Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Deep learning for the prediction of type 2 diabetes mellitus from neck-to-knee Dixon MRI in the UK biobank

Christian Wachinger[a,b,c,*], Tom Nuno Wolf[a,c], Sebastian Pölsterl[b]

[a] Department of Radiology, Technical University of Munich, Klinikum Rechts der Isar, Ismaningerstr. 22, 81675, München, Germany
[b] Lab for Artificial Intelligence in Medical Imaging, Department of Medicine, LMU Klinikum, Germany
[c] Munich Center for Machine Learning (MCML), Germany

ARTICLE INFO

Keywords:
Diabetes mellitus
Type 2
Deep learning
Whole-body magnetic resonance imaging
UK biobank imaging

ABSTRACT

Rationale and objectives: We evaluate the automatic identification of type 2 diabetes from neck-to-knee, two-point Dixon MRI scans with 3D convolutional neural networks on a large, population-based dataset. To this end, we assess the best combination of MRI contrasts and stations for diabetes prediction, and the benefit of integrating risk factors.
Materials and methods: Subjects with type 2 diabetes mellitus have been identified in the prospective UK Biobank Imaging study, and a matched control sample has been created to avoid confounding bias. Five-fold cross-validation is used for the evaluation. All scans from the two-point Dixon neck-to-knee sequence have been standardized. A neural network that considers multi-channel MRI input was developed and integrates clinical information in tabular format. An ensemble strategy is used to combine multi-station MRI predictions. A subset with quantitative fat measurements is identified for comparison to prior approaches.
Results: MRI scans from 3406 subjects (mean age, 66.2 years ± 7.1 [standard deviation]; 1128 women) were analyzed with 1703 diabetics. A balanced accuracy of 78.7 %, AUC ROC of 0.872, and an average precision of 0.878 was obtained for the classification of diabetes. The ensemble over multiple Dixon MRI stations yields better performance than selecting the individually best station. Moreover, combining fat and water scans as multi-channel inputs to the networks improves upon just using single contrasts as input. Integrating clinical information about known risk factors of diabetes in the network boosts the performance across all stations and the ensemble. The neural network achieved superior results compared to the prediction based on quantitative MRI measurements.
Conclusions: The developed deep learning model accurately predicted type 2 diabetes from neck-to-knee two-point Dixon MRI scans.

1. Introduction

Diabetes mellitus is a metabolic disorder with an estimated 463 million people suffering from diabetes in 2019 and a predicted increase to 700 million by 2045 [1]. Diabetes has a major impact on the lives and well-being of individuals and societies. In this study, we focus on type 2 diabetes mellitus (T2DM), which is the dominant form with roughly 90 % of cases and the driver of the worldwide

[*] Corresponding author. Department of Radiology, Technical University of Munich, Klinikum Rechts der Isar, Ismaningerstr. 22, 81675, München, Germany.
E-mail address: Christian.wachinger@tum.de (C. Wachinger).

https://doi.org/10.1016/j.heliyon.2023.e22239
Received 24 April 2022; Received in revised form 7 November 2023; Accepted 7 November 2023
Available online 10 November 2023

Abbreviations

CNN	convolutional neural network
LR	logistic regression
PDFF	proton density fat fraction
UKB	UK Biobank
ROC AUC	area under the receiver operating curve
SAT	subcutaneous adipose tissue
T2DM	type 2 diabetes mellitus
VAT	visceral adipose tissue

diabetes epidemic [2,3].

MRI provides new opportunities for developing diagnostic and prognostic tools for diabetes [4]. Pathological changes in body tissue can directly be displayed by MRI, and body composition can be assessed with the display of ectopic fat accumulation [5]. While the specific role of various local fat depots in the development of diabetes is not yet fully understood, previous studies have reported that abdominal visceral adipose tissue (VAT), as well as fat in the liver and pancreas, are linked to diabetes [6–9]. The potential to extract quantitative fat measurements enabled the classification of diabetes from MRI with VAT and hepatic fat measures [10] and with radiomics features from the liver [11]. Such diagnostic tools could be used to identify undiagnosed asymptomatic diabetics, monitor progression [10], and supplement existing tests, which have been reported to be sensitive to confounding factors [12,13]. If MRI is performed for screening purposes, the diabetes prediction could be run as an additional test, without additional costs, without fasting, and non-invasively. In such a scenario, whole-body MRI screening would not be performed solely for diagnosing diabetes, but it would be one test in a whole range of tests that can be run on such a comprehensive characterization of human morphology.

A common basis for imaging studies of diabetes is chemical shift-encoded MRI, which enables the separation of fat and muscle compartments. Thanks to advances in MRI acquisition, two-point Dixon imaging has recently been integrated into population-imaging studies like the UK Biobank (UKB) Image Enhancement with a targeted 100,000 subjects [14] and the German national cohort [15]. Neck-to-knee imaging captures the major structures affected by diabetes: endocrine organs like kidneys, liver and pancreas, adipose tissue, and the cardiovascular system. Previous water-fat studies in UKB and other datasets have focused on extracting specific quantitative measures relevant to diabetes, e.g., VAT and liver fat [6,16]. However, the direct application of the entire neck-to-knee Dixon sequence for studying diabetes has not yet been investigated.

In this work, we propose the prediction of diabetes from whole-body MRI scans with deep learning. Our end-to-end learning approach directly learns to extract features from the entire image sequence that help classify diabetes. To have enough data for training complex neural networks, we leverage a large number of images in UKB, providing a comprehensive training and evaluation resource. In a careful evaluation that avoids data leakage and bias, we demonstrate that the resulting network predicts diabetes with a balanced accuracy of 78.7 % and outperforms alternative approaches that rely on quantitative Dixon measures. Essential for the design of the network was the combination of multiple MRI contrasts from several stations and the integration of clinical information for a holistic patient characterization.

2. Materials and Methods

2.1. Data

The UK Biobank imaging enhancement is a population-based, prospective study that performs MRI of 100,000 subjects [14], where our current sample contains 40.046 subjects (Fig. 1). The abdominal MRI protocol includes a six station Dixon sequence that covers the body from neck-to-knee. The acquisition yields four contrasts: water (W), fat (F), in-phase (IN), and opposed-phase (OPP). Subjects were scanned with a Siemens Aera 1.5T scanner (Siemens, Erlangen, Germany) with a dual-echo Dixon Vibe protocol in a supine position with the arms along the sides. The six stations of the Dixon protocol cover a total of 1.1 m with axial 3D spoiled gradient dual-echo images. The water-fat Dixon images were reconstructed with the integrated scanner software. Typical parameters for all scans were: TR = 6.69 ms, TE = 2.39/4.77 ms, and bandwidth 440 Hz. The images were resampled to an isotropic resolution of 4.5 mm^3. Fig. 2 shows all six stations of abdominal water and fat scans. The image processing is described in supplement S.1.

We identified all subjects with T2DM following the procedure outlined in Ref. [6] using participant information from the imaging visit, yielding 1703 subjects (Fig. 1). To avoid confounding bias, we created a control group that matches age, age squared,[1] sex, and site of the diabetic group. For this purpose, we applied the function match. it in the software package R [17], which uses nearest neighbor matching on the propensity score estimated using a logistic regression of the treatment on the covariates [18]. The outlined procedure will lead to conservative estimates of prediction accuracy, as our matching procedure limits the possibilities of shortcut learning. We further excluded subjects with type 1 diabetes and cancer. Table 1 reports demographics and clinical measurements of the entire sample and both subgroups. For the comparison to alternative approaches based on visceral adipose tissue (VAT), abdominal

[1] Age-squared is used because the risk of T2D is not necessarily linearly associated with age.

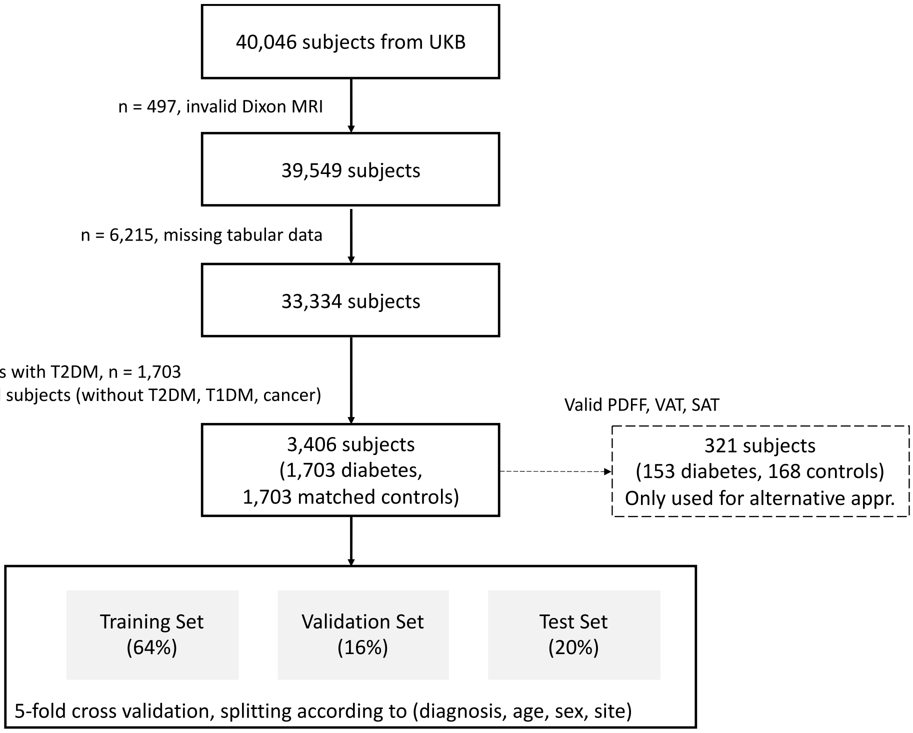

40,046 subjects from UKB
n = 497, invalid Dixon MRI
39,549 subjects
n = 6,215, missing tabular data
33,334 subjects
Patients with T2DM, n = 1,703
Control subjects (without T2DM, T1DM, cancer)
3,406 subjects
(1,703 diabetes,
1,703 matched controls)
Valid PDFF, VAT, SAT
321 subjects
(153 diabetes, 168 controls)
Only used for alternative appr.
Training Set
(64%)
Validation Set
(16%)
Test Set
(20%)
5-fold cross validation, splitting according to (diagnosis, age, sex, site)

Fig. 1. Flowchart diagram visualizes the organization of our dataset from the UKB. Subjects with invalid Dixon MRI scans and missing tabular variables of interest were excluded. Subjects with T2DM were selected and a matching control group was created to avoid confounding bias. The resulting dataset was split into training, validation, and test set using stratified sampling. The results from this cross-validation are the main results reported in the article. Only for the comparison with alternative approaches in Table 2, subjects with valid PDFF, VAT, and SAT values were selected.

subcutaneous adipose tissue (SAT), and liver proton density fat fraction (PDFF), we extracted a subset of participants that have those values available, see supplement S.2. This research has been conducted using the UKB resource, project ID 34479. The study was approved by the North West Multicenter Research Ethics Committee, UK (approval number: 11/NW/0382). Written informed consent was obtained from all subjects before study entry.

2.2. Neural network

We use a 3D convolutional neural network (CNN) to predict binary diabetes status. The multiple contrasts from the Dixon sequence form the multi-channel input. The network consists of convolution layers, followed by batch norm, ReLU, and downsampling. Fig. 3 illustrates the network architecture with details about each layer's parameters.

As clinical information to the network, we consider typical risk factors: age, sex, blood pressure, cholesterol medication, alcohol intake, smoking, and BMI. We consider two strategies to integrate clinical information that comes in tabular format. First, the concatenation of tabular data with image features after global average pooling (CONCAT). Second, applying the recently introduced Dynamic Affine Feature Map Transform (DAFT, Fig. 3) [19]. It is a general-purpose module for CNNs that incites or represses high-level concepts learned from a 3D image by conditioning feature maps of a convolutional layer on both a patient's image and tabular clinical information. The combination is achieved by using an auxiliary neural network that outputs a scaling factor and an offset to dynamically apply an affine transformation to the feature maps of a convolutional layer. The included risk factors can lead to anatomical changes visible in MRI scans. Hence, they can be confounders for the classification. By including them in the network, we are controlling for such confounding. The network has 1.384.861 parameters, and details about network training are reported in supplement S.3.

For the combination of MRI stations, we consider two strategies. First, we stitch the information from stations to create an image with an increased field-of-view, which is then used as input to the network (see S.1). Second, we use an ensemble strategy, where individual networks are trained for each station separately, and the prediction results are then combined. In particular, we consider a probability-weighted voting strategy across stations, where the probabilities are the output from the softmax layer.

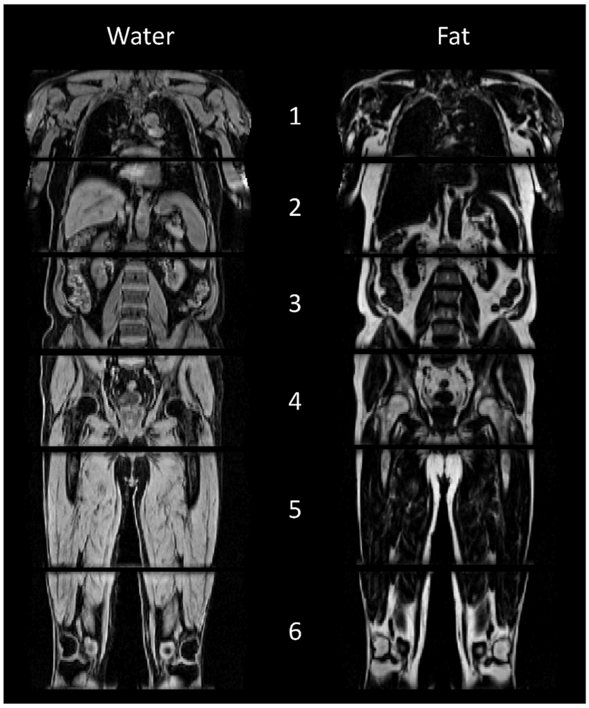


Fig. 2. Visualization of the six stations for the Dixon MRI scans for water (left) and fat (right) contrasts. Note that adjacent stations are overlapping and therefore partially display the same anatomy.

Table 1
Subject demographics and clinical measures (BP: blood pressure, LDL: low-density lipoprotein, HDL: high-density lipoprotein).

		All	Control	Diabetes	*P*-Value
n		3406	1703	1703	
Age, mean (SD)		66.2 (7.1)	66.2 (7.1)	66.2 (7.1)	1.000
Sex, n (%)	Female	1128 (33.1)	564 (33.1)	564 (33.1)	1.000
	Male	2278 (66.9)	1139 (66.9)	1139 (66.9)	
BMI, mean (SD)		27.5 (4.9)	25.9 (4.0)	29.1 (5.2)	<0.001
Smoking, n (%)	Never	1966 (57.7)	1038 (61.0)	928 (54.5)	0.001
	Previous	1320 (38.8)	607 (35.6)	713 (41.9)	
	Current	120 (3.5)	58 (3.4)	62 (3.6)	
Alcohol, n (%)	Never	305 (9.0)	113 (6.6)	192 (11.3)	<0.001
	Special occasions only	420 (12.3)	149 (8.7)	271 (15.9)	
	Once or twice a week	821 (24.1)	391 (23.0)	430 (25.2)	
	One to three times a month	420 (12.3)	179 (10.5)	241 (14.2)	
	Three or four times a week	902 (26.5)	560 (32.9)	342 (20.1)	
	Daily or almost daily	538 (15.8)	311 (18.3)	227 (13.3)	
Diastolic BP, mean (SD)		78.4 (9.8)	78.7 (9.9)	78.2 (9.6)	0.162
Systolic BP, mean (SD)		141.3 (17.8)	139.9 (18.1)	142.6 (17.3)	<0.001
LDL, mean (SD)		3.4 (0.9)	3.6 (0.8)	3.2 (0.9)	<0.001
HDL, mean (SD)		1.3 (0.3)	1.4 (0.4)	1.2 (0.3)	<0.001
Triglycerides, mean (SD)		2.0 (1.2)	1.7 (1.0)	2.3 (1.3)	<0.001
Cholesterol med, n (%)	No	1711 (50.2)	1220 (71.6)	491 (28.8)	<0.001
	Yes	1695 (49.8)	483 (28.4)	1212 (71.2)	
BP medication, n (%)	No	1933 (56.8)	1245 (73.1)	688 (40.4)	<0.001
	Yes	1473 (43.2)	458 (26.9)	1015 (59.6)	

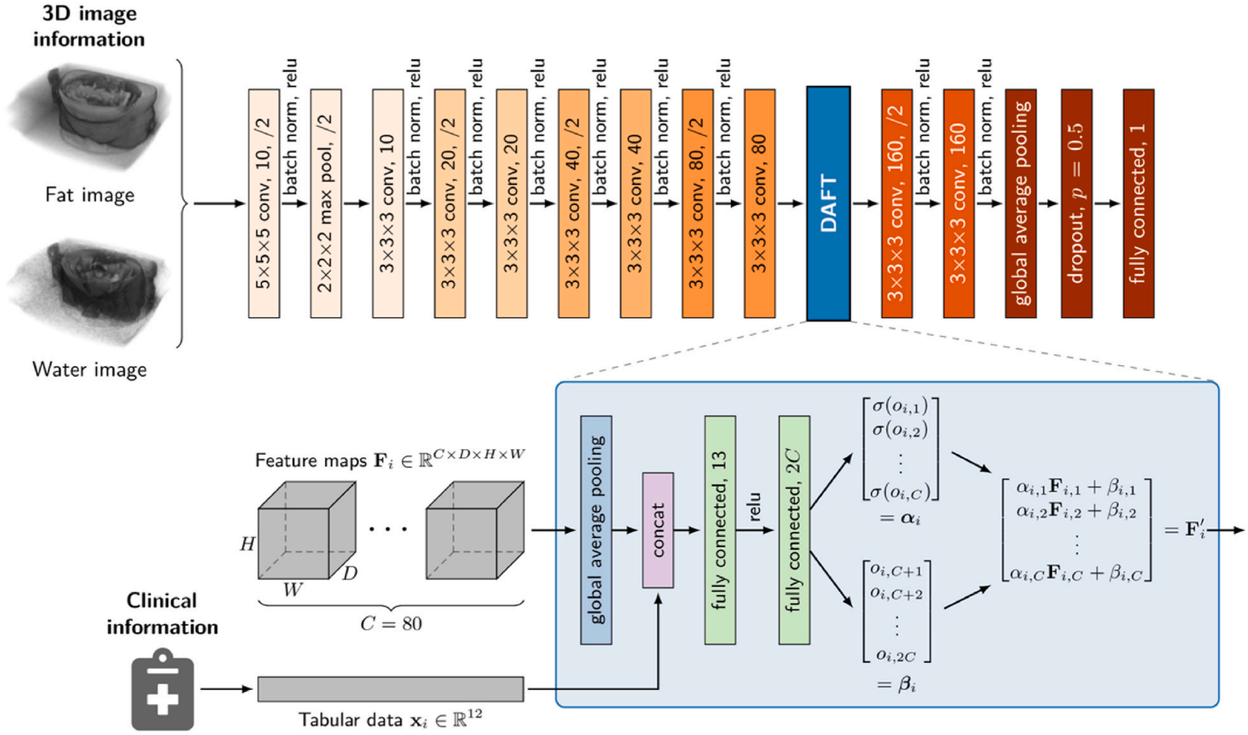


Fig. 3. Convolutional neural network (CNN) for the image-based prediction. Shown is the integration of clinical information as tabular data with the Dynamic Affine Feature Map Transform (DAFT). Feature maps are transformed subject to tabular data $\mathbf{x}_i$ from subject *i*. Multiple image contrasts are input as multi-channel, illustrated are fat and water scans. The clinical variables considered are age, sex, blood pressure, cholesterol medication, alcohol intake, smoking, and BMI.

2.3. *Statistical analysis*

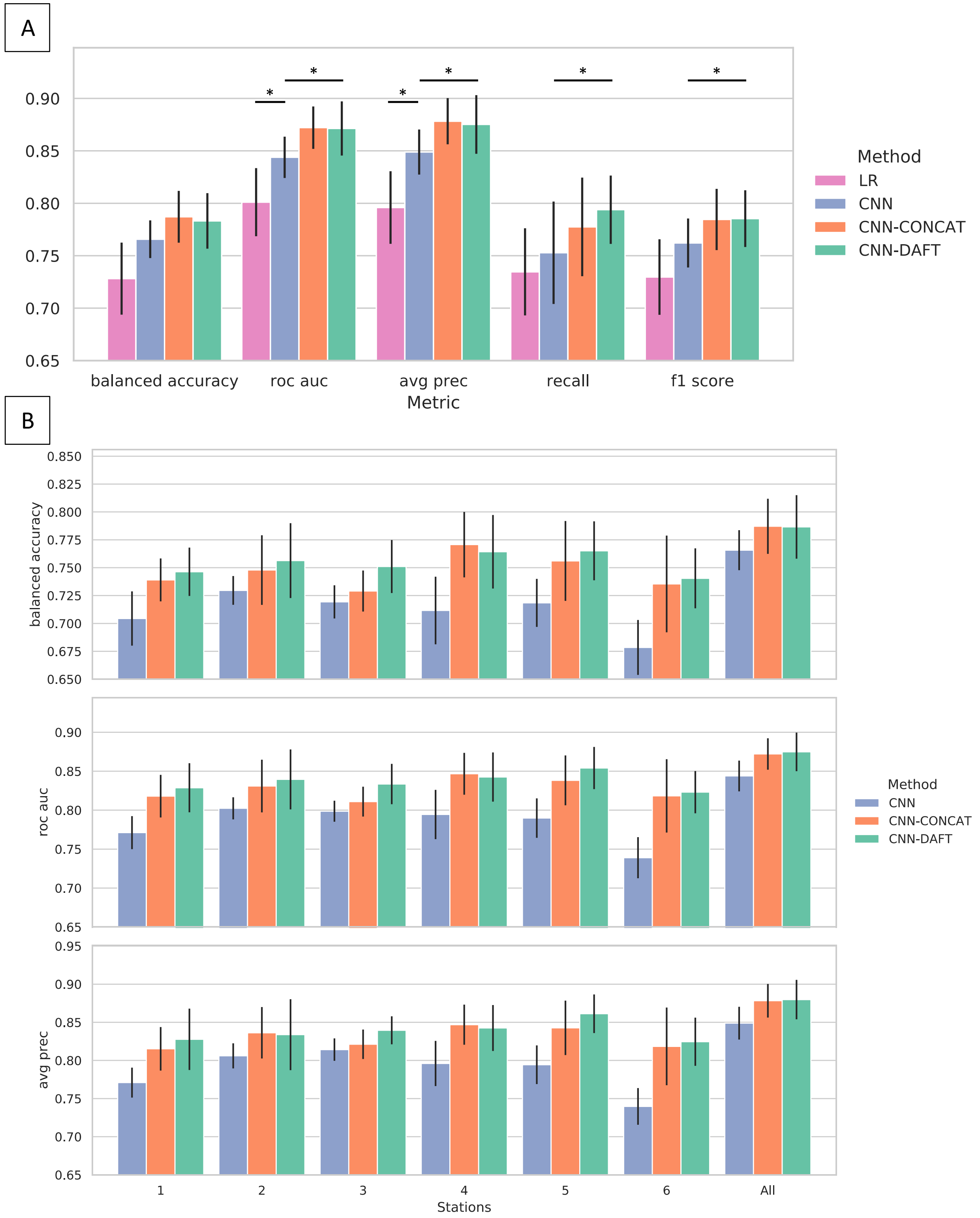
We compute balanced accuracy, the area under the receiver operating characteristic curve (ROC AUC),[2] average precision, F1-score, and recall to evaluate the classification performance. Data leakage and confounding effects due to age and sex must be considered carefully to avoid biased evaluation results. Hence, we use a modified five-fold cross-validation that considers confounders. To this end, we split the data into five non-overlapping folds such that diagnosis, age, sex, and site are balanced across folds. We assess the balance of a split by computing the propensity score, i.e., the probability of a sample belonging to the training data. Next, we compare the percentiles of the propensity score distribution in the training and test data and use the maximum deviation across all percentiles as a measure of the imbalance [18]. For each of the five splits, this process is repeated for 1000 randomly selected partitions, and the partition with the minimum imbalance is ultimately the selected split.[3] Once the five splits are created, we proceed with standard five-fold cross-validation, where 20 % of the data are used for testing in each round. The remaining 80 % of the data are again split 80/20 for training and validation, resulting in 64 % for training and 16 % for validation. The correlated *t*-test was used for performance comparison; a p-value of less than 0.05 was considered to indicate a significant difference.

3. Results

1703 MRI scans from 1703 diabetic subjects were included in this study (mean age 66.2 years ± 7.1 [standard deviation]; 564 women), see Table 1. A matched sample of non-diabetic subjects with identical demographic characteristics was selected, yielding a total sample of 3406 MRI scans from 3406 subjects (66.2 years ± 7.1; 1128 women). Fig. 1 shows the flowchart for subject selection. The results in sections 3.1–3.3 are from five-fold cross-validation. In section 3.4, the subset with PDFF, VAT, and SAT measurements was used as the test set.

[2] The receiver operating characteristic (ROC) curve illustrates the diagnostic ability of a binary classifier when its discrimination threshold is varied. AUC measures the area underneath the entire ROC curve.

[3] As the test set is balanced with respect to diagnosis, balanced accuracy is identical to accuracy.



(caption on next page)

Fig. 4. **A:** Comparison of performance for diabetes prediction. Logistic regression (LR) uses only clinical information for prediction. CNN uses only fat and water images for the prediction. CNN-CONCAT and CNN-DAFT use the combination of image and clinical data with concatenation and DAFT, respectively. All image-based methods use an ensemble across all six stations. **B:** Diabetes prediction performance from fat and water images (multi-channel). Results are shown for six individual stations and the ensemble across all stations. Reported metrics are balanced accuracy, the area under the receiver operating curve (ROC AUC), average precision (AVG PREC), recall and f1 score. Bars show the mean, lines show the standard deviation. Bars show the mean, lines show the standard deviation. * indicates a significant improvement ($p < 0.05$) in A for pairwise comparisons between LR – CNN, CNN–CNN-CONCAT, and CNN – CNN-DAFT.

3.1. Prediction of diabetes status

Fig. 4A shows the diabetes prediction results for logistic regression (LR) with clinical information only, image-based classification with CNNs, and the integration of clinical information in CNNs with CONCAT and DAFT, respectively. DAFT and CONCAT have an average balanced accuracy of 78.3 % and 78.7 %, respectively. The balanced accuracy for CNN is 76.6 % and 72.8 % for LR. The improvement of the CNN with respect to LR is significant for AUC (LR: 0.801, CNN: 0.844, $p=0.035$) and precision (LR: 0.796, CNN: 0.849, $p=0.010$). These results demonstrate that for the used classifiers, whole-body images contain more diabetes-related information than the selected clinical variables. Further, the combination of image and clinical data yields a further improvement, which is significant for DAFT for AUC (CNN:0.844, DAFT: 0.871, $p=0.042$), precision (CNN: 0.849, DAFT: 0.875, $p=0.046$), recall (CNN: 0.753, DAFT: 0.794, $p=0.027$), and F1-score (CNN: 0.762, DAFT: 0.785, $p=0.011$). This indicates that the clinical information is complementary to the diabetes-related image information. The improvement of DAFT to LR is significant for all five metrics.

In Fig. 4B, we compare the prediction results for the individual MRI stations and the ensemble strategy results. We observe that DAFT and CONCAT yield a clear improvement over CNN across all stations. The largest improvement in accuracy is for station 5, where image information alone has the weakest performance. The results for CONCAT and DAFT are comparable, where DAFT has a slight advantage on 5 out of 6 stations. The largest improvement of DAFT in comparison to CONCAT is for station 3. For the ensemble across all 6 stations, the apparent improvement over the purely image-based network persists, where the results for CONCAT and DAFT are comparable.

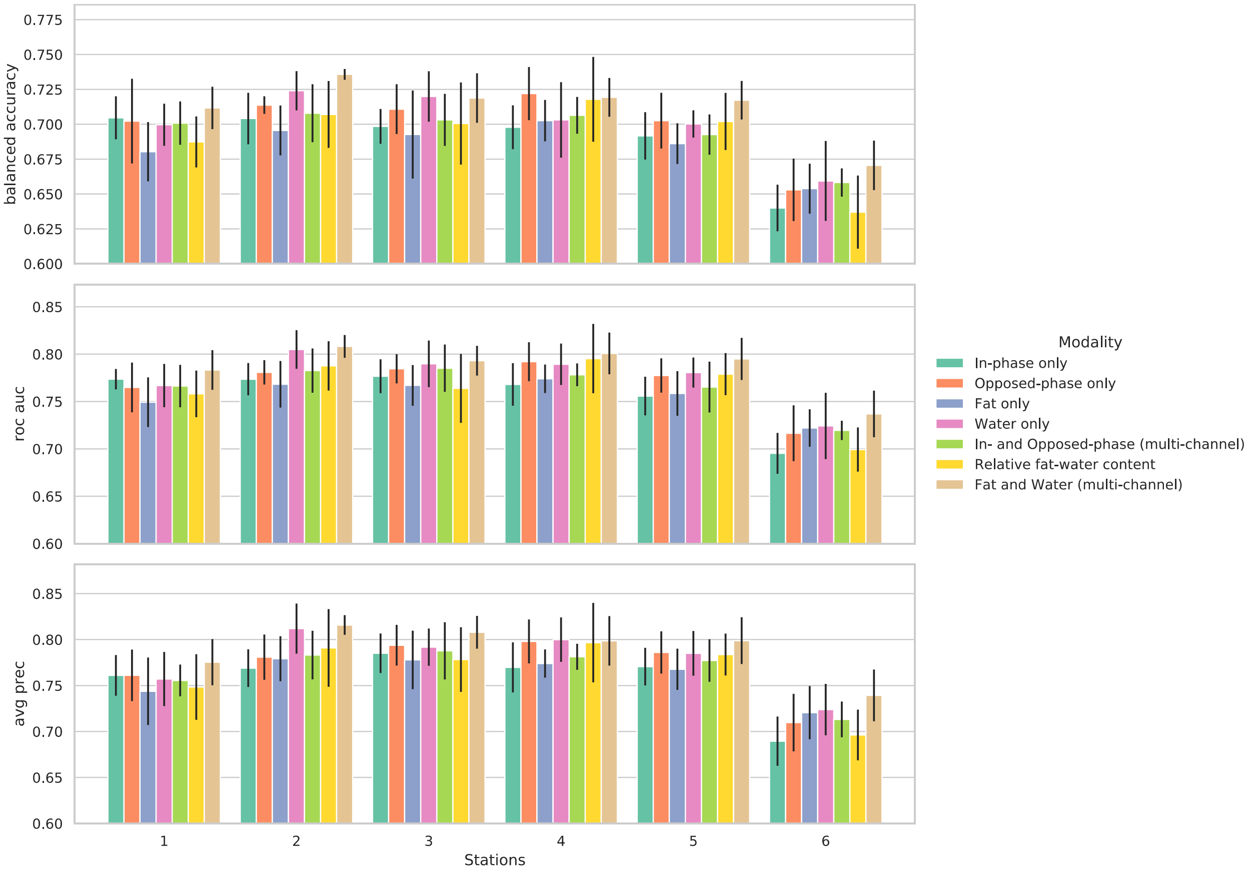


Fig. 5. Performance of the neural network for diabetes prediction by stations (1–6) and image contrasts. Reported metrics are balanced accuracy (top panel), the area under the receiver operating curve (ROC AUC; middle panel), and average precision (AVG PREC; bottom panel). Bars show the mean, and lines show the standard deviation.

In supplementary Figure SF1, we evaluate the impact of data augmentation during training. Augmentation yields a clear improvement in prediction performance across all metrics.

The inference time for one scan is 2 ms. Multiple contrasts are handled as multiple input channels, yielding no notable increase in runtime. For the multi-stage approach, the inference is performed for each stage, so the total runtime is about 12 ms.

3.2. Evaluation of imaging contrasts

Fig. 5 shows the diabetes prediction results for different image contrasts and MRI stations. Next to each of the four contrasts, we evaluate multi-channel input by combining in- and opposed-phase scans or fat and water scans. Further, we evaluate the performance using a normalized image representing relative fat-water content, as used in Ref. [11].

The performance for stations 1 to 5 is similar, with a balanced accuracy between 70 % and 73 % for most contrasts. The accuracy for station 6, which covers parts around the knee, is distinctively lower than for the other stations (around 65 %). The ROC AUC and average precision values show a similar tendency to balanced accuracy.

In comparing the individual image contrasts, the highest accuracy, AUC, and precision are obtained by water images from station 2. The generally good performance from water scans is followed by opposed-phase scans. Combining in- and opposed-phase scans as

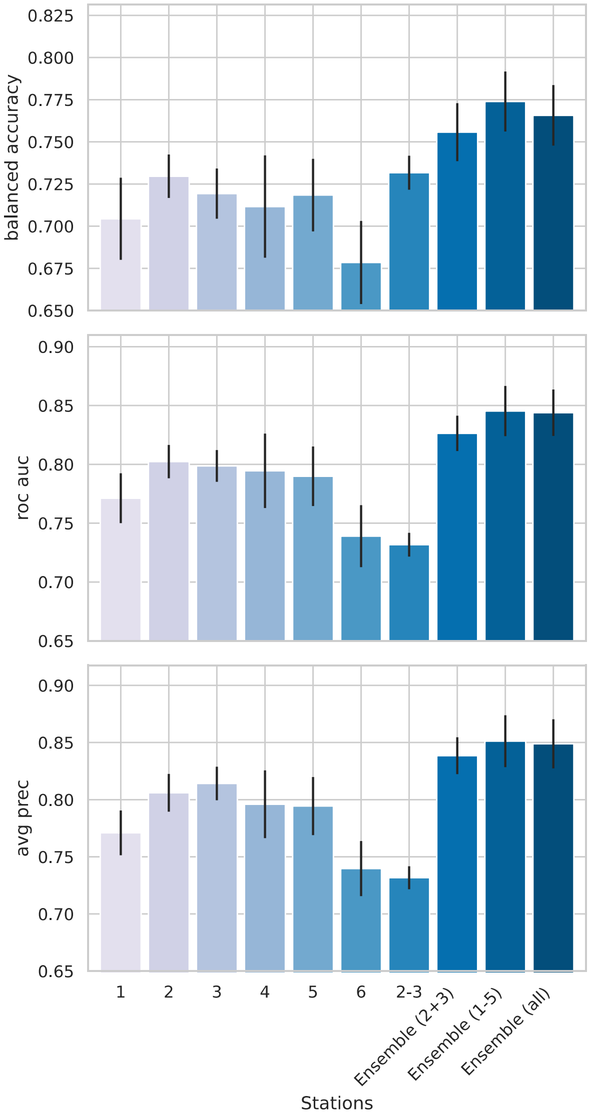


Fig. 6. Performance for diabetes prediction from fat and water images (multi-channel). Results for prediction from individual stations (1–6), for the stitching of stations 2 and 3 (2–3), for ensembling stations 2 and 3, for ensembling stations 1 to 5, and for ensembling all stations. Reported metrics are balanced accuracy (top panel), the area under the receiver operating curve (ROC AUC; middle panel), and average precision (AVG PREC; bottom panel). Bars show the mean, and lines show the standard deviation.

multi-channel inputs does not surpass the performance of solely using opposed-phase scans. Also, scans with relative fat-water content do not yield a performance improvement. In contrast, combining fat and water scans as multi-channel input improves the individual performances of fat and water scans. The combination of fat and water on station 2 achieves the best overall performance.

3.3. *Evaluation of combining multiple stations*

Fig. 6 shows the results of combining multiple stations of fat and water scans. We consider two strategies for combining stations. First, we stitch scans from stations 2 and 3 into an extended image. These stations achieve the highest prediction accuracy on fat and water in Fig. 5 and cover many abdominal organs (liver, kidneys, pancreas) as well as VAT and SAT depots. As a second strategy, we train independent networks for each station and then combine the results with an ensemble voting strategy. We evaluate the voting on stations 2 and 3 as a direct comparison to the stitching strategy. We further evaluate voting of stations 1 to 5, and voting across all 6 stations. Stitching across all 6 stations is not evaluated as it would result in very large images that would complicate network training. The results in Fig. 6 demonstrate that stitching stations 2 and 3 does not yield an increase in accuracy over station 2, and AUC, as well as precision, are decreasing. In contrast, the ensemble over stations leads to a clear improvement across all three metrics. The improvement is even more pronounced when computing the ensemble over more stations, as demonstrated by the results for voting across stations 1 to 5 and all 6 stations. Dropping station 6, the station with the lowest individual performance yields a slight performance increase over using all 6 stations.

3.4. *Comparison of prediction with quantitative measurements*

Table 2 reports diabetes classification results with logistic regression for several combinations of independent variables. As image-derived, quantitative measurements, we use PDFF, VAT, and SAT, which previously have been used for diabetes classification [10,11,20]. As these measurements were only available for some of the participants, we worked on a subset of data in this section, also shown in Fig. 1. Next to these variables, we also consider BMI as an anthropometric marker used in Ref. [11]. Furthermore, we evaluate the addition of demographic information, medication (cholesterol, blood pressure), smoking, and alcohol consumption. Finally, we report the performance of CNN-DAFT on the same test set but use a larger training set to have enough data for training.

Among the logistic regression models, the highest accuracy and AUC are achieved for model 7, which combines all variables. Surprisingly, PDFF alone (model 1) performs better than PDFF combined with VAT, SAT, BMI, and demographics (models 2–5). Further, PDFF performs similarly to model 6, which combines non-image related measures. Model 4 has the highest average precision. Model 6 is identical to the logistic regression model reported in previous results. Consistent with our previous results about the combination of image and clinical data, we note that the inclusion of image-based measures in model 7 (PDFF, VAT, SAT) yields an increase in accuracy and AUC compared to model 6.

The performance of CNN-DAFT leads to the overall best results with a large margin to the other methods across all metrics. The clinical variables in CNN-DAFT are identical to those in model 6.

We have further evaluated alternative classifiers than logistic regression. In Supplementary Table 1, we also report the results for a Random Forest classifier and an XGBoost classifier. Logistic regression achieved superior performance for the wide majority of models and metrics, justifying its use in Table 2.

4. Discussion

Our findings demonstrated that neural networks can accurately predict the diabetes status from Dixon whole-body MRI scans. The automatic feature extraction on the whole sequence with an end-to-end learning approach led to better results than operating on quantitative image measurements, as shown in Fig. 4. To train the CNN, we have used a large amount of MRI data available in UKB. In addition, the dataset provided a detailed characterization of the participants and a precise description of body fat distributions. For the

Table 2

Prediction results with logistic regression for different combinations of independent variables and CNN-DAFT. The experiments are performed on a dataset subset containing PDFF, VAT, and SAT values. PDFF values are log transformed to account for skewness of the variable. For medication, we consider medication for cholesterol and blood pressure. CNN-DAFT uses a larger dataset for training. Mean and standard deviation for balanced accuracy, ROC AUC, and average precision are reported. Best results are highlighted in bold face. PDFF: Proton Density Fat Fraction; VAT: Visceral Adipose Tissue; SAT: Subcutaneous Adipose Tissue; CNN: Convolutional Neural Network.

	Independent variables	Balanced Accuracy		ROC AUC		Average Precision	
		Mean	Std	Mean	Std	Mean	Std
1	PDFF	0.672	0.051	0.696	0.063	0.727	0.049
2	VAT, SAT	0.624	0.059	0.680	0.070	0.677	0.032
3	BMI	0.620	0.025	0.667	0.061	0.641	0.057
4	PDFF, VAT, SAT, BMI	0.661	0.070	0.716	0.068	**0.742**	0.048
5	PDFF, VAT, SAT, BMI, Age, Sex	0.665	0.077	0.713	0.069	0.736	0.047
6	BMI, Age, Sex, Medication, Smoking, Alcohol	0.673	0.062	0.733	0.066	0.691	0.051
7	PDFF, VAT, SAT, BMI, Age, Sex, Medication, Smoking, Alcohol	**0.695**	0.054	**0.763**	0.067	0.739	0.054
	CNN - DAFT	***0.762***	*0.065*	***0.847***	*0.030*	***0.857***	*0.049*

neural network, the type of MRI contrasts and stations impact the performance, as illustrated in Figs. 5 and 6. Including clinical information significantly improved the performance, where CONCAT and DAFT yielded comparable results, as seen in Fig. 4.

4.1. MRI contrasts

The results for the different Dixon contrasts in Fig. 5 showed the best performance for water and opposed-phase scans. Considering the combination of images as multi-channel inputs and the relative fat-water content (RFWC), we observe mixed results. The combination of IN and OPP images, as well as the RFWC scans, did not yield an improvement over the individual scans. For the multi-channel input of fat and water images, we see, however, an increase in performance for most stations, with the overall highest balanced accuracy, ROC AUC, and average precision on station 2. Notably, next to an increase in mean performance, also the variance decreased. These results suggest that the network can benefit from complementary information in these scans for diabetes prediction. The improvement of the combination of fat and water over the combination of IN and OPP images may be surprising, as the former ones are linear combinations of the latter ones, yet underlines the importance of the input to neural networks.

4.2. MRI stations

We noted a good predictive accuracy from stations 1 to 5 (see Figs. 5 and 6), with a distinctive drop for station 6, which is the last station reaching up to the knees. Remarkably, stations focusing on the upper lungs (1) or upper legs (5) result in such a high predictive performance. To improve the overall accuracy, we evaluated the combination of the stations. Surprisingly, stitching stations 2 and 3 did not result in an improvement over using station 2 alone. In contrast, the ensemble strategy of stations 2 and 3 yielded a clear improvement. The UKB imaging protocol uses the jugular notch to locate the first MRI station so that the anatomical region in the first station is more consistent across the sample than for the last stations due to variations in height. This also needs to be considered in the interpretation of the results. Finally, the best performances were achieved by ensembling over stations 1 to 5 and all 6 stations.

4.3. Alternative approaches

In the past, the analysis of fat content in the liver and visceral adipose tissue has been of high interest in T2DM, because of the accumulation of adipose tissue in the insulin-secreting organs with effects on endocrine function and the disruption of multiple pathways in gluconeogenesis due to excessive VAT [7,21]. Hence, previous studies for diabetes classification have extracted quantitative measures from MRI whole-body scans to measure hepatic fat content and VAT [10,20]. In comparing a subset of our sample, we noted a higher accuracy for predicting diabetes with deep learning on the entire MRI sequence, see Table 2. These results demonstrate the potential of end-to-end learning to extract features that capture the multi-factorial changes of diabetes in the body. Importantly, our approach does not rely on image segmentations of fat depots [10,20] or manual segmentations of the liver [11]. In parallel work [30], used a deep neural network for predicting diabetes from a whole-body MRI sequence, and [29] predicted various measures including body composition and subject characteristics such as T2D on UK biobank. In contrast to these approaches, our study focused on evaluating different MRI contrasts and stages, and integrating tabular data with DAFT.

4.4. Evaluation and translation

We have used balanced sets to obtain conservative estimates of classification accuracy for our research questions. Also, in creating the splits, we have used age, sex, and site, next to the diagnosis. Accounting for these potential confounders leads to conservative estimates of the classification accuracy as the negative impact of shortcut learning is mitigated. For the translation of the proposed approach to the general population, with lower rates of T2D, a different training strategy will likely yield better results. In transfer learning, methods have been proposed that can tailor the model to the target distribution. If the distribution of T2D in the target application is available, taking this information into account during training would be beneficial. Going in the opposite direction, the problem could be further narrowed down by adding BMI as an additional variable to create the control group. This would investigate the question of differentiating diabetics from non-diabetics in a population with higher BMI.

4.5. Limitations

Our study has some limitations. First, the MRI data were acquired on the same scanner models, where future work will need to validate our predictive performance on data from multiple centers to take the effect of different population characteristics and scanners into account. Second [22], reported evidence of a healthy volunteer selection bias in UKB. Nevertheless, the large sample size and heterogeneity in UKB generally support the generalization to the broader population. Third, we have not used the genetic data in the UKB, whose inclusion could be one step closer to a comprehensive approach to diabetic precision medicine. Fourth, the reported results depend on the hyper-parameters, for which we have used grid search on the validation set. A more exhaustive tuning could lead to improved absolute performances, but we do not expect that the relative performances of the different configurations would be affected by this. Fifth, since the data release used in this manuscript, more UK biobank data has become available, particularly PDFF, VAT, and SAT measurements.

5. Conclusions

Overall, our results demonstrate that it is beneficial to not only focus on the abdominal area of whole-body scans for diabetes prediction but also to take the entire neck-to-knee sequence into account and include multiple contrasts from the Dixon sequence. Including clinical information increases predictive performance, which may further be expanded by including relevant genetic markers [23]. When whole-body imaging is performed for screening purposes, diabetes prediction can be used to detect cases that have not yet been identified without additional costs. Unlike fasting glucose tests, no fasting is required; it is entirely non-invasive and potentially less sensitive to short-term lifestyle changes. Coming data releases from UKB will provide data about more subjects and follow-up examinations, which will support model training and enable the prediction of prognostic information, e.g., the risk of developing diabetes in the future or the occurrence of diabetes-related complications. Our study demonstrated that deep learning captured relevant information in Dixon sequences to predict diabetes, and we evaluated the best input configuration. The non-invasive and radiation-free acquisition of MRI scans facilitates the application of whole-body imaging in epidemiological studies and clinical practice, supporting the use of neural networks as diagnostic tools.

Data availability

Publicly available data from the UK Biobank study was analyzed in this study. The datasets are available to researchers through an open application via https://www.ukbiobank.ac.uk/register-apply/.

CRediT authorship contribution statement

Christian Wachinger: Writing – original draft, Supervision, Resources, Funding acquisition, Data curation, Conceptualization. **Tom Nuno Wolf:** Methodology, Investigation, Data curation. **Sebastian Pölsterl:** Writing – review & editing, Visualization, Supervision, Methodology, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Christian Wachinger reports financial support was provided by Bavarian State Ministry of Science and the Arts. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was conducted using the UK Biobank Resource under Application No. 34479. This research was supported by the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e22239.

References

[1] P. Saeedi, et al., Global and Regional Diabetes Prevalence Estimates for 2019 and Projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, ninth ed., 157, Diabetes Res. Clin. Pract., 2019, 107843.
[2] S. Chatterjee, K. Khunti, M.J. Davies, Type 2 diabetes, Lancet 389 (10085) (2017) 2239–2251.
[3] R. Unnikrishnan, R. Pradeepa, S.R. Joshi, V. Mohan, Type 2 diabetes: demystifying the global epidemic, Diabetes 66 (6) (2017) 1432–1442.
[4] S. Weckbach, S.O. Schoenberg, Whole body MR imaging in diabetes, Eur. J. Radiol. 70 (3) (Oct. 2009) 424–430.
[5] E.L. Thomas, J.A. Fitzpatrick, S.J. Malik, S.D. Taylor-Robinson, J.D. Bell, Whole body fat: Content and distribution, Prog. Nucl. Magn. Reson. Spectrosc. 73 (2013) 56–80.
[6] J. Linge, et al., Body Composition profiling in the UK biobank imaging study, Obesity 26 (11) (2018) 1785–1795.
[7] F. Bamberg, et al., Subclinical disease burden as assessed by whole-body MRI in subjects with prediabetes, subjects with diabetes, and normal control subjects from the general population: the KORA-MRI study, Diabetes 66 (1) (2017) 158–169.
[8] S.D. Heber, et al., Pancreatic fat content by magnetic resonance imaging in subjects with prediabetes, diabetes, and controls from a general population without cardiovascular disease, PLoS One 12 (5) (2017) 1–13.
[9] R. Lorbeer, et al., Association betweenMRI-derived hepatic fat fraction and blood pressure in participants without history of cardiovascular disease, J. Hypertens. 35 (4) (2017) 737–744.
[10] M. Wang, et al., Prediction of type 2 diabetes mellitus using non-invasive MRI quantitation of visceral abdominal adiposity tissue volume, Quant. Imag. Med. Surg. 9 (6) (2019).
[11] D.A.P. Gutmann, et al., MRI-derived radiomics features of hepatic fat predict metabolic states in individuals without Cardiovascular disease, Acad. Radiol. (2020).
[12] M.S. Hutchinson, et al., Effects of age and sex on estimated diabetes prevalence using different diagnostic criteria: the tromsø OGTT study, Internet J. Endocrinol. 2013 (2013).

[13] F. Guo, D.R. Moellering, W.T. Garvey, Use of HbA1c for diagnoses of diabetes and prediabetes: Comparison with diagnoses based on fasting and 2-Hr glucose values and effects of gender, race, and age, Metab. Syndr. Relat. Disord. 12 (5) (2014) 258–268.

[14] T.J. Littlejohns, et al., The UK Biobank imaging enhancement of 100.000 participants: rationale, data collection, management and future directions, Nat. Commun. 11 (1) (2020) 1–12.

[15] F. Bamberg, et al., Whole-body MR imaging in the German national Cohort: rationale, design, and technical background, Radiology (2015).

[16] H.R. Wilman, et al., Characterisation of liver fat in the UK Biobank cohort, PLoS One 12 (2) (2017) 1–14.

[17] R. C. Team, R: A Language and Environment for Statistical Computing, 2013.

[18] D.E. Ho, K. Imai, G. King, E.A. Stuart, Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference, Polit. Anal. 15 (3) (2007) 199–236.

[19] S. Pölsterl, T.N. Wolf, C. Wachinger, Combining 3D image and tabular data via the dynamic affine feature map Transform, Int. Conf. Med. Image Comput. Comput. Assist. Interv. (2021).

[20] J. Linge, B. Whitcher, M. Borga, O. Dahlqvist Leinhard, Sub-phenotyping metabolic disorders using body Composition: an individualized, nonparametric approach utilizing large data sets, Obesity 27 (7) (2019) 1190–1199.

[21] I.S. Idilman, et al., Quantification of liver, pancreas, kidney, and vertebral body MRI-PDFF in non-alcoholic fatty liver disease, Abdom. Imag. 40 (6) (2015) 1512–1519.

[22] A. Fry, et al., Comparison of sociodemographic and health-related Characteristics of UK biobank participants with those of the general population, Am. J. Epidemiol. 186 (9) (2017) 1026–1034.

[23] Y. Ji, et al., Genome-wide and abdominal MRI data provide evidence that a genetically determined favorable adiposity phenotype is characterized by lower ectopic liver fat and lower risk of type 2 diabetes, heart disease, and hypertension, Diabetes 68 (1) (2019) 207–219.

[29] B. Dietz, J. Machann, V. Agrawal, M. Heni, P. Schwab, J. Dienes, S. Reichert, A.L. Birkenfeld, H.U. Häring, F. Schick, N. Stefan, Detection of diabetes from whole-body MRI using deep learning, JCI insight 6 (21) (2021).

[30] T. Langner, A. Martínez Mora, R. Strand, H. Ahlström, J. Kullberg, MIMIR: deep regression for automated analysis of UK biobank MRI scans, Radiology: Artif. Intell. 4 (3) (2022), e210178.