Alzheimer's
Research & Therapy

RESEARCH

Open Access

Higher performance for women than men in MRI-based Alzheimer's disease detection

Malte Klingenberg[1,2], Didem Stark[1,2], Fabian Eitel[1,2], Céline Budding[3], Mohamad Habes[4], Kerstin Ritter[1,2*] for the Alzheimer's Disease Neuroimaging Initiative

Abstract

Introduction Although machine learning classifiers have been frequently used to detect Alzheimer's disease (AD) based on structural brain MRI data, potential bias with respect to sex and age has not yet been addressed. Here, we examine a state-of-the-art AD classifier for potential sex and age bias even in the case of balanced training data.

Methods Based on an age- and sex-balanced cohort of 432 subjects (306 healthy controls, 126 subjects with AD) extracted from the ADNI data base, we trained a convolutional neural network to detect AD in MRI brain scans and performed ten different random training-validation-test splits to increase robustness of the results. Classifier decisions for single subjects were explained using layer-wise relevance propagation.

Results The classifier performed significantly better for women (balanced accuracy 87.58 ± 1.14%) than for men (79.05 ± 1.27%). No significant differences were found in clinical AD scores, ruling out a disparity in disease severity as a cause for the performance difference. Analysis of the explanations revealed a larger variance in regional brain areas for male subjects compared to female subjects.

Discussion The identified sex differences cannot be attributed to an imbalanced training dataset and therefore point to the importance of examining and reporting classifier performance across population subgroups to increase transparency and algorithmic fairness. Collecting more data especially among underrepresented subgroups and balancing the dataset are important but do not always guarantee a fair outcome.

Keywords Alzheimer's disease, Deep learning, MRI, Sex, Bias

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

*Correspondence:
Kerstin Ritter
kerstin.ritter@charite.de
[1] Charité - Universitätsmedizin Berlin (corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health), Department of Psychiatry and Neurosciences, Berlin, Germany
[2] Bernstein Center for Computational Neuroscience, Berlin, Germany
[3] Eindhoven University of Technology, Eindhoven, Netherlands
[4] Neuroimage Analytics Laboratory and Biggs Institute Neuroimaging Core, Glenn Biggs Institute for Neurodegenerative Disorders, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

In recent years, a number of machine learning (ML) algorithms have been proposed to diagnose Alzheimer's disease based on structural magnetic resonance imaging (MRI) [1, 2], and classification accuracies on par or even exceeding that of human experts have been reported [3]. One class of ML algorithms, namely convolutional neural networks (CNNs), have been shown to be very powerful for this task because they are capable of operating directly on raw or minimally processed MRI data and do not need a previous feature extraction [2, 4].

With the increasing use of ML methods in both medical and other decision-making systems, algorithmic bias, such as sex, gender, or racial bias, has come into the focus of current research. One possible cause for these biases is an underlying imbalance in the dataset used for training the ML algorithms. Medical research has, for a long time, been conducted primarily on male patients [5]. Additionally, even if female subjects are included, the results are often not analysed separately by sex or gender [6, 7]. This has, for example, led to underdiagnosis of heart attacks in women, as their symptoms can differ from those of male patients [8]. When unbalanced datasets are used to train ML algorithms, this can result in biased classifiers, with a consistent decrease in performance for population groups underrepresented in the training data [9, 10]. Several methods have been developed to understand and mitigate these biases. One simple and straightforward method is to ensure that the training datasets are balanced and representative across all relevant population subgroups. This can lead to a performance increase for underrepresented groups, while not necessarily negatively affecting performance for the overrepresented group [9, 11]. However, using a balanced dataset alone has been shown to not always be sufficient to prevent biased classification results as shown in a chest X-ray classification task recently [12]. Here, the differences in true-positive rate (TPR) across different subgroups are not correlated with the subgroups' proportional disease membership and having the same portion of images within a label might not be enough to mitigate the resulting diagnostic bias [12].

From a clinical perspective, sex has an important impact on the presentation of AD. Women have a higher lifetime risk of developing AD, and also show faster ageing-related cognitive deterioration than men [13]. The atrophy rate both of the hippocampus and the overall brain matter is also higher for women than for men [14, 15]. Additionally, for women, pathological changes in the brain are more likely to result in clinical AD, with men being more resilient to the pathophysiological processes of AD [16, 17]. However, sex differences in the performance of ML-based classification of AD have so far not systematically been investigated.

Explainable artificial intelligence (AI) has become an important topic in recent years as more ML models are being implemented for medical applications [18]. For image data in combination with CNNs, the most promising approach refers to so-called heatmap or attribution methods that exploit the gradient or the architecture of the model to compute pixel- or voxel-wise explanations [19]. Notably, for each input image, a heatmap is generated that indicates the importance or relevance of each pixel or voxel for the final classification decision based on the respective model. In the medical context, it means that these methods provide a visual representation of the area that the model utilises for each individual patient, but do not provide any information on what is used within this area [4, 20].

In this study, we examine a state-of-the-art CNN classifier for MRI-based AD detection with respect to sex differences. To reduce the effects of possible biases in the training dataset, we balanced the training set to contain an equal number of women and men and used undersampling to equalise the female and male age distributions. We hypothesise that—in the case of balanced training data—there is no sex difference in detecting AD. To explain the classifier decisions, we use the Layer-Wise Relevance Propagation algorithm (LRP) [21], which produces an individual heatmap for each input image, showing the relevance of each voxel for the final classifier decision, and has been shown to give reasonable explanations in the context of AD [4, 22]. In particular, a significant correlation between local LRP relevance and atrophy in the hippocampus has been reported [4].

Methods

Data set

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

Inclusion criteria

For this analysis, we included subjects from all ADNI study phases which were, at the time of their baseline visit, either healthy controls (HC) or diagnosed with AD

Table 1 Population characteristics. This table gives information about the demographics and clinical measures of one of the ten datasets created from the base study population. As different subjects are selected for each split, the values for other splits can vary slightly. All values are given as mean ± SD

	Female		**Male**	
	HC	**AD**	**HC**	**AD**
#subjects	153	63	153	63
#images	418	175	479	182
Age (years)	73.8 ± 6.3	75.7 ± 7.3	73.8 ± 6.4	75.4 ± 6.9
CDR-SB	0.01 ± 0.07	4.63 ± 1.93	0.03 ± 0.15	4.68 ± 1.57
ADAS13	9.14 ± 4.49	30.58 ± 7.65	10.15 ± 4.53	31.56 ± 7.59
MMSE	29.17 ± 1.05	22.88 ± 2.54	28.99 ± 1.23	22.77 ± 2.04

using the official diagnoses provided by ADNI[1]. Subjects which were labelled as HC at their first visit, but at a later visit were diagnosed with either MCI or AD (or vice versa) were excluded from the analysis. This resulted in a population size of 573 subjects (406 HC, 167 AD). For each subject, up to three MRI scans from different time points were included in order to increase the sample size. To avoid data leakage between multiple scans originating from the same subject, we performed the splitting for the CNN training on the subject level and not on the image level (see below).

In the remaining population, younger female subjects are overrepresented [23]. To ensure that there is no significant difference in the age distributions of female and male subjects, we used undersampling based on subject age, diagnosis, and sex. To this end, we divided the population into bins containing subjects of a specific age range (in 5-year increments) and diagnosis (HC or AD). From each bin, we then randomly dropped subjects until the number of female and male subjects in the bin was equal. The resulting population contained 432 subjects (306 HC, 126 AD) with no significant differences in their age distributions (two-sample t-test: $p > 0.75$). Table 1 gives full information about the size, age distribution, and clinical measures of the resulting dataset. The precise values vary slightly depending on the specific subjects removed during the undersampling step.

Splitting

We then used a stratified split based on subject age range, sex, and diagnosis to divide the population into a training set (216 HC, 90 AD) and validation and test sets (45 HC, 18 AD each). Splitting on the subject level ensures independence between training and test data, as splitting on the image level may lead to data leakage and unreliable results [2].

For datasets of this size, the results can vary significantly depending on the specific dataset split [23]. We therefore repeated the undersampling and splitting process ten times with different random seeds and trained and evaluated the classifier on all dataset splits. Depending on the number of scans taken of the subjects remaining after the undersampling, the training set size ranged from 758 to 834 images. For the test sets, we used only the baseline scan of each subject, to prevent the presence of multiple scans of the same subject distorting the results.

Image preprocessing

For our analysis, we downloaded T1-weighted structural MRI scans of all selected subjects. The scans were acquired at multiple imaging sites at a magnetic field strength of 3 T (for scanner and sequence parameters, we refer the reader to the ADNI imaging protocols[2]). The images had already been preprocessed with gradient non-linearity correction (gradwarping) and intensity inhomogeneity correction and were scaled for gradient drift using the phantom data. We did not further harmonise the data.

We recently showed that, for a CNN trained on a relatively small dataset of MRI brain scans, using a non-linear registration method gives the best results compared to unregistered or linearly registered images [23]. Accordingly, we used the non-linear SyN algorithm [24] as implemented by Advanced Normalization Tools (ANTs)[3] to register all scans to the 1mm T1 version of the MNI-ICBM152 reference brain. We chose SyN over other non-linear registration algorithms because of its consistently good performance reported by Klein et al. [25]. After registration, we used the FSL Brain Extraction Tool (fsl-bet) [26, 27] to remove the skull and soft tissue from the images.

Network architecture and training

For this analysis, we used the convolutional neural network architecture proposed by Böhle et al. [4]. This is a standard CNN with four convolutional layers, each comprising 8, 16, 32, and 64 filters respectively, with a filter size of 3×3×3. Each of these layers is followed by batch normalisation and max pooling with window size 2, 3, 2, and 3. The convolutional layers are followed by two fully connected layers of sizes 128 and 2, with dropout

[1] https://clinicaltrials.gov/ct2/show/NCT00106899

[2] https://adni.loni.usc.edu/methods/documents/mri-protocols/

[3] https://stnava.github.io/ANTs/

($p = 0.4$) being applied before both of these layers. The 2-unit layer uses a softmax function and provides the model output, with the two units giving the class scores for HC and AD.

The network was trained using the Adam optimiser [28] and cross-entropy loss with a learning rate and weight decay of 10^{-4}. We used a batch size of 16 images, which was limited by the available GPU memory. During training, the data was augmented by flipping the images across the sagittal plane and translating along the sagittal axis by up to two voxels in either direction, with both methods being standard data augmentation methods for deep learning methods performing medical image analysis [29, 30].

Training was stopped once the balanced accuracy achieved by the model on the validation set did not improve over eight epochs, after which the model state with the highest validation accuracy was evaluated on the test set. To achieve robust results and reduce the impact of lucky or unlucky data splits, we repeated the training process five times for each of the ten dataset splits, giving a total of 50 different models.

Model evaluation and comparison to clinical measures

To evaluate whether there are statistically significant differences in the classifier performance for women and men, we used independent samples *t*-tests on the balanced accuracy, sensitivity, and specificity values. We also calculated the receiver operating characteristic (ROC) curves, which show the relationship between false-positive rate and true-positive rate, for women and men separately. This will allow to determine whether choosing different classification thresholds for women and men would help to achieve equal performance on the two subgroups.

We also examined the distributions of three different clinical measures of disease severity among men and women: the Clinical Dementia Rating (CDR) sum of boxes score [31], the Alzheimer's Disease Assessment Scale (ADAS13) [32], and the Mini-Mental State Examination (MMSE) [33]. These measures have been shown to be correlated with both the brain atrophy rate and the ventricular enlargement rate [34] and can therefore provide insight into the degree of AD evidence present in the brain scans of our population. If there were a significant difference in the distributions of these measures between women and men, this could also explain a difference in classifier performance, as the differing disease severity could manifest as different degrees of structural AD evidence. To achieve robust results for this analysis, we used data from women and men, but only those subjects which appeared in at least two of the ten dataset splits.

Layer-wise relevance propagation

For explaining the classifier's decision, we used the layer-wise relevance propagation (LRP) algorithm by Bach et al. [21, 35], which produces heatmaps showing the relevance of each individual input voxel for the final classification.

To achieve this, LRP considers how the activation of each node in the model contributes to the final output class score layer by layer. The initial relevance value for a specific class is simply the activation of the corresponding output node. This relevance is then distributed to all nodes in the preceding layer which contributed to the activation of the output node. The distribution follows the conservation rule

$$\sum_i R_{i\leftarrow j}^{(l,l+1)} = R_j^{(l+1)} \tag{1}$$

where $R_j^{(l+1)}$ is the relevance of node j in layer $l+1$, and $R_{i\leftarrow j}^{(l,l+1)}$ is the share of relevance that node i in layer l receives from node j. The total relevance of a node in a specific layer is then the sum of the relevance it acquires from all nodes in the following layer. This ensures that the total amount of relevance in the input layer is precisely the output class score.

There are several different ways in which the relevance can be distributed through the model. For our analysis, we chose the β-rule [36], given by

$$R_{i\leftarrow j}^{(l,l+1)} = \left((1+\beta)\frac{z_{ij}^{+}}{z_j^{+}} - \beta\frac{z_{ij}^{-}}{z_j^{-}} \right) R_j^{(l+1)} \tag{2}$$

which allows for separate treatment of the positive and inhibitory contributions $z_{ij}^{+/-}$ by changing the parameter β. A value of $\beta = 0$ produces heatmaps showing only positive contributions (meaning evidence for the presence of AD), whereas choosing $\beta > 0$ includes inhibitory effects produced by evidence against AD. For a full description of the LRP algorithm and the β-rule, we refer the reader to Bach et al. [21] and Binder et al. [36].

When using LRP to visualise the decisions of an AD-detecting CNN, the resulting heatmaps are relatively robust to the chosen β-value, with the only change being increasing sparseness for higher values [4]. Additionally, ignoring negative contributions might give more informative results, as AD can, especially in its early stages, affect the brain in a highly localised manner. Surrounding healthy tissue could therefore mask the positive contributions of small areas of evidence for AD. We have accordingly limited our analysis to using a value of $\beta = 0$, as this leads to heatmaps showing all positive contributions, regardless of possible surrounding negative evidence, giving the final distribution rule

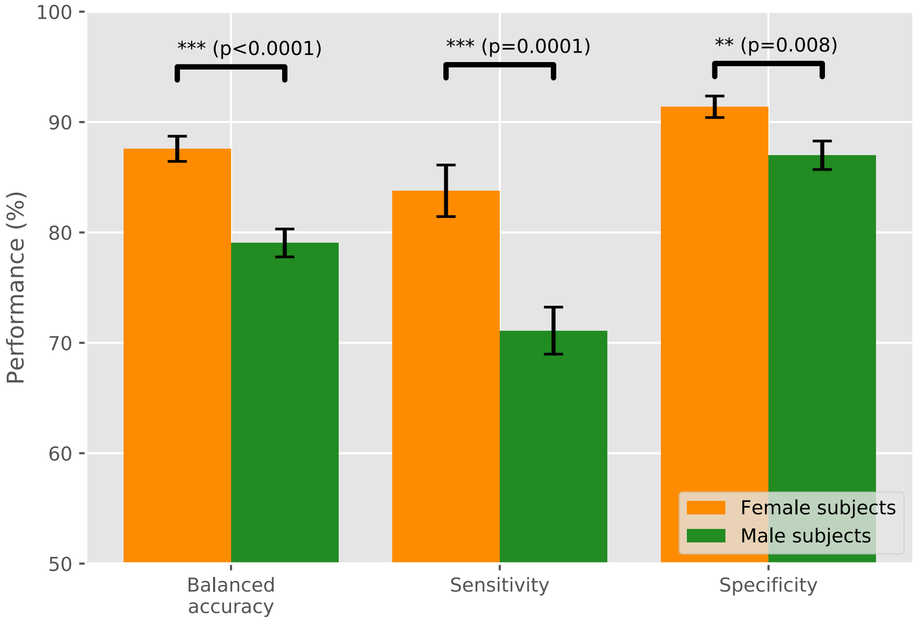

Fig. 1 Classifier performance. The balanced accuracy, sensitivity, and specificity of the classifier for women and men averaged over all runs for all splits. The error bars show the standard error of the mean

$$R_{i\leftarrow j}^{(l,l+1)} = \frac{z_{ij}^{+}}{z_j^{+}} R_j^{(l+1)} \tag{3}$$

Region-wise relevance analysis

To enable a quantitative analysis of the resulting heatmaps, we used the Neuromorphometrics Scalable Brain Atlas [37] to determine the amount of relevance present in the different brain regions. Simply summing the relevances of all voxels of a brain region naturally gives a measure that is strongly correlated to the region size [4]. We therefore normalised the relevance sum of each brain region by its number of voxels to determine its relevance density. This is a more informative measure, as a low amount of relevance spread out over a large area might be simply due to statistical noise, while a strongly localised cluster of voxels with high relevance could indicate the presence of structural changes as evidence for AD.

For the analysis, we selected a subset of brain regions which have been shown to have high susceptibility to structural changes due to AD. Among others, we included areas of the limbic system (such as the hippocampus, entorhinal area, and amygdala), the ventricles, and the cingulate gyrus. For comparison, we also included the motor cortex, which is among the last areas to be affected by AD [38].

Results

Classifier performance

The balanced accuracy, sensitivity, and specificity achieved by the classifier are shown in Fig. 1 for women and men separately. The given results are averaged over all 50 runs (five runs for each of the ten different dataset splits), with the error bars showing the standard error of the mean. A difference in performance for women and men is clearly visible, with a balanced accuracy of $87.58 \pm 1.14\%$ for women and $79.05 \pm 1.27\%$ for men. While the performance for women is better overall, the results seem to also be more robust, as determined by the lower standard error. A similar pattern holds for the sensitivity ($83.77 \pm 2.34\%$ for women, $71.10 \pm 2.13\%$ for men) and specificity ($91.38 \pm 0.97\%$ for women, $86.99 \pm 1.29\%$ for men). All differences are statistically significant, with *t*-test *p*-values of $p = 2.4 \cdot 10^{-6}$ for the balanced accuracy, $p = 1.2 \cdot 10^{-4}$ for the sensitivity, and $p = 7.8 \cdot 10^{-3}$ for the specificity.

Figure 2 shows the ROC curves separately for women and men, averaged over all trained models. Again, a clear difference is visible, with an area under curve of 0.950 ± 0.007 for women and 0.862 ± 0.014 for men.

A post hoc analysis revealed that when we train a model on a sex-balanced dataset of half the size to generate a comparison baseline for sex-specific models, the classifier suffered a large drop in accuracy. Thus, due to the limited sample size, a subgroup analysis is not feasible here.

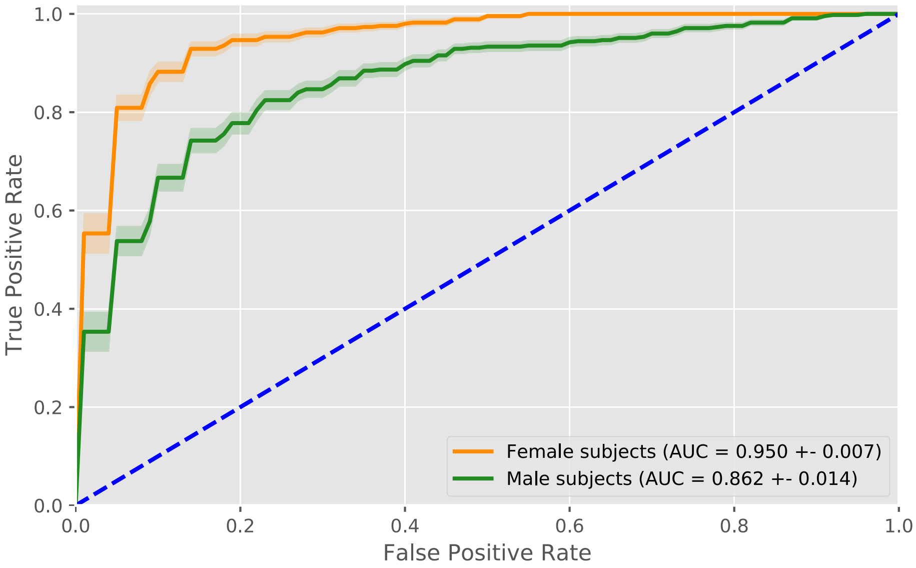

Fig. 2 Receiver operating characteristic curve. The average ROC curve of the classifier when separately evaluated on women and men. The ROC curve was averaged over all runs for all splits, with the shaded area showing the standard error of the mean. The area under curve (AUC) is given in the legend

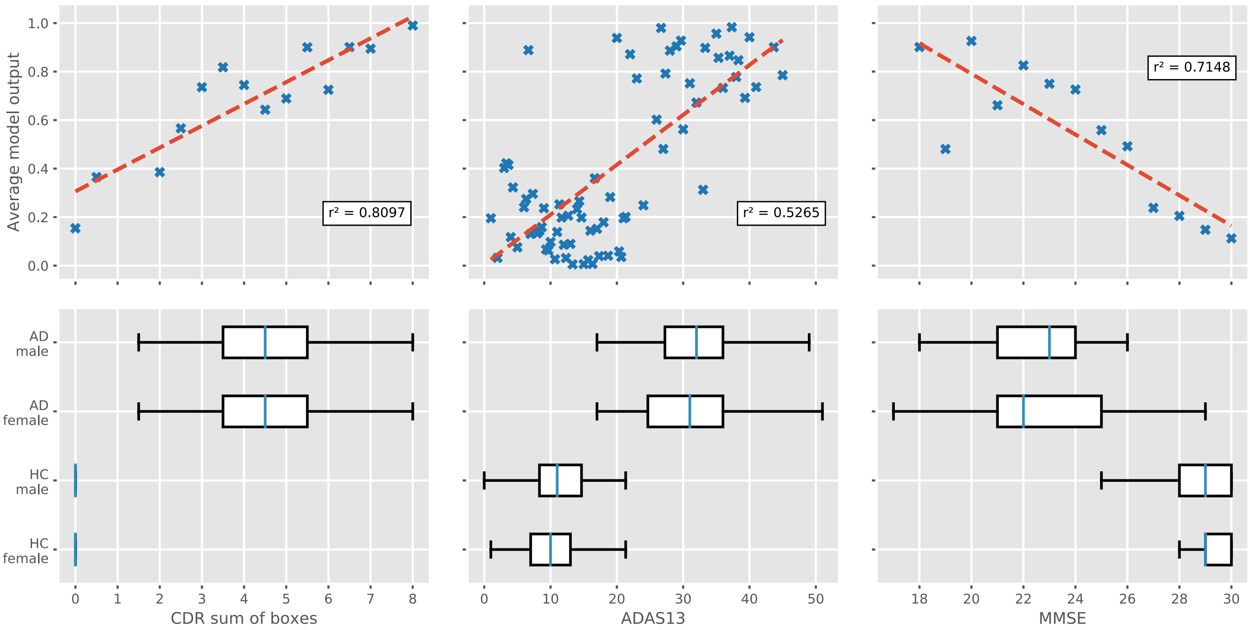

Fig. 3 Clinical measures. Shown in the top row of plots is the relationship between the average model output and each of the three clinical measures (CDR sum of boxes, ADAS-Cog-13 and MMSE). While calculating the average output, only subjects appearing in at least two splits were taken into account. Overlaid in red is a linear regression, with the correlation coefficient also given. The plots on the bottom row show the distribution of the three clinical measures in the dataset. Note that, because this includes all subjects, the boxplot whiskers can extend past the values visible in the top plots

Clinical measures

The distributions of the clinical AD measures in our dataset are presented in Fig. 3. The top row of plots show the relationship between the three clinical measures and the raw output of the classifier, i.e. after applying the softmax function. A clear correlation can be seen for all three measures, with scores indicating more severe disease leading to a higher model output.

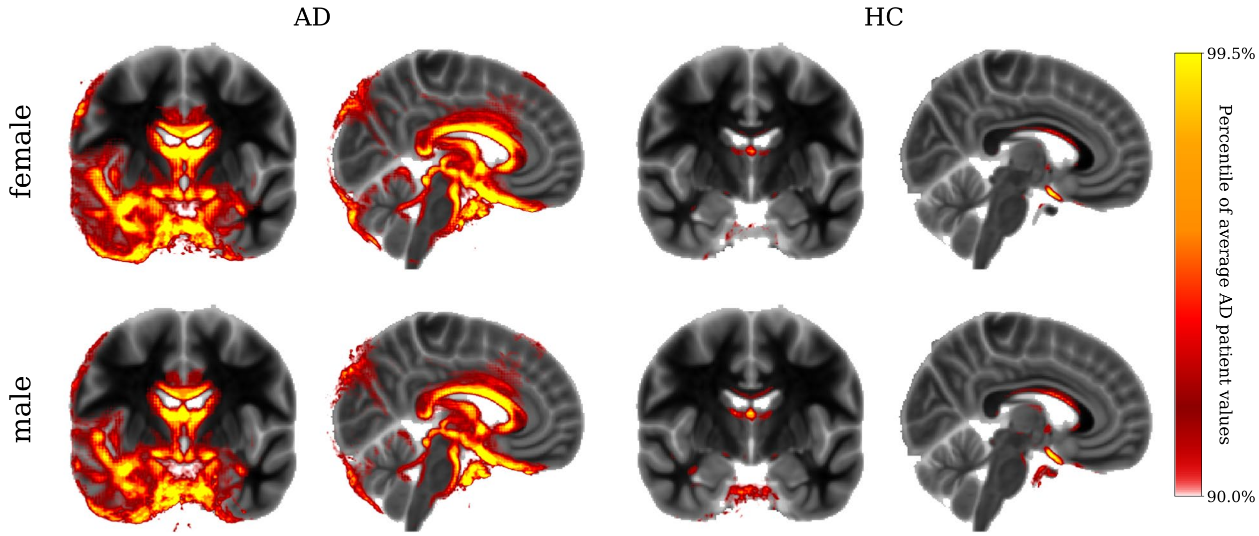

Fig. 4 Average relevance heatmaps. The average relevance heatmaps across all subjects and all classifiers are shown separately for women (top row) and men (bottom row), as well as Alzheimer's patients (left column) and healthy subjects (right column). The colour bar was chosen according to the relevance values of the average AD subject, with only the top 10% of values being shown to highlight the most relevant areas. For reference, the heatmaps are shown over the MNI-ICBM152 reference brain we used for registering the input images

The bottom row of plots gives the distribution of the clinical measures among women and men after balancing the dataset for age and sex. There are no significant differences between the CDR sum of boxes and MMSE scores of women and men for any of the ten dataset splits. The same is true for the ADAS13 scores of healthy subjects. For AD subjects, we found a significant difference in six out of the ten splits, but these differences were weak, with a p-score of $0.05 > p > 0.01$ in all of those cases but one ($p = 0.006$).

Visualisation

In Fig. 4, we show the average heatmaps of all classifier decisions, separately for healthy and affected women and men. The heatmaps are overlaid over the MNI-ICBM152 template, only showing the top 10% of relevance values compared to the average AD heatmap. A coronal slice at $y = 120$ shows areas of the frontal and temporal lobes, including the hippocampus and parahippocampal gyrus, and a sagittal slice at $x = 85$ gives a view of, among others, the lateral ventricles, brain stem, and cerebellum.

For AD subjects, large amounts of relevance can be seen in and around the hippocampus and other areas of the temporal lobe. Significant relevance is also present around the lateral ventricle. There is only little relevance present in healthy subjects. No clear difference is visible between the average heatmaps for women and men.

To give a sense for the inter-patient variability of the heatmaps, we show the results for four individual subjects in Fig. 5. We selected four AD subjects, two women (68 and 88 years) and two men (67 and 87 years), and a single model which correctly classified all four subjects with high confidence (AD class score > 0.97). It can be seen that the heatmaps are highly individual to each subject, although the general pattern of strong relevance in the temporal lobe and around the lateral ventricles still holds. The classifier also places significant amounts of relevance on individual cortices, which is well visible in all subjects but especially pronounced in the younger male brain, which shows severe atrophy. The younger male brain also shows a strong enlargement of the lateral ventricles. The classifier correctly identifies this, as can be seen by the relevance accurately placed on the border of the enlarged ventricle.

Relevance analysis

Figure 6 shows the relevance attributed by the LRP algorithm to several brain areas for female and male AD patients.

The highest relevance densities are found in areas that are part of the limbic system, such as the entorhinal area, the hippocampus, and the amygdala. The motor cortex, which is only affected in later stages of AD, has one of the lowest relevance densities among the examined regions.

The results for female and male patients are similar, with a close match in the order of area relevance density. However, the relevance density for women is consistently higher than for men.

We also show the results for two individual subjects, specifically the young female and male subjects from

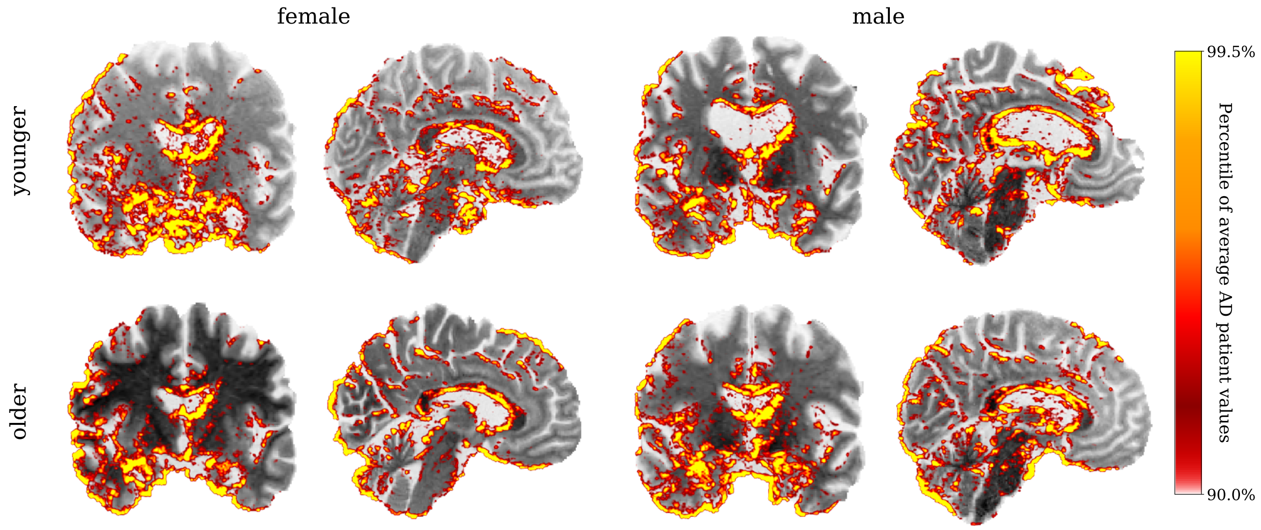

Fig. 5 Individual relevance heatmaps. The relevance heatmaps for four individual AD patients are shown, each overlaid over the corresponding brain scan. All scans were classified by the same model, to enable a comparison of inter-subject differences. We selected two female (68 and 88 years) and two male subjects (67 and 87 years), which were correctly diagnosed by the classifier with high confidence (AD class score > 0.97). The colour bar was chosen as in Fig. 4, based on the relevance values of the average AD subject heatmap

Fig. 5. This illustrates that the distribution of relevance over different brain areas can vary significantly between subjects. For the female subject, the relevance density is at the top of its range for hippocampus, parahippocampal gyrus, and amygdala, matching the visual impression of the heatmap.

The male patient has values around or slightly above average for most areas, except for the cingulate gyrus, which is among the highest values we found.

The bottom plot shows the coefficient of variation of the relevance density, i.e. its standard deviation divided by its mean. This measures the inter-patient variability of the relevance of the different brain areas. The coefficient is consistently larger for men by 10–30%; in other words, the distribution of relevance is more uniform for women than for men.

Discussion

Despite balancing our dataset for subject sex and age, we found a statistically significant difference in classifier performance for women and men, with women having a higher balanced accuracy, sensitivity, and specificity. The classifier also achieves a higher area under the ROC curve for women, showing that the difference is not just due to a suboptimal threshold for men. There is no threshold value that would lead to equal performance for women and men, as their ROC curves do not intersect. Additionally, choosing individual thresholds for each population subgroup may not be a feasible solution in other, more complicated datasets with multiple intersectional subgroups and small subpopulations [10].

While an imbalance in the training data has been identified to be a possible cause of sex bias [9], other research has shown that the inverse is not true, with no significant correlation between classifier performance disparity and data imbalance ratio [12]. Our findings mirror this, as even a perfectly balanced dataset containing the same number of women and men does not lead to equal performance.

We have also examined a possible imbalance between women and men of disease severity as measured by several cognitive scoring systems. However, after equalising the age distributions, we did not find significant differences in disease severity for CDR and MMSE. The observed differences in ADAS13 scores for some splits indicated more severe cases in men, which would seem to indicate an easier diagnosis rather than the observed worse performance. While we have therefore excluded

(See figure on next page.)

Fig. 6 Relevance by area for AD subjects. The top plot shows the size-normalised relevance for selected brain areas for female and male AD subjects. The mean values are displayed as dots, with the shaded areas showing the relevance density distribution across all AD subjects. The dotted and dashed lines show the values for two individual subjects, namely the young female (Patient 1) and young male (Patient 2) subjects for which the heatmaps are shown in Fig. 5. The bottom plot gives the coefficient of variation, i.e. the standard deviation divided by the mean of the relevance density for the same brain areas

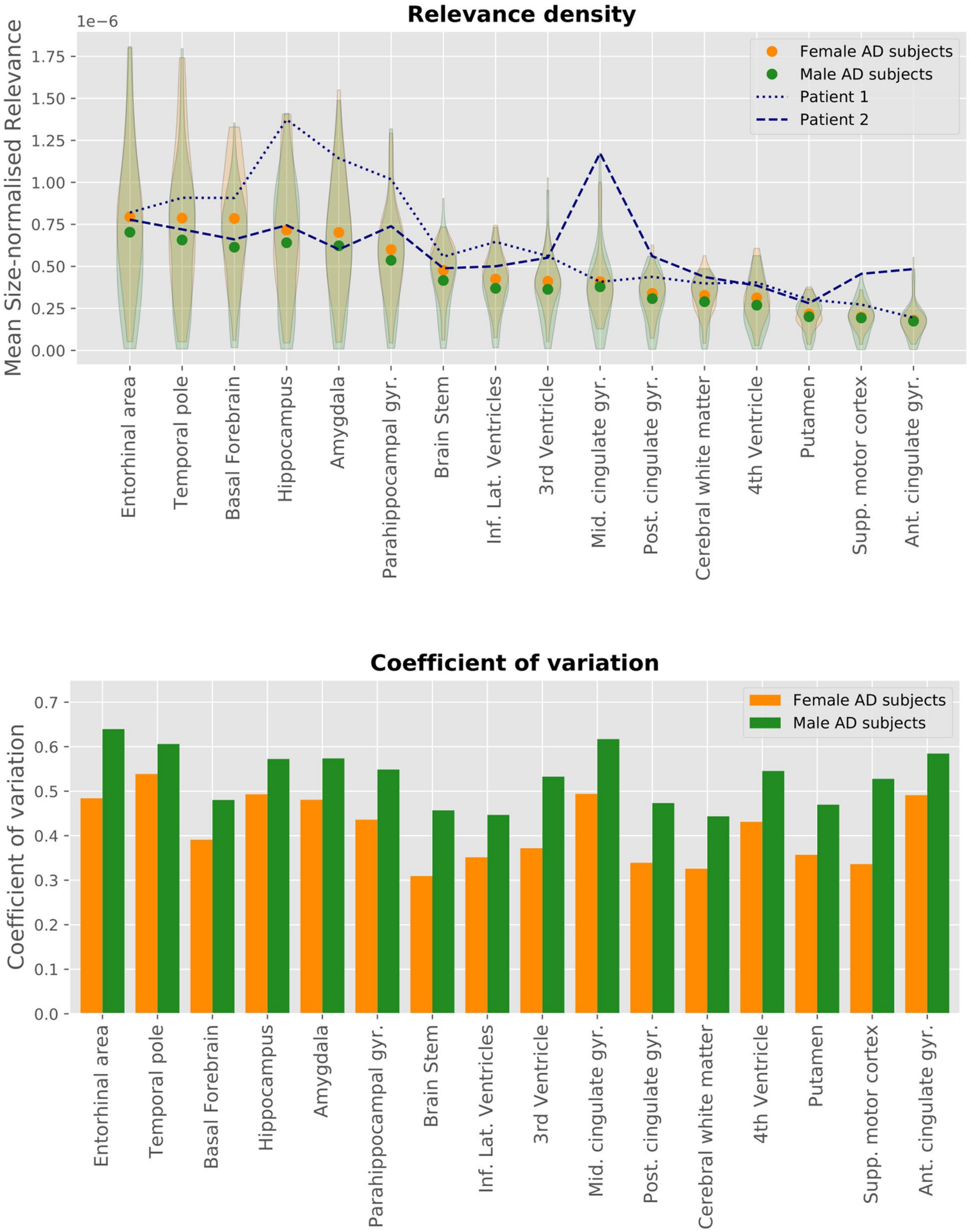

Fig. 6 (See legend on previous page.)

several possible sources of bias, we cannot rule out the presence of other confounding variables. Future research should investigate this further, looking for example at biomarkers of disease progression, such as the levels of amyloid beta and tau proteins. For example, Gamberger et al. have shown that when clustering the ADNI population based on biological and clinical descriptors, female AD patients form one coherent group, but male AD patients can be divided into two distinct clusters [39].

The observed performance difference may also be rooted in the different ways in which AD affects women and men. Female AD patients show both faster rates of cognitive decline and larger atrophy rates in limbic system areas specifically as well as overall brain matter when compared to male AD patients of the same age [14, 15, 40]. This implies that a perfect age-matching between women and men in the training data might actually have a detrimental effect in terms of classifier bias, as consistently more severe cases of AD in women would make them easier to diagnose. However, this hypothesis is contradicted by our analysis of the clinical measures of our study population, which showed no significant difference in disease severity between women and men.

On the other hand, men have been shown to have a higher resilience to the pathophysiological processes of AD compared to women [16, 17]. This means, that in cognitively intact older men, the presence of biomarkers for AD such as amyloid load has a lower influence on the clinical presentation than for women. As a consequence, the healthy male subjects in our study population might have more structural AD evidence than healthy female subjects, making the distinction between male healthy and AD subjects harder for the classifier. Because the increased amount of AD biomarkers does not lead to more severe cognitive symptoms in this case, this effect would not be visible in our analysis of clinical measures. Additionally, previous research has shown that men are at higher risk for cerebrovascular disease [41]. This could lead to higher prevalence in vascular cognitive impairment in men [42]. Even though the ADNI study inclusion criterion was AD-related dementia, dementia of vascular origin could be more present in the male cohort and might explain greater ventricle enlargement in men compared to women in our findings.

Regarding the heatmap results, our findings are in line with previous studies showing that LRP heatmaps are noisy and asymmetric [4, 22]. Moreover, a recent study showed that CNNs are susceptible to spatial bias, mainly due to architectural choices such as padding, resulting in activation blind spots in feature maps [43]. The stochastic nature of model training is expected to cause some randomness in the results; hence, the lack of symmetry is not surprising.

The heatmaps show that the classifier focuses on areas where structural changes due to AD are expected. Significant relevance being seen in the lateral ventricles in AD patients is reasonable as ventricular enlargement is a reliable measure of AD progression [44]. Additionally, the individual heatmaps agree with research showing that measures such as sulcal width and cortical thickness around sulci are good neuroanatomical markers of AD [45, 46]. Changes in periventricular white matter were shown to correlate with elevated cerebral amyloid [47] and cognitive decline [48]. Our findings that the relevance heatmaps focus on periventricular white matter could be capturing those clinical markers of AD. Our results showing relevance in the brainstem regions are in line with previous research showing brainstem atrophy in the early stages of AD [49, 50]. Overall, a visual inspection of the relevance heatmaps reveals no clear reason for the performance difference between women and men.The variability across subjects of different ages and between specific classifier models seems to eclipse any possible systematic difference between female and male heatmaps.

In the quantitative relevance analysis, the overall results are as expected from clinical research and match the visual impression of the presented heatmaps. The consistently higher relevance density for women compared to men indicates that the network was generally more confident when classifying women, as a higher model output results in a larger amount of relevance being distributed backwards through the network and thus a larger total relevance in the heatmap.

We found a larger variation in how the relevance is distributed among the brain areas for male AD subjects compared to female AD subjects. Additionally, the standard error of the performance metrics as given in Figs. 1 and 2 is slightly larger for men than for women, indicating a larger variation between different runs for male than for female subjects. This seems to reflect studies showing that there are general structural differences in the brains of women and men. These differences cover overall brain size, cortical thickness, and grey matter volume in specific brain areas, with men generally having a greater variance in these structural measures [51, 52]. However, other studies have questioned this, showing that, when corrected for brain size, sex accounts only for a small percentage of the structural variance, and that female and male brains are structurally very similar [53–55]. Men also have been shown to have more heterogeneous patterns of AD presentation compared to women. For instance, recent research has shown that the hippocampal sparing subtype of AD was more prevalent in men, which was associated with more white matter lesions [56].

We would like to point out the following limitations. First, we limited our analysis to only one specific network architecture, and thus, it is not clear to what extent these results will generalise to other classifiers, with either small changes such as an increase in the number of layers, or larger alterations to the entire network architecture. However, we used here a standard CNN that has been shown to be useful for AD classification before [4], and sex differences in classification accuracy have not yet been investigated.

Second, our dataset was quite small in terms of typical deep learning applications, with the training set size of around 800 images being limited by the available neuroimaging studies. Repeating our analyses on a larger dataset might alleviate the larger variance in male subjects. However, the ADNI study is currently the largest available dataset of brain MRI scans for AD, and the effect of larger male variance is already reduced due to our use of multiple dataset splits and several training runs per split. Given that deep learning analyses are very sensitive to the amount of training data, it was also not feasible to train sex-specific models (and models with different ratios of women/men) to additionally explore the influence of training set biases onto accuracy. Future studies might address this point using larger (not yet available) datasets.

Third, the preprocessing pipeline including skull stripping as well as the chosen brain atlas for quantitative analyses could be better adapted to the sample at hand by using age- and disease-specific templates [57, 58]. For instance, the high relevance density in the cingulate gyrus of the male patient from Fig. 5 is likely to be misattributed. As this patient has enlarged lateral ventricles, the boundaries of these ventricles do not match the brain atlas and instead extend into surrounding atlas areas. While the heatmap shows that the network correctly identifies the ventricle boundaries, the area-wise relevance analysis is based on the brain atlas and therefore can not take patient specifics into regard. However, disease-specific templates have the disadvantage to introduce prior information into the classification task, and thus might reduce the clinical significance of results. We therefore decided to keep the processing pipeline as general as possible.

And lastly, we did not employ an external validation dataset, instead generating both the training and validation datasets from the same study population (where we focused on clear AD and HC cases). This is common practice, as only few studies are testing the generalisability of classification models to external validation data, with results varying from only minor differences to comparatively large performance decreases [2, 59]. However, if the studies that collect the validation data adhere to the same inclusion criteria, the classifier performance would be similar to that on the initial dataset [2]. While independent and external validation of any classification and prediction models used in a healthcare setting is important before they are applied in clinical decision-making, the goal of our study was to point out performance differences between population subgroups. We believe that for this purpose, the cross-validation that we used by creating several different dataset splits was an appropriate choice. Yet, future studies should investigate how our findings apply across independent datasets, and with respect to changes in population (including for example MCI patients).

Conclusion

In this study, we trained a CNN to detect AD on 3D MRI brain scans. Despite carefully balancing the training data for subject sex and age, we found that the classifier performs significantly better on women than on men. The difference was neither explainable by a suboptimal cutoff point, nor a difference in disease severity between women and men, as measured by cognitive assessments. We found some evidence indicating a higher variability among men, suggesting that controlling for subject sex and age might not be enough to ensure a truly balanced dataset. Even when accounting for other confounding variables, differing clinical manifestations of diseases for different population subgroups may make equal performance for all subgroups an unreachable goal. Collecting more data across subgroups and having a balanced dataset are important measures towards fairness of the algorithms; however, those measures are not always enough to provide a fair outcome. Therefore, when using ML methods in medical applications, care should be taken to evaluate and report bias in the resulting classifier. This would strengthen the transparency and fairness of the chosen methods and increase their chance of being adopted as diagnostic tools.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (https://www.fnih.org). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Authors' contributions

MK, DS, FE, CB, and KR designed the study. The Alzheimer's Disease Neuroimaging Initiative (ADNI) provided all data. MK performed the analyses under the supervision of DS, FE, CB, and KR. MK, DS, MH, and KR wrote and revised the manuscript. The authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. We acknowledge the support from the German Research Foundation (DFG, 389563835; 402170461-TRR 265; 414984028-CRC 1404; 442075332-RU 5187; 459422098-RU 5363), the Brain & Behavior Research Foundation (NARSAD grant) and the Manfred and Ursula-Müller Stiftung. This study was supported in part by the National Institute of Health (NIH) grant numbers 1R01AG080821, P30AG066546, 1U24AG074855, and the San Antonio Medical Foundation grant SAMF-1000003860.

Availability of data and materials

The ADNI data base is public for researchers and can be downloaded upon request at https://adni.loni.usc.edu (see also the acknowledgements). The subject and image IDs for all used dataset splits and the code are available at https://github.com/malteklingenberg/AD-sex-bias.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Received: 15 August 2022 Accepted: 3 April 2023
Published online: 20 April 2023

References

1. Payan A, Montana G. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. arXiv preprint. 2015. ArXiv:1502.02506.
2. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, et al. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Med Image Anal. 2020;63:101694.
3. Klöppel S, Stonnington CM, Barnes J, Chen F, Chu C, Good CD, et al. Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. Brain. 2008;131(11):2969–74.
4. Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front Aging Neurosci. 2019;11:194.
5. Nowogrodzki A. Inequality in medicine. Nature. 2017;550(7674):S18–9.
6. Howard LM, Ehrlich AM, Gamlen F, Oram S. Gender-neutral mental health research is sex and gender biased. Lancet Psychiatr. 2017;4(1):9–11.
7. Mansukhani NA, Yoon DY, Teter KA, Stubbs VC, Helenowski IB, Woodruff TK, et al. Determining if sex bias exists in human surgical clinical research. JAMA Surg. 2016;151(11):1022–30.
8. Mosca L, Banka CL, Benjamin EJ, Berra K, Bushnell C, Dolor RJ, et al. Evidence-based guidelines for cardiovascular disease prevention in women: 2007 update. Circulation. 2007;115(11):1481–501.
9. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci. 2020;117(23):12592–4.
10. Seyyed-Kalantari L, Zhang H, McDermott M, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176–82.
11. Yu Z, Chakraborty J, Menzies T. Fair balance: mitigating machine learning bias against multiple sensitive attributes with data balancing. arXiv preprint. 2021. ArXiv:2107.08310.
12. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In: Pac Symp Biocomput 2021;26:232-243; 2020. p. 232–243.
13. Gao S, Hendrie HC, Hall KS, Hui S. The relationships between age, sex, and the incidence of dementia and Alzheimer disease: a meta-analysis. Arch Gen Psychiatr. 1998;55(9):809–15.
14. Hua X, Hibar DP, Lee S, Toga AW, Jack CR Jr, Weiner MW, et al. Sex and age differences in atrophic rates: an ADNI study with n=1368 MRI scans. Neurobiol Aging. 2010;31(8):1463–80.
15. Koran MEI, Wagener M, Hohman TJ, et al. Sex differences in the association between AD biomarkers and cognitive decline. Brain Imaging Behav. 2017;11(1):205–13.
16. Barnes LL, Wilson RS, Bienias JL, Schneider JA, Evans DA, Bennett DA. Sex differences in the clinical manifestations of Alzheimer disease pathology. Arch Gen Psychiatry. 2005;62(6):685–91.
17. Cavedo E, Chiesa PA, Houot M, Ferretti MT, Grothe MJ, Teipel SJ, et al. Sex differences in functional and molecular neuroimaging biomarkers of Alzheimer's disease in cognitively normal older adults with subjective memory complaints. Alzheimers Dement. 2018;14(9):1204–15.
18. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82–115.
19. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR. Explainable AI: interpreting, explaining and visualizing deep learning. vol. 11700. Springer Nature; 2019.
20. Eitel F, Soehler E, Bellmann-Strobl J, Brandt AU, Ruprecht K, Giess RM, et al. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. NeuroImage Clin. 2019;24:102003.
21. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE. 2015;10(7):E0130140.
22. Wang D, Honnorat N, Fox PT, Ritter K, Eickhoff SB, Seshadri S, et al. Deep neural network heatmaps capture Alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies. NeuroImage. 2023;269:119929.
23. Klingenberg M, Stark D, Eitel F, Ritter K, et al. MRI Image Registration Considerably Improves CNN-Based Disease Classification. In: Abdulkadir A, et al., editors. International Workshop on Machine Learning in Clinical Neuroimaging, vol. 13001. Cham: Springer; 2021. p. 44–52.
24. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal. 2008;12(1):26–41.
25. Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang MC, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage. 2009;46(3):786–802.
26. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. Neuroimage. 2012;62(2):782–90.
27. Smith SM. Fast robust automated brain extraction. Hum Brain Mapp. 2002;17(3):143–55.
28. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: 3rd International Conference for Learning Representations. San Diego: ICLR; 2015.
29. Abdollahi B, Tomita N, Hassanpour S. Data augmentation in training deep learning models for medical image analysis. In: Nanni L, Brahnam S, Brattin R, Ghidoni S, Jain L, editors. Deep learners and deep learner descriptors for medical applications, vol. 186. Cham: Springer; 2020. p. 167–80.
30. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. J Med Imaging Radiat Oncol. 2021;65(5):545–63.
31. Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology. 1993;43(11):2412–4.

32. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatry. 1984;141(11):1356–64.
33. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J Psychiatric Res. 1975;12(3):189–98.
34. Ridha BH, Anderson VM, Barnes J, Boyes RG, Price SL, Rossor MN, et al. Volumetric MRI and cognitive measures in Alzheimer disease. J Neurol. 2008;255(4):567–74.
35. Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: an overview. In: Samek W, Montavon G, Vedaldi A, Hansen L, Müller KR, editors. Explainable AI: interpreting, explaining and visualizing deep learning. vol. 11700. Cham: Springer; 2019. p. 193–209.
36. Binder A, Montavon G, Lapuschkin S, Müller KR, Samek W. Layer-wise relevance propagation for neural networks with local renormalization layers. In: Villa A, Masulli P, Pons Rivero A, editors. International Conference on Artificial Neural Networks, vol. 9887. Cham: Springer; 2016. p. 63–71.
37. Bakker R, Tiesinga P, Kötter R. The scalable brain atlas: instant web-based access to public brain atlases and related content. Neuroinformatics. 2015;13(3):353–66.
38. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta neuropathologica. 1991;82(4):239–59.
39. Gamberger D, Ženko B, Mitelpunkt A, Shachar N, Lavrač N. Clusters of male and female Alzheimer's disease patients in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Brain Inform. 2016;3(3):169–79.
40. Skup M, Zhu H, Wang Y, Giovanello KS, Lin Ja, Shen D, et al. Sex differences in grey matter atrophy patterns among AD and aMCI patients: results from ADNI. Neuroimage. 2011;56(3):890–906.
41. Habes M, Erus G, Toledo JB, Zhang T, Bryan N, Launer LJ, et al. White matter hyperintensities and imaging patterns of brain ageing in the general population. Brain. 2016;139(4):1164–79.
42. Gannon O, Robison L, Custozzo A, Zuloaga K. Sex differences in risk factors for vascular contributions to cognitive impairment & dementia. Neurochem Int. 2019;127:38–55.
43. Alsallakh B, Kokhlikyan N, Miglani V, Yuan J, Reblitz-Richardson O. Mind the Pad–CNNs can Develop Blind Spots. 2020. arXiv preprint arXiv:2010.02178.
44. Nestor SM, Rupsingh R, Borrie M, Smith M, Accomazzi V, Wells JL, et al. Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. Brain. 2008;131(9):2443–54.
45. Bertoux M, Lagarde J, Corlier F, Hamelin L, Mangin JF, Colliot O, et al. Sulcal morphology in Alzheimer's disease: an effective marker of diagnosis and cognition. Neurobiol Aging. 2019;84:41–9.
46. Cai K, Xu H, Guan H, Zhu W, Jiang J, Cui Y, et al. Identification of early-stage Alzheimer's disease using Sulcal morphology and other common neuroimaging indices. PLoS ONE. 2017;12(1):E0170875.
47. Marnane M, Al-Jawadi OO, Mortazavi S, Pogorzelec KJ, Wang BW, Feldman HH, et al. Periventricular hyperintensities are associated with elevated cerebral amyloid. Neurology. 2016;86(6):535–43.
48. van Straaten EC, Harvey D, Scheltens P, Barkhof F, Petersen RC, Thal LJ, et al. Periventricular white matter hyperintensities increase the likelihood of progression from amnestic mild cognitive impairment to dementia. J Neurol. 2008;255:1302–8.
49. Ji X, Wang H, Zhu M, He Y, Zhang H, Chen X, et al. Brainstem atrophy in the early stage of Alzheimer's disease: a voxel-based morphometry study. Brain Imaging Behav. 2021;15:49–59.
50. Simic G, Stanic G, Mladinov M, Jovanov-Milosevic N, Kostovic I, Hof P. Does Alzheimer's disease begin in the brainstem? Neuropathol Appl Neurobiol. 2009;35(6):532–54.
51. Lotze M, Domin M, Gerlach FH, Gaser C, Lueders E, Schmidt CO, et al. Novel findings from 2,838 adult brains on sex differences in gray matter brain volume. Sci Rep. 2019;9(1671).
52. Ritchie SJ, Cox SR, Shen X, Lombardo MV, Reus LM, Alloza C, et al. Sex differences in the adult human brain: evidence from 5216 UK biobank participants. Cereb Cortex. 2018;28(8):2959–75.
53. Eliot L, Ahmed A, Khan H, Patel J. Dump the "dimorphism": Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. Neurosci Biobehav Rev. 2021;125:667–97.
54. Jäncke L, Mérillat S, Liem F, Hänggi J. Brain size, sex, and the aging brain. Human Brain Mapp. 2015;36(1):150–69.
55. Joel D, Persico A, Salhov M, Berman Z, Oligschläger S, Meilijson I, et al. Analysis of human brain structure reveals that the brain "types" typical of males are also typical of females, and vice versa. Front Hum Neurosci. 2018;12:399.
56. Toledo JB, Liu H, Grothe MJ, Rashid T, Launer L, Shaw LM, et al. Disentangling tau and brain atrophy cluster heterogeneity across the Alzheimer's disease continuum. Alzheimers Dement Transl Res Clin Interv. 2022;8(1):e12305.
57. Fonov V, Coupe P, Eskildsen S, Collins L. Atrophy specific MRI brain template for Alzheimer's disease and mild cognitive impairment. In: Alzheimer's Association International Conference. France. Vol. 7. 2011. p. S58. hal-00645521.
58. Rorden C, Bonilha L, Fridriksson J, Bender B, Karnath HO. Age-specific CT and MRI templates for spatial normalization. Neuroimage. 2012;61(4):957–65.
59. Bron EE, Klein S, Papma JM, Jiskoot LC, Venkatraghavan V, Linders J, et al. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. NeuroImage Clin. 2021;31:102712.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC