Magnetic Resonance Imaging-based Deep Learning Predictive Models of Brain Disorders: A Systematic Review of Modelling Practices, Transparency, and Interpretability

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, National University of Ireland, Galway, Ireland
[2]Clinical Neuroimaging Laboratory, Galway Neuroscience Centre, College of Medicine Nursing and Health Sciences, National University of Ireland, Galway, Ireland

Abstract

Brain disorders are characterised by impaired cognition, mood alteration, psychosis, depressive episodes, and neurodegeneration, and comprise several psychiatric and neurological disorders. 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 consortiums (such as the Alzheimer's Disease Neuroimaging Initiative) facilitating a greater number of MRI-based classification studies, convolutional neural networks (CNNs), which are multi-layer representation-based models particularly well suited to image processing, have become increasingly popular for research into brain conditions. Despite this, modelling practices, the degree of transparency, and considerations of interpretability vary widely across studies, making them difficult to both compare and/or reproduce. Modelling practices here refers to issues surrounding the data splitting procedure, the presence or absence of repeat experiments, the critical appraisal of performance metrics, and the overall reliability of the modelling approach. Transparency refers to how detailed the authors' methodological descriptions are, and the availability of code. Finally, interpretability refers to the attempt made by the authors to identify structural brain alterations driving model predictions – this is particularly important as the application of deep learning systems becomes more widespread in clinical settings. Here, we conduct a systematic literature review of 55 studies carrying out CNN-based predictive modelling of brain disorders using MRI data and critique their modelling practices, transparency, and considerations of interpretability; we furthermore propose several practical recommendations aimed at promoting comprehensive, clear, and reproducible research into brain disorders using MRI-based deep learning models.

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 pathophysiological neuroanatomical changes, with considerable collective public health burden 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. In particular, there is significant interest in biomarkers for the differentiation of conditions and their subtypes, which could yield greater mechanistic understanding of symptomatic presentations [3, 4]. Neurobiological markers, such as differential neuroanatomical variation, have been extensively studied for discriminative and descriptive purposes [5, 6, 7]. This research is usually facilitated by magnetic resonance imaging (MRI) data modalities, which can offer non-invasive measures of brain structure [8]. The increasing availability of MRI data, and the long term goal to incorporate biological information into diagnostic systems, have enabled a wealth of research in this domain focused on predictive modelling [9]. For example, machine learning and classical statistical learning algorithms have previously highlighted differential neuroanatomical patterns across several conditions, including subcortical structure volume reduction in bipolar disorder and Alzheimer's disease [10, 11]. However, incorporating such information into clinical systems is non-trivial, as the precise dynamics and limitations of a particular biomarker must be known and addressed prior to use [12, 13]. Additionally, the methods used to describe these discriminative features have their own considerations, such as requiring preprocessing tools to derive tabular brain summary information [14, 15]. These tools can produce variable results depending on the parameters chosen, even when applied to the same dataset, partly owing to the large number of parameter choices available per tool – this means that domain expertise is often necessary to justify decisions [16]. Finally, statistical modelling techniques often require formal specification of the expected variable

*Corresponding author, email: s.oconnell29@nuigalway.ie

relationships with the output, and generally are unsuited to high-dimensional data structures such as structural MRI scans and/or pattern discovery. Machine learning approaches are also limited by their inability to consider spatial relationships between groups of pixels in imaging data structures, making it necessary to utilise the aforementioned tabular summary derivation tools. With these factors in mind, deep learning algorithms – and particularly those well-suited to imaging data structures – have become popular. This is because of their ability to consider arbitrarily complex relationships without tabular summary derivation, meaning that researchers are afforded greater model flexibility and do not need to specify expected variable relationships. Convolutional neural networks (CNNs), which have shown impressive predictive performances in generic image classification tasks, have been applied to the medical imaging field more broadly for segmentation and prediction tasks, and are becoming increasingly popular for predictive modelling of the brain in terms of aging and psychiatric/neurological disorders [17, 18, 19, 20, 21, 22, 23]. These models are designed specifically to detect and leverage spatial patterns in image data structures, making them well suited for these applications [24].

These recent developments have been further enabled by access to large standardised neuroimaging data colelctions available for general research use, such as the Alzheimer's Disease Neuroimaging Initiative and the UK Biobank [25, 26]. In effect, this means that researchers can train complex predictive models with relatively straightforward open-source frameworks (such as Tensorflow [27] and Pytorch [28]) on large amounts of data without requiring domain expertise. While this makes the application of these approaches more accessible, there are a few caveats that must be considered. Firstly, deep learning models have a number of unique limitations, such as their high parameter dimensionality, lack of interpretability, stochasticity in weight initialisation, lack of uncertainty, and difficulty to train [29, 30, 24, 31]. Secondly, clinical decision systems require rigorous validation and reporting frameworks for more interpretable models, meaning that the use of opaque deep learning algorithms make these principles of transparency and validation even more difficult to achieve [32, 33]. Clinical decision systems that offer no indication as to why they have made a particular classification are unsuitable for use in real world applications. These factors combine to make the application of deep learning to clinical settings challenging.

As the number of studies applying deep learning to brain disorder neuroimaging data increases, research highlighting the potential clinical utility of these methodologies must be proactive in addressing these issues for the sake of patients and the scientific integrity of the field, with transparent reporting, critical examination of performance metrics, and thorough considerations of interpretability. In this paper, we seek to assess the state of 1) modelling practices, 2) transparency, and 3) interpretability in imaging-based deep learning predictive models applied to brain disorders. We systematically review 55 papers and analyse their methodologies in the context of these important concepts.
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 focusing on these three topics in particular; we then identify key challenges of the selected papers in the context of the aforementioned principles and posit several recommendations for future studies based on our analysis.

1.1 Convolutional Neural Networks

Glossary

- ***Node***: Sum-weighted combination of inputs at a particular layer.
- ***Layer***: Collection of nodes, the outputs of which can act as the input to the next layer of nodes.
- ***Activation function***: Arbitrary function applied to nodes which confers nonlinearity.
- ***Convolution***: Matrix multiplication of input window by weights window of the same size.
- ***Filter***: Weights window that forms part of convolution operation defined above.
- ***Feature map***: Output image given by convolution of filter across every window of a specified size.
- ***Neural network***: Model consisting of an arbitrary number of nodes and layers, which are thresholded by activation functions.
- ***Convolutional Neural Network***: Model consisting of an aribtrary number of feature maps with respective filters, thresholded by activation functions, which recognise spatial data patterns via convolutions. A standard neural network is usually placed at the end of a Convolutional Neural Network.

CNNs are a popular deep learning image algorithm in many areas of research, particularly in studies making use of MRI data [17, 18, 19, 20, 21, 22, 23]. Their structure is explicitly 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, which for example, may be $2x2$ pixels large) and an input image patch of the same size. The convolution of the same filter over every patch of the input image outputs the entire 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. Because filters are convolved across entire input images, the exact spatial location of a pattern is unimportant, allowing the model to detect and leverage data patterns regardless of exact image coordinates; this is termed *shift invariance*. The convolutional operation can be repeated multiple times and is usually accompanied by downsampling operations which reduce the dimensionality of the output. This flattened list based vector

is then usually passed to a typical neural network model, which consists of a series of layers with an arbitrary number of nodes. Each node at a given layer is the sum-weighted output of the all previous layer input values and their output is thresholded by an *activation function*, typical examples of which include the rectified linear unit ($max(0, x)$) or the logistic function ($\frac{1}{1+e^{-x}}$). After an arbitrary number of layers, the sum-weighted output of all nodes at the penultimate layer can be passed through a sigmoidal function which transforms the result into the probability space. Mathematically, a general neural network can be defined as:

$$f(x, W) = \sigma\left(\sum_{i=1}^{M} w_{ij}^{k} h + b\right) \tag{1}$$

Here, $f(x, W)$ is the output of the network, M is the number of nodes at the previous layer, w_{ij}^{k} represents the connection between node/input i and j at layer k, h refers to the output of the previous layer nodes, b refers to the bias term, similar to ϵ in a linear regression model, and σ refers to an arbitrary activation function.
Weights are updated through backpropagation using the gradient of the output with respect to the input which allows minimisation of an objective function (which is often the log-loss of the prediction vs. the ground truth label in binary classification settings), and can be broadly defined as:

$$W_{i+1} = W_i - \alpha \frac{\delta f(x, W)}{\delta x} \tag{2}$$

Here, α is a hyperparameter controlling the severity of the weight update at a particular epoch/time point, denoted by i, δ represents the gradient of a specified quantity, and $f(x, W)$ is the output of the network where W is the weights vector and x is the input. The chain rule allows for weight updates to be applied to all layers via partial derivatives, and a more in-depth consideration of neural network training can be found in LeCun et al., 2012 [31].

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 of this set of procedures. Preprocessing is usually applied to skull strip, linearly or non-linearly register raw input images to a template, 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 [34, 35]. Some studies also train custom architectures on 2D data [36, 37, 38]. The prediction output is usually presented as a probability, with the final layer of outputs transformed via a sigmoidal/softmax function. This probability is then used to calculate performance metrics such as the area under the receiver operating characteristic curve (AUC) and accuracy. Interpretation of the results can be carried out with gradient based saliency metrics, or by visualising feature maps at specific layers [39]. Counterfactuals can also be employed to further understand what relationships have been captured by the model by generating plausible instances to act as the input – often, saliency metrics reduce non-linear relationships into single-number measures of 'importance', which can be difficult to interpret in isolation [40]. Weights are often randomly initialised according to a specified statistical distribution, such as the Gaussian distribution, making the training procedure sensitive to the starting conditions. Additionally, owing to the large number of parameters (often in the millions) there is no strict definition of algorithmic convergence, meaning that there may be multiple optimal or suboptimal sets of weights that minimise the objective function. Weights can be updated according to an optimisation function, such as adaptive moment estimation, with each update strategy having their own set of hyperparameters such as the learning rate (α from Equation 2) [41]. The overall architecture specifics can also differ, with the number of layers, feature maps, and nodes per layer varying widely according to the precise method. Additionally, several specific architectural variants of CNNs are commonly applied, including DenseNet and Recurrent Neural Networks, which reorganise the structure and operations of standard networks by altering how information is passed from one layer to the next [42, 43].

In the following sections, we define and attempt to justify the need for good modelling practices, transparency, and interpretability.

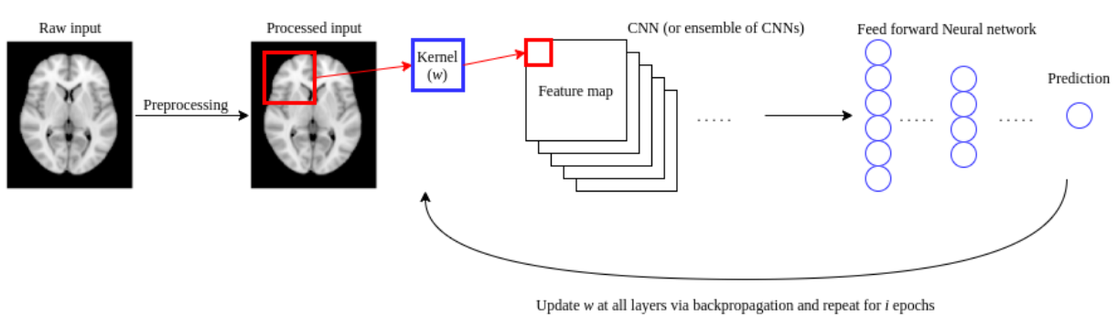

Figure 1: General experimental workflow. The preprocessed input image, either in 2 or 3 dimensional format is passed to a CNN model (or ensemble of CNN models) for training and prediction, The weights vector w is updated via backpropagation at each epoch, minimising the error of the objective function.

1.3 Modelling Practices

Good modelling practices broadly refers to the robustness of the methodology used – as previously mentioned, deep learning models have a number of unique limitations that must be considered. For example, we can examine the presence of repeat experiments, the data splitting procedure, and whether or not there was information leakage as indicators of whether or not good modelling practices have been observed. Information leakage describes situations whereby model testing and training sets are not kept entirely independent during training, which may lead to inflated model performance estimation; this can impact reproducibility and ultimately the reliability of clinical applications. A useful type of repeat experiment that can be carried out is k-fold cross validation, whereby the data is split into k folds, and $k-1$ folds are used to train the model. The remaining fold is used as the testing set, and this procedure is repeated k times, until every fold has served as the testing set. Because neural networks have stochastic weight initialisations and no numeric definition of convergence, repeat experiments can ensure these limitations are mitigated by averaging over multiple random start points. An additional benefit of k-fold cross validation in particular is its ability to estimate model performance on multiple data splits, which can yield a more robust prediction metric compared to experiments with single data splits.

1.4 Transparency

Transparency here refers to how thorough the study's methodological discussions are. As previously detailed, there are many hyperparameters associated with deep learning models, which can have effects on the overall performance and utility of the system. Therefore, we can assess the degree of code sharing, discussions of preprocessing and limitations, and model availability when considering the transparency of a given study. These factors can influence the ability of researchers to reproduce reported experimental findings, which can effect the overall confidence in the methodological approach. This is especially important in the context of integrating deep learning models into clinical settings.

1.5 Interpretability

Interpretability here refers to the efforts made to explain model predictions. Understanding why a model made a particular classification is important for patient trust, biomarker discovery, and the validation of existing clinical knowledge. Deep learning models are not usually well suited to interpretative frameworks, but studies underlining the potential clinical utility of a model must attempt to explain the model's decisions when considering healthcare applications. We evaluate the use of saliency or other methods to explain model decisions and the discussion of interpretation findings to assess the attention given to interpretability.

2 Methods

We conducted a systematic literature review according to PRISMA guidelines, the details of which are provided below [44].

2.1 Inclusion/exclusion criteria

We limited our search to consider studies making use of CNN architectures, as CNN-based architectures are the most popular deep learning modelling approach for medical imaging data structures. 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 difficult to compare relative to structural studies.

2.2 Search details

We performed a Web of Science (all databases) and Pubmed search with the following keywords:
((((structural) AND (imaging)) AND (MRI)) AND ((CNN) OR (convolutional neural network) OR (3D-CNN))) AND (psychiatric OR depression OR autism OR bipolar OR Alzheimer's OR neurological).
For Web of Science, 71 results were returned, and 110 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 66 papers. 11 studies were excluded for various reasons, including functional MRI data being the main focus of the study, and the use of hybrid models where CNNs were not the primary modelling method; this resulted in a total of 55 papers remaining for review.

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 examine implementation details and transparency in a qualitative framework. A quantitative analysis of performance metric variation across studies was not the focus of this work.

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 a sizeable fraction of studies (20/55) represented data in 2D; while this is more computationally efficient than 3D representation, it can introduce potential sources of information leakage (Table 2). Furthermore, accuracy calculation can be carried out per slice or per patient, introducing issues surrounding the optimal majority voting method for clinical settings. Of the 20 studies making use of 2D slices, only one explicitly referred to voting methods, and 15 studies suffered from potential leakage [45]. Several studies made use of single slices, or the same slice indices across different patients [46, 47, 37]. Given the often minimal preprocessing protocols that accompany deep learning papers, there is no guarantee that the same biological information is considered per patient when taking this approach. Additionally, relevant spatial information can be lost when modelling in 2D, even if multiple slices are taken as they may not be considered in unison during training. One paper making use of 2D slices provided code [48]. Additionally, $\approx 44\%$ of studies (24/55) made use of multiple models for training and prediction, which in some cases translated to stacking, whereby the output of one trained model is passed to another for training as the input [49, 50, 51, 52, 53, 54]. In a number of papers, statistical tests or accuracy thresholds were used to pre-select informative image patches [54, 53, 46]. This can introduce bias via focusing the model on 'informative' regions that meet certain criteria which may not translate directly to that region's biological relevance or utility in a full model, which risk missing mechanistically relevant revelations and could result in false negative studies.
In several studies, one model was trained and the weights from that model were used for transfer learning of a subsequent model, or the predictive/statistical utility of individual patches was used to focus attention on specific regions prior to testing [55, 56, 45, 46, 38]. This can be classed as a specific form of variable selection bias that means the model is focusing on specific features highlighted by previous methods, which would have major implications for biomarker discovery.
Thirty two out of 55 studies employed repeat experiments through cross validation or other means. A third of studies carrying out repeat experiments (10/32) reported only point estimates for their results, and 5 provided code [57, 48, 58, 55, 59]. Of the 14 studies that had both repeat experiments and considerations of interpretability, none detailed whether or not their saliency method was applied per fold or on a hold out test set, and this information was also not detailed where code was provided. This suggests that while most studies carried out repeat experiments, there remained issues in their methodologies and reporting.

3.2 Transparency

We found that $\approx 90\%$ of papers (49/55) did not provide any code or model weights to supplement their methodological descriptions, meaning that the majority of studies relied on textual summaries for their methods sections. This highlights a lack of methodological transparency across the considered research, especially considering the many different subjective choices required during model construction that can have misleading effects on the overall performance of the system. Of the studies that did make code available, no paper provided detailed tutorials of preprocessing and model construction – understandably, the quality and thoroughness of reported code is another important aspect of reproducibility and transparency which is not solved by making code available [55, 57, 58, 59, 60, 48]. Additionally, most papers considered were from journals, meaning that the majority of studies underwent some form of peer review (43/55).

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) [20]	3D	Yes	No	No	No	Conference
Çitak-ER et al. (2017) [61]	2D	No	Yes	No	No	Journal
Taheri Gorji and Kaabouch (2019) [62]	3D	No	No	No	No	Conference
Spasov et al. (2018) [63]	2D	No	Yes	No	No	Conference
Wang et al. (2018) [64]	3D	Yes	Yes	No	No	Journal
Li and Liu (2019) [65]	3D	Yes	Yes	No	No	Journal
Liu et al. (2020) [66]	3D	Yes	Yes	No	No	Journal
Hosseini-Asl et al. (2016) [67]	3D	Yes	Yes	No	No	Conference
Li et al. (2017) [68]	3D	Yes	Yes	No	No	Conference
Folego et al. (2020) [55]	3D	Yes	Yes	Yes	Yes	Journal
Marzban et al. (2020) [69]	3D	No	Yes	No	No	Journal
Hosseini-Asl et al. (2018) [22]	3D	Yes	No	No	No	Journal
Gunawardena et al. (2017) [70]	2D	No	Yes	No	No	Conference
Basaia et al. (2019) [71]	3D	No	Yes	No	No	Journal
Tufail et al. (2020) [72]	2D	Yes	Yes	No	No	Journal
Hu et al. (2020) [73]	3D	Yes	No	No	No	Journal
Cheng et al. (2017) [74]	3D	No	Yes	No	No	Conference
Nanni et al. (2020) [75]	3D	No	No	No	No	Journal
Lin et al. (2018) [56]	2D	Yes	Yes	No	No	Journal
Billones et al. (2016) [35]	2D	No	No	No	No	Conference
Barbaroux et al. (2020) [36]	2D	Yes	No	No	No	Conference
Yigit and Işik (2020) [76]	2D	No	Yes	No	No	Journal
Pan et al. (2020) [77]	2D	Yes	Yes	No	No	Journal
Ahmed et al. (2020) [45]	2D	No	Yes	No	No	Journal
Ortiz-Suárez et al. (2017) [78]	2D	Yes	Yes	No	Yes	Conference
Aderghal et al. (2017) [37]	2D	No	No	No	No	Conference
Lian et al. (2020) [79]	3D	No	No	No	Yes	Journal
Li et al. (2021) [80]	3D	Yes	No	No	Yes	Journal
Cui and Liu (2018) [81]	3D	Yes	No	No	No	Conference
Aderghal et al. (2020) [82]	2D	No	No	No	No	Journal
Böhle et al. (2019) [57]	3D	Yes	No	Yes	Yes	Journal
Sarraf et al. (2019) [48]	2D	Yes	Yes	Yes	Yes	Journal
Liu et al. (2018) [83]	3D	Yes	Yes	No	Yes	Journal
Zhang et al. (2020) [21]	3D	Yes	No	No	Yes	Journal
Lee et al. (2019) [84]	2D	Yes	Yes	No	No	Journal
Qiu et al. (2020) [58]	3D	Yes	Yes	Yes	Yes	Journal
Spasov et al. (2019) [59]	3D	Yes	No	Yes	No	Journal
Sun et al. (2020) [85]	3D	Yes	No	No	No	Journal
Oh et al. (2020) [86]	3D	Yes	No	No	No	Journal
Lian et al. (2020) [49]	3D	No	No	No	Yes	Journal
Cui and Liu (2019) [87]	3D	Yes	No	No	Yes	Journal
Raju et al. (2020) [50]	3D	Yes	No	No	Yes	Journal
Mendoza-Léon et al. (2020) [46]	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) [38]	2D	Yes	No	No	Yes	Journal
Li and Liu (2018) [51]	3D	Yes	No	No	Yes	Journal
Bae et al. (2021) [88]	3D	No	No	No	Yes	Journal
Cui and Liu (2019) [52]	3D	Yes	No	No	No	Journal
Liu et al. (2018) [53]	3D	No	No	No	No	Journal
Liu et al. (2018) [54]	3D	Yes	No	No	No	Journal
Al-Khuzaie et al. (2021) [89]	2D	No	Yes	No	No	Journal
Zhang et al. (2021) [90]	3D	No	No	No	Yes	Journal
Hu et al. (2021) [60]	3D	No	No	Yes	Yes	Journal
Herzog and Magoulas (2021) [47]	2D	No	Yes	No	No	Journal
Yee et al. (2021) [91]	3D	Yes	Yes	No	No	Journal
Mukhtar and Farhan (2020) [92]	2D	No	Yes	No	No	Journal

Question	Answer
How are data represented?	2D(n=20), 3D(n=35)
Is code available?	No(n=49), Yes(n=6)
Conference or Journal?	Conference(n=12), Journal(n=43)
Is saliency considered?	No(n=36), Yes(n=19)
Are there repeat experiments?	No(n=23), Yes(n=32)
Is there potential information leakage?	No(n=28), Yes(n=27)

Table 2: Numeric summary of study attributes from the pool of 55 selected papers.

3.3 Interpretability

We noted that 19 out of 55 studies considered interpretability by applying a saliency method (such as Gradient-based Class Activation Mapping [93]) or visualising feature maps [19]. Of the 19 studies considering saliency, 4 papers dedicated sections of their discussion to the interpretation of saliency outputs, with the remainder reporting saliency outputs without further commentary [57, 83, 58, 50]. All 4 of these papers with discussions of interpretability made use of a single saliency method and furthermore assumed the relationship captured by the model was the same as previously reported neuroanatomical patterns, without carrying out experiments to confirm or refute this assumption. Two of the 4 studies provided code [57, 58]. Despite welcome considerations of saliency, the 19 studies considering interpretability had broad variability in the quality and detail afforded to model interrogation. For instance, 9 studies presented the results of a saliency method or an attempt at interpretability of their respective models with little to no critical discussion of regions highlighted or methods employed [65, 66, 79, 80, 48, 83, 87, 51, 88]. The 9 aforementioned studies did not provide any information on implementation details. Of the remaining 10 studies, 5 provided code, but as previously mentioned did not include detailed walkthroughs of their interpretation pipelines [57, 48, 58, 55, 60].
A subsection of studies also specifically underscored that extensive, expert driven preprocessing is not required with deep learning studies, and almost all studies alluded to this fact in their introductions [60, 69, 79, 51, 52]. This position downplays the importance of expert opinion in model interpretation and preprocessing decisions for medical imaging studies. Without explicit knowledge concerning what image aspects to exclude, researchers can include irrelevant information during model training, as demonstrated by the inclusion of skull and neck information in several studies [60, 91, 38, 84, 67, 77]. In practice, this would mean the models may have picked up on irrelevant information about neck size or skull thickness, that, if used in clinical applications, could lead to misclassifications of patients with those specific physical characteristics, which may have nothing to do with the condition of interest. Furthermore, certain studies avoided considering interpretability in greater detail due to their lack of expertise, which implies that expert knowledge is a requirement for thorough considerations of saliency [60]. These examples illustrate the crucial importance of both domain knowledge and interpretability, which in this case would have highlighted potentially spurious relationships the model may have been using during classification.

4 Discussion

Our results demonstrated issues in modelling practices, transparency, and interpretability across a selected pool of 55 papers concerned with CNN-based predictive modelling of brain disorders with MRI data. We found that 20 out of 55 papers considered made use of 2D data structures, 49 did not provide code, 36 did not consider saliency, 32 employed repeat experiments, and 27 may have suffered from data leakage. We discuss these findings below and propose several recommendations to improve the quality of studies concerned with CNN-based predictive modelling of brain disorders using structural neuroimaging data.

4.1 Data representation

A majority of papers in this set of literature made use of 3-dimensional data representations, which is computationally intensive but robust. Modelling data in 3D ensures that all biological information is used during training, as opposed to individual 2D slices whose spatial inter-dependencies may not be considered. 3D modelling also ensures the same biological information is utilised per patient, which may not be the case for 2D experiments that work via indexing. There was still a significant minority of papers making use of 2D data structures (20/55), which may pose issues for downstream clinical applications.

As highlighted previously, 2D-data based models have a number of limitations, including not considering the same biological information per patient, multiple potential majority voting strategies, and possible information leakage. These factors in combination with a lack of thorough interpretability make these models unsuitable for application to real-world clinical settings, a drawback not addressed seriously by any work in the presented literature. We therefore recommend that researchers think critically about these limitations before deciding to use 2D-based models, and take practical steps to ensure the reliability of the methodology. For example, where multiple per-slice voting schemes are available, researchers should examine how performance metrics change relative to the strategy implemented, and what potential caveats each approach could have in practice. To proactively address information leakage, researchers should take care to split data into 2D-based representations after data has been split into train, test, and validation sets at the patient level, and should provide code to prove they followed this procedure. We furthermore recommend that single-slice-based studies consider alternative modelling approaches, as the implications of using one 2D slice from a 3D stack, along one dimension, with no guarantee the same biological information is being considered per patient, may lead to performance estimation inflation which will ultimately hamper reproducibility.

4.2 Repeat experiments

Most studies implemented repeat experiments, which ensures that stochasticity in weight initialisation and/or performance estimation inflation due to particular fold splitting is mitigated. Averaging over multiple random weight start points is a useful strategy to obtain a robust performance estimation, and cross validation schemes are useful model diagnostics; these strategies can help to ensure the resultant performance metrics are reliable. Despite this, 23 of the 55 considered papers did not employ repeat experiments. Cross validation as a diagnostic is a useful way to assess model performance across various splits of the data even when weight initialisation is not random – factoring in this added stochasticity associated with deep learning models, repeat experiments become all the more important. Additionally, a number of studies that did employ repeat experiments only reported point estimates, which undermines the utility of carrying out this modelling practice to begin with. Code inaccessibility exacerbates this issue further, leaving the reader unclear as to what procedure was followed. We recommend that researchers employ repeat experiments via cross-validation or repeated model fitting where data is not split multiple times and report their results as a spread of points with standard deviations. This can provide further confidence in the ability of the model to generalise to differing data splits, and although in itself it is not a remedy for overfitting, it remains a useful model diagnostic.

4.3 Code availability

The majority of studies in this set of literature did not provide any code. Wen et al. (2020) [94] previously underlined the importance of fairness, accountability, and transparency in deep learning modelling studies, but many studies fall short of fulfilling these principles. The construction of deep learning systems requires many hyperparameter and algorithmic decisions which can influence overall model performance, introduce bias, and impact reproducibility. Deep learning models are essentially systems that optimise an objective function over a specified set of arguments, meaning that any decisions taken in preprocessing and model construction, such as the choice of learning rate, loss function, and number of layers, can affect the capabilities of the system as a whole, and as a result, propagate subjective choices throughout 'objective' models [95]. For instance, several studies have examined algorithmic biases against underrepresented and/or marginalised groups, which can persist even if code is freely available [96, 97, 98]. Aside from deep learning-specific benefits to code sharing, the larger scientific community has recently shifted towards open science frameworks, with several high-profile journals requiring thorough methodological transparency [99, 100, 101]. Therefore, we recommend that thorough documentation of code and methodological details is at the minimum an essential aspect of deep learning experiments in this domain. Patient privacy concerns, while valid, are no impediment to model weight and code sharing, and minimal testing datasets could be provided via anonymisation procedures [102]. Of the studies that did make code available, no paper provided tutorials of preprocessing and model construction through, for example, minimal Jupyter notebook/Google Colab implementations, with justifications provided for algorithmic decisions. We further recommend such practical steps, which could facilitate greater methodological transparency and allow researchers to understand the experimental decisions that gave rise to the results. This would also have the useful properties of allowing researchers to examine what pipelines gave rise to successful experiments, and allow them to spot potential 'blind spots' that the model authors may have overlooked. Making entire pipelines easily accessible and examinable facilitates accountability and aids reproducibility efforts overall.

4.4 Saliency and interpretability

We found a lack of adequate model interrogation in the presented studies. As previously stated, algorithmic biases in predictive settings against marginalised or underrepresented groups is of serious concern, particularly for clinical settings, and saliency methods can help researchers to identify sources of potential bias. Additionally, biomarker discovery could be greatly assisted by meticulous interrogation of model predictions in various scenarios. Thus, regardless of predictive performance, without at least identifying what neuroanatomical patterns are driving decisions, studies not considering interpretability are unsuitable for use in clinical settings. Additionally, the 'importance' per pixel, the quantity most often returned by saliency methods, has no direct interpretation that can relate region relevance back to

human-interpretable neuroanatomical pathologies. In many instances, it represents the degree of change in the output relative to a small change in the input pixel value, collapsing a potentially non-linear relationship to a collection of single values without units. While an empirical measure, it offers little interpretative value in comparison to the coefficients returned by classical statistical methods, which can provide immediate explanatory insight. This is partly due to a focus of the deep learning field in general on prediction as opposed to inference, meaning that the mechanistic understanding of relationship dynamics is of less importance than the final predictive performance. For previously specified reasons, thorough examination of models, and their inferential properties, is crucial where patient care is concerned. We therefore recommend that saliency methods be employed in future studies. Furthermore, individual saliency methods have their own unique limitations which must be considered, with various implementation strategies (such as aggregation of local examples, what saliency outputs to report where repeat experiments have been carried out, etc.), which should be carefully considered through consultation of the relevant literature [39]. Where possible, multiple saliency methods should be employed. Additionally, we recommend that researchers use counterfactuals to confirm whether or not their findings are the same as previously reported using different methods, as all studies in the considered set of literature make this assumption without experimental confirmation [40]. Given the structure of neural networks and their ability to highlight spatially invariant patterns, there is no guarantee that a salient group of pixels in, for example, the amygdala, translates directly to volumetric reductions in that area previously reported by logistic regression models. Another alternative would be to examine the correlation between, for example, Freesurfer-derived tabular summary data for a particular region and the average saliency for the same region across all patients, which could allow researchers to confirm or refute their hypotheses. This could also mitigate the issues previously highlighted with interpreting 'importance' metrics by using a combination of explanatory methods.

4.5 Information leakage and model stacking

We observed a fraction of studies that may have been prone to information leakage through their data splitting procedures. As previously mentioned, this could be mitigated by ensuring slice-level data is derived after patient-level data splitting. Several studies also made use of model stacking, whereby the output of one trained model, whose objective is to discriminate between classes, is fed as the input to another trained model. This may have implications for predictive accuracy, leading to inflated model performance estimation, as well as posing issues for biomarker discovery. This may also lead to further model overfitting. We therefore recommend that researchers avoid model stacking where possible, especially considering that model interpretation is such an important aspect of predictive studies in this field. Model stacking and variable selection may complicate interpretative efforts further.

4.6 Peer review

The majority of studies in this set of literature underwent peer review either through their journal or conference submission process. This indicates that the issues in the methodologies and transparency of the presented research may have been missed by reviewers. We encourage conferences and journals to hold researchers accountable to the aforementioned recommendations by being critical of methodological details, requiring code transparency, questioning unsubstantiated claims, and expecting well-detailed interpretability.

4.7 Future perspectives and commentary

The findings from this systematic literature review highlight long-standing differences between deep learning and classical statistics. Deep learning has historically been concerned with minimising the loss of objective functions without making or testing formal assumptions of the data generating processes, or discovering the inferential dynamics of the considered system. This has led to numerous advances in image processing, with several state-of-the-art approaches developed to address tasks not suited to classical statistical modelling [19, 24]. In such cases where inferential dynamics are not the main focus of the model, neural networks have clear advantages in their depth and ability to consider non-linearities. However, as deep learning becomes more readily applied to medical imaging applications, with high-stake consequences for patients, the previously-mentioned dichotomy of prediction versus inference must be abandoned. It is no longer sufficient to have flexible predictive machines returning 'black-box' decisions with no indication as to what relationships have been captured or are being leveraged. In order to achieve the highest standard of care for patients, it is crucial to understand what input features deep learning models are basing their decisions on, in order to increase confidence in these approaches. Considering the widespread ramifications of diagnoses for brain disorders, and the aforementioned examples of deep learning-amplified modelling biases, it is essential to ensure deep learning models are making use of salient information; a focus on inference would also have clear benefits for biomarker discovery.

We posit that the successful application of deep learning to the diagnoses of brain disorders in clinical settings using structural neuroimaging data is contingent upon adherence to the principles of good modelling practices, interpretability, and transparency. We therefore encourage researchers carrying out experiments in the field, as well as readers and reviewers of published research, to carefully consider the recommendations outlined in this paper which are summarised in Table 3.

Key Recommendations	Benefits	Risk from non-adherence
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, which would be a worthwhile endeavour. This work also did not include considerations of studies making use of functional neuroimaging data, and it would be interesting to examine whether or not the same trends exist in research using a different data modality.

6 Conclusion

In summation, we conducted a systematic literature review of 55 studies carrying out CNN-based predictive modelling of brain disorders using structural brain imaging data and found issues with modelling practices, transparency, and interpretability. We strongly recommend that researchers place greater emphasis on these principles in their experiments for the sake of patients, and that in combination journal/conference editors be mindful of the importance of the outlined concepts. Careful consideration of these principles will inform a clinical framework that can effectively incorporate deep learning into diagnostic and prognostic systems, furthering our physiological understanding of these disorders and 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] Nicholas J Bray and Michael C O'Donovan. The genetics of neuropsychiatric disorders. *Brain and neuroscience advances*, 2:2398212818799271, 2018.

[6] Weichen Song, Wei Qian, Weidi Wang, Shunying Yu, and Guan Ning Lin. Mendelian randomization studies of brain mri yield insights into the pathogenesis of neuropsychiatric disorders. *BMC genomics*, 22(3):1–11, 2021.

[7] Amit Etkin. A reckoning and research agenda for neuroimaging in psychiatry. *American Journal of Psychiatry*, 176(7):507–511, 2019.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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`.

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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`.

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[42] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[43] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[44] 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.

[45] 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.

[46] 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.

[47] 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/`.

[48] 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.

[49] 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.

[50] 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`.

[51] 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`.

[52] 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.

[53] 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/.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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/.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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/.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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`.

[83] 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.

[84] 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.

[85] 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.

[86] 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/`.

[87] 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`.

[88] 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.

[89] 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.

[90] 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`.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

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

[98] 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.

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

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

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

[102] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International Symposium on Visual Computing*, pages 565–578. Springer, 2019.