Clever Hans effect found in a widely used brain tumour MRI dataset

David Wallis, Irène Buvat

Medical Image Analysis 77 (2022) 102368

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Clever Hans effect found in a widely used brain tumour MRI dataset

David Wallis*, Irène Buvat

Laboratory of Translational Imaging, Institut Curie, Université Paris Saclay, Orsay, France

ARTICLE INFO

Article history:
Received 26 April 2021
Revised 19 December 2021
Accepted 10 January 2022
Available online 12 January 2022

MSC:
41A05
41A10
65D05
65D17

Keywords:
Deep learning
Model interpretation
Clever Hans

ABSTRACT

Machine learning is revolutionising medical image analysis, and clearly the future of the field lies in this direction. However, with increasing automation there is a danger of misunderstanding or misinterpreting models. In this paper, we expose an underlying bias in a commonly used publicly available brain tumour MRI dataset. We propose that this is due to implicit radiologist input in the selection of the 2D slices. Through several experiments we show how this bias allows us to achieve a high tumour classification accuracy, even with no information regarding the tumour itself. No other papers that use the dataset mention this bias. These findings demonstrate the importance of understanding machine learning models and their medical context, and the perils of not doing so.

1. Introduction

The last 20 years have seen an explosion in the use of machine learning (ML) algorithms in the medical imaging domain. Models have been built to automate tasks across the gamut of clinical oncology, from tumour detection and segmentation to therapy decisions (Skourt et al., 2018; Giger, 2018; Esteva et al., 2017; Hosny et al., 2018). In many cases these models are beginning to outperform human experts. With more data, higher quality images, and more powerful computers, machine learning-based automation is predicted to fundamentally change the way clinical medicine operates (Litjens et al., 2017; Hosny et al., 2018; Ghafoorian et al., 2017).

However, we should proceed with caution. As models become increasingly sophisticated and complex, they become increasingly difficult to interpret (Hosny et al., 2019; Reyes et al., 2020; Oren et al., 2020). Blindly employing models can lead to erroneous results and false conclusions, highlighting the importance of domain-specific methodological knowledge (Wen et al., 2020; Barrett, 2019; Raimondi et al., 2021). As designers of models with potentially life-changing impacts on people's lives, we need to be vigilant of these issues. This means not only building a model with a good performance, but also understanding why the performance is good.

One way a misunderstanding of data can bias a result is if there are spurious correlations in the training data, known as the Clever Hans effect (Pfungst, 1911). In Lapuschkin et al. (2019) this effect is illustrated for several commonly used non-medical ML problems. In one example a neural network was trained to classify images (into categories e.g. 'person', 'train', 'car', 'horse'), but closer inspection of the data revealed that all the 'horse' images had text in the corners. Saliency maps showed that the network was using this text, rather than the horses, to classify the images. Clearly the network would not generalise to images without this text present. Other studies have shown models that learnt to classify images based on the backgrounds of images, rather than the objects themselves. Zhu et al. (2017) found that a model trained with only background context could still achieve reasonable performance on an image classification task. Beery et al. (2018) showed that their model was good at classifying objects in common contexts (e.g. a cow in an alpine pasture), but performed poorly when objects were placed in unusual settings (e.g. a cow on a beach was labelled as 'seahorse'). Several other papers give examples of networks taking shortcuts to achieve high classification accuracy (Geirhos et al., 2020; Heuer et al., 2016; Rosenfeld et al., 2018; Dawson et al., 2019). Similar effects have also been demonstrated in medical contexts. In the KDD CUP breast cancer identification challenge, it was found that the patient IDs (which had not been removed from the data) were highly correlated with the malignancy of the patients' tumours (Perlich et al., 2008). Raimondi et al. (2021) showed that many methods used to distinguish 'driver' mutations from 'passenger' mutations in cancer genome analysis use datasets with a

* Corresponding author.
E-mail address: wallisphd@gmail.com (D. Wallis).

https://doi.org/10.1016/j.media.2022.102368

Table 1
Summary of published PubMed studies that cite the original article and use the dataset to build models. Any studies that used the dataset for classification with whole slice inputs will have been affected by the bias.

Reference	Task	Description
Cheng et al. (2015)	Classification (segmented region)	Original study. Used radiomic features and tested dilating the tumour region (to include surrounding tissue). Best results were obtained using some surrounding tissue
Gu et al. (2021)	Classification (whole slice)	Used convolutional dictionary learning with local constraints. Ran tests on this dataset and a second brain MRI dataset, using whole slice inputs. Classification results will have been affected by the bias
Gunasekara et al. (2021)	Segmentation and classification	Performed segmentation and classification on a subset of the images (using axial gliomas and meningiomas only). Classification results will have been affected by the bias
Díaz-Pernas et al. (2021)	Segmentation and classification	Used a multi-scale CNN to both classify and segment the tumours. Classification results will have been affected by the bias
Kutlu and Avci (2019)	Classification (whole slice)	Used a CNN for feature extraction from whole slices, discrete wavelet transforms for signal processing, and long short-term memory for signal classification. Classification results will have been affected by the bias

construction bias. This bias allows ML models to take a 'short-cut' and solve a much easier task. Building a dataset without this bias resulted in an 8–28 percentage point drop in the performance (in terms of area under the receiver operating characteristic curve). Several recent studies (DeGrave et al., 2021; Maguolo and Nanni, 2021; López-Cabrera et al., 2021; Teixeira et al., 2021) have shown that many datasets constructed to identify COVID-19 from X-ray images are biased, with positive and negative images taken from different sources. DeGrave et al. (2021) showed that ML models learnt to identify the sources of the images rather than COVID-relevant features. Performance was much worse when models were tested on external test sets from different sources. Saliency maps revealed that the models often focused on laterality markers that originate during the radiograph acquisition process. Maguolo and Nanni (2021) exposed a similar bias in some COVID-19 X-ray studies. They showed that a similar classification performance could be achieved placing large zero-intensity rectangles on the lung regions (thus training their model on only the outer parts of the images). Again, this demonstrated that the ML models were exploiting biases in the datasets to cheat, rather than truly learning to distinguish COVID-19 from non COVID-19 cases.

In this paper, we demonstrate a Clever Hans effect affecting classification performance in a commonly used publicly available medical image dataset. The dataset consists of T1-weighted contrast-enhanced MRI images with tumours segmented and categorised as one of three types (meningiomas, gliomas, and pituitary tumours). Made available in 2015, the dataset was originally used by Cheng et al. (2015) to build a radiomics-based tumour classification model. It has since been used in numerous publications to benchmark different machine learning, deep learning, segmentation, and classification models. Table 1 lists published PubMed articles which use the dataset. In addition, the study is cited by 177 articles according to Google Scholar (Google, 2021), with the number of citations increasing year on year since publication. Any studies that used the dataset to classify the tumours using whole slice inputs will have been affected by the bias. Some studies simultaneously segmented and classified the tumours; whether these were affected by the bias depends on the exact experimental setup. A thorough investigation of all cited studies was performed and none mentioned the bias.

By running the experiments described below, we clearly show the source of the effect and its impact on classification models. In publishing these results we hope to alert others to the dangers of hidden data biases.

2. Material and methods

2.1. Dataset

The data were acquired from Nanfang Hospital, Guangzhou, China, and General Hospital, Tianjing Medical University, China, from 2005 to 2010. The dataset was originally used in Cheng et al. (2015) to automatically identify tumour type and after made publicly available at [dataset]Jun Cheng (2017). It consists of 2D T1-weighted contrast-enhanced MRI images containing one of three brain tumour types (meningiomas, gliomas, or pituitary tumours). There are 233 cancer patients, but multiple images for each patient, giving a total of 3064 images. In total, there are 82 patients with meningiomas (708 slices), 89 with gliomas (1426 slices), and 61 with pituitary tumours (930 slices). The images are a mixture of orientations (axial, coronal, and sagittal), and there are not the same number of images per patient. All images have a resolution of 512 $\times$ 512 pixels and a pixel size of 0.49 $\times$ 0.49 mm^2. Further information such as scanner type and imaging protocol is not detailed. Tumours have been manually delineated by three experienced radiologists, though it is not clear if this delineation was individual or consensus-based. Some example images are shown in Fig. 1, and a full dataset breakdown is shown in Table 2.

2.2. Preprocessing

We ran several experiments to test for the presence of underlying biases. Before running the experiments, we preprocessed the images. Images were symmetrically cropped to 450 $\times$ 450 pixels. This removed some of the background space surrounding the head. Given that MRI intensity values are not absolute, we also rescaled each image to between 0 and 1000 using the 99$^{\text{th}}$ percentile. These preprocessing steps are consistent with other studies in Table 1.

2.3. Feature set creation

To test for the presence of a bias, we created sets of features using six different preprocessing procedures as follows:

1. Extracting radiomic features from the full 450 $\times$ 450 pixel slice without any tumour segmentation
2. Extracting radiomic features from the segmented tumour region (as in conventional radiomic studies)
3. Placing an 80 $\times$ 80 pixel square of intensity zero centred on the tumour (i.e. with its centre at the tumour centre of mass), then extracting radiomic features from the whole slice
4. Thresholding the image by setting pixels less than 100 to zero and pixels greater than 100 to one. This left a binary image of the skull outline. Radiomic features were then extracted from this binary image (with two discretisation levels)
5. Using only three features: the number of zero-intensity pixels (i.e. the size of the background region), the maximum intensity, and the orientation of the slice (sagittal, axial, or coronal)
6. Combining the features from 1. and 2.

Examples of these procedures are shown in Fig. 2. Radiomic features were extracted using PyRadiomics (van Griethuysen et al., 2017) with a fixed bin width of 25 (except for experiment 4, which only had two intensity levels). Default PyRadiomics settings

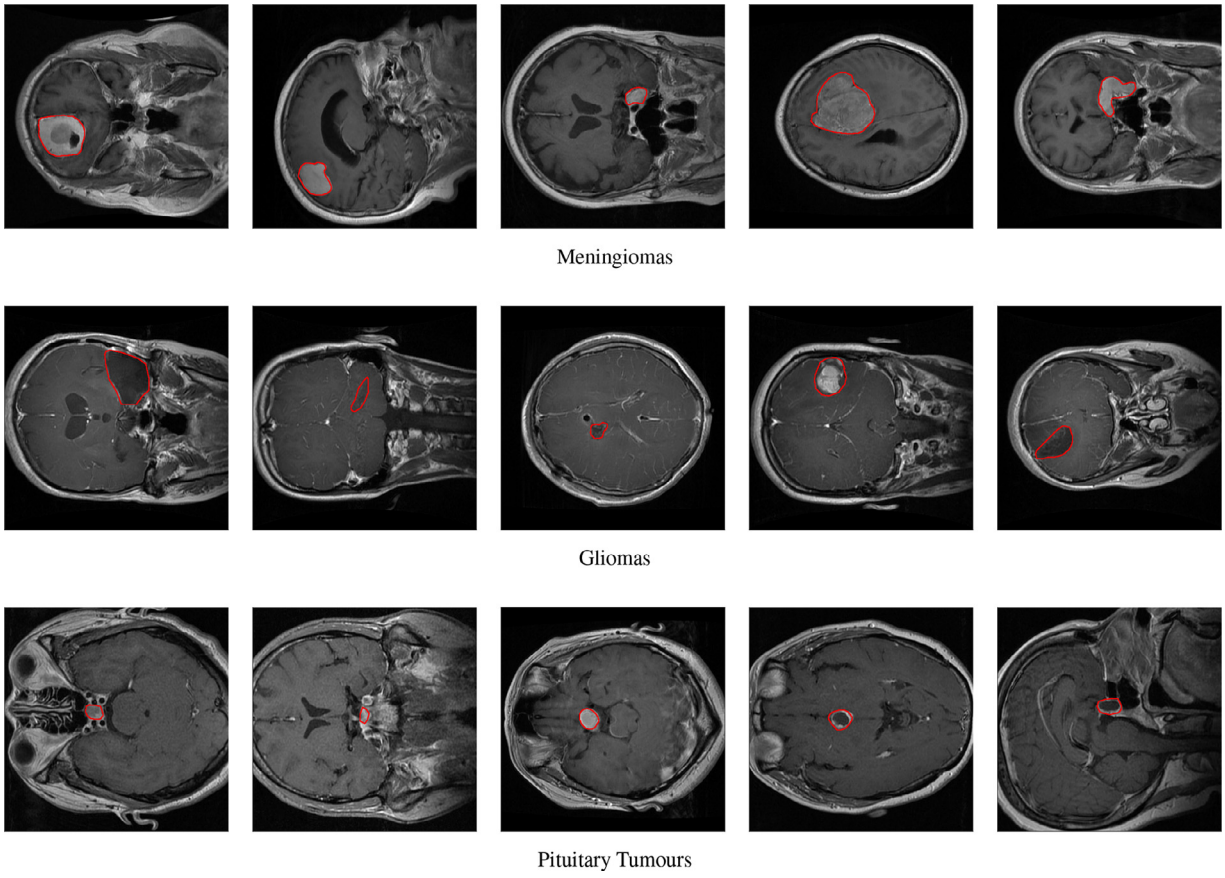
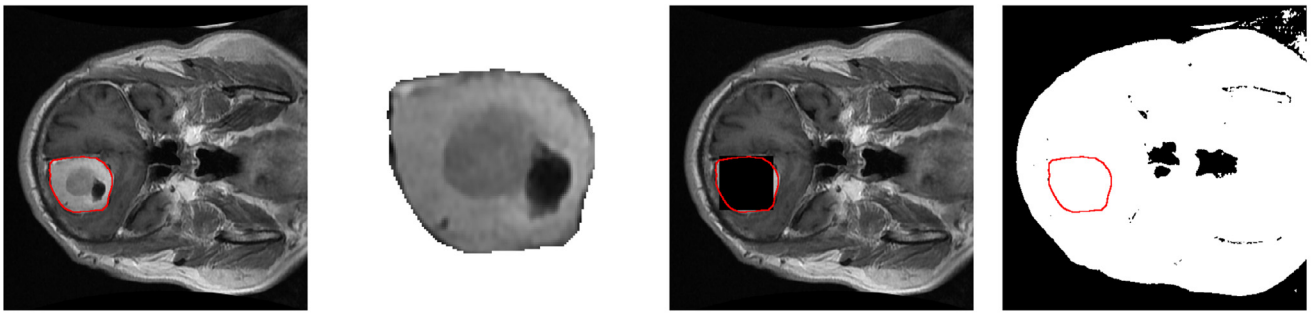


Fig. 1. Examples of T1-weighted contrast-enhanced MRI images with the tumour segmentations in red. The three classes (meningiomas, gliomas, and pituitary tumours) are shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Full image without tumour segmentation (experiment 1)

(b) Segmented tumour region (experiment 2)

(c) Image with an 80 × 80 pixel zero-intensity square covering the tumour (experiment 3)

(d) Image with pixels <100 set to zero and >100 set to one, meaning only the skull outline is shown (experiment 4)

Fig. 2. Examples of images used in the different experiments. The tumour outline is shown in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Dataset Summary.

Tumour Type	Number of Patients	Number of Images	MRI View	Number of Images
Meningioma	82	708	Axial	209
			Coronal	268
			Sagittal	231
Glioma	89	1426	Axial	494
			Coronal	437
			Sagittal	495
Pituitary Tumour	62	930	Axial	291
			Coronal	319
			Sagittal	320
Total	233	3064		3064

were used in all cases. We extracted first-order, grey level cooccurence matrix (GLCM), neighbouring grey tone difference matrix (NGTDM), grey level run length matrix (GLRLM), grey level size zone matrix (GLSZM), grey level dependence matrix (GLDM), and first-order features based on gradient, exponential, and wavelet filters (with coiflet 1 wavelets and high and low-pass filters in both dimensions). For experiment 2, we also extracted shape-based features. This gave a total of 229 features per image (239 for experiment 2 as this also included shape-based features). See the Supplementary Material for full details of the radiomic features used.

2.4. Training and evaluation of the ML model

We then built support vector machines (SVMs) to compare the performances of these feature sets. The experimental setup was identical in each case. We used a linear kernel with a C-value of 1 and a one-vs-all setup, balancing class weights by their frequencies in the training set. Excepting experiment 5, all feature sets contained a large number of features. We reduced the number of features using a two-stage feature reduction process; the 140 best features were chosen using a univariate ANOVA F-test, then any features above a correlation threshold of 0.95 were removed (for two correlated features the mean absolute correlation of each feature was calculated and the feature with the largest mean absolute correlation was removed). Before analysis, features were scaled to have zero mean and unit variance. For both of these steps the training data was used for calibration.

The dataset was split into 190 patients in the training set and 43 patients in the validation set. The selection of training and validation set was done patient-by-patient (rather than image-by-image) so that images from the same patient were not present in the training and validation sets. Performance was evaluated using a 200-iteration random holdout method, with the training and validation sets selected randomly each time. The accuracy and its associated standard deviation were calculated by averaging the performances obtained on the validation sets during this process.

2.5. Model interpretation

We then wanted to explain the performance of models built using radiomic features extracted from full slices (experiment 1 in 2.3). To do this, we found the most important radiomic features for each of the 200 models using the SVM feature coefficients (each feature had three coefficients corresponding to the three classes, we ranked them using the mean of these three coefficients). By examining these, we could determine which features within the images were being used for classification and whether these features were related to the tumours or to the bias.

3. Results

3.1. Performance of different feature sets

The performances of the SVMs built using the different features sets are listed in Table 3.

Table 3
Accuracy and standard deviation for SVMs built using the different feature sets described in 2.3.

Preprocessing Procedure	Accuracy
1. Full 450 × 450 slice	0.86 ± 0.03
2. Segmented tumour region	0.86 ± 0.03
3. Occluded tumour region	0.87 ± 0.04
4. Skull outline	0.74 ± 0.04
5. Three features	0.77 ± 0.04
6. Combining (1) and (2)	0.94 ± 0.02

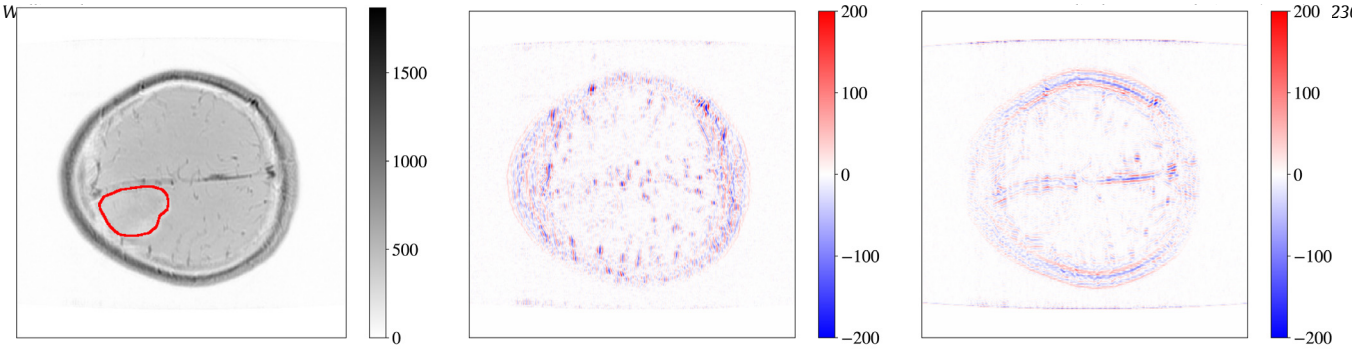
3.2. Model interpretation

The two most important features from experiment 1's models were the interquartile range of low/high pass wavelet-transformed images and the robust mean absolute deviation of high/low pass wavelet-transformed images (the robust mean absolute deviation is defined as the mean distance of all intensity values from the mean value calculated on pixels between the 10^{th} and 90^{th} percentile), with one of these features being the most important in 146 out of 200 random holdout models. Here and in Figs. 3 and 4, we denote a wavelet-transformed image created using a low-pass filter in the x-dimension and a high-pass filter in the y-dimension as low/high and vice versa. Because the dataset consists of slices of different orientations, these do not always refer to the same dimension on the MRI scan. Rather, they just refer to the horizontal or vertical dimensions of the images.

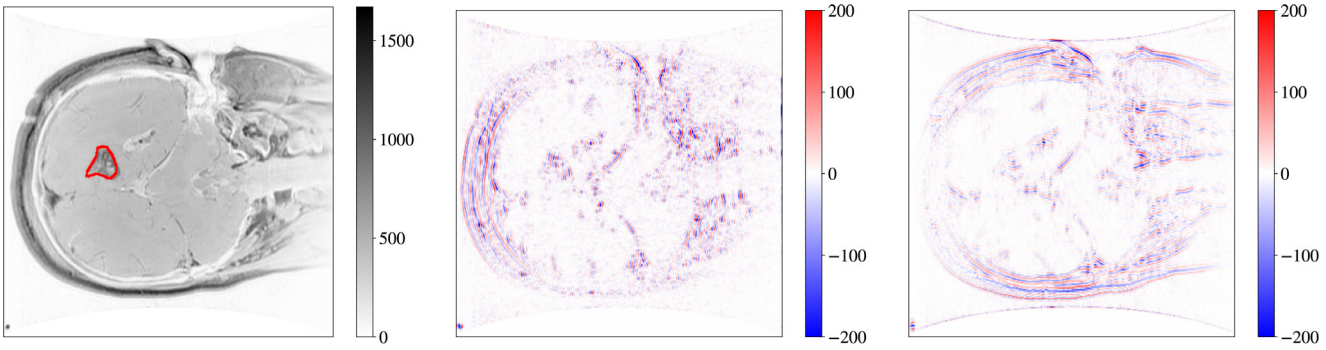
To understand these wavelet-based features, in Fig. 3 we plotted histograms of the pixel intensities of wavelet-transformed images across all patients for each class separately. In Fig. 4, we plotted some example wavelet-transformed images to link the histogram distributions to physical characteristics of the MRI scans. The intensities have been cropped at -200/+200 to show the extreme values more clearly.

4. Discussion

Experiment 1 (extracting radiomic features from whole slices) achieved a comparable accuracy to experiment 2 (using a precise tumour segmentation). Given these two results, we could conclude that the radiomic features used in experiment 1 were capable of extracting useful information from the tumour, even with this whole slice view. However, we thought this initial hypothesis unlikely. The tumours were small compared to the slice size, and there is a lot of heterogeneity in the non-tumour regions which we thought would obscure any tumour-based differences. In experiment 3, we achieved an accuracy similar to that of experiment 1 (0.87 ± 0.04 vs 0.86 ± 0.03) while occluding the tumour with a zero-intensity square. Evidently the tumour itself was not essential for classification. We may still hypothesise that there was sufficient information in the surrounding (non-occluded) tissue to classify the images, or that the occluded region itself was being used by the radiomic features to locate the tumour and therefore aid in classification. In experiment 4, we erased all tissue-based information in the brain, leaving just the outline of the skull. We were still able to achieve a high classification accuracy (0.74 ± 0.04). This result clearly shows that there is another force at play. The three types of brain tumour classified here tend to grow in specific areas of the brain (meningiomas on membranes surrounding the brain and spinal chord, gliomas in the cerebrum or cerebellum, and pituitary tumours on the pituitary gland (Johns Hopkins Medicine (a); Johns Hopkins Medicine (b); Johns Hopkins Medicine (c))), so positional information and information outside the tumour are useful for classification. However, with 2D slices this information is implicitly contained within the slice, which has been preselected by a radiologist because it contains the tumour. The position and orientation of the slice can therefore be used as a proxy for tumour type (the Clever Hans effect). There was still a performance difference between experiments 1 and 4, but this is likely due to the extra information contained in experiment 1's full images (such as the amount of high-intensity bone) which allows more precise positioning and orientation of the slice. In experiment 5, we used three features designed to position and orientate the slice without any brain tissue information (slice orientation, number of zero-intensity pixels, and maximum intensity). The slice orienta-

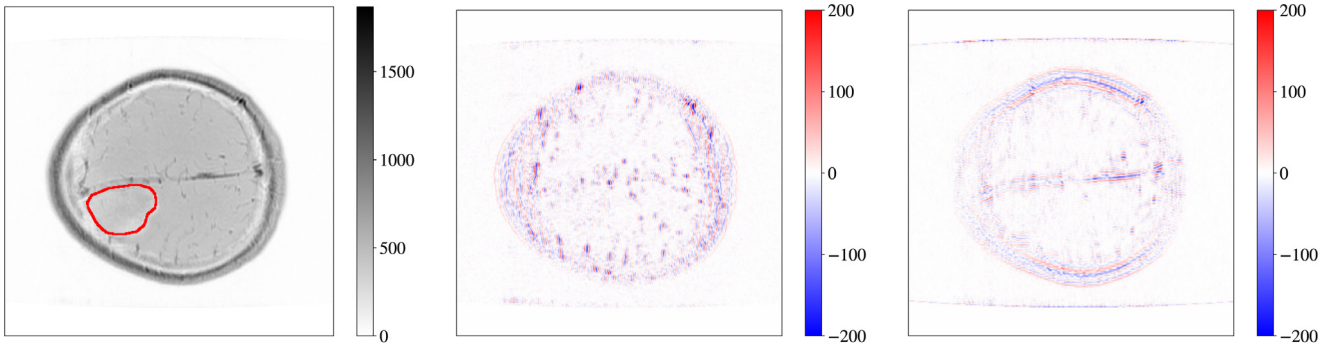


(a) Original image (b) High/low pass wavelet-transformed image (c) Low/high pass wavelet-transformed image

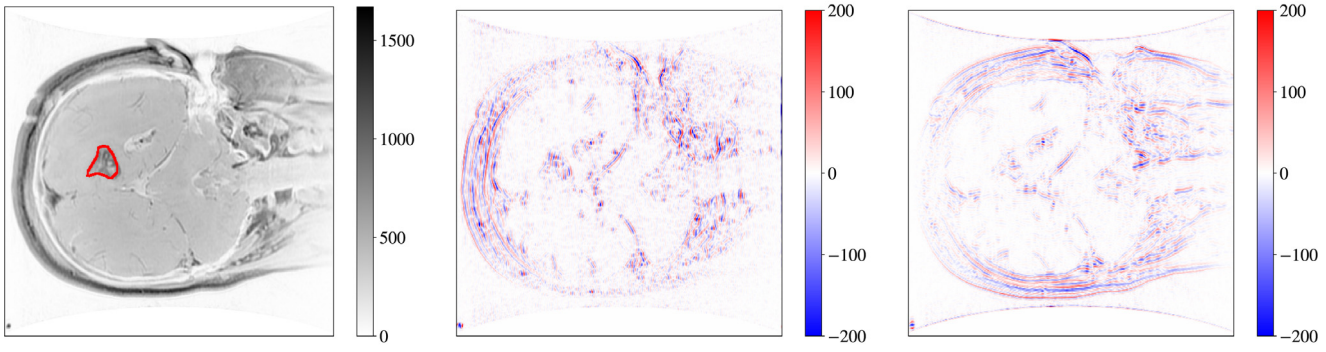


(d) Original image (e) High/low pass wavelet-transformed image (f) Low/high pass wavelet-transformed image

Fig. 3. Histograms of the intensity values of high/low pass and low/high pass wavelet-transformed images for all patients, separated by class. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Original image (b) High/low pass wavelet-transformed image (c) Low/high pass wavelet-transformed image



(d) Original image (e) High/low pass wavelet-transformed image (f) Low/high pass wavelet-transformed image

Fig. 4. Examples of high/low pass and low/high pass wavelet-transformed images for three scans. The tumour outline is shown in red on the original image. The scales have been cropped at -200/+200 on the wavelet-transformed images to better highlight high intensity regions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tion gave the type of slice (sagittal, axial, or coronal). To find where the slice was located within the image (i.e. fix the final coordinate), we used the number of zero-intensity pixels and the maximum intensity in the image. The number of zero-intensity pixels counted the number of pixels outside the skull. This is related to the size of the skull and therefore the position of the slice (an axial slice near the top of a scan will not contain much of the head so will have a large number of zero-intensity pixels). The maximum intensity of each image was in the bone in all cases (never in the tumour) and helped further fix the position of the slice, as different slice positions will have different types/densities of bone Peterson and Dechow, 2003. Evidently there is some ambiguity using these final two features to locate a slice (there will be several slice positions with the same number of zero-intensity pixels), but we still achieved an accuracy of 0.77 ± 0.04. We also tried using the mean coordinate of the tumour to position the slice, but this did not work, probably because the patients were not all in the exact same position and their heads were not the same size. Finally, in experiment 6 we combined the feature sets from experiments 1 and 2. This gave a higher accuracy, again suggesting that the features in experiment 1 are not based on the tumour itself.

The results in 3.2 help us to understand the high accuracy of experiment 1. In Fig. 3, we see that wavelet-transformed images of the different classes have different pixel value distributions, especially at the more extreme (high and low) intensities. The two most important features used in experiment 1's models, low/high pass interquartile range and high/low pass robust mean absolute deviation, both depend on the intensity range of pixel values, corroborating what we see in the histograms. However, looking at the example wavelet-transformed images (Fig. 4), we see that the tumour is not highlighted in any of the cases. Instead, the extreme values mostly highlight the skull outline (this is unsurprising, as wavelet transformations highlight edges in images). This is further evidence that the model in experiment 1 is using this skull outline (which is indicative of slice position), rather than tumour-based information, to classify the images.

These results demonstrate a clear bias in this dataset. Quoting from the original study: "In clinical settings, usually only a certain number of slices of brain contrast-enhanced MRI (CE-MRI) with a large slice gap are acquired and available. Therefore, the development of a 2D image-based CBIR system for clinical applications is more practical ... Representative slices that have large lesion sizes are selected to construct the dataset". The selection of these *representative slices* is a human-introduced prior that ML models can exploit. The original study was radiomics-based and used segmented tumour regions, meaning the bias did not invalidate the results. However, many of the subsequent deep learning-based publications classified the tumours using whole-slice images directly input into CNNs. None made reference to this bias. This is probably because CNNs are routinely used without segmentation, so suspicions were not raised. Using radiomic features, as in our experiments, the problem is much clearer. We cannot know whether the CNNs in these studies truly used information within the tumour or the positional and orientational information hard-coded into the images. They may have used a combination of the two. To state that these models are automatically classifying the tumours is therefore untrue. The slices have been manually preselected by a radiologist, and this preselection is already sufficient to classify the images correctly in most cases.

These findings do not mean that this dataset is of no use. As mentioned, many studies used the dataset for segmentation tasks or classified the tumours using only the segmented regions. These investigations are still valid. In the future we expect that 3D analyses, which can exploit full volumetric information, will result in better performing models. However, a lot of studies are still in 2D. As evidenced by our literature search, a lot of groups are actively working on this dataset. These groups should take care to not bias their findings.

There are some limitations to our study. We optimised our SVM by systematically testing different kernels and C-values. Radial basis function and polynomial kernels with C-values ranging from 0.1 to 5 were tested, but these did not significantly improve the results. We also systematically tested different feature reduction F-test and correlation thresholds, optimising the parameters for each experiment separately using the training data. The optimum parameters were the same for each experiment. Our setup may still not be optimal; there are of course innumerable different ML and feature reduction methods (Fatima and Pasha, 2017; Cai et al., 2018). By using the same setup for each experiment we tried to ensure that this choice did not bias the results. Additionally, as with any cross-validation or holdout-based method, the standard deviation may be biased (Bengio and Grandvalet, 2004).

5. Conclusions

To conclude, this study exposes a bias in a commonly used publicly available dataset. This work demonstrates why it is important to understand ML models, not just blindly deploy them. We hope that this work will alert others to the dangers of black box automated models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

David Wallis: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft. **Irène Buvat:** Conceptualization, Writing – review & editing, Supervision.

Acknowledgements

This project has received funding from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement (no. 764458).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2022.102368.

References

Barrett, H.H., 2019. Is there a role for image science in the brave new world of artificial intelligence? J Med Imaging 7 (1), 1–6.
Beery, S., et al., 2018. Recognition in terra incognita. In: Computer Vision - ECCV 2018, pp. 472–489.
Bengio, Y., Grandvalet, Y., 2004. No unbiased estimator of the variance of k-fold cross-validation. J Mach Learn Res 5, 1089–1105.
Cai, J., et al., 2018. Feature selection in machine learning: a new perspective. Neurocomputing 300, 70–79.
Cheng, J., et al., 2015. Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLoS ONE 10, e0140381.
Dawson, M., et al., 2019. From same photo: Cheating on visual kinship challenges. In: Computer Vision – ACCV 2018, pp. 654–668.
DeGrave, A.J., et al., 2021. AI For radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 3, 610–619.
Díaz-Pernas, F.J., et al., 2021. A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. Healthcare 2, 153.
Esteva, A., et al., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118.
Fatima, M., Pasha, M., 2017. Survey of machine learning algorithms for disease diagnostic. J Intell Syst Appl 9, 1–16.

Geirhos, R., et al., 2020. Shortcut learning in deep neural networks. Nat Mach Intell 2, 665–673.
Ghafoorian, M., et al., 2017. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Sci Rep 7, 5110.
Giger, M., 2018. Machine learning in medical imaging. J Am Coll Radiol 15, 512–520.
Google, 2021. Google Scholar citations for original study. https://scholar.google.fr/scholar?cites=2281002593004500320&as_sdt=2005&sciodt=0,5&hl=en. (Accessed 22 March 2021).
Gu, X., et al., 2021. Brain tumor MR image classification using convolutional dictionary learning with local constraint. Front Neurosci 28, 679847.
Gunasekara, S.R., et al., 2021. A systematic approach for MRI brain tumor localization and segmentation using deep learning and active contouring. J Healthc Eng 11, 6695108.
Heuer, H., et al., 2016. Generating captions without looking beyond objects. ECCV 2016 2nd Workshop on Storytelling with Images and Videos.
Hosny, A., et al., 2018. Artificial intelligence in radiology. Nat Rev Cancer 18, 500–510.
Hosny, A., et al., 2019. Handcrafted versus deep learning radiomics for prediction of cancer therapy response. Lancet Digit Health 1, e106–e107.
Johns Hopkins Medicine, a. Conditions and diseases: Gliomas. URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/gliomas. [Accessed 2021-01-28].
Johns Hopkins Medicine, b. Conditions and diseases: Meningioma. URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/meningioma. [Accessed 2021-01-28].
Johns Hopkins Medicine, c. Conditions and Diseases: Pituitary Tumours. URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/pituitary-tumors. [Accessed 2021-01-28].
[dataset Jun Cheng], 2017. Brain Tumour Dataset. https://figshare.com/articles/dataset/brain_tumor_dataset/1512427. (Accessed 17 Jan 2021).
Kutlu, H., Avci, E., 2019. A novel method for classifying liver and brain tumors using convolutional neural networks, discrete wavelet transform and long short-Term memory networks. Sensors 19, 1992.
Lapuschkin, S., et al., 2019. Unmasking clever hans predictors and assessing what machines really learn. Nat Commun 10, 1096.
Litjens, G., et al., 2017. A survey on deep learning in medical image analysis. Med Image Anal 42, 60–88.
López-Cabrera, J.D., et al., 2021. Current limitations to identify COVID-19 using artificial intelligence with chest X-ray imaging. Health Technol 11, 411–424.
Maguolo, G., Nanni, L., 2021. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. Information Fusion 76, 1–7.
Oren, O., et al., 2020. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. Lancet Digit Health 2, e486–e488.
Perlich, C., et al., 2008. Breast cancer identification: kdd cup winner's report. SIGKDD Explorations 10, 39–42.
Peterson, J., Dechow, P.C., 2003. Material properties of the human cranial vault and zygoma. Anat Rec 274A (1), 785–797.
Pfungst, O., 1911. Clever hans (the horse of mr. von osten): contribution to experimental animal and human psychology. J Philos Psychol Sci 8, 663–666.
Raimondi, D., et al., 2021. Current cancer driver variant predictors learn to recognize driver genes instead of functional variants. BMC Biol 19, 3.
Reyes, M., et al., 2020. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiol Artif Intell 2 (3), e190043.
Rosenfeld, A., et al., 2018. The elephant in the room. arXiv:1808.03305.
Skourt, B.A., et al., 2018. Lung CT image segmentation using deep neural networks. Procedia Comput Sci 127, 109–113.
Teixeira, L.O., et al., 2021. Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. Sensors 21, 21.
van Griethuysen, J.J.M., et al., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer Res 77, e104–e107.
Wen, J., et al., 2020. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med Image Anal 63, 101694.
Zhu, Z., Xie, L., Yuille, A.L., 2017. Object recognition with and without objects. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pp. 3609–3615.