Predictive Modelling of Brain Disorders with Magnetic Resonance Imaging: A Systematic Review of Modelling Practices, Transparency, and Interpretability in the use of Convolutional Neural Networks

Shane O'Connell *1, Dara M Cannon †2, and Pilib Ó Broin†1

[1]School of Mathematical and Statistical Sciences, College of Science and Engineering, University of Galway, Galway, Ireland
[2]Clinical Neuroimaging Laboratory, Galway Neuroscience Centre, College of Medicine Nursing and Health Sciences, University of Galway, Galway, Ireland

Abstract

Brain disorders comprise several psychiatric and neurological disorders which can be characterised by impaired cognition, mood alteration, psychosis, depressive episodes, and neurodegeneration. Clinical diagnoses primarily rely on a combination of life history information and questionnaires, with a distinct lack of discriminative biomarkers in use for psychiatric disorders. Given that symptoms across brain conditions are associated with functional alterations of cognitive and emotional processes, which can correlate with anatomical variation, structural magnetic resonance imaging (MRI) data of the brain are an important focus of research studies, particularly for predictive modelling. With the advent of large MRI data consortia (such as the Alzheimer's Disease Neuroimaging Initiative) facilitating a greater number of MRI-based classification studies, convolutional neural networks (CNNs) – deep learning models suited to image processing – have become increasingly popular for research into brain conditions. This has resulted in a myriad of studies reporting impressive predictive performances, demonstrating the potential clinical value of deep learning systems. However, modelling practices, transparency, and interpretability vary widely across studies, making them difficult to compare and/or reproduce, thus potentially limiting clinical applications. Here, we conduct a qualitative systematic literature review of 60 studies carrying out CNN-based predictive modelling of brain disorders using MRI data and evaluate them based on three principles – modelling practices, transparency, and interpretability. We furthermore propose several recommendations aimed at maximising the potential for the integration of CNNs into clinical frameworks.

1 Introduction

Brain disorders, which include bipolar disorder, alzheimer's disease, and schizophrenia, are a collection of debilitating neurological and psychiatric conditions characterised by a variety of features including impaired cognition, altered mood states, psychosis, neurodegeneration, and memory loss [1]. These phenotypes, each with varied clinical presentations, are all associated with neuroanatomical changes, incurring public and personal health burdens through reduced quality of life, social stigma, and increased mortality [1, 2]. As such, these conditions are the focus of intense research across multiple disciplines. There is significant interest in building predictive models designed to differentiate conditions and their subtypes, which could incorporate biological information into current clinical frameworks and yield mechanistic insights via the biomarkers used [3, 4]. Additionally, biomarker-informed diagnoses could offer the potential of early intervention and management, a concept well understood in general medicine [5]. Magnetic resonance imaging (MRI) provides non-invasive measures of brain structure and the increasing availability of large-scale collections of MRI data has enabled a wealth of predictive modelling studies [6, 7].

Previously, machine learning and classical statistical approaches have been used to highlight differential neuroanatomical patterns across several conditions, including subcortical structure volume reduction in bipolar disorder and alzheimer's disease [8, 9]. However, incorporating such information into clinical systems is non-trivial, as the dynamics and limitations of a particular biomarker must be addressed prior to use [10, 11]. Additionally, the methods used to identify discriminative features have their own considerations, such as requiring preprocessing tools to derive tabular brain summary information [12, 13]. These tools can produce variable results depending on the parameters chosen, even when applied to the same dataset, highlighting the importance of domain expertise to justify decisions [14]. Additionally, statistical modelling often requires formal specification of expected variable relationships, and generally are unsuited to high-dimensional imaging data structures. Traditional machine learning approaches are also limited by their inability to consider spatial dependencies between groups of pixels, making it necessary to use tabular summary data. With these factors in mind, deep learning algorithms – and particularly those well-suited to imaging – have become a popular methodology. This is because of their ability to

*Corresponding author, email: s.oconnell29@nuigalway.ie
†Provided equal supervision

consider arbitrarily complex relationships, meaning greater model flexibility without specification of expected variable relationships. Convolutional neural networks (CNNs) are deep learning models designed to detect spatial patterns in imageing data and have shown impressive predictive performances in various classification tasks. They have also been widely applied in the field of medical imaging for segmentation and prediction, particularly in the context of aging and psychiatric/neurological disorder diagnosis [15, 16, 17, 18, 19, 20, 21, 22].

These recent developments have been enabled by access to large standardised neuroimaging data collections, such as the alzheimer's Disease Neuroimaging Initiative [23] and the UK Biobank [24]. The predictive capabilities of these approaches is promising in the context of potential clinical applications; brain disorder classifications are usually based on life history information and questionnaires, and thus leveraging models making use of neuroanatomical measures could supplement existing diagnostic frameworks. However, there are a few caveats that bear consideration; firstly, deep learning models have a number of limitations, such as high parameter dimensionality, lack of interpretability, weight stochasticity, lack of uncertainty, and difficulty to train [25, 26, 22, 27]. Secondly, clinical decision systems require rigorous validation and reporting frameworks for more interpretable models; the use of opaque deep learning algorithms make validation and transparency more difficult to achieve [28, 29]. Clinical decision systems that offer no explanation of a classification are less likely to be incorporated into patient care frameworks. These factors combine to make the application of deep learning to clinical settings challenging, particularly where medical imaging is concerned.

As the number of studies applying deep learning to brain disorder prediction using neuroimaging data increases, the opportunity arises to examine factors that may limit their potential for clinical application. In this work, we systematically review 60 papers which report on such approaches. While many of the studies examined have been designed to demonstrate the predictive capabilities, we sought to assess the existing literature with the aim of identifying key principles that can maximise the potential clinical value of future work; these principles are: 1) modelling practices, 2) transparency, and 3) interpretability.
Below, we first provide a brief overview of CNNs and their workflow in the context of brain disorder imaging-based models, and subsequently detail our motivation behind these three principles; we then analyse the selected papers in the context of these principles and suggest several recommendations for future studies based on our results.

1.1 Convolutional Neural Networks

CNNs are a popular deep learning algorithm for many areas of research, particularly those utilising MRI data [15, 16, 17, 18, 19, 20, 21]. Their structure is designed to account for spatial data patterns; this is accomplished through the use of filters and feature maps. A feature map is derived via *convolutional operations*, which are a matrix multiplication between a weights vector of an arbitrary size (the filter, for example, may be 2×2 pixels large in a 2D example) and an input image patch of the same size. The result, which is every number in the input multiplied by every number in the filter and summed together, is then the pixel of a new feature map. The convolution of the same filter over every patch of the input image generates the entire output feature map, which is usually the same size as the input image. Multiple feature maps are used in CNN architectures, each with their own filters, which, throughout model training, can detect distinct data patterns such as shapes and/or edges. The goal of CNNs is to build increasingly abstract representations of data through iterative downsampling and transformations until such a time as linear seperation of classes is possible in predictive tasks. Weight initilisation is often random and training is carried out via backpropagation. More in-depth considerations of neural networks and their training can be found in LeCun et al., 1995 and 2012 [22, 27].

1.2 CNN Implementations

MRI-based predictive modelling of brain conditions with deep learning models generally follows the pipeline presented in Figure 1, or a variant thereof. Preprocessing is usually applied to skull strip, register raw input images, crop, resize, and/or contrast normalise. The preprocessed inputs are then used as training data for a CNN (or an ensemble of CNNs). Owing to the fact that many existing CNN models have been applied to 2D data domains, studies in the medical imaging field can adapt their data to fit existing architectures via transfer learning or train new models in the 3D space, as structural MRI scans are usually 3D [30, 31]. Some studies also train custom architectures on 2D data [32, 33, 34]. The output is usually presented as a probability, which is then used to calculate performance metrics such as the area under the receiver operating characteristic curve (AUC) and accuracy.

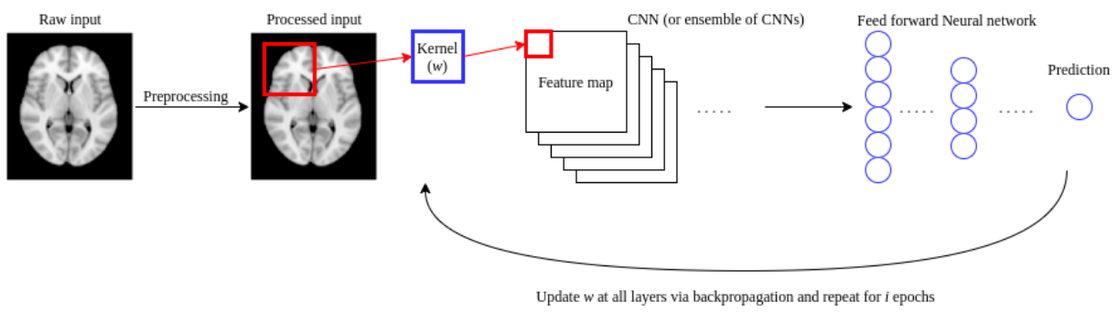


Figure 1: General experimental workflow. The preprocessed input image, either in 2 or 3 dimensional format is passed to a CNN model (or ensemble of CNN models) for training and prediction, The weights vector w is updated via backpropagation at each epoch, minimising the error of the loss function.

In the following sections, we define and justify our emphasis on modelling practices, transparency, and interpretability in the context of brain disorder classification using neuroimaging data for potential clinical benefit. We note that these principles bear domain-agnostic importance for predictive studies and overlap with recent recommendations for improving the clinical and biological translational potential of machine learning and deep-learning models [35].

1.3 Modelling Practices

Modelling practices here refers to the reliability of the methodology used; methodologies optimising reproducibility efforts via approaches demonstrating the reliability of experimental results are increasing the potential clinical value of a study. Generally, studies that can be reproduced and that have attempted to mitigate factors that can influence the reliability of results are more likely to witness clinical integration. As previously mentioned, deep learning models have a number of unique features that make this task difficult, but several procedures can be observed. We examine the presence of repeat experiments, the data splitting procedure, potential information leakage, and the data representation strategy to evaluate this principle. Repeat experiments ensure that the reported performance metrics are trustworthy across multiple random weight initialisations and that the system as a whole can be expected to perform well if retrained. This is pertinent given that CNNs are paramater-dense, making them more prone to overfitting. A useful type of repeat experiment includes k-fold cross validation, whereby data is split into k folds, and $k-1$ folds are used to train the model and the k-th fold serves as the testing set. This procedure is repeated k times, until every fold has served as the testing set, providing an estimate of model performance on multiple data splits.

The data splitting procedure is important as training/testing separation must be ensured for robust performance estimation. Information leakage describes situations whereby model testing and training sets are not kept entirely independent during model tuning, which may lead to inflated performance metrics; this can occur where model performance, usually estimated from entirely independent data, is partly estimated on examples already used to optimise weights. This can limit reproducibility and ultimately the potential clinical value of an approach.

The data representation strategy is of specific importance for CNN models in this domain, as structural MRI data is 3D, whereby each number is represented by a pixel. Thus, modelling entire volumes can be computationally expensive, and some studies may opt to split data into individual 2D slices. This comes with a set of caveats: firstly, each 2D slice is treated as an individual instance during conventional training procedures, meaning that performance metrics can either be reported per slice or combined to derive patient-level quantities, prompting consideration of voting strategies; secondly, 2D data are more prone to information leakage if train-test splitting is carried out after 2D slice derivation. Together, these issues can contribute to inflated estimates of performance.

1.4 Transparency

Transparency refers to how clearly the study's methods are reported, including code and model sharing. This principle bears general importance, particularly for models with clinical potential [36]. Several important advantages to code sharing having been described previously, including facilitating greater understanding of experiments and facilitating reproducibility [37, 38]. There are many hyperparameters associated with deep learning models which can effect performance, making transparent reportage necessary. Descriptions of model architectural choices and training schedules can help to increase potential for clinical translation through increased reproducibility and understanding of studies. Furthermore, model weight sharing can mitigate the computational overhead of model training.

1.5 Interpretability

Interpretability refers to the efforts made to explain features driving model predictions. Deep learning systems are not well suited to interpretation, but efforts can be made to examine image regions that are used during prediction to determine whether that information is relevant. This is particularly important as CNNs are prone to overfitting and can make use of any image feature, in turn making algorithmic biases more likely if not examined [39, 40]. Ensuring CNNs are using relevant information can increase clinical potential and confidence in the system. Models can be interpreted by saliency methods such as gradient-based class activation mapping [41, 17], which rely on deriving the gradient of model output with respect to input and weighting that quantity by the input – the final metric is then overlaid on the input for visualisation. This can indicate what regions are most 'important', but they do not offer the same explanatory power as coefficients from regression models. Another approach to understanding model behaviour is counterfactuals, which involve measuring the changes in predictive performance of models when they are exposed to inputs with known qualities; for example, noting the change in model output when a patient image with a thicker amygdala is used as the input [42].

2 Methods

We conducted a systematic literature review according to PRISMA guidelines [43], the details of which are provided below.

2.1 Inclusion/exclusion criteria

We limited our search to consider studies making use of CNNs exclusively (end-to-end), and convolutional layer outputs, or other model outputs, are not used to train separate machine learning models. This is because they are the most common architecture. We also focused our attention on studies that use structural MRI data, as functional MRI data structures can often have different modelling requirements, including the use of time series methodologies that make them more difficult to compare.

2.2 Search details

We performed a Web of Science (all databases) and Pubmed search with the following keywords:

((((structural) or (T1-weighted)) AND (imaging)) AND ((MRI) OR (T1 MRI)) AND ((CNN) OR (convolutional neural network) OR (3D-CNN))) AND (psychiatric OR depression OR autism OR bipolar OR Alzheimer's OR neurological) NOT (segmentation)

For Web of Science, 77 results were returned, and 114 results were returned from Pubmed. Titles and abstracts were screened for relevance to the research question, and duplicates across both databases were removed, leaving a total of 76 papers. 16 studies were excluded for using functional MRI data and using hybrid models where CNNs were not the primary modelling method; this resulted in a total of 60 papers remaining for review. The flowchart of this process is presented in Figure 2.

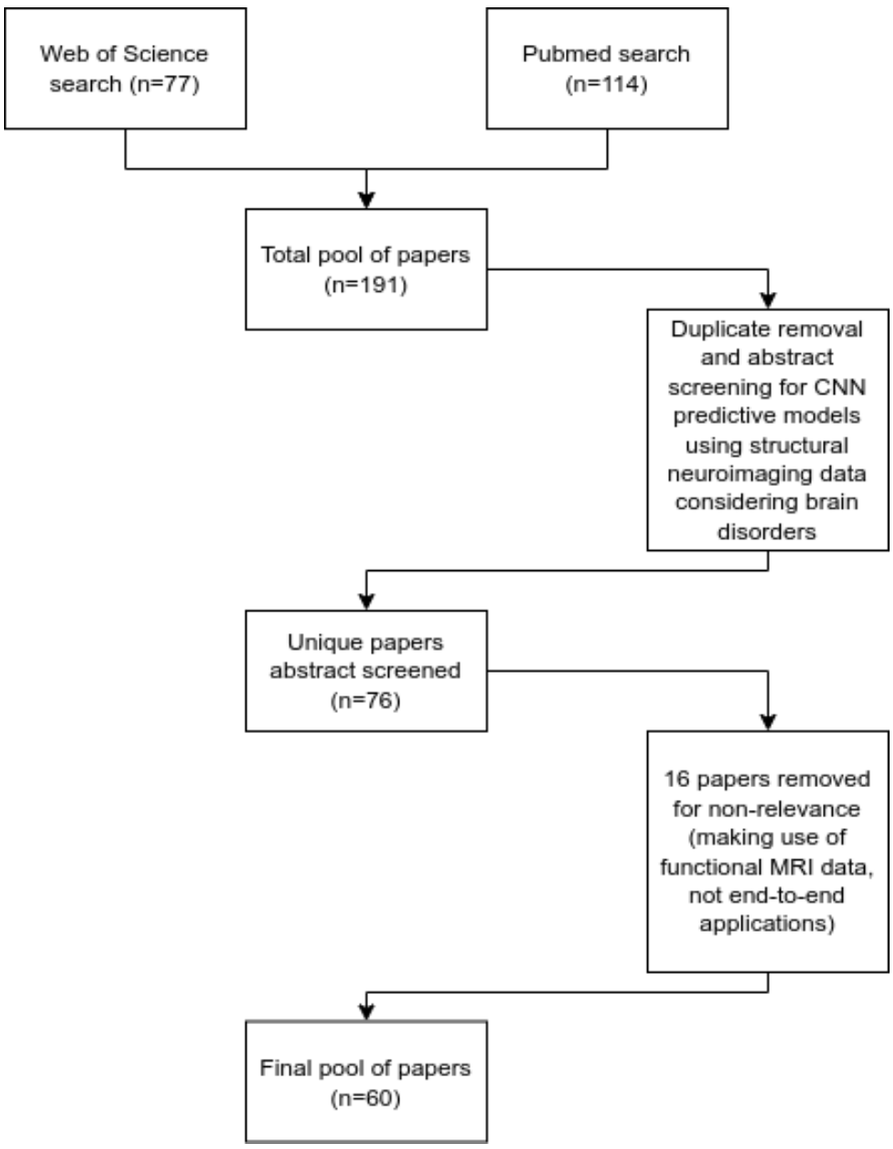


Figure 2: Flowchart detailing the paper selection process.

2.3 Desired variables

A standardised questionnaire was designed to evaluate the methodological details of the studies considered. No numerical variables were sought as this work aimed to qualitatively examine the outlined

principles.

3 Results

We organise our findings according to our three principles: modelling practices, transparency, and interpretability. The selected papers and their attributes can be found in Table 1, and a numerical summary of the results can be found in Table 2.

3.1 Modelling practices

We found that 24 out of 60 papers represented data in 2D (Table 2). While this is more computationally efficient than 3D, it can make information leakage more likely. Accuracy calculation can be carried out per slice or per patient, introducing issues surrounding optimal voting strategies. Of the 24 studies making use of 2D slices, only one referred to voting methods, and 15 studies suffered from potential leakage [44]. Leakage was flagged where it was clear that training/test splitting occurred per slice and not per patient or where it could not be ruled out based on methodological descriptions. Several studies made use of single slices per patients [45, 46, 33]. Even with registration, there is no guarantee that the same biological information is considered per patient with this approach. Spatial dependencies between regions are ignored when using 2D data, which may have implications for mechanistic understanding and predictive capabilities. One paper making use of 2D slices provided code [47]. We noted that 25 out of 60 studies made use of multiple models for training and prediction, with some papers using the output of one trained CNN as the input to another [48, 49, 50, 51, 52, 53]. This may facilitate overfitting by impacting generalisation. A number of studies used statistical tests to pre-select informative image patches which can introduce bias by focusing the model on regions which may not be informative in full models. [53, 52, 45]. Furthermore, pre-selecting regions based on accuracy metrics in one population may influence generalisability in another, which is pertinent for clinical potential. In several studies, one model was trained on the whole dataset the weights from that model were used for transfer learning of another model, leading to potential leakage or overfitting [54, 55, 44, 45, 34].
Thirty four out of 60 studies employed repeat experiments. Ten such papers reported point estimates, and 5 provided code [56, 47, 57, 54, 58]. Of the 15 studies with repeat experiments and interpretability efforts, none detailed whether saliency methods were applied per fold or on test sets.

3.2 Transparency

We found that 54 out of 60 papers did not provide code or model weights, meaning that the majority of studies relied on textual methods summaries. This implies limited methodological transparency; this is an issue when considering how modelling choices can impact system performance. Studies reporting code facilitate clear, reproducible experimental practices [54, 56, 57, 58, 59, 47].

Table 1: Tabular presentation of the studies considered for this systematic literature review.

Authors and citation	Modelling Practices			Transparency	Interpretability	Publication Status
	Data representation	Repeat experiments	Data leakage	Code availability	Saliency	
Zou et al. (2017) [18]	3D	Yes	No	No	No	Conference
Çitak-ER et al. (2017) [60]	2D	No	Yes	No	No	Journal
Taheri Gorji and Kaabouch (2019) [61]	3D	No	No	No	No	Conference
Spasov et al. (2018) [62]	2D	No	Yes	No	No	Conference
Wang et al. (2018) [63]	3D	Yes	Yes	No	Yes	Journal
Li and Liu (2019) [64]	3D	Yes	Yes	No	Yes	Journal
Liu et al. (2020) [65]	3D	Yes	Yes	No	No	Journal
Hosseini-Asl et al. (2016) [66]	3D	Yes	Yes	No	No	Conference
Li et al. (2017) [67]	3D	Yes	Yes	No	No	Conference
Folego et al. (2020) [54]	3D	Yes	Yes	Yes	Yes	Journal
Marzban et al. (2020) [68]	3D	No	Yes	No	No	Journal
Hosseini-Asl et al. (2018) [20]	3D	Yes	No	No	No	Journal
Gunawardena et al. (2017) [69]	2D	No	Yes	No	No	Conference
Basaia et al. (2019) [70]	3D	No	Yes	No	No	Journal
Tufail et al. (2020) [71]	2D	Yes	Yes	No	No	Journal
Hu et al. (2020) [72]	3D	Yes	No	No	No	Journal
Cheng et al. (2017) [73]	3D	No	Yes	No	No	Conference
Nanni et al. (2020) [74]	3D	No	No	No	No	Journal
Lin et al. (2018) [55]	2D	Yes	Yes	No	No	Journal
Billones et al. (2016) [31]	2D	No	No	No	No	Conference
Barbaroux et al. (2020) [32]	2D	Yes	No	No	No	Conference
Yigit and Işik (2020) [75]	2D	No	Yes	No	No	Journal
Pan et al. (2020) [76]	2D	Yes	Yes	No	No	Journal
Ahmed et al. (2020) [44]	2D	No	Yes	No	No	Journal
Ortiz-Suárez et al. (2017) [77]	2D	Yes	Yes	No	Yes	Conference
Aderghal et al. (2017) [33]	2D	No	No	No	No	Conference
Lian et al. (2020) [78]	3D	No	No	No	Yes	Journal
Li et al. (2021) [79]	3D	Yes	No	No	Yes	Journal
Cui and Liu (2018) [80]	3D	Yes	No	No	No	Conference
Aderghal et al. (2020) [81]	2D	No	No	No	No	Journal
Böhle et al. (2019) [56]	3D	Yes	No	Yes	Yes	Journal
Sarraf et al. (2019) [47]	2D	Yes	Yes	Yes	Yes	Journal
Liu et al. (2018) [82]	3D	Yes	Yes	No	Yes	Journal
Zhang et al. (2020) [19]	3D	Yes	No	No	Yes	Journal
Lee et al. (2019) [83]	2D	Yes	Yes	No	No	Journal
Qiu et al. (2020) [57]	3D	Yes	Yes	Yes	Yes	Journal
Spasov et al. (2019) [58]	3D	Yes	No	Yes	No	Journal
Sun et al. (2020) [84]	3D	Yes	No	No	No	Journal
Oh et al. (2020) [85]	3D	Yes	No	No	No	Journal
Lian et al. (2020) [48]	3D	No	No	No	Yes	Journal
Cui and Liu (2019) [86]	3D	Yes	No	No	Yes	Journal
Raju et al. (2020) [49]	3D	Yes	No	No	Yes	Journal
Mendoza-Léon et al. (2020) [45]	2D	No	Yes	No	No	Journal

Table 1: Tabular presentation of the studies considered for this systematic literature review.

Authors and citation	**Modelling Practices**			**Transparency**	**Interpretability**	**Publication Status**
	Data representation	**Repeat experiments**	**Data leakage**	**Code availability**	**Saliency**	
Pelka et al. (2020) [34]	2D	Yes	No	No	Yes	Journal
Li and Liu (2018) [50]	3D	Yes	No	No	Yes	Journal
Bae et al. (2021) [87]	3D	No	No	No	Yes	Journal
Cui and Liu (2019) [51]	3D	Yes	No	No	No	Journal
Liu et al. (2018) [52]	3D	No	No	No	No	Journal
Liu et al. (2018) [53]	3D	Yes	No	No	No	Journal
Al-Khuzaie et al. (2021) [88]	2D	No	Yes	No	No	Journal
Zhang et al. (2021) [89]	3D	No	No	No	Yes	Journal
Hu et al. (2021) [59]	3D	No	No	Yes	Yes	Journal
Herzog and Magoulas (2021) [46]	2D	No	Yes	No	No	Journal
Yee et al. (2021) [90]	3D	Yes	Yes	No	No	Journal
Mukhtar and Farhan (2020) [91]	2D	No	Yes	No	No	Journal
Bae et al. (2020) [92]	2D	Yes	No	No	No	Journal
Nanni et al. (2020) [93]	2D + 3D	No	No	No	No	Journal
Nigri et al. (2020) [94]	2D	No	No	No	Yes	Conference
Li et al. (2021) [95]	2D	Yes	Yes	No	No	Journal
Kiryu et al. (2019) [96]	2D	No	No	No	No	Journal

Question	Answer
How are data represented?	2D(n=24), 3D(n=35), Both(n=1)
Is code available?	No(n=54), Yes(n=6)
Conference or Journal?	Conference(n=13), Journal(n=47)
Is saliency considered?	No(n=40), Yes(n=20)
Are there repeat experiments?	No(n=26), Yes(n=34)
Is there potential information leakage?	No(n=32), Yes(n=28)

Table 2: Numeric summary of study attributes from the 60 papers satisfying selection criteria.

3.3 Interpretability

Twenty of 60 studies considered interpretability by applying a saliency method [41] or visualising feature maps [17]. Of these 20, 5 papers discussed their interpretation of saliency outputs in their findings, with the remainder providing little to no commentary [56, 82, 57, 49, 94]. Four such papers with discussions made use of single saliency methods, with two out of five providing code [56, 57]. Nine studies presented interpretability results with little to no commentary and no code [64, 65, 78, 79, 47, 82, 86, 50, 87]. Of the remaining 11 studies, 5 provided code, with a combination of saliency and transparency [56, 47, 57, 54, 59]. A subset of studies underscored that expert-driven preprocessing is not required with deep learning studies, and almost all studies alluded to this fact in their introductions [59, 68, 78, 50, 51].

4 Discussion

We conducted a systematic literature review of 60 studies carrying out CNN-based predictive modelling of brain disorders using MRI data and evaluated their modelling practices, transparency, and considerations of interpretability in the context of their potential clinical value. Our results identified several areas for potential improvement across three principles that we believe will maximise the potential for clinical integration. Below, we discuss the findings summarised in Table 2 in and propose several recommendations to maximise the potential clinical value of future studies.

4.1 Data representation

A majority of papers made use of 3D data representations, which is sensitive to aforementioned CNN-specific limitations of 2D data. This ensures that all biological information is used during training, as opposed to individual slices where spatial inter-dependencies are ignored. There was a significant minority of papers making use of 2D data structures (24/60), which may pose issues for downstream clinical applications. Two-dimensional models have multiple caveats discussed previously, including multiple majority voting strategies and potential information leakage. The high computational cost of modelling on 3D data structures impedes their implementation. Thus, leveraging 2D model weights for transfer learning is attractive, and we recommend cognizance of the limitations associated with 2D modelling and attempt to mitigate these issues. For example, researchers can examine how performance metrics change relative to different voting strategies. Regarding information leakage, researchers can ensure slice conversion post-data splitting at the patient level; providing code can ensure supplement this via transparency. Utilising single slices may lead to performance estimation inflation owing to the fact that there is no guarantee the same biological information is being considered per patient. Several studies also made use of model stacking, whereby the input of a model is the output of another trained model. This may impact the model's ability to generalise to different data by increasing the chances of overfitting. This is because the first model in stacking situations has already derived a representation of the data, meaning the next model's internal representation of the problem domain is built upon a an initial abstraction of the input data. This is distinct from using traditional dimensionality reduction techniques before training a predictive model in two ways – firstly, model stacking in this domain is often not unsupervised step and hence the first model's representation is based upon knowledge of test labels. Secondly, deep learning systems are opaque, making it difficult to understand both the predictive model's data representation and the specifics of the reduced dimensionality space used as the input for the predictive model.

4.2 Repeat experiments

Most studies implemented repeat experiments, which demonstrate robust modelling practices. Such procedures account for weight initialisation stochasticity and potential performance inflation arising from fold splitting. Averaging over multiple random weight start points can return accurate performance estimates, with cross validation schemes being a reliable model diagnostic. Twenty-six of the 60 considered papers did not employ repeat experiments, which reduces confidence in reported results. A number of repeat experiment studies reported point estimates, which does not convey the spread of performance variability. Code inaccessibility exacerbates this issue, leaving the reader unclear as to the procedure followed. Reporting an average as a point estimate versus the range may cause the reader to assume the range was not large when performance may have been variable across repeats. We recommend that researchers employ repeat experiments and report their results with means and standard deviations.

4.3 Code availability

Most studies did not provide code. Wen et al. (2020) [97] underlined the importance of fairness, accountability, and transparency in deep learning modelling studies, and code inaccessibility runs contrary to these principles. The construction of deep learning systems requires many algorithmic decisions which can influence performance, introduce bias, and impact reproducibility. Deep learning models optimise an objective function over a set of arguments, meaning that any decisions taken in preprocessing and model construction can affect the capabilities of the system as a whole, and propagate subjective choices throughout ostensibly objective models [40]. For instance, several studies have examined algorithmic biases against underrepresented and/or marginalised groups [98, 99, 100]. Aside from domain-specific benefits to code sharing, the larger scientific community has recently shifted towards open science frameworks, with several high-profile journals requiring methodological transparency [101, 102, 103, 37]. Therefore, we believe that code availability and transparent methodological descriptions are an important aspect of deep learning experiments in this domain independent of potential clinical applications. Within a patient-care context, we underscore the importance of constructing reproducible systems to increase trust, both from a clinician and patient perspective. Studies making code available are proactively embracing these essential principles. We further encourage that minimal Jupyter/Google Colab notebooks, and other literate programming tools, be explored to enhance understanding and reproducibility [104, 105, 37]. This would also have the useful properties of allowing researchers to examine pipelines and identify potential 'blind spots' that the model authors may have overlooked in their modelling decisions, encouraging accountability [97]. Additionally, model training is incredibly computationally intensive; having access to models trained in similar domains could enable transfer learning approaches and mitigate data representation issues. Therefore, we recommend that authors share model weights and code to increase the potential of clinical translation.

4.4 Saliency and interpretability

We found many studies did not interrogate their presented models to ensure that relevant information is being used. Where irrelevant information is included, such as skull thickness when examining alzheimer's disease neurodegeneration, without confirmation that the model is utilising brain information, attempts at patient care integration will have limited success. Even in cases where known irrelevant information can be removed by preprocessing, visual maps can draw attention to previously-unknown irrelevant information. Opaque black box models are less likely to be implemented in patient care settings, making the use of interpretability efforts of crucial importance. As previously stated, algorithmic biases in predictive settings is concerning, and saliency methods can help researchers to identify sources of bias where they occur. Twenty studies investigated neuroanatomical features driving model predictions via saliency methods, thus demonstrating relevant information is used during model training. While there exists variability in application, especially in terms of region understanding, we nonetheless recommend that future studies apply saliency methods to highlight brain regions being used. Saliency experiments leveraging multiple methodologies and those offering transparency through code availability can further increase potential clinical utility.

However, interpretability methods have a number of limitations that may hamper their application, prompting a discussion around how best to understand opaque models. Most existing methods return an 'importance' per pixel, which has no direct link to human-interpretable neuroanatomy. Usually, it represents the degree of change in the output relative to a small perturbation in the input pixel value, collapsing a potentially non-linear relationship to single values. While it provides an empirical assessment of captured patterns, it offers little interpretative value compared to coefficients returned by classical statistics. The deep learning field in general is focused on prediction as opposed to inference, meaning that the mechanistic understanding of relationship dynamics is often secondary to test accuracy. This is challenging in the context of discovery and clinical settings. Furthermore, saliency methods have their own limitations arising from their algorithmic derivation of 'importance', which can effect interpretation [106]. Similarly, counterfactuals, while promising, are difficult to empiricise and require significant computational overhead. Nonetheless, interpretability efforts allow researchers to visually evaluate model attention, which, for clinical translation, can serve to increase confidence and reduce bias, a topic of concern with respect to models applied to society at large [99, 40].

These methods may also be used to generate new hypotheses for downstream experiments. Highlighting neuroanatomical regions discriminative for particular conditions may suggest they have mechanistic relevance. Biomarker categorisation is an important step towards expanding current clinical care practices.

4.5 Future perspectives and commentary

This systematic literature review highlights areas of focus across modelling practices, transparency, and interpretability in the context of maximising the potential for clinical utility. These points underscore long-standing differences between deep learning and classical statistics, whereby the former is usually concerned with predictive performance and the latter with making inferential statements. The predictive imperative has led to numerous advances in image processing, with several state-of-the-art approaches developed to address tasks not suited to classical statistics [17, 22]. Neural networks have clear advantages where inferential dynamics are not a concern.
However, as deep learning becomes more readily applied to medical imaging domains, with potential consequences for patients, dichotomies of prediction versus inference should be retired. Researchers can maximise potential clinical benefit and potentially increase the quality of patient care by embracing the principles of reproducibility, transparency, and interpretability for predictive models. This can increase the confidence in such methods and consequently increase the potential for future clinical integration.
We summarise our key recommendations in Table 3.

Key Recommendations	**Benefits**	**Risk(s) mitigated**
Make well-annotated code freely available	- Improve chances of reproducibility – Readers can better understand workflow – Encourage accountability and transparency	- Limit reproducibility efforts – Models remain opaque
Employ repeat experiments	- Improve confidence in model estimation – Mitigate random weight initialisation	- Risk reporting overfitted results – Performance estimation inflation – Diminished confidence in system overall
Use saliency metrics and counterfactuals	- Validate that model is using relevant information – Potential biomarker discovery – Improve confidence in system overall	- Models remain opaque – Diminished confidence in system overall – Unsure what information is being used by models
Avoid 2D data structures where possible	- Ensure spatial information between slices is considered – Lessen chance of information leakage – No requirement of multiple voting strategies	- Information leakage is more likely to occur – No guarantee spatial information between slices is considered – Must consider multiple voting strategies

Table 3: Key recommendations arising from the results of this systematic literature review, their benefits, and the risks associated with non-adherence.

5 Limitations

This work reviewed studies from 2 database sources, but is not guaranteed to have evaluated all available relevant research. This study also did not undertake a quantitative review of reported accuracy metrics. This work also did not include considerations of studies making use of functional neuroimaging data, which contains a large corpus of research. We note that this review highlights issues with modelling 2D data structures; we acknowledge that this may not be feasible with respect to limited computational power. The extent of information leakage across these studies may be lower than reported; we classified a study as having potential information leakage where it was not possible to rule out its occurrence in studies that may have been prone (where data is not represented with one quantity per patient). Additionally, we acknowledge the drawbacks of saliency methods, whereby they primarily offer a visual check of model focus; nonetheless. Finally, we have endeavoured to ensure that our evaluation of studies is neither reflective of overall study quality nor reductive with respect to three nuanced principles; we

acknowledge that the nature of the questionnaire results may present our findings as such – our binary descriptors are intended to serve as a vehicle to explore nuanced concepts.

6 Conclusion

In summation, we conducted a systematic literature review of 60 studies carrying out CNN-based predictive modelling of brain disorders using structural brain imaging data and evaluated them in the context of their modelling practices, transparency, and interpretability. We set forth recommendations that we believe will increase the future potential clinical value of deep learning systems in this domain. Careful consideration of these concepts can help to inform a clinical framework that can effectively incorporate deep learning into diagnostic and prognostic systems, enhancing our ability to improve patient care.

7 Declaration of Competing Interest

All authors report no competing interests.

8 Acknowledgements

This work was conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6214.

9 Data availability

All studies in this systematic literature review are accessible via PubMed and Web of Science.

References

[1] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. Autor, Washington, DC, 5th ed. edition, 2013.

[2] Spencer L. James, Degu Abate, Kalkidan Hassen Abate, Solomon M. Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, Ibrahim Abdollahpour, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 392(10159):1789–1858, November 2018. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(18)32279-7. URL `https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)32279-7/abstract`. Publisher: Elsevier.

[3] David J Kupfer, Michael B First, and Darrel A Regier. *A research agenda for DSM V*. American Psychiatric Pub, 2008.

[4] Katherine H Taber, Robin A Hurley, and Stuart C Yudofsky. Diagnosis and treatment of neuropsychiatric disorders. *Annual review of medicine*, 61:121–133, 2010.

[5] Jai Shah and Jan Scott. Concepts and misconceptions regarding clinical staging models. *Journal of psychiatry & neuroscience: JPN*, 41(6):E83, 2016.

[6] Vijay PB Grover, Joshua M Tognarelli, Mary ME Crossey, I Jane Cox, Simon D Taylor-Robinson, and Mark JW McPhail. Magnetic resonance imaging: principles and techniques: lessons for clinicians. *Journal of clinical and experimental hepatology*, 5(3):246–255, 2015.

[7] Michael P Milham, R Cameron Craddock, and Arno Klein. Clinically useful brain imaging for neuropsychiatry: How can we get there? *Depression and anxiety*, 34(7):578–587, 2017.

[8] DP Hibar, Lars T Westlye, Theo GM van Erp, J Rasmussen, Cassandra D Leonardo, J Faskowitz, Unn K Haukvik, Cecilie Bhandari Hartberg, Nhat Trung Doan, Ingrid Agartz, et al. Subcortical volumetric abnormalities in bipolar disorder. *Molecular psychiatry*, 21(12):1710–1716, 2016.

[9] Jee Hoon Roh, Anqi Qiu, Sang Won Seo, Hock Wei Soon, Jong Hun Kim, Geon Ha Kim, Min-Jeong Kim, Jong-Min Lee, and Duk L Na. Volume reduction in subcortical regions according to severity of alzheimer's disease. *Journal of neurology*, 258(6):1013–1020, 2011.

[10] Bernard J Carroll. Biomarkers in dsm-5: lost in translation. *Australian & New Zealand Journal of Psychiatry*, 47(7):676–678, 2013.

[11] Gisele Silvaa Karen Furiea and S Gisele. Biomarkers in neurology. *Frontiers of neurology and neuroscience*, 25:55–61, 2009.

[12] Martin Reuter, Nicholas J. Schmansky, Herminia Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, 2012. doi: 10.1016/j.neuroimage.2012.02.084. URL `http://dx.doi.org/10.1016/j.neuroimage.2012.02.084`.

[13] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

[14] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, 2020.

[15] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[16] Masaru Ueda, Koichi Ito, Kai Wu, Kazunori Sato, Yasuyuki Taki, Hiroshi Fukuda, and Takafumi Aoki. An age estimation method using 3d-cnn from brain mri images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 380–383. IEEE, 2019.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[18] Liang Zou, Jiannan Zheng, Chunyan Miao, Martin J. Mckeown, and Z. Jane Wang. 3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI. *IEEE Access*, 5:23626–23636, 2017. ISSN 2169-3536. doi: 10.1109/ACCESS.2017.2762703. Conference Name: IEEE Access.

[19] Jianing Zhang, Xuechen Li, Yuexiang Li, Mingyu Wang, Bingsheng Huang, Shuqiao Yao, and Linlin Shen. Three dimensional convolutional neural network-based classification of conduct disorder with structural MRI. *Brain Imaging and Behavior*, 14(6):2333–2340, December 2020. ISSN 1931-7565. doi: 10.1007/s11682-019-00186-5.

[20] Ehsan Hosseini-Asl, Mohammed Ghazal, Ali Mahmoud, Ali Aslantas, Ahmed M. Shalaby, Manual F. Casanova, Gregory N. Barnes, Georgy Gimel'farb, Robert Keynton, and Ayman El-Baz. Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. *Frontiers in Bioscience (Landmark Edition)*, 23:584–596, January 2018. ISSN 1093-4715. doi: 10.2741/4606.

[21] Nicola K Dinsdale, Emma Bluemke, Stephen M Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana IL Namburete. Learning patterns of the ageing brain in mri using deep convolutional networks. *Neuroimage*, 224:117401, 2021.

[22] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[23] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[24] Thomas J Littlejohns, Jo Holliday, Lorna M Gibson, Steve Garratt, Niels Oesingmann, Fidel Alfaro-Almagro, Jimmy D Bell, Chris Boultwood, Rory Collins, Megan C Conroy, et al. The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications*, 11(1):1–12, 2020.

[25] Zhongheng Zhang, Marcus W Beck, David A Winkler, Bin Huang, Wilbert Sibanda, Hemant Goyal, et al. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of translational medicine*, 6(11), 2018.

[26] Jim YF Yam and Tommy WS Chow. A weight initialization method for improving training speed in feedforward neural network. *Neurocomputing*, 30(1-4):219–232, 2000.

[27] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[28] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) the tripod statement. *Circulation*, 131(2):211–219, 2015.

[29] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S Greene, et al. Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829):E14–E16, 2020.

[30] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[31] Ciprian D Billones, Olivia Jan Louville D Demetria, David Earl D Hostallero, and Prospero C Naval. Demnet: a convolutional neural network for the detection of alzheimer's disease and mild cognitive impairment. In *2016 IEEE region 10 conference (TENCON)*, pages 3724–3727. IEEE, 2016.

[32] Hugo Barbaroux, Xinyang Feng, Jie Yang, Andrew F. Laine, and Elsa D. Angelini. Encoding Human Cortex Using Spherical CNNs - A Study on Alzheimer's Disease Classification. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1322–1325, April 2020. doi: 10.1109/ISBI45749.2020.9098353. ISSN: 1945-8452.

[33] Karim Aderghal, J. Benois-Pineau, K. Afdel, and G. Catheline. FuseMe: Classification of sMRI images by fusion of Deep CNNs in 2D+e projections. *CBMI*, 2017. doi: 10.1145/3095713.3095749.

[34] Obioma Pelka, Christoph M. Friedrich, Felix Nensa, Christoph Mönninghoff, Louise Bloch, Karl-Heinz Jöckel, Sara Schramm, Sarah Sanchez Hoffmann, Angela Winkler, Christian Weimar, Martha Jokisch, and Alzheimer's Disease Neuroimaging Initiative. Sociodemographic data and APOE-ϵ4 augmentation for MRI-based detection of amnestic mild cognitive impairment using deep learning systems. *PloS One*, 15(9):e0236868, 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0236868.

[35] Ian Walsh, Dmytro Fishman, Dario Garcia-Gasulla, Tiina Titma, Gianluca Pollastri, Jennifer Harrow, Fotis E Psomopoulos, and Silvio CE Tosatto. Dome: recommendations for supervised machine learning validation in biology. *Nature methods*, 18(10):1122–1127, 2021.

[36] Ben Goldacre, Caroline E Morton, and Nicholas J DeVito. Why researchers should share their analytic code, 2019.

[37] Stephen J Eglen, Ben Marwick, Yaroslav O Halchenko, Michael Hanke, Shoaib Sufi, Padraig Gleeson, R Angus Silver, Andrew P Davison, Linda Lanyon, Mathew Abrams, et al. Toward standard practices for sharing computer code and programs in neuroscience. *Nature neuroscience*, 20(6): 770–773, 2017.

[38] Florian Markowetz. Five selfish reasons to work reproducibly. *Genome biology*, 16(1):1–4, 2015.

[39] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4): 611–627, 2018.

[40] Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4): 100241, 2021. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2021.100241. URL `https://www.sciencedirect.com/science/article/pii/S2666389921000611`.

[41] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[42] Mark T. Keane and Barry Smyth. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). *CoRR*, abs/2005.13997, 2020. URL `https://arxiv.org/abs/2005.13997`.

[43] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, 372, 2021.

[44] Samsuddin Ahmed, Byeong C. Kim, Kun Ho Lee, Ho Yub Jung, and for the Alzheimer's Disease Neuroimaging Initiative. Ensemble of ROI-based convolutional neural network classifiers for staging the Alzheimer disease spectrum from magnetic resonance imaging. *PLOS ONE*, 15(12):e0242712, December 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0242712. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0242712`. Publisher: Public Library of Science.

[45] Ricardo Mendoza-Léon, John Puentes, Luis Felipe Uriza, and Marcela Hernández Hoyos. Single-slice Alzheimer's disease classification and disease regional analysis with Supervised Switching Autoencoders. *Computers in Biology and Medicine*, 116:103527, January 2020. ISSN 1879-0534. doi: 10.1016/j.compbiomed.2019.103527.

[46] Nitsa J. Herzog and George D. Magoulas. Brain Asymmetry Detection and Machine Learning Classification for Diagnosis of Early Dementia. *Sensors (Basel, Switzerland)*, 21(3):778, January 2021. ISSN 1424-8220. doi: 10.3390/s21030778. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7865614/`.

[47] Saman Sarraf, Danielle D. Desouza, John Anderson, Cristina Saverino, and Alzheimer's Disease Neuroimaging Initiative. MCADNNet: Recognizing Stages of Cognitive Impairment through Efficient Convolutional fMRI and MRI Neural Network Topology Models. *IEEE access: practical innovations, open solutions*, 7:155584–155600, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2949577.

[48] Chunfeng Lian, Mingxia Liu, Yongsheng Pan, and Dinggang Shen. Attention-Guided Hybrid Network for Dementia Diagnosis With Structural MR Images. *IEEE Transactions on Cybernetics*, pages 1–12, 2020. ISSN 2168-2275. doi: 10.1109/TCYB.2020.3005859. Conference Name: IEEE Transactions on Cybernetics.

[49] Manu Raju, Varun P. Gopi, V. S. Anitha, and Khan A. Wahid. Multi-class diagnosis of Alzheimer's disease using cascaded three dimensional-convolutional neural network. *Physical and Engineering Sciences in Medicine*, 43(4):1219–1228, December 2020. ISSN 2662-4737. doi: 10.1007/s13246-020-00924-w. URL https://doi.org/10.1007/s13246-020-00924-w.

[50] Fan Li and Manhua Liu. Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. *Computerized Medical Imaging and Graphics*, 70:101–110, December 2018. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2018.09.009. URL https://www.sciencedirect.com/science/article/pii/S089561111830199X.

[51] Ruoxuan Cui and Manhua Liu. Hippocampus Analysis by Combination of 3-D DenseNet and Shapes for Alzheimer's Disease Diagnosis. *IEEE journal of biomedical and health informatics*, 23 (5):2099–2107, September 2019. ISSN 2168-2208. doi: 10.1109/JBHI.2018.2882392.

[52] Mingxia Liu, Jun Zhang, Dong Nie, Pew-Thian Yap, and Dinggang Shen. Anatomical Landmark Based Deep Feature Representation for MR Images in Brain Disease Diagnosis. *IEEE journal of biomedical and health informatics*, 22(5):1476–1485, September 2018. ISSN 2168-2194. doi: 10.1109/JBHI.2018.2791863. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6238951/.

[53] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Landmark-based deep multi-instance learning for brain disease diagnosis. *Medical Image Analysis*, 43:157–168, January 2018. ISSN 1361-8415. doi: 10.1016/j.media.2017.10.005. URL https://www.sciencedirect.com/science/article/pii/S1361841517301524.

[54] Guilherme Folego, Marina Weiler, Raphael F. Casseb, Ramon Pires, and Anderson Rocha. Alzheimer's Disease Detection Through Whole-Brain 3D-CNN MRI. *Frontiers in Bioengineering and Biotechnology*, 8, 2020. ISSN 2296-4185. doi: 10.3389/fbioe.2020.534592. URL https://www.frontiersin.org/articles/10.3389/fbioe.2020.534592/full. Publisher: Frontiers.

[55] Weiming Lin, Tong Tong, Qinquan Gao, Di Guo, Xiaofeng Du, Yonggui Yang, Gang Guo, Min Xiao, Min Du, Xiaobo Qu, and The Alzheimer's Disease Neuroimaging Initiative. Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer's Disease Prediction From Mild Cognitive Impairment. *Frontiers in Neuroscience*, 12, 2018. ISSN 1662-453X. doi: 10.3389/fnins.2018.00777. URL https://www.frontiersin.org/articles/10.3389/fnins.2018.00777/full. Publisher: Frontiers.

[56] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. *Frontiers in Aging Neuroscience*, 11, 2019. ISSN 1663-4365. doi: 10.3389/fnagi.2019.00194. URL https://www.frontiersin.org/articles/10.3389/fnagi.2019.00194/full. Publisher: Frontiers.

[57] Shangran Qiu, Prajakta S. Joshi, Matthew I. Miller, Chonghua Xue, Xiao Zhou, Cody Karjadi, Gary H. Chang, Anant S. Joshi, Brigid Dwyer, Shuhan Zhu, Michelle Kaku, Yan Zhou, Yazan J. Alderazi, Arun Swaminathan, Sachin Kedar, Marie-Helene Saint-Hilaire, Sanford H. Auerbach, Jing Yuan, E. Alton Sartor, Rhoda Au, and Vijaya B. Kolachalama. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain: A Journal of Neurology*, 143(6):1920–1933, June 2020. ISSN 1460-2156. doi: 10.1093/brain/awaa137.

[58] Simeon Spasov, Luca Passamonti, Andrea Duggento, Pietro Liò, and Nicola Toschi. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage*, 189:276–287, April 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.01.031. URL https://www.sciencedirect.com/science/article/pii/S105381191930031X.

[59] Jingjing Hu, Zhao Qing, Renyuan Liu, Xin Zhang, Pin Lv, Maoxue Wang, Yang Wang, Kelei He, Yang Gao, and Bing Zhang. Deep Learning-Based Classification and Voxel-Based Visualization of Frontotemporal Dementia and Alzheimer's Disease. *Frontiers in Neuroscience*, 14, 2021. ISSN 1662-453X. doi: 10.3389/fnins.2020.626154. URL https://www.frontiersin.org/articles/10.3389/fnins.2020.626154/full. Publisher: Frontiers.

[60] Füsun Çitak-ER, Dionysis Goularas, and Burcu Ormeci. A novel convolutional neural network model based on voxel-based morphometry of imaging data in predicting the prognosis of patients with mild cognitive impairment. *Journal of Neurological Sciences*, 34(1), 2017.

[61] Hamed Taheri Gorji and Naima Kaabouch. A Deep Learning approach for Diagnosis of Mild Cognitive Impairment Based on MRI Images. *Brain Sciences*, 9(9):217, August 2019. ISSN 2076-3425. doi: 10.3390/brainsci9090217. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6770590/.

[62] Simeon E. Spasov, Luca Passamonti, Andrea Duggento, Pietro Lio, and Nicola Toschi. A Multimodal Convolutional Neural Network Framework for the Prediction of Alzheimer's Disease. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2018:1271–1274, July 2018. ISSN 2694-0604. doi: 10.1109/EMBC.2018.8512468.

[63] Yan Wang, Yanwu Yang, Xin Guo, Chenfei Ye, Na Gao, Yuan Fang, and Heather T. Ma. A Novel Multimodal MRI Analysis for Alzheimer's Disease Based on Convolutional Neural Network. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 754–757, July 2018. doi: 10.1109/EMBC.2018.8512372. ISSN: 1558-4615.

[64] Fan Li and Manhua Liu. A hybrid Convolutional and Recurrent Neural Network for Hippocampus Analysis in Alzheimer's Disease. *Journal of Neuroscience Methods*, 323: 108–118, July 2019. ISSN 0165-0270. doi: 10.1016/j.jneumeth.2019.05.006. URL https://www.sciencedirect.com/science/article/pii/S0165027019301463.

[65] Manhua Liu, Fan Li, Hao Yan, Kundong Wang, Yixin Ma, Li Shen, and Mingqing Xu. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *NeuroImage*, 208:116459, March 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.116459. URL https://www.sciencedirect.com/science/article/pii/S105381191931050X.

[66] Ehsan Hosseini-Asl, Robert Keynto, and Ayman El-Baz. Alzheimer's Disease Diagnostics by Adaptation of 3D Convolutional Network. *2016 IEEE International Conference on Image Processing (ICIP)*, pages 126–130, September 2016. doi: 10.1109/ICIP.2016.7532332. URL http://arxiv.org/abs/1607.00455. arXiv: 1607.00455.

[67] F. Li, D. Cheng, and Manhua Liu. Alzheimer's disease classification based on combination of multi-model convolutional networks. *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2017. doi: 10.1109/IST.2017.8261566.

[68] Eman N. Marzban, Ayman M. Eldeib, Inas A. Yassine, Yasser M. Kadah, and for the Alzheimer's Disease Neurodegenerative Initiative. Alzheimer's disease diagnosis from diffusion tensor images using convolutional neural networks. *PLOS ONE*, 15(3): e0230409, March 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0230409. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230409. Publisher: Public Library of Science.

[69] K A N N P Gunawardena, R N Rajapakse, and N D Kodikara. Applying convolutional neural networks for pre-detection of alzheimer's disease from structural MRI data. In *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pages 1–7, November 2017. doi: 10.1109/M2VIP.2017.8211486.

[70] Silvia Basaia, Federica Agosta, Luca Wagner, Elisa Canu, Giuseppe Magnani, Roberto Santangelo, and Massimo Filippi. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical*, 21:101645, January 2019. ISSN 2213-1582. doi: 10.1016/j.nicl.2018.101645. URL https://www.sciencedirect.com/science/article/pii/S2213158218303930.

[71] Ahsan Bin Tufail, Qiu-Na Zhang, and Yong-Kui Ma. Binary Classification of Alzheimer Disease using sMRI Imaging modality and Deep Learning. *Journal of Digital Imaging*, 33(5): 1073–1090, October 2020. ISSN 0897-1889, 1618-727X. doi: 10.1007/s10278-019-00265-5. URL http://arxiv.org/abs/1809.06209. arXiv: 1809.06209.

[72] Mengjiao Hu, Kang Sim, Juan Helen Zhou, Xudong Jiang, and Cuntai Guan. Brain MRI-based 3D Convolutional Neural Networks for Classification of Schizophrenia and Controls. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2020:1742–1745, July 2020. ISSN 2694-0604. doi: 10.1109/EMBC44109.2020.9176610.

[73] Danni Cheng, Manhua Liu, Jianliang Fu, and Yaping Wang. Classification of MR brain images by combination of multi-CNNs for AD diagnosis. 10420:1042042, July 2017. doi: 10.1117/12.2281808. URL https://ui.adsabs.harvard.edu/abs/2017SPIE10420E..42C. Conference Name: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series.

[74] Loris Nanni, Matteo Interlenghi, Sheryl Brahnam, Christian Salvatore, Sergio Papa, Raffaello Nemni, Isabella Castiglioni, and The Alzheimer's Disease Neuroimaging Initiative. Comparison of Transfer Learning and Conventional Machine Learning Applied to Structural Brain MRI for the Early Diagnosis and Prognosis of Alzheimer's Disease. *Frontiers in Neurology*, 11, 2020. ISSN 1664-2295. doi: 10.3389/fneur.2020.576194. URL https://www.frontiersin.org/articles/10.3389/fneur.2020.576194/full. Publisher: Frontiers.

[75] Altug Yigit and Zerrin Işik. Applying deep learning models to structural MRI for stage prediction of Alzheimer's disease. *Turkish J. Electr. Eng. Comput. Sci.*, 2020. doi: 10.3906/elk-1904-172.

[76] Dan Pan, An Zeng, Longfei Jia, Yin Huang, Tory Frizzell, and Xiaowei Song. Early Detection of Alzheimer's Disease Using Magnetic Resonance Imaging: A Novel Approach Combining Convolutional Neural Networks and Ensemble Learning. *Frontiers in Neuroscience*, 14:259, May 2020. ISSN 1662-4548. doi: 10.3389/fnins.2020.00259. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7238823/.

[77] Juan M. Ortiz-Suárez, Raúl Ramos-Pollán, and Eduardo Romero. Exploring Alzheimer's anatomical patterns through convolutional networks. In *12th International Symposium on Medical Information Processing and Analysis*, volume 10160, page 101600Z. International Society for Optics and Photonics, January 2017. doi: 10.1117/12.2256840.

[78] Chunfeng Lian, Mingxia Liu, Jun Zhang, and Dinggang Shen. Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural MRI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):880–893, April 2020. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2889096. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[79] Aojie Li, Fan Li, Farzaneh Elahifasaee, Manhua Liu, Lichi Zhang, and Alzheimer's Disease Neuroimaging Initiative. Hippocampal shape and asymmetry analysis by cascaded convolutional neural networks for Alzheimer's disease diagnosis. *Brain Imaging and Behavior*, January 2021. ISSN 1931-7565. doi: 10.1007/s11682-020-00427-y.

[80] Ruoxuan Cui and Manhua Liu. Hippocampus analysis based on 3D CNN for Alzheimer's disease diagnosis. In *Tenth International Conference on Digital Image Processing (ICDIP 2018)*, volume 10806, page 108065O. International Society for Optics and Photonics, August 2018. doi: 10.1117/12.2503194.

[81] Karim Aderghal, Karim Afdel, Jenny Benois-Pineau, and Gwénaëlle Catheline. Improving Alzheimer's stage categorization with Convolutional Neural Network using transfer learning and different magnetic resonance imaging modalities. *Heliyon*, 6(12): e05652, December 2020. ISSN 2405-8440. doi: 10.1016/j.heliyon.2020.e05652. URL `https://www.sciencedirect.com/science/article/pii/S2405844020324956`.

[82] Manhua Liu, Danni Cheng, Kundong Wang, Yaping Wang, and Alzheimer's Disease Neuroimaging Initiative. Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer's Disease Diagnosis. *Neuroinformatics*, 16(3-4):295–308, October 2018. ISSN 1559-0089. doi: 10.1007/s12021-018-9370-4.

[83] Bumshik Lee, W. Ellahi, and J. Choi. Using Deep CNN with Data Permutation Scheme for Classification of Alzheimer's Disease in Structural Magnetic Resonance Imaging (sMRI). *IEICE Trans. Inf. Syst.*, 2019. doi: 10.1587/TRANSINF.2018EDP7393.

[84] Jingwen Sun, Shiju Yan, Chengli Song, and Baosan Han. Dual-functional neural network for bilateral hippocampi segmentation and diagnosis of Alzheimer's disease. *International Journal of Computer Assisted Radiology and Surgery*, 15(3):445–455, March 2020. ISSN 1861-6429. doi: 10.1007/s11548-019-02106-w.

[85] Jihoon Oh, Baek-Lok Oh, Kyong-Uk Lee, Jeong-Ho Chae, and Kyongsik Yun. Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm. *Frontiers in Psychiatry*, 11:16, February 2020. ISSN 1664-0640. doi: 10.3389/fpsyt.2020.00016. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7008229/`.

[86] Ruoxuan Cui and Manhua Liu. RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, 73:1–10, April 2019. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2019.01.005. URL `https://www.sciencedirect.com/science/article/pii/S0895611118303987`.

[87] Jinhyeong Bae, Jane Stocks, Ashley Heywood, Youngmoon Jung, Lisanne Jenkins, Virginia Hill, Aggelos Katsaggelos, Karteek Popuri, Howie Rosen, M. Faisal Beg, Lei Wang, and Alzheimer's Disease Neuroimaging Initiative. Transfer learning for predicting conversion from mild cognitive impairment to dementia of Alzheimer's type based on a three-dimensional convolutional neural network. *Neurobiology of Aging*, 99:53–64, March 2021. ISSN 1558-1497. doi: 10.1016/j.neurobiolaging.2020.12.005.

[88] Fanar E. K. Al-Khuzaie, Oguz Bayat, and Adil D. Duru. Diagnosis of Alzheimer Disease Using 2D MRI Slices by Convolutional Neural Network. *Applied Bionics and Biomechanics*, 2021:e6690539, February 2021. ISSN 1176-2322. doi: 10.1155/2021/6690539. URL `https://www.hindawi.com/journals/abb/2021/6690539/`. Publisher: Hindawi.

[89] Jie Zhang, Bowen Zheng, Ang Gao, Xin Feng, Dong Liang, and Xiaojing Long. A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. *Magnetic Resonance Imaging*, 78: 119–126, May 2021. ISSN 0730-725X. doi: 10.1016/j.mri.2021.02.001. URL `https://www.sciencedirect.com/science/article/pii/S0730725X21000138`.

[90] Evangeline Yee, Da Ma, Karteek Popuri, Lei Wang, Mirza Faisal Beg, and for the Alzheimer's Disease Neuroimaging Initiative, and and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing. Construction of MRI-Based Alzheimer's Disease Score Based on Efficient 3D Convolutional Neural Network: Comprehensive Validation on 7,902 Images from a Multi-Center Dataset. *Journal of Alzheimer's disease: JAD*, 79(1):47–58, 2021. ISSN 1875-8908. doi: 10.3233/JAD-200830.

[91] Gulshan Mukhtar and Saima Farhan. Convolutional neural network based prediction of conversion from mild cognitive impairment to alzheimer's disease: A technique using hippocampus extracted from mri. *Advances in Electrical and Computer Engineering*, 20(2):113–122, 2020.

[92] Jong Bin Bae, Subin Lee, Wonmo Jung, Sejin Park, Weonjin Kim, Hyunwoo Oh, Ji Won Han, Grace Eun Kim, Jun Sung Kim, Jae Hyoung Kim, et al. Identification of alzheimer's disease using a convolutional neural network model based on t1-weighted magnetic resonance imaging. *Scientific Reports*, 10(1):1–10, 2020.

[93] Loris Nanni, Matteo Interlenghi, Sheryl Brahnam, Christian Salvatore, Sergio Papa, Raffaello Nemni, Isabella Castiglioni, Alzheimer's Disease Neuroimaging Initiative, et al. Comparison of transfer learning and conventional machine learning applied to structural brain mri for the early diagnosis and prognosis of alzheimer's disease. *Frontiers in neurology*, page 1345, 2020.

[94] Eduardo Nigri, Nivio Ziviani, Fabio Cappabianco, Augusto Antunes, and Adriano Veloso. Explainable deep cnns for mri-based diagnosis of alzheimer's disease. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[95] Zhuangzhuang Li, Wenmei Li, Yan Wei, Guan Gui, Rongrong Zhang, Haiyan Liu, Yuchen Chen, and Yiqiu Jiang. Deep learning based automatic diagnosis of first-episode psychosis, bipolar disorder and healthy controls. *Computerized Medical Imaging and Graphics*, 89:101882, 2021.

[96] Shigeru Kiryu, Koichiro Yasaka, Hiroyuki Akai, Yasuhiro Nakata, Yusuke Sugomori, Seigo Hara, Maria Seo, Osamu Abe, and Kuni Ohtomo. Deep learning to differentiate parkinsonian disorders separately using single midsagittal mr imaging: a proof of concept study. *European radiology*, 29 (12):6891–6899, 2019.

[97] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*, 63:101694, 2020.

[98] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32:15479–15488, 2019.

[99] Nicholas Diakopoulos. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism*, 3(3):398–415, 2015.

[100] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[101] Victoria C Stodden. Trust your science? open your data and code. 2011.

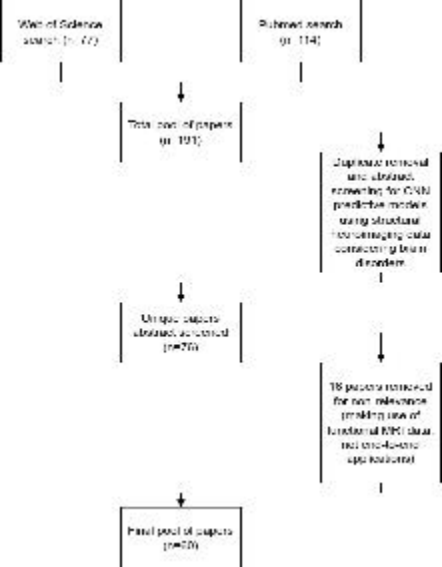
[102] Nature editorial policies. https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards. Accessed: 27-10-2021.

[103] Science editorial policies. https://www.science.org/content/page/science-journals-editorial-policies. Accessed: 27-10-2021.

[104] Google. Colaboratory: Frequently asked questions. 2018.

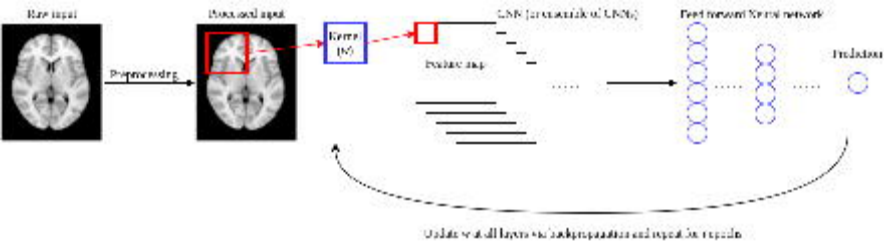
[105] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.

[106] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018.


Web of Science search (n = 77)
Pubmed search (n = 114)
Total pool of papers (n = 191)
Duplicate removal and abstract screening for CNN predictive models using structural neuroimaging data connectomics brain disorders
Unique papers abstract screened (n=70)
78 papers removed for non relevance (missing use of functional MRI data, not cross-sectional, opinion)
Final pool of papers (n=20)

Key Recommendations	**Benefits**	**Risk(s) mitigated**
Make well-annotated code freely available	- Improve chances of reproducibility – Readers can better understand workflow – Encourage accountability and transparency	- Limit reproducibility efforts – Models remain opaque
Employ repeat experiments	- Improve confidence in model estimation – Mitigate random weight initialisation	- Risk reporting overfitted results – Performance estimation inflation – Diminished confidence in system overall
Use saliency metrics and counterfactuals	- Validate that model is using relevant information – Potential biomarker discovery – Improve confidence in system overall	- Models remain opaque – Diminished confidence in system overall – Unsure what information is being used by models
Avoid 2D data structures where possible	- Ensure spatial information between slices is considered – Lessen chance of information leakage – No requirement of multiple voting strategies	- Information leakage is more likely to occur – No guarantee spatial information between slices is considered – Must consider multiple voting strategies

Table 1: Key recommendations arising from the results of this systematic literature review, their benefits, and the risks associated with non-adherence.


Raw input
Preprocessing
Processed input
Kernel (N)
CNN (a cascade of CNNs)
Feature map
Fully connected Neural network
Prediction
Update weights in all layers via backpropagation and repeat for epochs

Question	**Answer**
How are data represented?	2D(n=24), 3D(n=35), Both(n=1)
Is code available?	No(n=54), Yes(n=6)
Conference or Journal?	Conference(n=13), Journal(n=47)
Is saliency considered?	No(n=40), Yes(n=20)
Are there repeat experiments?	No(n=26), Yes(n=34)
Is there potential information leakage?	No(n=32), Yes(n=28)

Table 1: Numeric summary of study attributes from the 60 papers satisfying selection criteria.

Table 1: Tabular presentation of the studies considered for this systematic literature review.

Authors and citation	**Modelling Practices**			**Transparency**	**Interpretability**	**Publication Status**
	Data representation	**Repeat experiments**	**Data leakage**	**Code availability**	**Saliency**	
Zou et al. (2017) [18]	3D	Yes	No	No	No	Conference
Çitak-ER et al. (2017) [60]	2D	No	Yes	No	No	Journal
Taheri Gorji and Kaabouch (2019) [61]	3D	No	No	No	No	Conference
Spasov et al. (2018) [62]	2D	No	Yes	No	No	Conference
Wang et al. (2018) [63]	3D	Yes	Yes	No	Yes	Journal
Li and Liu (2019) [64]	3D	Yes	Yes	No	Yes	Journal
Liu et al. (2020) [65]	3D	Yes	Yes	No	No	Journal
Hosseini-Asl et al. (2016) [66]	3D	Yes	Yes	No	No	Conference
Li et al. (2017) [67]	3D	Yes	Yes	No	No	Conference
Folego et al. (2020) [54]	3D	Yes	Yes	Yes	Yes	Journal
Marzban et al. (2020) [68]	3D	No	Yes	No	No	Journal
Hosseini-Asl et al. (2018) [20]	3D	Yes	No	No	No	Journal
Gunawardena et al. (2017) [69]	2D	No	Yes	No	No	Conference
Basaia et al. (2019) [70]	3D	No	Yes	No	No	Journal
Tufail et al. (2020) [71]	2D	Yes	Yes	No	No	Journal
Hu et al. (2020) [72]	3D	Yes	No	No	No	Journal
Cheng et al. (2017) [73]	3D	No	Yes	No	No	Conference
Nanni et al. (2020) [74]	3D	No	No	No	No	Journal
Lin et al. (2018) [55]	2D	Yes	Yes	No	No	Journal
Billones et al. (2016) [31]	2D	No	No	No	No	Conference
Barbaroux et al. (2020) [32]	2D	Yes	No	No	No	Conference
Yigit and Işik (2020) [75]	2D	No	Yes	No	No	Journal
Pan et al. (2020) [76]	2D	Yes	Yes	No	No	Journal
Ahmed et al. (2020) [44]	2D	No	Yes	No	No	Journal
Ortiz-Suárez et al. (2017) [77]	2D	Yes	Yes	No	Yes	Conference
Aderghal et al. (2017) [33]	2D	No	No	No	No	Conference
Lian et al. (2020) [78]	3D	No	No	No	Yes	Journal
Li et al. (2021) [79]	3D	Yes	No	No	Yes	Journal
Cui and Liu (2018) [80]	3D	Yes	No	No	No	Conference
Aderghal et al. (2020) [81]	2D	No	No	No	No	Journal
Böhle et al. (2019) [56]	3D	Yes	No	Yes	Yes	Journal
Sarraf et al. (2019) [47]	2D	Yes	Yes	Yes	Yes	Journal
Liu et al. (2018) [82]	3D	Yes	Yes	No	Yes	Journal
Zhang et al. (2020) [19]	3D	Yes	No	No	Yes	Journal
Lee et al. (2019) [83]	2D	Yes	Yes	No	No	Journal
Qiu et al. (2020) [57]	3D	Yes	Yes	Yes	Yes	Journal
Spasov et al. (2019) [58]	3D	Yes	No	Yes	No	Journal
Sun et al. (2020) [84]	3D	Yes	No	No	No	Journal
Oh et al. (2020) [85]	3D	Yes	No	No	No	Journal
Lian et al. (2020) [48]	3D	No	No	No	Yes	Journal
Cui and Liu (2019) [86]	3D	Yes	No	No	Yes	Journal
Raju et al. (2020) [49]	3D	Yes	No	No	Yes	Journal
Mendoza-Léon et al. (2020) [45]	2D	No	Yes	No	No	Journal

Table 1: Tabular presentation of the studies considered for this systematic literature review.

Authors and citation	**Modelling Practices**			**Transparency**	**Interpretability**	**Publication Status**
	Data representation	**Repeat experiments**	**Data leakage**	**Code availability**	**Saliency**	
Pelka et al. (2020) [34]	2D	Yes	No	No	Yes	Journal
Li and Liu (2018) [50]	3D	Yes	No	No	Yes	Journal
Bae et al. (2021) [87]	3D	No	No	No	Yes	Journal
Cui and Liu (2019) [51]	3D	Yes	No	No	No	Journal
Liu et al. (2018) [52]	3D	No	No	No	No	Journal
Liu et al. (2018) [53]	3D	Yes	No	No	No	Journal
Al-Khuzaie et al. (2021) [88]	2D	No	Yes	No	No	Journal
Zhang et al. (2021) [89]	3D	No	No	No	Yes	Journal
Hu et al. (2021) [59]	3D	No	No	Yes	Yes	Journal
Herzog and Magoulas (2021) [46]	2D	No	Yes	No	No	Journal
Yee et al. (2021) [90]	3D	Yes	Yes	No	No	Journal
Mukhtar and Farhan (2020) [91]	2D	No	Yes	No	No	Journal
Bae et al. (2020) [92]	2D	Yes	No	No	No	Journal
Nanni et al. (2020) [93]	2D + 3D	No	No	No	No	Journal
Nigri et al. (2020) [94]	2D	No	No	No	Yes	Conference
Li et al. (2021) [95]	2D	Yes	Yes	No	No	Journal
Kiryu et al. (2019) [96]	2D	No	No	No	No	Journal