>Ax or Sag T1 &amp; T1 + C &lt;br&gt;Ax T2 &lt;br&gt;Ax T2-FLAIR &lt;br&gt;Ax ADC &amp; DWI &lt;br&gt;Ax T2 * GRE</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Nayak et al., 2020 [42]</td>
<td>Harvard Medical School Whole Brain Atlas &amp; 2D single slice of: &lt;br&gt;Ax T2</td>
<td>275 images</td>
<td>165</td>
<td>N/A</td>
<td>110</td>
<td>20% normal &lt;br&gt;20% stroke &lt;br&gt;20% neurodegenerative &lt;br&gt;20% infectious &lt;br&gt;20% neoplasm</td>
<td>2D single slice of: &lt;br&gt;Ax T2</td>
<td>1.5 T</td>
</tr>
<tr>
<td>Nayak et al., 2020 [43]</td>
<td>Harvard Medical School</td>
<td>200 images</td>
<td>120</td>
<td>N/A</td>
<td>80</td>
<td>20% normal &lt;br&gt;20% stroke &lt;br&gt;20% neurodegenerative &lt;br&gt;20% infectious &lt;br&gt;20% neoplasm</td>
<td>2D single slice of: &lt;br&gt;Ax T2</td>
<td>1.5 T</td>
</tr>
<tr>
<td>Pereira et al., 2019 [44]</td>
<td>BRATS 2013 &lt;br&gt;BRATS 2017 &lt;br&gt;ISLES 2017</td>
<td>471 patients</td>
<td>358</td>
<td>10</td>
<td>103</td>
<td>89% tumor (high grade glioma, low grade glioma) &lt;br&gt;11% acute &amp; subacute infarct</td>
<td>Ax T1 &amp; T1 + C &lt;br&gt;Ax T2 &lt;br&gt;Ax T2-FLAIR &lt;br&gt;Ax DWI, ADC, perfusion</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Rauschecker et al., 2020 [23]</td>
<td>Private</td>
<td>178 patients</td>
<td>86</td>
<td>N/A</td>
<td>92</td>
<td>19 different abnormalities incl. MS, high grade glioma, and vascular (acute or subacute ischemia)</td>
<td>Ax T1 + C &lt;br&gt;Ax T2 &lt;br&gt;Ax T2-FLAIR &lt;br&gt;Ax T2 * GRE &lt;br&gt;Ax DWI &amp; ADC</td>
<td>1.5 T 3 T</td>
</tr>
<tr>
<td>Wood et al., 2022 [45]</td>
<td>Private</td>
<td>71,206 patients</td>
<td>53,409</td>
<td>9425</td>
<td>7372</td>
<td>Normal and 90+ different abnormalities including vascular disease, neoplasms, demyelination, and atrophy</td>
<td>Ax T2-FLAIR &lt;br&gt;Ax DWI</td>
<td>1.5 T 3 T</td>
</tr>
</tbody>
</table>

Abbreviations: WMH = White Matter Hyperintensity Challenge; BRATS = Brain Tumor Segmentation Challenge Data; HGG = High Grade Glioma; LGG = Low Grade Glioma; ISLES = Ischemic Stroke Lesion Segmentation; OASIS-3 = Open Access Series of Imaging Studies; Ax. = axial. Private = local in-house dataset not publicly available. * Train/test distribution only reported for data pertaining subpart of study performing case-level classification of stroke/not stroke. Disease frequencies in the study populations were also varied with diverse samples of cerebrovascular disease, brain tumors, inflammatory disease, neurodegenerative diseases, neuroinfectious diseases, dementia, traumatic brain injury, and various less significant pathologies. All study populations had samples of brain tumors, while 11 (58%) studies had samples of both normal scans, brain infarcts, hemorrhages, and tumors. The brain pathologies were identified on MR sequences found in typical routine acquisitions with T1, T2, and T2-FLAIR being the most prevalent. Eight studies (42%) used only a single scan sequence, either T2 or T2-FLAIR, for disease inference.
All studies developed algorithms focusing on brain disease identification using either classification or segmentation tasks. Seven studies (37%) focused on a binary classification of images into either normal/abnormal or differentiation between two diseases, five studies (26%) focused on a multiclass classification of images into specific disease categories, and seven studies (37%) focused on a multiclass segmentation of specific diseases. Most algorithmic tasks employed deep discriminative models, with 14 (74%) studies using convolutional neural networks (CNN). Three (16%) studies employed deep generative models, with two (11%) studies using variational autoencoders (VAE) and one (5%) study using generative adversarial networks (GANs). Reference tests were mostly labels and delineations made by neuroradiologists. Exceptions were found in the study by Wood et al. [46] where reference labels were generated using natural language processing (NLP) of radiological reports and in the study by Ahmadi et al. [28], where reference delineations were constructed using principal component analysis (PCA). Details about the test setup and performance metrics for binary classification, multiclass classification, and segmentation algorithms are summarized in Tables 3 and 4, respectively.

Different performance measures were reported for each study. For classification studies, the most frequently reported performance metrics were AUC, accuracy, sensitivity, and specificity. AUC ranged from 0.765 to 0.997 while accuracy, sensitivity, and specificity ranged from 80% to 100%, 72% to 100%, and 65% to 100%, respectively. Positive predictive and negative predictive values were reported in nine studies (47%) and ranged from 12% to 94% and 48% to 99%, respectively. The higher performance values were predominantly observed in binary classification studies with a smaller study population, while the lower values were seen when identifying brain tumors. For segmentation studies, the Dice Score Coefficient was the most reported measure ranging from 0.300 for infarct segmentations to 0.912 for glioma and multiple-sclerosis segmentations. Sensitivity and specificity were observed to range from 13% to 99.9% and 87% to 99.8%, respectively, with the lower sensitivity values attributed to brain infarct segmentations.

3.2. Applicability to Workflow Improvement

The applicability of the included ML algorithms for improvements in scan acquisition and interpretation workflows was evaluated based on the satisfaction of the three requirements of (A) testing environments reflecting clinical practice, (B) test on external out-of-distribution datasets, and (C) acceptable algorithm performance results; see Section 2. Evaluation results for each requirement are summarized in Tables 3 and 4 as well. Ten (53%) of nineteen studies were assessed as having acceptable performance; however, only one (5%) satisfied the requirement of being tested in a clinical environment reflecting clinical practice and three (15%) satisfied the requirement of testing on an external out-of-distribution dataset. Three studies (15%) satisfied two main requirements. These studies used privately acquired datasets. No studies satisfied all three main requirements for successful workflow integrations.

3.3. Quality Assessment

The Quality Assessment of Diagnostic Accuracy Studies 2 tool was applied to all included studies in this review. The results of the risk of bias/concern for applicability analysis are presented in Table 5 and summarized in Figure 2.
### Table 3. (a) Performance results of binary classification algorithms. (b) Performance results of multiclass classification algorithms.

<table>
<thead>
<tr>
<th>Author</th>
<th>Aim of Algorithm</th>
<th>Type of Algorithm</th>
<th>Ground Truth</th>
<th>Testing Strategy</th>
<th>Performance Results</th>
<th>Workflow Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fayaz et al., 2021 [31]</td>
<td>Binary classification of normal and abnormal</td>
<td>CNN + DWT</td>
<td>Expert labels</td>
<td>Train-test split</td>
<td>AUC 0.997</td>
<td>(a) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: S</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Acc (%) N/A</td>
<td>Note: High performance observed on single 2D MR slices</td>
</tr>
<tr>
<td>Felipe Fattori Alves et al., 2020 [32]</td>
<td>Binary classification of inflammatory lesions and brain tumors</td>
<td>RF SVM k-NN</td>
<td>Expert delineation</td>
<td>Train-test split</td>
<td>AUC 0.906</td>
<td>(a) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: S</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sens (%) 82.7</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spec (%) 91.2</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F1 (%) 87.5</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPV (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NPV (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Gauriau et al., 2021 [33]</td>
<td>Binary classification of normal and abnormal</td>
<td>CNN</td>
<td>Radiological report</td>
<td>Train-test split incl. external test set</td>
<td>AUC 0.800</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sens (%) 77.0</td>
<td>[75; 80]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spec (%) 65.0</td>
<td>[61; 69]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F1 (%) 78.0</td>
<td>[76; 80]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPV (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NPV (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Gilanie et al., 2018 [34]</td>
<td>Binary classification of normal and abnormal</td>
<td>Gabor filter SVM</td>
<td>Expert labels</td>
<td>Train-test split</td>
<td>AUC 0.970</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: S</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sens (%) 96.5</td>
<td>98.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spec (%) 92.0</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F1 (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPV (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NPV (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Lu et al., 2021 [39]</td>
<td>Binary classification of stroke/non-stroke patients</td>
<td>CNN + Gating attention mechanism ranking of multi-contrast MRI</td>
<td>Expert labels</td>
<td>Train-test split</td>
<td>AUC 0.881</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: S</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sens (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spec (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F1 (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPV (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NPV (%) N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Lu, Lu et Zhang., 2019 [40]</td>
<td>Binary classification of normal and abnormal</td>
<td>CNN + transfer learning</td>
<td>Expert labels</td>
<td>Train-test split</td>
<td>AUC 0.830</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: S</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sens (%) Abn: 80.0</td>
<td>Inf.: 90.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spec (%) Abn: 87.0</td>
<td>Inf.: 82.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F1 (%) Abn: 91.0</td>
<td>Inf.: 94.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPV (%) Abn: 90.0</td>
<td>Inf.: 92.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NPV (%) Abn: 72.0</td>
<td>Inf.: 32.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
<td>ME: 92.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
<td>ME: 98.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
<td>ME: 99.0</td>
</tr>
<tr>
<td>Wood et al., 2022 [45]</td>
<td>Binary classification of normal and abnormal</td>
<td>Ensemble CNN</td>
<td>NLP labelled radiological report</td>
<td>Train-test split incl. external test set</td>
<td>AUC 0.948</td>
<td>(B) External validation: S (C) Performance: S</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sens (%) Dem: 91.9</td>
<td>BM: 89.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spec (%) Dem: 84.2</td>
<td>BM: 82.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F1 (%) Dem: 92.3</td>
<td>BM: 90.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPV (%) Dem: 90.0</td>
<td>BM: 94.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NPV (%) Dem: 88.0</td>
<td>BM: 86.2</td>
</tr>
<tr>
<td>Han et al., 2020 [35]</td>
<td>Multiple binary classification of normal/clinical dementia (Dem), normal/brain metastasis (BM), and normal/various diseases (VO) incl. small infarct and hemorrhage</td>
<td>Unsupervised GAN + 7 Self-attention (SA) modules</td>
<td>Expert label</td>
<td>Train-test split</td>
<td>AUC 0.765</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: S</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sens (%) Dem: 89.9</td>
<td>BM: 93.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spec (%) Dem: 82.2</td>
<td>BM: 86.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F1 (%) Dem: 90.3</td>
<td>BM: 94.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPV (%) Dem: 90.0</td>
<td>BM: 94.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NPV (%) Dem: 91.0</td>
<td>BM: 92.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
<td>BM: 32.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
<td>BM: 12.0</td>
</tr>
<tr>
<td>Nael et al., 2021 [41]</td>
<td>Multiple binary classification of normal/any abnormalities (abn), infarct (inf)/non-infarct, hemorrhage (hem)/non-hemorrhage, and mass effect (ME)/non-mass effect</td>
<td>CNN</td>
<td>Radiological report Expert image delineation</td>
<td>Train-test split incl. external test set</td>
<td>AUC 0.880</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: S</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sens (%) Abn: 80.0</td>
<td>Inf.: 95.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spec (%) Abn: 87.0</td>
<td>Inf.: 81.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F1 (%) Abn: 89.0</td>
<td>Inf.: 94.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPV (%) Abn: 87.0</td>
<td>Inf.: 82.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NPV (%) Abn: 79.0</td>
<td>Inf.: 91.0</td>
</tr>
</tbody>
</table>

Note: High performance observed on single 2D MR slices.
Table 3. Cont.

<table>
<thead>
<tr>
<th>Author et al., 2020 [42]</th>
<th>Aim of Algorithm</th>
<th>Type of Algorithm</th>
<th>Ground Truth</th>
<th>Testing Strategy</th>
<th>Performance Results</th>
<th>Workflow Applicability</th>
</tr>
</thead>
</table>
| Nayak et al., 2020 [42] | Multiclass classification of normal, stroke, tumor, infectious, degenerative | CNN | Expert labels | Train-test split | N/A *** 97.5 | N/A N/A N/A N/A N/A N/A | (A) Reflecting clinical practice: NS  
(B) External validation: NS  
(C) Performance: S  
Note: High test performance result on small test set of 2D MR slices |
| Nayak et al., 2020 [43] | Multiclass classification of normal, stroke, tumor, infectious, degenerative | CNN + ELM | Expert labels | Train-test split | 93.8 | N/A N/A N/A N/A N/A N/A | (A) Reflecting clinical practice: NS  
(B) External validation: NS  
(C) Performance: S  
Note: High test performance result on small test set <100 2D MR slices |
| Rauschecker et al., 2020 [23] | Multiclass classification of 19 brain diseases incl multiple sclerosis (MS), high grade glioma, and vascular infarct defined as correctly classified within top 3 differential diagnosis | CNN + Bayesian inference | Expert labels | Train-test split | 0.920 [0.880; 0.950] 0.910 [0.84; 0.96] | N/A N/A N/A N/A N/A N/A | (A) Reflecting clinical practice: NS  
(B) External validation: NS  
(C) Performance: S  
Note: High test performance result on small test set available in the study |

(a) Abbreviations: AUC = area under receiver operatic characteristics curve; Acc = accuracy; Sens = sensitivity; Spec = specificity; PPV = positive predictive value; NPV = negative predictive value; S = satisfied; NS = Not Satisfied (see Section 2); CNN = convolutional neural network; VAE = variational auto-encoder; ELM = extreme learning machine; GAN = generative adversarial network; RF = Random forest; SVM = Support vector machine; k-NN = k-Nearest Neighbor; DWT = Deep Wavelet Transform.  
* Metrics are reported for a random forest classifier on only T1 images  
** Results reported for subpart of study performing case-level classification of stroke/not stroke  
*** Metrics reported for the largest data subset available in the study  
(b) Abbreviations: ELM = extreme learning machine; GAN = generative adversarial network.
### Table 4. Performance results of segmentation algorithms.

<table>
<thead>
<tr>
<th>Author</th>
<th>Aim of Algorithm</th>
<th>Type of Algorithm</th>
<th>Ground Truth Testing Strategy</th>
<th>Performance Results</th>
<th>Workflow Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahmadi et al., 2021 [28]</td>
<td>Multiclass segmentation incl. neoplasm and neurodegenerative disease</td>
<td>CNN</td>
<td>Synthetic labels via robust PCA</td>
<td>DSC: 0.912, Sens: 99.9, Spec: 99.8, PPV: N/A, NPV: N/A</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: NS</td>
</tr>
<tr>
<td>Baur et al., 2021 [29]</td>
<td>Multiclass segmentation of normal, MS, glioblastoma (GBM), glioma, microangiopathy (MA), and WMH</td>
<td>Unsupervised VAE</td>
<td>Radiological report, Expert image delineation</td>
<td>Train-test split</td>
<td>MS: 0.650, Sens: 62.0, Spec: 99.9, PPV: N/A, NPV: N/A, MA: 0.730, Sens: 36.0, Spec: 13.0, PPV: N/A, NPV: N/A</td>
</tr>
<tr>
<td>Duong et al., 2019 [30]</td>
<td>Multiclass segmentation of 19+ different abnormalities incl. MS, high grade glioma, and infarcts</td>
<td>CNN</td>
<td>Expert image delineation</td>
<td>DSC: 0.789, Sens: 76.7, Spec: 99.9, PPV: 76.9, NPV: N/A</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: NS</td>
</tr>
<tr>
<td>Hu et al., 2020 [36]</td>
<td>Multiclass segmentation of infarct and glioma</td>
<td>CNN</td>
<td>Expert image delineation</td>
<td>Infarct: 0.300, Sens: 43.0, Spec: 87.0, PPV: N/A, NPV: N/A, Glioma: 0.520, Sens: 33; 87, Spec: 47; 81, PPV: N/A, NPV: N/A</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: NS</td>
</tr>
<tr>
<td>Kamnitsas et al., 2017 [37]</td>
<td>Multiclass segmentation of infant, traumatic brain injury (TBI), and glioma</td>
<td>Ensemble CNN</td>
<td>Expert image delineation</td>
<td>Infarct: 0.590, Sens: 60.0, Spec: 87, PPV: N/A, NPV: N/A, TBI: 0.645, Sens: 63.9, Spec: 87, PPV: N/A, NPV: N/A, Glioma: 0.849, Sens: 87, Spec: 87, PPV: N/A, NPV: N/A</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: NS</td>
</tr>
<tr>
<td>Kim et al., 2021 [38]</td>
<td>Multiclass segmentation of infarct and glioma</td>
<td>Unsupervised VAE</td>
<td>Expert image delineation</td>
<td>Infarct: 0.278, Sens: 42.9, Spec: 67.5, PPV: N/A, NPV: N/A, Penumbra: 0.283, Sens: 42.2; 43.6, Spec: 65.1; 69.9, PPV: N/A, NPV: N/A, Glioma: 0.692, Sens: 65; 89, Spec: 67; 75, PPV: N/A, NPV: N/A</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: NS</td>
</tr>
<tr>
<td>Pereira et al., 2019 [44]</td>
<td>Multiclass segmentation of infarct incl. penumbra and glioma</td>
<td>CNN</td>
<td>Expert image delineation</td>
<td>Infarct: 0.340, Sens: 55.0, Spec: 84.6, PPV: N/A, NPV: N/A, Penumbra: 0.540, Sens: 55; 85, Spec: 84, PPV: N/A, NPV: N/A, Glioma: 0.046, Sens: 11; 61, Spec: 89, PPV: N/A, NPV: N/A</td>
<td>(A) Reflecting clinical practice: NS (B) External validation: NS (C) Performance: NS</td>
</tr>
</tbody>
</table>

**Abbreviations:** VAE = variational auto-encoder; CNN = convolutional neural network; DSC = Dice-score coefficient; Sens = sensitivity; Spec = specificity; PPV = positive predictive value; NPV = negative predictive value; S = satisfied; NS = Not Satisfied; CNN = convolutional neural network; VAE = variational autoencoder; PCA = principal component analysis; MS = multiple sclerosis; WMH = white matter hyperintensities.
<table>
<thead>
<tr>
<th>Source</th>
<th>Risk of Bias</th>
<th>Concern for Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patient Selection</td>
<td>Index Test</td>
</tr>
<tr>
<td>Ahmadi et al., 2021 [28]</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Baur et al., 2021 [29]</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Duong et al., 2019 [30]</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Fayaz et al., 2021 [31]</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Felipe Fattori Alves et al., 2020 [32]</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Gauriau et al., 2021 [33]</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Gilanie et al., 2018 [34]</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Han et al., 2020 [35]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hu et al., 2020 [36]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kamnitsas et al., 2017 [37]</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td>Kim et al., 2021 [38]</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Lu et al., 2021 [39]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lu, Lu et Zhang., 2019 [40]</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Nael et al., 2021 [41]</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>Nayak et al., 2020 [42]</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td>Nayak et al., 2020 [43]</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Pereira et al., 2019 [44]</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td>Rauschecker et al., 2020 [23]</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Wood et al., 2022 [45]</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Summary of risk of bias and concern for applicability for all included studies. 
○ = low risk of bias and concern for applicability; ■ = high risk of bias and concern for applicability; □ = unclear risk of bias and concern for applicability.
Significant risks of bias and concern for applicability were seen in the domain of patient selection, index test, and reference. Reasons for these include a lack of consecutive patient populations in eighteen studies, arbitrary classification of equivocal diseases in two studies, large threshold values and the exclusion of smaller lesions in two studies, and automatically generated reference labels in two studies. Only two studies were assessed with low or unclear risk of bias and concern for applicability in all domains.

No meta-analysis was conducted due to inherent heterogeneities in study tasks, population characteristics, and performance metrics.

4. Discussion

In this systematic review, we found that the included algorithms varied considerably in terms of tasks, data requirements, and applicability to workflow improvement. With respect to patient selection, index tests, and reference tests, a significant risk of bias was seen. Most (63%) surveyed algorithms were developed using public datasets derived from ML development challenges. This largely explains the following observed patterns: data size restricted to a few hundred patients, specific disease distribution in multiple datasets, and algorithm inference capabilities based on a limited amount of MR scan sequences. T2 and T2-FLAIR were the most frequently used sequences for disease identification of multiple brain diseases. However, this observation might be confounded by the usage of public datasets. Deep neural networks and derivatives thereof were the most frequently applied ML algorithms, which might be due to their proven high performance and robust feature input methods [47]. All studies published in clinical journals used private datasets with larger patient populations. All studies that satisfied more than one workflow applicability requirement likewise used private datasets. This observed pattern of private dataset usage fits into the general trend, where promising ML algorithms are validated and regulatorily approved for clinical usage based on retrospective, unpublished, and often proprietary data from a single institution [48].

Performance results varied considerably as well. About half of the algorithms exceeded the pre-defined threshold of 85% of their respective performance metrics. Disease segmentation performance was generally lower due to the complexity of this task. These results are corroborated by similar reviews performed by Zhang et al. [49] and van Kempen et al. [50] focusing on ischemic stroke and glioma segmentation, respectively. Similar performance levels in relation to triaging performance were also observed across other imaging modalities. Hickman et al. for instance demonstrated pooled AUC, sensitivity, and specificity of 0.89, 75.4%, and 90.6%, respectively, for screening and triaging mammography using machine learning techniques [51], which are in line with what is observed in this review. Hence, consistent performance results are reported across multiple imaging modalities when using similar methods. [19,52]. However, large performance gaps were seen across clinical settings and study designs, partially owing to the well-documented effect of domain shift [53]. For example, Gauriau et al. [33] tested an algorithm with moderately low sensitiv-
ity and specificity of 77% and 65%, respectively. These results were, however, attained on a large out-of-distribution dataset with a comprehensive representation of almost all diseases seen in everyday clinical practice. On the other hand, the algorithm developed by Lu, Lu et Zhang [40] achieved a binary classification accuracy, sensitivity, and specificity of 100%, but this was achieved on a very small subset of 87 2D MR-slices split out from the same data source as the training data and not reflecting clinical practice. These findings support the approach of considering multiple requirements for study design, study population, testing strategies, and performance when assessing benefits and limitations of applying ML algorithms into existing workflows.

4.1. Potential Benefits of Integrating ML into Existing Scan- and Interpretation Workflows

ML algorithms for clinical workflow integrations have been studied extensively in the past years with multiple authors suggesting different applications [11,12,52,54,55]. Olthof et al. suggest that radiologist workflows could be supported, extended, or replaced by ML functionalities [56].

Based on the findings in this review, scan acquisitions workflows could be supported by multiclass classification and segmentation algorithms. These algorithms, using only a few scan sequences acquired at the beginning of the scan acquisition process, could help classify initial scan images into different disease categories while the patient is still in the scanner and subsequently direct further scan acquisition based on real-time findings. This could prevent the excessive scanning of patients with no significant findings while ensuring fast scan acquisition for stroke patients and appropriate scan acquisition for tumor patients. The fact that 42% of the ML algorithms included in this review could successfully perform multiclass classification and segmentation based on a single MR sequence supports the feasibility of this concept.

Scan interpretation workflows, on the other hand, could be supported by all algorithms in this review. In fact, some of the surveyed binary classification studies aimed explicitly to support interpretation workflows by doing worklist prioritization of critical findings [33,41,45] and, hence, ensure faster reporting times and improved patient outcomes. Multi-class classification and segmentation algorithms could extend this further by offering potential automated diagnosis reporting, biomarker quantification, and even disease progression predictions. None of the surveyed algorithms, however, satisfied all requirements for successful workflow improvements due to key limitations.

4.2. Limitations of Included Studies and Future Directions

Important limitations pertaining to study design, data source, model development, and testing methodologies were uncovered using an analysis of the risk of bias/concern for applicability and applicability assessment for workflow improvements. First, patient selections were not consecutive and largely based on public datasets that consisted of imaging cases with high signal-to-noise ratios and selection biases. This is especially true for the BRATS challenge dataset, which is known to have handpicked and well-processed representations of brain gliomas that are very characteristic and visually recognizable, thus resulting in many algorithms achieving good performance when being developed and tested on it [57]. This could potentially introduce an overestimation of model performance and limit integration into clinical practices that face more heterogeneous images of brain diseases. Secondly, index tests were limited by insufficient reporting of model thresholds or deliberately large thresholds chosen for favorable performance reporting. Nael et al. [41], for instance, demonstrated that their model performance dropped significantly when detecting a smaller infarction volume of <0.25 mL compared to volumes of 1 mL. Because the accurate delineation of size, location and development of ischemic lesions have great prognostic implications [58], this trend of size-dependent accuracy could pose challenges to performing accurate recovery predictions and, hence, overall stroke management. Thirdly, reference tests similarly introduced critical biases, especially in the included studies that used 2D-image datasets with handpicked 2D-images and labels as ground truth. This selection
of representative images could have introduced priors that are easily exploitable by ML algorithms, as has previously been demonstrated in similar datasets [59]. Fourthly, about half of the surveyed ML algorithms had unacceptably low sensitivity and specificity, which could increase scan acquisition workloads and more worryingly decrease patient safety. Finally, only a minor proportion reported on the clinically relevant metrics of positive and negative predictive values. This, combined with the lack of testing on out-of-distribution datasets, might have presented skewed performance impressions not accounting for all relevant conditions in the intended target population [13].

Future studies developing ML algorithms applicable for workflow improvements should ensure the possession of a consecutive patient population reflecting the desired target population, transparent reporting of patient population characteristics and thresholds for index tests, and performance levels reported through metrics that incorporate different aspects of positive and negative findings. Low false-negative rates should be prioritized, thus ensuring adequate patient safety by having the fewest possible missed findings. Disease prevalence must be considered so as to account for positive and negative predictive values. To alleviate some of these limitations, standardized reporting guidelines tailored for AI in radiology [60], prospective study designs with consecutive patient sampling, and multi-site testing with clinical partners must be considered. The challenges of low sensitivity and specificity might be addressed by rethinking existing data acquisition strategies and model architectures. For instance, temporal information from follow-up scans or contrast-enhancement kinetics can be taken into account. Similar strategies are being used on PET-CT scans resulting in improved tumor classification specificity [61].

4.3. Limitations of This Review

This review should be read in view of limitations including publication and reporting bias. We limited our inclusion criteria to studies that could identify multiple brain diseases including brain infarcts, hemorrhages, or tumors, and we further restricted our limitation to studies that have tested their algorithms on data separate from training data. Next, we assessed the applicability of ML algorithms for improving workflows based on a set of requirements not previously validated. All of this might have limited the overview and the impression of this research field. As these criteria were selected based on clinical relevance, the results nonetheless present clinically useful representations of how state-of-the-art ML algorithms could be applied to improve existing scan acquisition and interpretation workflows.

5. Conclusions

The surveyed algorithms could potentially support and extend existing workflows. However, limitations pertaining to study design, study data, reference standards, and performance reporting prevent clinical integration. No study satisfied all requirements for successful workflow integration. Standardized reporting guidelines tailored for ML in radiology, prospective study designs, and multi-site testing could help alleviate this. The findings from this review could aid future researchers and healthcare providers by allowing them to critically assess relevant ML studies for workflow improvements and by enabling them to better design studies that validate the benefits of deploying ML in scan acquisition and interpretation workflows.

Author Contributions: Conceptualization, K.S., C.M.O., A.P., J.F.C., T.C.T. and M.B.N.; methodology, K.S., C.M.O., A.P., J.F.C., T.C.T. and M.B.N.; investigation, K.S., C.M.O., J.M. and J.J.; data acquisition, K.S.; writing—original draft preparation, K.S.; writing—review and editing, K.S., C.M.O., A.P., J.J., J.M., J.F.C., T.C.T. and M.B.N.; supervision, A.P., J.F.C., T.C.T. and M.B.N.; project administration, K.S.; funding acquisition, M.B.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Search strings for databases

**MEDLINE (PubMed)**

  AND
  AND
  AND

**EMBASE**

- exp magnetic resonance imaging/
- exp brain disease/
- exp machine learning/
- exp classification/or detection.mp.
- (abnormal or patholog$ or multi-class or critical finding$ or triag$ or automat$).mp.
- [mp = title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword heading word, floating subheading word, candidate term word]
- 1 and 2 and 3 and 4 and 5

**Scopus**

- (TITLE-ABS-KEY (“magnetic resonance imaging” OR “Multiparametric Magnetic Resonance Imaging” OR “MRI”)) AND (TITLE-ABS-KEY (“Artificial Intelligence” OR “Machine Learning” OR “Deep Learning” OR “Neural Network” OR “Convolutional neural network”))
- (TITLE-ABS-KEY (brain AND disease OR brain AND infarct OR brain AND hemorrhage OR brain AND neoplasm* OR brain AND tumor OR brain AND anom* OR brain AND abnormal* OR brain AND patholog* OR brain “multi-class” OR brain “critical finding*” OR brain AND triag* OR brain AND automat*))
- (TITLE-ABS-KEY (classification OR detection))
- (LIMIT-TO (PUBYEAR, 2022) OR LIMIT-TO (PUBYEAR, 2021) OR LIMIT-TO (PUBYEAR, 2020) OR LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017)) AND (LIMIT-TO (DOCTYPE, “ar”) OR LIMIT-TO (DOCTYPE, “re”))

**Web of Science**

- ALL = (“magnetic resonance imaging” OR “Multiparametric Magnetic Resonance Imaging” OR “MRI”) AND (“Artificial Intelligence” OR “Machine Learning” OR “Deep
Deep Learning in Neuroradiology: A Systematic Review of Current Algorithms and Approaches

**References**


Diagnostics 2022, 12, 1878


