Identifying Morphological Indicators of Aging With Neural Networks on Large-Scale Whole-Body MRI

Taro Langner, Johan Wikström, Tomas Bjerner, Håkan Ahlström, and Joel Kullberg

***Abstract*— A wealth of information is contained in images obtained by whole-body magnetic resonance imaging (MRI). Studying the link between the imaged anatomy and properties known from outside sources has the potential to give new insights into the underlying factors that manifest themselves in individual human morphology. In this work we investigate the expression of age-related changes in the whole-body image. A large dataset of about 32,000 subjects scanned from neck to knee and aged 44-82 years from the UK Biobank study was used for a machine-based analysis. We trained a convolutional neural network based on the VGG16 architecture to predict the age of a given subject based on image data from these scans. In 10-fold cross-validation on 23,000 of these images the network reached a mean absolute error (MAE) of 2.49 years ($R^2 = 0.83$) and showed consistent performance on a separate test set of another 8,000 images. On a second test set of 100 images the network outperformed the averaged estimates given by three experienced radiologists, which reached an MAE of 5.58 years ($R^2 = 0.08$), by more than three years on average. In an attempt to explain these findings, we employ saliency analysis that opens up the image-based criteria used by the automated method to human interpretation. We aggregate the saliency into a single anatomical visualization which clearly highlights structures in the aortic arch and knee as primary indicators of age.**

***Index Terms*— Magnetic resonance imaging (MRI), whole-body, machine learning, neural network, age.**

I. Introduction

THE UK Biobank study is a large-scale health resource involving data from 500,000 volunteer participants [1]. The collected data range from questionnaires to physical measurements as well as biological samples, genotyping, and medical imaging with a long-term follow-up for individual health outcomes. The UK Biobank Imaging Study plans to collect image data from 100,000 of these participants, including volumetric water-fat separated magnetic resonance imaging (MRI) scans of the body from neck to knee, which contain comprehensive information about the anatomy of each subject. The large scope of this dataset has the potential to allow for the identification of new associations between the image-based information from the scans and the wealth of biomarkers and other measurements from the listed sources. One such association that is of special interest consists in the changes that the human body undergoes as a result of aging. Age is a major risk factor for conditions such as cardiovascular, neurodegenerative, and metabolic disease as well as cancer [2]. However, chronological age is not a reliable predictor for burden of disease and there is considerable variation in health outcomes and frailty between individuals of the same chronological age [3]. Attempts have been therefore made to identify an aging biomarker that describes a biological rather than chronological age and which could thereby track the aging process more accurately [4], [5].

Age-related changes in the appearance of the face have been previously shown to enable a machine learning strategy similar to the one presented in this work to estimate the age of a person by a mean absolute error of less than 3.5 years based on a facial photograph [6]. Medical imaging techniques enable assessments of chronological age based on various other anatomical sites such as the teeth, clavicle, and wrist [7]. The recent RSNA Pediatric Bone Age Machine Learning Challenge included more than 14,000 radiographs of the hand, which allowed automated methods based on neural networks to estimate the age of children and adolescent subjects by mere months based on skeletal maturation [8]. With the help of MRI, the age can furthermore be assessed with images of the knee [9] and brain [10], which is also viable in older populations. For the brain, the application of machine learning and deep learning in particular has established the concept of brain-predicted age as a biomarker [10], which may deviate from the chronological age and thereby indicate premature brain aging as a symptom of conditions such as Alzheimers disease, traumatic brain injury, and diabetes [11] and has been shown to be associated with increased risk of mortality [12] and mental disorders [13]. The UK Biobank dataset presents

Manuscript received October 1, 2019; revised October 15, 2019; accepted October 18, 2019. Date of publication October 30, 2019; date of current version April 30, 2020. This work was supported in part by a research grant from the Swedish Heart-Lung Foundation and in part by the Swedish Research Council under Grant 2016-01040. *(Corresponding author: Taro Langner.)*

T. Langner, J. Wikström, and T. Bjerner are with the Section of Radiology, Department of Surgical Sciences, Uppsala University, 751 85 Uppsala, Sweden (e-mail: taro.langner@surgsci.uu.se).

H. Ahlström and J. Kullberg are with the Section of Radiology, Department of Surgical Sciences, Uppsala University, 751 85 Uppsala, Sweden, and also with Antaros Medical, BioVenture Hub, 431 53 Mölndal, Sweden.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2950092

the opportunity to apply a similar strategy for age estimation to scans of the body.

In this work, we explore how age manifests itself in whole-body MRI scans of subjects aged 44-82 years and to what degree these effects are accessible to human observers. The underlying anatomical factors must be independent from the established patterns in the brain, teeth, and wrist, none of which are contained in the field of view. We propose an efficient representation of the whole-body scan for the training of a convolutional neural network in a regression task. The method is expected to capture age-related changes and allow for an automated age estimation which is compared to the assessment by three experienced radiologists. Finally, we examine how morphological indicators of age can be identified and made accessible to human interpretation by saliency analysis.

II. Methods

Data from the whole-body scans of the UK Biobank study were used both for the training of a neural network and for manual age assessment by trained radiologists. The neural networks were furthermore used to generate saliency maps that could explain which patterns enabled the automated assessment.

A. UK Biobank Imaging Protocol

At the time of writing 32,323 of 100,000 planned whole-body scans of male and female subjects have been made available as part of the UK Biobank Imaging Study. Subjects were contacted by letter for voluntary participation based on records of the National Health Service. Those who consented may withdraw from the study at any time and their provided data is anonymized before being shared with researchers. These scans were acquired at three different imaging centers (Cheadle, Reading, and Newcastle) in the United Kingdom. Each scan consists of three-dimensional dual-echo Dixon images obtained with a Siemens Aera 1.5T device with varying TR (6.69 or 6.67 ms), TE 2.39/4.77 ms, flip angle 10deg, in-slice voxel dimensions of 2.23 mm × 2.23 mm and varying slice thicknesses (3 mm, 3.5 mm or 4.5 mm, depending on the station) [14]. The scans are available in DICOM format and typically cover the body in supine position from about neck to knee level in six separate stations. The head and feet are therefore usually not included, whereas the arms and wrists likewise extend into the edges of the magnetic field where they are outside of the field of view or strongly distorted by artifacts. For each station, a volume has been acquired for in- and opposing phase and is available together with a water and fat signal as reconstructed on the scanner. Stations 2-4 were acquired with a 17 second breath hold and the total imaging time amounted to about six minutes.

B. Image Data

In order to form a single representation of the image data for each subject, the subvolumes for each station were grouped by type and fused by resampling into a common space at the highest voxel resolution of 2.23 × 2.23 × 3 mm. The voxel intensities were interpolated along overlapping areas of adjacent stations in order to smoothen out the station borders. This resulted in a three-dimensional volume with a typical resolution of 370 × 224 × 174 voxels. These volumes were then represented in a two-dimensional format showing the normalized mean intensity both for all coronal and all sagittal slices next to each other. This was done for both the water signal and the fat signal, as seen in Fig. 1. As a final processing step, these images were downsampled to a resolution of 256 × 256 pixels, combined as channels of a single image and encoded in 8bit format as a trade-off for faster processing. The choice of such a low-dimensional format reduces both the memory consumption and calculation times and moreover yielded superior results in preliminary experiments as compared to more complex representations.

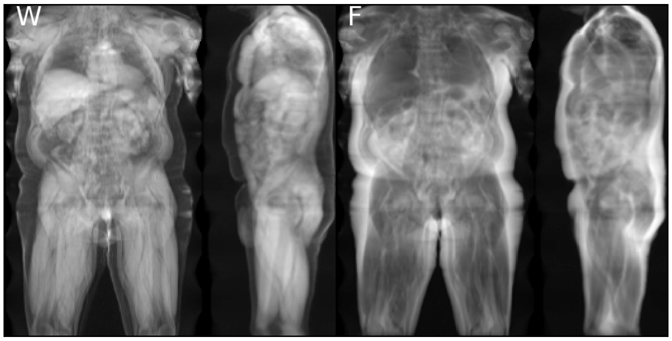

Fig. 1. As input for the network, the water signal (W) and fat signal (F) were compressed by mean intensity projection along the coronal and sagittal directions. The projections were normalized, concatenated and downsampled to 256 × 256 pixels before being combined as channels of a color image. Here the two channels are seen side by side.

The total of 32,323 available scans was manually inspected and some were excluded based on apparent problems that were visible in the projected view. These problems ranged from water-fat swaps (2%) to untypical fields of view, prosthetics, nonstandard leg positioning or voxel spacing as well as missing or corrupted data (2%).

C. Label Data

In addition to the image data, the UK Biobank study collected a wide range of information for each subject, including the year and month of birth. The recorded study date in the DICOM metadata therefore makes it possible to calculate the age of a subject at scan time. Since the exact day of birth is not available, all subjects were assumed to be born in the middle of the recorded month, so that the assigned age is accurate to about half a month. A histogram for the age at scan time for all imaged subjects is shown in Fig. 2.

Three separate datasets were formed from the total of 32,323 scans. The 23,905 scans that were made available in an early release of UK Biobank were divided into a training set **A** and a small test set **B** which was used for comparison to human performance. The subsequently released additional 8,418 subjects were grouped into a second, larger test set **C**. Whereas the dataset **A** was used for training and cross-validation, both test sets **B** and **C** were never used for training.

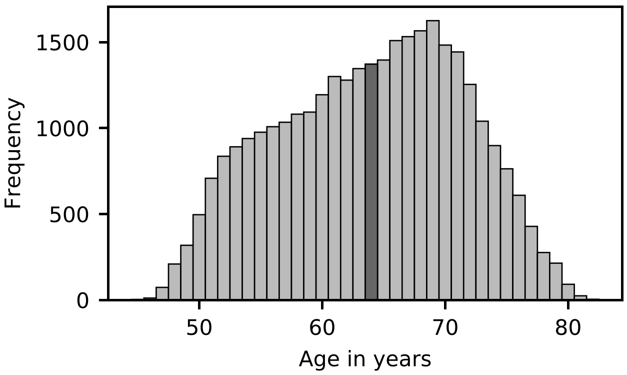

Fig. 2. Distribution of age among the 32,323 subjects with image data. The marked column at 64 years represents the mean age.

TABLE I
DATASETS AND AGE DISTRIBUTION

Dataset	N	mean ± SD	min	max	sex ratio
Training set **A**	23,120	63.44 ± 7.51	44.60	81.87	0.48
Test set **B**	100	64.01 ± 7.14	46.15	76.30	0.50
Test set **C**	8,064	65.18 ± 7.42	46.15	82.32	0.48

* SD: Standard deviation of age. The sex ratio was calculated as number of men divided by number of women. Those subjects that were excluded due to faulty images are not included in the listed numbers.

Additional information on the age distribution in the individual datasets is given in Table I.

Several other relevant properties were extracted together with the age, such as height, sex, and blood pressure measurements. Dependencies between these values and both the predicted and ground truth age were evaluated with Pearson's coefficient of correlation (r) in order to identify possible confounding factors in the dataset that could bias the training process.

D. Neural Network Configuration

Several convolutional neural network architectures were trained to predict the recorded age based on these images. The prediction was modeled as a regression task rather than a classification, so that each architecture was modified by removing the softmax layer and changing the number of output values of the final fully connected layer to one. The score predicted for this output was directly used as the estimated age. In a series of preliminary experiments leading up to the presented design choices, various combinations of architectures, hyperparameters and data formats (2d, 3d and hybrids) were explored. In the scope of those attempts, the VGG16 architecture with batch normalization [15] for regression outperformed corresponding versions of the ResNet50 [16] and DenseNet161 [17]. The VGG16 architecture was therefore used for all reported results. Volume-based architectures that were also examined proved to be prohibitively slow and were outperformed both in speed and accuracy by the chosen configuration.

Transfer learning by pretraining of neural networks on a large dataset such as ImageNet [18] is a common strategy to achieve a faster and better convergence. In this case, not all weights of a network pretrained in this way could be copied, since the ImageNet challenge requires the prediction of 1000 scores in the final layer rather than a single one, as well as a three-channel input image. Nonetheless, copying the weights from all remaining layers as well as the first two input channels led to a clearly improved result, even though the time to convergence remained roughly the same. Further design choices consisted in using the mean squared error as a loss function, which outperformed the mean absolute error loss, as well as an online augmentation strategy in which all images were randomly translated in the coronal plane by up to 16 pixels. Splitting the dataset into men and women was attempted at the early stages but led to worse results, so that no further attempts at a gender-specific prediction were made.

In this configuration the network was trained on the training dataset **A** in 10-fold cross-validation. Three 10-fold splits were generated for this purpose and the reported results are the average of the network performance on each of these splits. Repeated cross-validation was performed in this way for the full image and two ablation experiments in which either the lower or upper half of the images was cropped out. The network configuration was furthermore trained (without any cropping) on the entire training set **A** and then evaluated on the test sets **B** and **C**. On set **B** this experiment was repeated multiple times with a decreasing amount of unique training images (16000, 8000, 4000, 2000, 1000, and 500) in order to examine the relationship between the amount of training data, the network performance and the saliency maps.

All experiments were conducted in the framework PyTorch on a Nvidia GTX 1080 Ti 11 GB graphics card. The training was run with the Adam optimizer at a learning rate of 0.0001 and a batch size of 32.[1] Each training run in this configuration was run for a fixed number of 80,000 iterations and typically took about eight hours. The performance was evaluated with the mean absolute error (MAE) and its standard deviation (SD). We also list Pearson's coefficient of correlation (r) for the strength of the linear relationship between estimated and true age, as well as the coefficient of determination (R^2). Since our regression models are non-linear, the latter may be negative when the model provides a worse fit to the data than a consistent prediction of the mean value.

E. Saliency Analysis

The saliency analysis was based on guided gradient-weighted class activation maps [19] using an implementation for PyTorch [20]. For each input image this technique generates a pixel-wise map of positive and negative gradients that are linked to increased or decreased predicted age. It thereby becomes possible to highlight anatomical patterns and shapes that contribute to a higher predicted age. The saliency map generated in this way is only valid for the corresponding input image and often challenging to interpret due to noise and intricate arrangements that are highly specific to the training process. The interpretation of the saliency in the following sections is therefore predominantly based on examining recurring patterns in a large number of individual saliency maps. It is

[1] https://github.com/tarolangner/mri-biometry

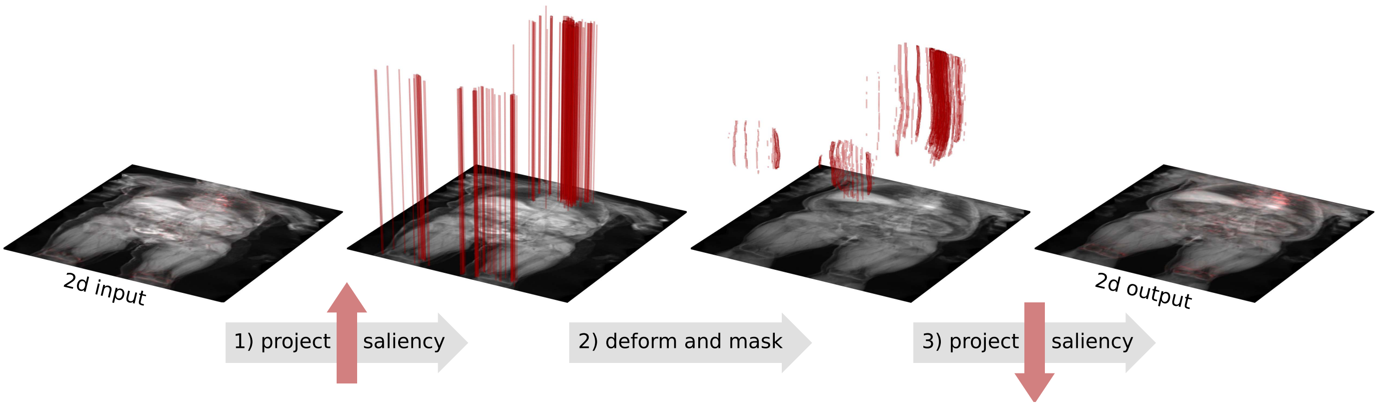

Fig. 3. The saliency maps were aggregated by projecting an individual two-dimensional saliency map into a volume (1), applying a volumetric deformation field (2) and projecting the resulting saliency volume back (3) into the two-dimensional space of the template subject.

nonetheless possible to also form a combined visualization by aggregating the saliency maps. Since all subjects share the same basic human anatomy, they can be registered to a template person, so that the aggregated saliency highlights corresponding anatomical structures in a common space. This was achieved by co-registering the MRI volumes [21] and applying the resulting deformation fields to the saliency maps. This procedure requires several steps:

1) The two-dimensional saliency map must first be mapped to the volumetric signal image. We simply stack the saliency map along the axis that the underlying view was originally projected from, separately for the coronal and the sagittal projection.
2) The resulting volumes must then be resampled to match the resolution of the original signal image and can consequently be transformed with the deformation field.
3) The deformed saliency volume could then be masked with a body mask to reduce blurring and finally be downsampled and projected back to match the original two-dimensional format.

A schematic view of these steps is seen in Fig. 3. The resulting aggregated saliency is often less exact than the individual saliency maps. We therefore use the magnitude of the gradients only, in contrast to the signed gradients in the individual saliency maps. The aggregation removes most of the noise as well as patient-specific features, however, and thereby allows for a comprehensive visualization that still retains the most common relevant anatomical structures.

F. Human Predictions

In order to examine human performance on this task as a point of reference, three experienced radiologists (25, 27, and 35 years of clinical experience) were asked to give age estimates for the 100 subjects of the test set **B**. These estimates are based on the extracted coronal slices of each subject for both the fat- and water signal image in full resolution. The image intensities were normalized and fixed for all images, so that no manual adjustment of the contrast was performed. Since age estimation based on these images is not typically performed in clinical practice, the radiologists were first shown a histogram of the distribution of ages among all scanned subjects and additionally viewed a set of 50 randomly selected subjects from the training set together with their recorded ground truth age. As part of this preparatory phase they reported and exchanged their observations and were then shown samples of the saliency maps generated by the network, so that they had information about both the decision criteria of each other and the patterns highlighted by the proposed automated method. The radiologists predicted the age on this test set in five-year increments and reported that a more fine-grained manual assessment was not feasible.

TABLE II
RESULTS OF 10-FOLD CROSS-VALIDATION ON DATASET A

Method	MAE ± SD	r	R^2
Main result	2.49 ± 1.90	0.91	0.83
Upper body only	2.70 ± 2.05	0.89	0.80
Lower body only	2.96 ± 2.26	0.87	0.75

* MAE: Mean absolute error, SD: Standard deviation of absolute error, r: Pearson's coefficient of correlation, R^2: Coefficient of determination. Vertical cropping of half the image separates upper and lower body.

III. Results

A. Prediction Performance

A detailed overview over the evaluation of the cross-validation for both the main configuration and the ablation experiments on the training dataset **A** is given in Table II. The repeated runs of the main cross-validation yielded nearly identical results on average, differing by a mean absolute error of less than four days. Table III shows the results on test dataset **B** together with the comparison to human performance and varying amounts of training data. The evaluation on the larger test set **C** is shown in Table IV. Fig. 4 shows a correlation plot in which the individual predictions for the network in cross-validation are compared to the ground truth age. No significant correlation was observed on the test data or in cross-validation between prediction error and subject gender or ethnicity label.

However, the error of the prediction is negatively correlated to the recorded ground truth age, so that there is a bias towards estimating the mean value of the training data. The experiments in which separate instances of the network were trained

TABLE III
Results on Test Dataset B

Method	MAE ± SD	r	R^2
Radiologist A	6.23 ± 4.47	0.47	-0.16
Radiologist B	7.68 ± 5.89	0.20	-0.85
Radiologist C	8.09 ± 5.65	0.29	-0.92
Radiologists Averaged	5.58 ± 3.95	0.45	0.08
Main result (23,120)	2.23 ± 1.60	0.92	0.85
Subset 16,000	2.55 ± 1.84	0.91	0.80
Subset 8,000	2.75 ± 1.98	0.91	0.77
Subset 4,000	2.93 ± 2.06	0.89	0.75
Subset 2,000	2.79 ± 1.86	0.89	0.78
Subset 1,000	3.41 ± 2.40	0.82	0.66
Subset 500	4.73 ± 3.27	0.68	0.35

* MAE: Mean absolute error, SD: Standard deviation of absolute error, r: Pearson's coefficient of correlation, R^2: Coefficient of determination

TABLE IV
Results on Test Dataset C

Method	MAE ± SD	r	R^2
Main result	2.47 ± 1.91	0.91	0.82

* MAE: Mean absolute error, SD: Standard deviation of absolute error, r: Pearson's coefficient of correlation, R^2: Coefficient of determination

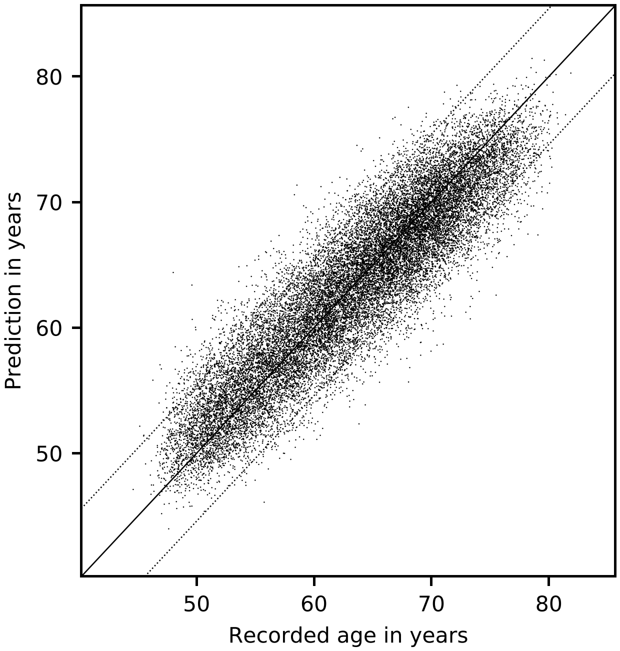

Fig. 4. Neural network predictions (y-axis) as compared to ground truth age (x-axis) in one run of the 10-fold cross-validation. The diagonal line represents a hypothetical perfect match whereas the dotted lines are placed at an offset of two standard deviations of the absolute prediction error.

on subsets of the training data of varying size and evaluated on the test dataset **B** show a clear relationship between the number of unique training samples and the predictive performance. A graph showing the individual performance for these results is shown in Fig. 5. This relationship can be observed both for the fixed evaluation after 80,000 iterations as well as when manually choosing the best-performing iteration, so that it is not merely a side-effect of the network overfitting on smaller training sets. The highest manual predictive performance was obtained when averaging the estimates by the radiologists for each subject, but still did not reach the automated prediction quality.

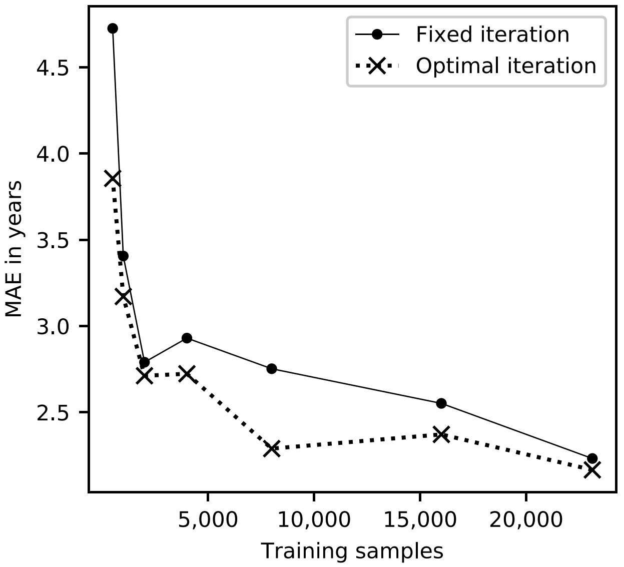

Fig. 5. For seven separately trained instances of the network, the mean absolute error (MAE) on the test set **B** of one hundred subjects clearly decreases as more unique training samples are used. This is seen both when evaluating the network after a fixed amount of 80,000 iterations and when manually choosing the snapshot with optimal performance to avoid confounding effects of overfitting. The shape of the curve indicates that the training process is not saturated even with 23,120 unique training images.

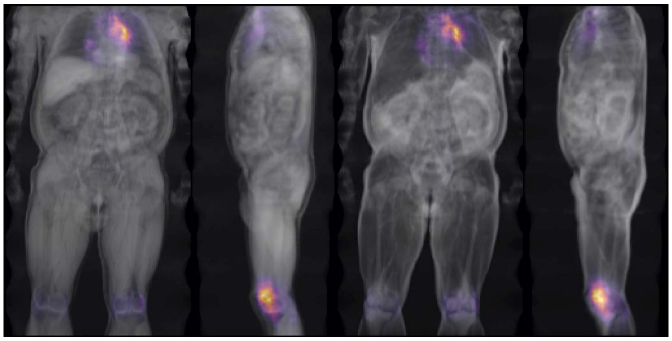

Fig. 6. An aggregated saliency map for the test set **B** clearly highlights the aortic arch and knees. The water (left) and fat (right) signal channel are shown side-by-side.

B. Saliency

The saliency analysis highlighted several patterns corresponding to anatomical structures that affect the prediction of the network. An aggregated saliency map for the test set **B** can be seen in Fig. 6, which was generated by the network trained on all training samples of dataset **A** and thereby achieved the highest performance. The associated one hundred individual saliency maps are not visualized here, but form the basis of the saliency analysis.[2] There are several patterns that reliably

[2] Further visualizations, including examples of individual saliency maps, are included in Fig. 8-10 (Media/Supplementary Material)

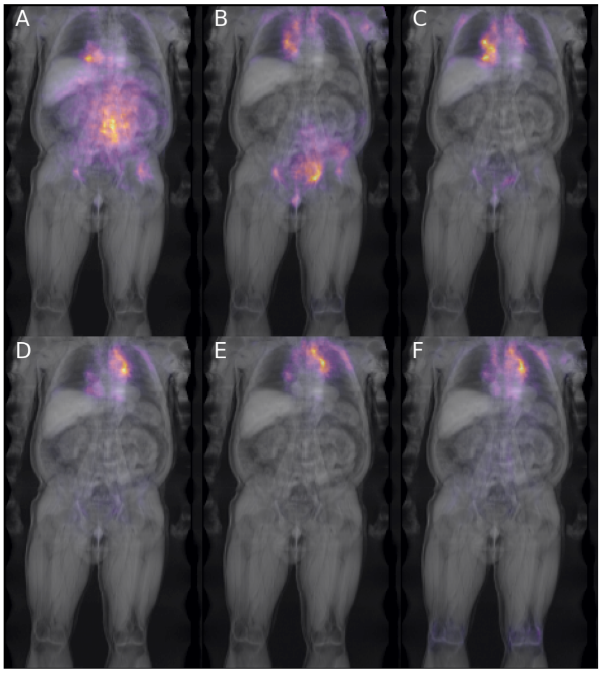

Fig. 7. Aggregated saliency maps for the test set **B** as generated by instances of the network trained on subsets amounting to 500 (A), 1,000 (B), 2,000 (C), 4,000 (D), 8,000 (E), and 16,000 (F) images show how the saliency increasingly targets the aortic arch and knee as the amount of training data and predictive performance increases. Only the coronal projection of the water image is shown.

occur in almost all examined cases and therefore allow for an interpretation on a more general level. The most consistently highlighted structures consist of the knee joint (highlighted in 98 of 100 images) and the aortic arch (in 89 of 100 images) as well as more diffuse streaks that surround the heart and extend into the lungs. In all of those eleven subjects for which the saliency analysis did not highlight any part of the chest there were nonetheless strong activations in the knee. On average, the eleven subjects for whom this is the case were about one standard deviation younger than the mean of the test set **B** as a whole (58 vs 64 years). In all of those images that contain part of the calf (about one in three), the highlighted region extended further down from the knees to include the outline of the tibia and the outer edge of the calf muscle.

Just like the prediction performance itself, the quality of the saliency maps also improved along with the amount of training data. Fig. 7 shows how the aggregated saliency map clears up as more training data is used and initially diffuse markings in the abdominal region give way to increasingly specific highlights around the aortic arch and eventually also the knees. The saliency analysis for the networks that were trained in cross-validation showed a similar pattern, but was only partially evaluated. Registration results for the aggregation of all saliency maps are not yet available at the time of writing and their number was too high for a thorough screening of all individual cases. The inspected saliency maps nevertheless often outlined both the aortic arch and knee. In the ablation experiment when the lower or upper half of the body was cropped out, the networks still showed a strong tendency to mark these regions where included. As compared to the results on the test data the saliency maps within cross-validation were often more diffuse, perhaps as a side-effect of a less accurate prediction.

C. Additional Findings

An examination of the non-image data yielded a weak but noteworthy correlation between systolic blood pressure and age ($r = 0.310$) in the scanned subjects, even without correcting for medication. At the same time, no such correlation was found for diastolic blood pressure ($r = -0.031$). Other correlations of obviously visible properties with age that may have acted as confounding factors, such as height ($r = -0.058$), weight ($r = -0.067$), and sex ($r = 0.090$, with males encoded as 1 and females as 0), proved to be weak and are therefore unlikely to have allowed for an indirect age estimation by the network.

IV. Discussion

The network reached a robust and high performance in cross-validation on the dataset **A**, with consistent quality on the larger test set **C**, and clearly outperformed both the individual and combined estimates by the radiologists on the smaller test dataset **B**. These results show that age causes structural changes in the UK Biobank population in a way that can be both automatically detected and used for an accurate assessment even in the projected and downsampled representation of the whole-body MRI scan. The comparison to the radiologists also indicates that the underlying patterns are not self-evident even to human specialists.

The mean absolute error of 2.23 years on dataset **B** is actually lower than the 2.49 years in cross-validation. This is likely to be influenced by the sample size of just 100 randomly chosen subjects in the test set **B**. However, the error of 2.47 years on the second test **C** is also slightly lower. This indicates that the larger amount of training data available when training on all samples of dataset **A** as compared to only using 90% for each split in cross-validation may play a role as well. This is supported by the proportional relationship between the amount of unique training samples and predictive performance seen in Fig. 5 and it seems likely that the prediction at this level is only possible due to the high amount of training data. The slope of the curve moreover suggests that the training process is not yet saturated and that even better results and more concise saliency maps could be achieved with more training samples. No significant difference in performance was observed between male and female subjects and preliminary experiments that included attempts to train gender-specific networks yielded inferior results. This is likely because of the reduced amount of training data, and it moreover appears that the network is able to differentiate between male and female subjects automatically.

In the manual assessment by the radiologists a wide range of different features was considered relevant. All of these results were achieved without using any of the known age-related changes that occur in the brain, teeth, and wrist, which are not included in the images. As part of the preparatory phase the radiologists reported that their assessments were mostly based on features of the muscle tissue, organs, and skeleton. Higher thigh muscle volume was expected to indicate a more youthful subject, whereas atrophy and fat infiltrations in the thigh as well as the gluteus muscles were assumed to indicate a higher age. In the upper body muscle this relationship appeared less clear. In women, a reduced density of breast tissue followed a similar pattern. Size differences in the kidneys and heart were not considered reliable, but elongated and curved blood vessels such as the aorta were considered a sign of higher age. In the skeleton, bone spurs and degenerative changes in the joints of the hip and spine as well as a thinner bone cortex were considered as probable but less reliable indicators. There was a general tendency to underestimate the age of the oldest patients, who may have an undiminished volume of muscle and kidneys, whereas obesity posed a considerable confounding factor.

The saliency analysis makes it possible to also trace back the criteria used for the automated prediction by the network. The highlighted patterns show a small overlap with the features chosen by the human operators and focus on other structures that manifest age-related changes, most of which have been previously reported in the literature. Not all of these patterns were anticipated by the radiologists, but the neural network was nonetheless able to leverage these features for a prediction that was more accurate than the human assessment. The saliency maps indicate that patterns in the aortic arch and knee play a key role in the automated prediction on the given data. There are supporting findings in the literature for similar populations, in which a systematic increase of defects and thinning of the knee cartilage on subjects aged up to 61 years has been shown [9]. Likewise, the aortic arch has been previously reported to both elongate and widen [22] as well as to undergo changes in its branching pattern [23], both correlated to age in similar demographic groups. The diffuse streaks observed in the saliency maps near the heart may correspond to changes in the blood vessels within the lung and could be detected by the network as a symptom of pulmonary hypertension. This effect has also been previously linked to aging [24] and furthermore matches the observed correlation between age and systolic blood pressure.

The ablation experiments show that the performance suffers when the upper or lower body are isolated by cropping out half of the image. This indicates that complementary information is contained in both halves and using only the aortic arch or the knees alone is not sufficient for an accurate assessment. Whenever visible in the scanned area, the calves also light up strongly in the saliency maps. The highlights around the knee joint typically extend further along the outline of the tibia and also form the outline of the calf muscle. This could mean that the network is taking the volume of the calf muscle and tibia bone cortex into account wherever possible. Previous work has described increasing fat infiltrations in MR images of the calf in an elderly population, but was conducted on only a small number of subjects [25]. The sporadic occurrence of the calf in the field of view makes this pattern difficult to examine. How much information is lost by the exclusion of the lower leg and feet remains an open question.

In eleven of the one hundred test subjects of set **B** no saliency was assigned to the area around the aortic arch. This subgroup is on average one standard deviation younger than the mean, so that this might indicate that age-related changes in this area are specific for higher age. It would be interesting to examine this effect on a larger sample size, such as in the cross-validation. A manual evaluation of tens of thousands of scans is prohibitive, however. Although not attempted here, the aggregation of the saliency maps for the entire study population would open up the potential to robustly quantify effects of this kind, for example by segmenting the template subject and measuring the relative amount of saliency in specific anatomical regions. At the time of writing the full study population has not yet been registered, so that the presented results of the saliency analysis are restricted to the test set **B** only.

When compared to age prediction based on brain images, the presented method accordingly uses not a single organ but complementary information of structures from neck to knee. The relationship between brain-predicted age and the body-predicted age as determined here has yet to be studied. It is possible that the error of both methods could be heavily correlated as a predictor of systemic, premature aging. It would be surprising if the mental disorders associated with brain-age [13] could also be predicted from the body, but there might be equal or greater potential for detection of metabolic and cardiovascular conditions. Another possibility is that premature aging effects in individual anatomical structures are effectively averaged out in the body-predicted age. The error might then be lower, but also less relevant as a potential biomarker.

There are several limitations to the presented results that are worth noting. Most importantly, the prediction was only trained and evaluated for the chosen demographic with an age range of 44-82 years. The prediction is biased towards estimating the mean value of 64 years, which is to be expected due to the choice of the mean squared error as a loss function. However, this also means that for subjects outside of this range, such as adolescents, the predictions would be increasingly inaccurate. Since all underlying data was gathered by the UK Biobank study, there may also be confounding effects unique to British populations that may have affected the prediction. It can not be ruled out that environmental factors correlated with age are represented in the image data. The network may therefore have learned to detect these patterns, which may not occur in populations from other environments. Another limitation consists in the reliability of the recorded human performance. Age assessment based on MRI scans is not a typical part of the work of radiologists and it is possible that with further preparation for this task the accuracy of the manual estimates would have been higher. The radiologists were informed about the patterns in the aortic arch and knee that were identified by the saliency analysis after the preparatory phase. But when examining the test images they reported that these criteria were difficult to use as part of their own assessment. Two possible explanations for this are that either more practice and feedback would have allowed for a better understanding of these visual changes, or alternatively that these patterns are so subtle or complex that they are only suitable for machine-based processing. In either case the fact that the knee appears in the saliency maps only after training on at least 16,000 samples suggests that detecting these changes requires familiarity with a large quantity images for context.

When considering the formatting of the image data, there are countless ways to represent the volumes obtained from the MRI scan. The number of pixel values in the chosen two-dimensional images used as input for the network is less than one percent of the original number of voxels. Likewise, the conversion to an 8bit format strongly compresses the image information and may be a bottleneck for the prediction quality. Even though much of the actual image information may still be preserved and we achieved our best results in this way within acceptable calculation times, there is no guarantee that this configuration and choice of architecture is optimal, and that the images obtained from the scan are used to their full potential. Future hardware and processing methods may be able to reach superior assessments on the same data.

The fact that these results were achieved with less than a third of the total amount of scans planned by the UK Biobank Imaging study also indicates that there is a large potential for future improvements. When taking into account the proportional relationship between the amount of training data and the prediction quality, an age assessment with a mean absolute error below two years might become possible in the future. The basis for this prediction can be visualized and made interpretable to human observers with the help of the saliency analysis. In this work it identified anatomical patterns that are plausible criteria for the estimation of age and support findings previously reported in the literature. However, there is no reason to assume that this strategy is restricted to age prediction alone. Future work will consist in leveraging the growing dataset to study links between the imaged anatomy and other non-image-based properties such as automated measurements, biomarkers, and future outcomes.

V. Conclusion

The presented machine learning approach was able to capture age-related changes in the UK Biobank MRI scans of the body and use them for an accurate, automated age assessment. For the examined subjects it achieved a mean absolute error below two and a half years and clearly outperformed three human radiologists on a small test dataset. The saliency analysis used to explain these findings highlights the aortic arch and knees as primary indicators of age. This is especially visible in the aggregated saliency map which allows for an interpretation on a larger scale that could cover an entire study population. This type of assessment and analysis is not restricted to age estimation and could help to uncover new associations between morphology and non-image-based information when applied to other labels. As the amount of data and predictive power increases, interpretable saliency can contribute to human understanding of these associations.

Acknowledgment

This research has been conducted using the U.K. Biobank Resource under application no. 14237. The authors would like to thank Dr. F. Malmberg for his support in configuring the whole-body registration pipeline.

References

[1] C. Sudlow *et al.*, "UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.*, vol. 12, Mar. 2015, Art. no. e1001779.

[2] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, "The hallmarks of aging," *Cell*, vol. 153, pp. 1194–1217, Jun. 2013.

[3] D. J. Lowsky, S. J. Olshansky, J. Bhattacharya, and D. P. Goldman, "Heterogeneity in healthy aging," *J. Gerontol. A, Biol. Sci. Med. Sci.*, vol. 69, pp. 640–649, Jun. 2014.

[4] J. Jylhävä, N. L. Pedersen, and S. Hägg, "Biological age predictors," *EBioMedicine*, vol. 21, pp. 29–36, Jul. 2017.

[5] S. Horvath and K. Raj, "DNA methylation-based biomarkers and the epigenetic clock theory of ageing," *Nature Rev. Genet.*, vol. 19, pp. 371–384, Jun. 2018.

[6] R. Rothe, R. Timofte, and L. Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Santiago, Chile, Dec. 2015, pp. 252–257.

[7] M. B. Bjørk and S. I. Kvaal, "CT and MR imaging used in age estimation: A systematic review," *J. forensic Odonto-Stomatol.*, vol. 36, no. 1, pp. 14–25, 2018.

[8] S. S. Halabi *et al.*, "The RSNA pediatric bone age machine learning challenge," *Radiology*, vol. 290, pp. 498–503, Feb. 2019.

[9] C. Ding, F. Cicuttini, F. Scott, H. Cooley, and G. Jones, "Association between age and knee structural change: A cross sectional MRI based study," *Ann. Rheumatic Diseases*, vol. 64, pp. 549–555, Apr. 2005.

[10] K. Franke, G. Ziegler, S. Klöppel, C. Gaser, and A. D. N. Initiative, "Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters," *Neuroimage*, vol. 50, no. 3, pp. 883–892, Apr. 2010.

[11] J. H. Cole *et al.*, "Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker," *NeuroImage*, vol. 163, pp. 115–124, Dec. 2017.

[12] J. H. Cole *et al.*, "Brain age predicts mortality," *Mol. Psychiatry*, vol. 23, pp. 1385–1392, May 2018.

[13] S. Shahab *et al.*, "Brain structure, cognition, and brain age in schizophrenia, bipolar disorder, and healthy controls," *Neuropsychopharmacology*, vol. 44, no. 5, pp. 898–906, 2019.

[14] J. West *et al.*, "Feasibility of MR-based body composition analysis in large scale population studies," *PLoS ONE*, vol. 11, Sep. 2016, Art. no. e0163332.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[18] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[20] U. Ozbulak, "PyTorch CNN visualizations," in *GitHub Repository*. San Francisco, CA, USA: GitHub, 2019. [Online]. Available: https://github.com/utkuozbulak/pytorch-cnn-visualizations

[21] S. Ekström, F. Malmberg, H. Ahlström, J. Kullberg, and R. Strand, "Fast graph-cut based optimization for practical dense deformable registration of volume images," Oct. 2018, *arXiv:1810.08427*. [Online]. Available: https://arxiv.org/abs/1810.08427

[22] A. Redheuil *et al.*, "Age-related changes in aortic arch geometry," *J. Amer. College Cardiol.*, vol. 58, pp. 1262–1270, Sep. 2011.

[23] A. Kojima and I. Saga, "Effect of aging on the configurational change of the aortic arch," *Geriatric Care*, vol. 2, no. 1, pp. 1–4, Apr. 2016.

[24] G. Berra, S. Noble, P. M. Soccal, M. Beghetti, and F. Lador, "Pulmonary hypertension in the elderly: A different disease?" *Breathe*, vol. 12, pp. 43–49, Mar. 2016.

[25] N. F. Schwenzer *et al.*, "Aging effects on human calf muscle properties assessed by MRI at 3 Tesla," *J. Magn. Reson. Imag.*, vol. 29, no. 6, pp. 1346–1354, 2009.