NeuroImage 260 (2022) 119505

Contents lists available at ScienceDirect

NeuroImage

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

DAFT: A universal module to interweave tabular data and 3D images in CNNs

Tom Nuno Wolf[a,b,1], Sebastian Pölsterl[a,1,*], Christian Wachinger[a,b,*], the Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging Biomarkers and Lifestyle flagship study of ageing

[a] The Lab for Artificial Intelligence in Medical Imaging (AI-Med), Department of Child and Adolescent Psychiatry, Ludwig-Maximilians-Universität, Nussbaumstraße 5, Munich 80336, Germany
[b] Technical University of Munich, School of Medicine, Department of Radiology, Ismaninger Straße 22, Munich 81675, Germany

ARTICLE INFO

Keywords:
Alzheimer's disease
Convolutional neural networks
Deep learning
Disease prediction
Magnetic resonance imaging
Tabular data
Time-to-event analysis

ABSTRACT

Prior work on Alzheimer's Disease (AD) has demonstrated that convolutional neural networks (CNNs) can leverage the high-dimensional image information for diagnosing patients. Beside such data-driven approaches, many established biomarkers exist and are typically represented as tabular data, such as demographics, genetic alterations, or laboratory measurements from cerebrospinal fluid. However, little research has focused on the effective integration of tabular data into existing CNN architectures to improve patient diagnosis. We introduce the Dynamic Affine Feature Map Transform (DAFT), 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. This is achieved by using an auxiliary neural network that outputs a scaling factor and offset to dynamically apply an affine transformation to the feature maps of a convolutional layer. In our experiments on AD diagnosis and time-to-dementia prediction, we show that the DAFT is highly effective in combining 3D image and tabular information by achieving a mean balanced accuracy of 0.622 for diagnosis, and mean *c*-index of 0.748 for time-to-dementia prediction, thus outperforming all baseline methods. Finally, our extensive ablation study and empirical experiments reveal that the performance improvement due to the DAFT is robust with respect to many design choices.

1. Introduction

Over the last decade, deep convolutional neural networks (CNNs) have become a staple for classification of Alzheimer's Disease (AD) from brain images acquired by magnetic resonance imaging (MRI) (Ebrahimighahnavieh et al., 2020). While CNNs excel at extracting abstract high-level representations of neuroanatomy from MRI, brain MRI only offers a limited view on the underlying changes causing cognitive decline (Jack et al., 2013). Thus, reliably diagnosing AD requires incorporating tabular data such as patient demographics, family history, or laboratory measurements from cerebrospinal fluid. Typically, tabular data are low-dimensional and individual variables capture rich clinical knowledge. However, the statistical properties of tabular variables can vary: laboratory measurements are continuous, the family history is a binary indicator, genetic alterations are counts, and a person's smoking habits are ordinal (e.g. frequent versus occasional smoking). In contrast, image information is continuous-valued, high-dimensional, and carries little information on a per-voxel level.

Due to image and tabular data often describing complementary aspects of the disease, we want to conjoin the two in a single neural network such that redundancies are avoided. For example, if a CNN extracts a feature from a patient's brain MRI that corresponds to the patient's age, the CNN extracted information that is readily available through tabular data. Ideally, a network should utilize both sources of information in a way that one source can inform the other and the network's prediction capability is enhanced. However, the capacity required to extract high-level information of image data exceeds the capacity required to summarize the tabular data by several orders of magnitude. Hence, training such a network implicitly encourages the network to prioritize changes in image-related parameters. Ultimately, models taking into account heterogeneous data perform only marginally better than unimodal CNNs (Pelka et al., 2020).

* Corresponding authors to: Technical University of Munich, School of Medicine, Department of Radiology; Ismaninger Straße 22; 81675 Munich, Germany
E-mail addresses: sebastian.poelsterl@med.uni-muenchen.de (S. Pölsterl), christian.wachinger@tum.de (C. Wachinger).
[1] T.N. Wolf and S. Pölsterl contributed equally to this work.

https://doi.org/10.1016/j.neuroimage.2022.119505.
Received 30 May 2022; Received in revised form 12 July 2022; Accepted 21 July 2022
Available online 22 July 2022.

Deep learning approaches most commonly integrate image and tabular data naively by concatenation of latent feature vectors. To this end, tabular data is concatenated with a high-level image descriptor produced by a CNN, and fed through the final (fully-connected) layers of the network (Esmaeilzadeh et al., 2018; Hao et al., 2019; Kopper et al., 2021; Liu et al., 2019; Mobadersany et al., 2018; Pölsterl et al., 2020). This approach limits the way the image-specific part of the network can interact with the tabular-specific part of the network and vice versa. Instead, we seek an architecture where tight interaction between image and tabular data enables the network to truly view image information in the context of the tabular information, and a two-way exchange of information is initiated.

In this work, we propose to fuse information from a patient's 3D brain MRI and tabular data via the Dynamic Affine Feature Map Transform (DAFT). The DAFT dynamically scales and shifts the feature maps in a convolutional layer via an auxiliary neural network that amalgamates image and tabular information. It is a generic module that can be easily incorporated in any type of CNN to establish a bidirectional exchange of information between data types. In our extensive experiments on AD diagnosis and time-to-dementia analysis, we compare the DAFT to three unimodal baselines, and five competing deep neural networks that fuse image and tabular information, and are evaluated with two different backbone architectures. The results demonstrate that DAFT outperforms all competing methods by a large margin: $+0.021$ balanced accuracy in the AD diagnosis task, and $+0.019$ concordance index in the time-to-dementia task. Finally, our ablation study and empirical experiments show that the DAFT is robust to various architectural changes and leads to improved prediction accuracy over baseline methods without the need of extensive hyper-parameter tuning.

The remainder of this paper is organized as follows. In Section 2, we discuss related work on deep neural networks to combine image and tabular data. In Section 3, we propose the DAFT for improved integration of image and tabular data, describe the data from the Alzheimer's Disease Neuroimaging Initiative used in our experiments, and present our evaluation scheme. The results and discussion of our experimental results on AD diagnosis and time-to-dementia diagnosis are presented in Section 4. Finally, we end with concluding remarks in Section 5.

2. Related work

A straight-forward way to combine image and tabular data is a two stage approach where one first trains a CNN on the image data alone, and then concatenates its predictions or latent representations with the tabular data to finally fit a separate linear model on the combined feature vectors. The authors of Li et al. (2019) followed this approach to fuse information from brain MRI and routinely acquired clinical markers to predict progression to AD. However, their two-step process does not fully utilize the CNN, because it does not consider that the learned image descriptor could contain information that is redundant to the tabular information.

This motivated others to propose end-to-end models, where the latent image representation is concatenated with the clinical information before the last fully connected (FC) layer. Hao et al. (2019) followed such an approach to train a single CNN that fuses histopathology images, genomic data, and demographics for survival prediction. The works in Kopper et al. (2021); Pölsterl et al. (2020) followed a similar approach to fuse a point cloud representation of the hippocampus with clinical markers for time-to-dementia prediction. The downside of such an approach is that the tabular data is limited to a linear contribution to the final prediction, unless non-linearities are modeled explicitly (e.g. via B-spline transformations).

This issue can be overcome by replacing the single FC layer, which combines image and tabular data, with a multilayer perceptron (MLP) such that non-linear relationships can be learned implicitly. Such a network was proposed by Mobadersany et al. (2018) to predict overall survival of patients diagnosed with glioma from digital pathology images and genomic data, and by Esmaeilzadeh et al. (2018); Liu et al. (2019) for AD diagnosis from brain MRI and clinical markers. A minor modification of this approach uses an additional MLP that is applied solely on the tabular data before concatenation with the latent image representation (El-Sappagh et al., 2020; Li et al., 2020; Spasov et al., 2019). However, adding additional FC layers increases the number of trainable parameters considerably, which can render the network more susceptible to over-fitting. During inference, the tabular information only interacts with the global image descriptor in concatenation-based approaches, which does not allow for fine-grained interactions on the voxel- or patch level. To our assumption, interactions on the voxel- or patch level improve the quality of image-features, which is evident from our results.

Braman et al. (2021) proposed to tackle aforementioned issues by fusing latent representations of three deep modality-specific subnetworks via an attention-gated tensor fusion process, and applying a penalty term that encourages the modality-specific latent representations to be orthogonal. They fuse MRI, histopathology images, genomic data, and clinical information to predict overall survival of glioma patients. Duanmu et al. (2020) proposed an alternative approach. They fuse information in a multiplicative manner to predict response to chemotherapy. Their approach uses an auxiliary network that takes the tabular data and outputs a scalar scaling factor to rescale the feature maps of every other convolutional layer of a CNN. This results in an amplification or repression of latent image feature maps that is conditional on the patient's tabular information. However, the size of their auxiliary network scales quadratically with the depth of the CNN, which in turn increases the runtime and memory requirements dramatically.

Perez et al. (2018) introduced the Feature-wise Linear Modulation (FiLM) layer for visual question answering in computer vision. Similar to the work by Duanmu et al. (2020), FiLM uses an auxiliary network, but in addition to the scaling factor, also outputs an offset to shift feature maps by. Therefore, the auxiliary network in FiLM can affinely transform each feature map of a convolutional layer, similar to our approach. The only application of FiLM to the medical domain was presented in Jacenków et al. (2020) to segment the myocardium and ventricular cavities conditional on the slice position and the phase of the cardiac cycle. Note that the premise of the two approaches above is very different from ours. In visual question answering and image segmentation, the meta information (question or cardiac cycle) is fundamentally related to the image content, because in both settings the meta information refers to a property of the image. Therefore, conditioning the CNN on the meta information, but not vice versa, is a reasonable approach. In contrast, in our work, the interrelation of image and tabular data is much weaker, which is why we argue that a bidirectional flow of information is preferred. In the proposed DAFT, feature maps are scaled and shifted, as in FiLM (Perez et al., 2018), but we make this transformation conditional on image *and* tabular data such that one source of information can inform the other.

A preliminary version of this work has been presented at a conference (Pölsterl et al., 2021). Here, we extend this work by providing more details on the technical aspects of DAFT, and extending the experimental evaluation with more metrics, an analysis of the contribution of individual tabular features on the predictive performance, and additional experiments to showcase that DAFT generalizes better than previous approaches.

3. Materials and methods

3.1. The dynamic affine feature map transform

By design, CNNs excel at extracting task-specific descriptors of high-dimensional 3D image data. Here, our aim is to use an existing CNN architecture and augment it with a versatile module to achieve seamless integration of low-dimensional tabular data such that the CNN can leverage this complementary information for improved prediction. Com-

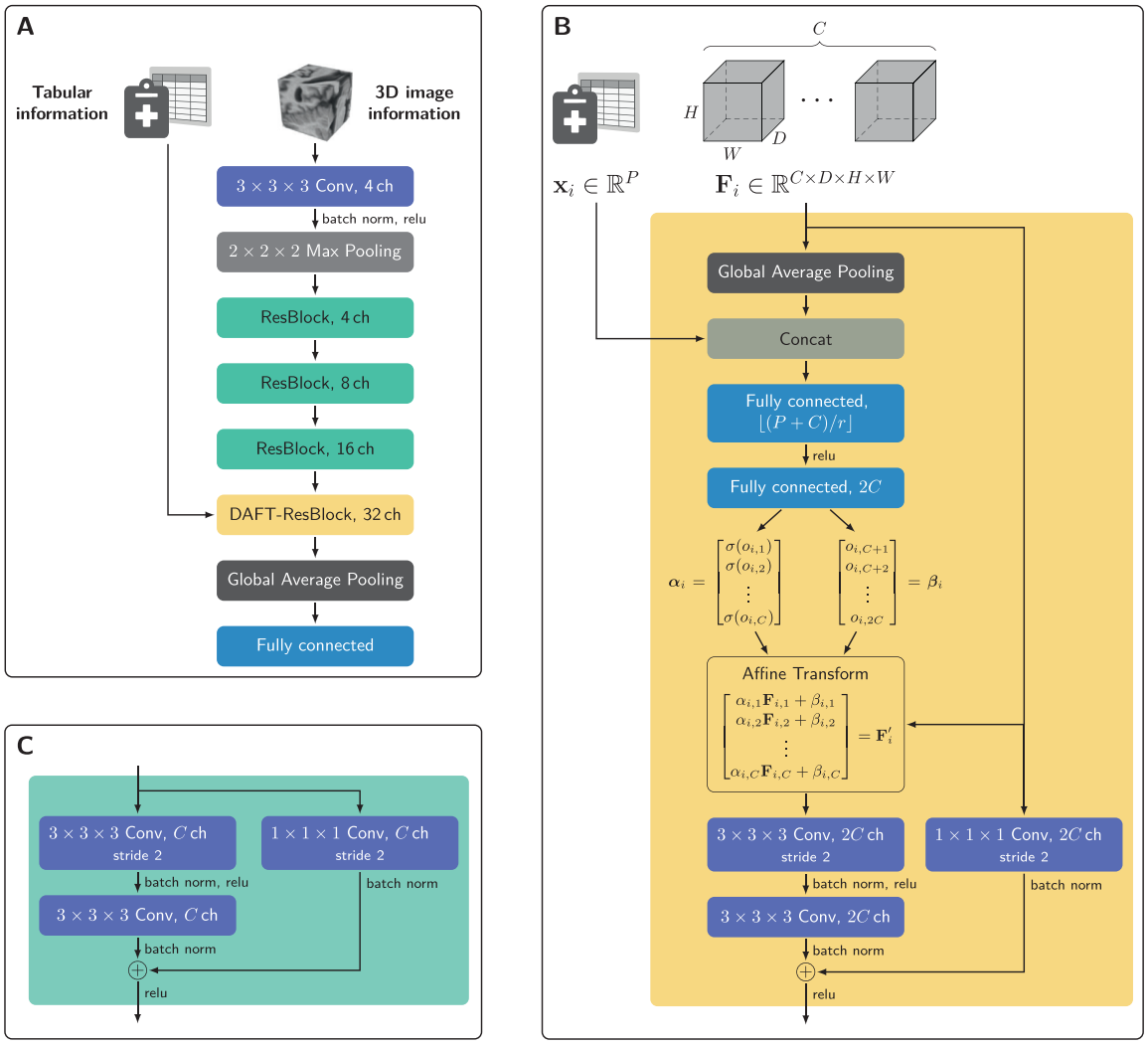

Fig. 1. Overview of the proposed network architecture. A: The backbone of our proposed network architecture is a ResNet, where the Dynamic Affine Feature Map Transform (DAFT) is applied in the last residual block. B: For each instance i in a batch, the DAFT first squeezes the spatial dimensions of a feature map $\mathbf{F}_i$ of size $C \times D \times H \times W$ via global average pooling. Next, the resulting vector is concatenated with the vector $\mathbf{x}_i \in \mathbb{R}^P$ of tabular information. The result is fed to a set of fully-connected layers with intermediate ReLU activation that compress the feature vector by a factor r (throughout this work, we use $r = 7$). The output is a vector of scales $\boldsymbol{\alpha}_i \in \mathbb{R}^C$, and offsets $\boldsymbol{\beta}_i \in \mathbb{R}^C$, which are used to affinely transform the input feature maps $\mathbf{F}_{i,c}$, yielding $\mathbf{F}'_{i,c}$ $(c = 1, \ldots, C)$. C: A standard ResBlock with down-sampling via strided convolutions. If the number of input feature maps equals the number of output feature maps, no downsampling is performed and the residual connection contains no convolutional nor batch norm layer.

mon tabular information, such as demographics or aggregate statistics, describe the patient's state holistically, thus, a level exchange of information between image and tabular data is required. Since early layers of a CNN typically describe rather primitive concepts (e.g. edges, blobs), we propose to transform the feature maps of a 3D convolutional layer that appears late in the CNN and captures broad concepts in the image. We select a ResNet (He et al., 2016) as the backbone of our approach, as it is the most common backbone architecture used in previous work. We propose to dynamically scale and shift the feature maps of a 3D convolutional layer in the last residual block, conditional on a patient's brain MRI and clinical tabular information. Our full network design is summarized in Fig. 1.

Formally, let $\mathbf{x}_i \in \mathbb{R}^P$ denote the tabular clinical information for the ith instance in the dataset and $\mathbf{F}_{i,c} \in \mathbb{R}^{D \times H \times W}$ the cth output (feature map) of a 3D convolutional layer based on the ith volumetric image ($c \in \{1, \ldots, C\}$ and P denotes the number of tabular features, D, H, W the depth, height, and width of the feature map). The ability to incite or repress high-level concepts learned from the image is achieved by conditioning the outputs $\mathbf{F}_{i,c}$ of a convolutional layer on the image and tabular data. For this purpose, the Dynamic Affine Feature Map Transform (DAFT) learns to predict scale $\alpha_{i,c}$ and offset $\beta_{i,c}$:

$$\mathbf{F}'_{i,c} = \alpha_{i,c}\mathbf{F}_{i,c} + \beta_{i,c}, \tag{1}$$

$$\alpha_{i,c} = f_c(\mathbf{F}_{i,c}, \mathbf{x}_i), \qquad \beta_{i,c} = g_c(\mathbf{F}_{i,c}, \mathbf{x}_i), \tag{2}$$

where f_c, g_c are arbitrary mappings from image and tabular space to a scalar. In our work, a single auxiliary neural network h_c models f_c, g_c and outputs a single

$$\boldsymbol{\alpha} - \boldsymbol{\beta}$$

-pair, which is referred to as DAFT, visually represented in Fig. 1B.

First, DAFT creates a bottleneck via global average pooling of the spatial dimensions of the image feature map. Next, the resulting vector is concatenated with the tabular data. This combined vector is squeezed by an FC layer with a bottleneck and fed through a ReLU non-linearity (Nair and Hinton, 2010). A second FC layer expands the squeezed vector and yields the output vectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$; motivated

by Hu et al. (2020), we do not add bias terms to FC layers in the auxiliary network. In addition, we allow applying a non-linear activation function $\sigma(\cdot)$ to $\alpha_{i,c}$, such that the scaling factors can be restricted to a particular domain (e.g. $[0; 1]$ for sigmoid activation). In our experiments, we explore three options: linear, sigmoid and tanh.

As the proposed DAFT does not depend on the number of instances in the dataset, nor the spatial resolution of the feature map, it is computationally efficient. Due to parameter sharing, DAFT is able to dynamically scale (via $\alpha_{i,c}$) and shift (via $\beta_{i,c}$) feature maps of a convolutional layer, conditional on the specific image and tabular information of the ith patient. Moreover, our proposed DAFT is a versatile module that can be effortlessly applied to any type of CNN to fuse tabular information, not just the CNN in Fig. 1A, which we demonstrate in Section 4.6.

3.2. Network training

In this work, DAFT is evaluated on two tasks using T1 brain MRI. The first one is to diagnose patients as cognitively normal (CN), mild cognitively impaired (MCI), or demented (AD). The second task is to predict time of dementia onset for patients of the MCI cohort. The diagnosis task can be formulated as a multi-class classification problem, which means we can minimize the standard cross-entropy loss during training.

In the time-to-dementia task, we have to account for the fact that only a subset of patients has been observed to convert from MCI to AD. For the remaining patients, we did not observe the time of conversion. Instead, the time of their last follow-up visit is a lower bound on the time of conversion: their time of conversion is *right censored*. Let $t_i > 0$ denote the time of conversion to AD and $c_i > 0$ the time of right censoring for the ith patient. In practice, we can only observe patients that converted while participating in the study ($t_i < c_i$). Therefore, the observable time is defined as $y_i = \min(t_i, c_i)$, and $\delta_i = I(t_i \leq c_i)$ is a binary event indicator. We account for right censored conversion times, by minimizing the negative partial log-likelihood of Cox's model (Faraggi and Simon, 1995) – traditionally used in survival analysis. Let $M(\mathbf{I}, \mathbf{x} \mid \boldsymbol{\Theta})$ denote the predicted risk score of conversion, based on image $\mathbf{I}$ and tabular data $\mathbf{x}$, then we update the network's parameters $\boldsymbol{\Theta}$ by solving

$$\min_{\boldsymbol{\Theta}} \sum_{i=1}^{B} \delta_i \left[M(\mathbf{I}_i, \mathbf{x}_i \mid \boldsymbol{\Theta}) - \log \left(\sum_{j \in \mathcal{R}_i} \exp(M(\mathbf{I}_j, \mathbf{x}_j \mid \boldsymbol{\Theta})) \right) \right], \tag{3}$$

where B is the batch size, and $\mathcal{R}_i = \{j \mid y_j \geq t_i\}$ denotes the set of patients who remained MCI shortly before time point t_i.

In both tasks, we minimize the respective loss with mini-batch stochastic gradient descent using the AdamW optimizer, which has been shown to be superior to plain Adam (Loshchilov and Hutter, 2019). Diagnosis and progression tasks are trained for 30 and 80 epochs, respectively. Additionally, we apply a learning rate scheduler: The learning rate is decreased by a factor of 10 after 60% of epochs are finished. After 90% of epochs, the initial learning rate is decreased by a factor of 20. We carry out a grid search on the validation set to select the best configuration of learning rate and weight decay. For each model, we evaluate a total of 5×3 configurations: Learning rate $\in \{0.03, 0.013, 0.0055, 0.0023, 10^{-3}\}$ and weight decay $\in \{0, 10^{-4}, 10^{-2}\}$ and $\in \{10^{-3}, 10^{-2}, 0.1\}$ for the diagnosis and time-to-dementia task, respectively.

3.3. Dataset

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

For the time-to-dementia task, only patients that were classified as MCI during their baseline visits are included. Additionally, patients with bi-directional change in diagnosis over time are excluded, because their diagnoses can be considered unreliable (Wen et al., 2020).

Table 1 summarizes the data used in our experiments.

3.4. Data processing

3.4.1. Image data

T1-weighted MRI are obtained from the ADNI study (Jack et al., 2008). Brain MRI scans are first normalized with the minimal preprocessing pipeline introduced in Wen et al. (2020). Next, images are segmented with FreeSurfer 5.3 (Fischl, 2012), yielding an extracted region of interest of size 64^3 around the left hippocampus, which is known to be strongly affected by AD (Frisoni et al., 2008).

3.4.2. Tabular data

For tabular data, the nine selected variables are: ApoE4, cerebrospinal fluid biomarkers $A\beta_{42}$, P-tau181 and T-tau, the demographic variables age, gender, education, and two measures derived as a summary from 18F-fluorodeoxyglucose (FDG) and florbetapir (AV45) PET scans. As some biomarkers were not acquired at all times, missing values are accounted for by adopting an approach similar to Jarrett et al. (2020): we append binary variables that indicate if a feature is missing for each tabular feature, with the exception of age, gender, and education, which have no missing values. This enables the network to learn from incomplete data and patterns of missingness. The resulting tabular data consists of $P = 15$ features. For the experiments in Section 4.5, we select the tabular features that are available in both ADNI and AIBL, i.e. ApoE4, age, gender. An additional binary missing variable indicator for ApoE4 is added, resulting in a total of four tabular features.

3.4.3. Splitting of data

As demonstrated by Wen et al. (2020), data leakage and confounding effects due to age and sex must be considered carefully to avoid biased evaluation results. Hence, we split the data into five non-overlapping folds using only baseline visits such that diagnosis, age and sex are balanced across folds. To this end, we assess the balance of a split by computing the propensity score, i.e. the probability of a sample belonging to the training data, based on a logistic regression model comprising the known confounders age, sex, and education (Barnes et al., 2010; Stern et al., 2020). 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 imbalance (Ho et al., 2007). For each of the five folds, this process is repeated for 1000 randomly selected partitions and the partition with the minimum imbalance is ultimately the selected split.

Each of the five folds serves as a test set once. The remaining folds are again partitioned into five balanced chunks, out of which one is randomly selected as the validation set and the remaining data as the training set. Thus, the resulting size of data splits is 20% test set, 16% validation set and 64% training set. For the diagnosis task, the training set is extended by including each patient's longitudinal data as in Wen et al. (2020) (3.49 ± 2.56 visits per patient; validation and test sets remain unchanged).

3.5. Baseline methods

We compare against two unimodal baselines: (i) a ResNet (He et al., 2016) that uses image information only – using the architecture shown in Fig. 1A, but where the DAFT-ResBlock is replaced with a standard ResBlock – and (ii) a linear model that uses tabular information only. Additionally, we compare against a two-stage approach, where we first extract the latent image representations from the aforementioned ResNet,

Table 1
Dataset statistics at initial visit. ± indicates the standard deviation within the dataset.

Task	Subjects	Age	Sex (male)	Education	MMSE	Diagnosis
Diagnosis	1341	73.9 ± 7.2	51.8%	15.9 ± 2.9	27.2 ± 2.7	Dementia (19.6%), MCI (40.1%), CN (40.3%)
Progression	755	73.5 ± 7.3	60.4%	15.9 ± 2.9	27.5 ± 1.8	Progressor (37.4%), median follow-up time 2.01 years
AIBL	653	72.9 ± 6.6	43.6%	N/A	27.5 ± 3.5	Dementia (11.6%), MCI (15.5%), CN (72.9%)

Table 2
Number of parameters for each model. We report the additional number of parameters for neural networks with respect to ResNet, indicated with '+'.

	Parameters	
Model	Diagnosis	Progression
Linear Model	48	14
ResNet	56,535	56,469
Concat-1FC	+45	+14
Concat-2FC	+108	+160
1FC-Concat-1FC	+76	+64
Duanmu et al. (2020)	+328	+320
FiLM (Perez et al., 2018)	+188	+184
DAFT	+252	+248

which is subsequently combined with the tabular data in a linear model (Linear model /w ResNet features). In the diagnosis task, the linear model is a multinomial logistic regression, in the time-to-dementia task, Cox's proportional hazards model (Cox, 1972). Note that we performed experiments with gradient boosted models too, but they did not show any improvement over a linear model and are therefore not considered in this work.

Moreover, we evaluate three networks as baselines that fuse image and tabular data by concatenation and are derived from the architecture of the ResNet in Fig. 1A by replacing the DAFT-ResBlock with a standard ResBlock. In the first network, Concat-1FC, the latent image feature vector, which is the output of the global average pooling layer after the last ResBlock, is concatenated with the tabular data vector and fed directly to the final classification layer. It only models tabular data linearly, thus it is related to the linear model baseline with the advantage that it simultaneously learns an image descriptor. In the second baseline network, Concat-2FC, the concatenated vector is created the same as in Concat-1FC, but it is fed to a two-layer FC bottleneck with intermediate ReLU non-linearity (similar to Esmaeilzadeh et al., 2018; Liu et al., 2019). The third concatenation-based baseline, 1FC-Concat-1FC, is inspired by Spasov et al. (2019) and feeds the tabular data to a two-layer FC bottleneck layer before concatenating it with the latent image representation, as in Concat-1FC.

Finally, we compare against two approaches that fuse tabular data by inciting or repressing feature maps of a convolutional layer conditional on tabular information, that means they only establish a one-way exchange of information between data types. One follows the network design introduced by Duanmu et al. (2020), the other the one by Perez et al. (2018). The former scales feature maps of every other convolutional layer via multiplication, the latter scales and shifts the feature maps of one convolutional layer in the last residual block via Feature-wise Linear Modulation (FiLM). Four of the networks (Concat-2FC, 1FC-Concat-1FC, FiLM and DAFT) contain a bottleneck layer. We compress the input vector to 4 dimensions, approximately a fourth of the number of tabular features. We use the identity function $\sigma(x) = x$ in the auxiliary network for the scale $\alpha_{i,c}$ in FiLM and DAFT. The implementation of all models is available at https://github.com/ai-med/DAFT and Table 2 reports the number of parameters for each of the different models.

3.6. Evaluation metrics

For each model, we report the performance on the test set after early stopping on the validation set. In the diagnosis task, we evaluate models' performance using the balanced accuracy (bACC; Brodersen et al., 2010), which accounts for class imbalance. Additionally, we report the micro- and macro-averaged F1 score, and the true positive fraction per class (TPF), which extends sensitivity and specificity to multi-class classification (Bron et al., 2015). The class prediction in the multi-class classification is achieved via extraction of the max probability. In the time-to-dementia task, we evaluate models in terms of discrimination and calibration. We assess discrimination using an inverse probability of censoring weighted estimator of the concordance index (*c*-index; Uno et al., 2011). The *c*-index is identical to the area under the receiver operating characteristics curve in case of a binary outcome without censoring. Moreover, we compute the integrated time-dependent Brier score (IBS; Graf et al., 1999), which is a measure of both discrimination and calibration. The time-dependent Brier score at time t is an extension of the mean squared error to right censored data. It is complementing the *c*-index, because it measures the average difference between the true progression status and the estimated risk of progression, whereas the *c*-index only assesses whether the ordering by predicted risk scores is concordant with the ordering by true progression times. We report the Brier score integrated over 31 time points – roughly 1 month apart – between the 6th and 36th month from the first MCI diagnosis. The *c*-index and IBS are estimated using their implementation in scikit-survival 0.13.1 (Pölsterl, 2020).

4. Results and discussion

4.1. Predictive performance

The predictive performance for the diagnosis task (balanced accuracy, true positive fraction, micro- and macro-averaged F1 score) is summarized in Tables 3 and 4, and for the time-to-dementia task (c-index, IBS) in Table 5.

The results demonstrate that a linear model, that only uses tabular data, outperforms the unimodal, image-based ResNet across all tasks and metrics. This matches our expectation, because tabular data contain amyloid-specific measures that are derived from cerebrospinal fluid and PET imaging, which typically become abnormal before atrophy becomes visible in MRI (Jack et al., 2013). Furthermore, the predictive performance does not increase significantly if image descriptors are learned independently and combined subsequently in a second model (third row in Tables 3–5). This result is evidence that the image descriptor comprises information that is at least partially redundant to the clinical information.

In contrast, all concatenation-based networks successfully extract complementary image information in the diagnosis task, with an increase in average bACC by at least 0.024 over the Linear Model. Moreover, predicting the MCI class correctly remains difficult, with only Concat-1FC achieving a higher TPF than the linear model with ResNet features (0.037 higher TPF_{MCI}). This is in line with results on the CAD-Dementia challenge, were the winning entry in terms of accuracy only obtained a TPF_{MCI} of 0.287 (Bron, et al., 2015). The worst performing network that models the tabular data in a non-linear manner, is the one by Duanmu et al. (2020). In terms of bACC, it performs worse than all models, except Concat-1FC. The TPF values reveal that the network can only adequately identify healthy controls ($TPF_{CN} = 0.774$), but struggles with the remaining classes ($TPF_{MCI} = 0.470$, $TPF_{AD} = 0.490$). However, the model of Duanmu et al. (2020) outperforms the FiLM-based model

Table 3
Predictive performance for the diagnosis task (true positive fraction and balanced accuracy). We report the mean and standard deviation across five folds. Higher values are better. The use of data of each model is indicated in columns 2–3 (I = the use of image data; T = use of tabular data). L/NL denote the linearity of the model for the tabular data transform (linear/non-linear).

			True positive fraction (testing) ↑			Balanced accuracy ↑	
	I	T	CN	MCI	AD	Validation	Testing
Linear Model	✗	L	0.721 ± 0.048	0.533 ± 0.054	0.403 ± 0.040	0.571 ± 0.024	0.552 ± 0.020
ResNet	√	✗	0.597 ± 0.117	0.370 ± 0.121	0.544 ± 0.116	0.568 ± 0.015	0.504 ± 0.016
Linear Model /w ResNet Features	√	L	0.659 ± 0.088	0.532 ± 0.025	0.490 ± 0.072	0.585 ± 0.050	0.560 ± 0.055
Concat-1FC	√	L	0.701 ± 0.128	0.569 ± 0.165	0.490 ± 0.110	0.630 ± 0.043	0.587 ± 0.045
Concat-2FC	√	NL	0.727 ± 0.098	0.440 ± 0.052	0.560 ± 0.117	0.633 ± 0.036	0.576 ± 0.036
1FC-Concat-1FC	√	NL	0.721 ± 0.083	0.501 ± 0.114	0.552 ± 0.118	0.632 ± 0.020	0.591 ± 0.024
Duanmu et al. (2020)	√	NL	0.774 ± 0.025	0.470 ± 0.070	0.490 ± 0.068	0.634 ± 0.015	0.578 ± 0.019
FiLM (Perez et al., 2018)	√	NL	0.734 ± 0.101	0.410 ± 0.163	0.660 ± 0.160	0.652 ± 0.033	0.601 ± 0.036
DAFT	√	NL	0.767 ± 0.080	0.449 ± 0.154	0.651 ± 0.144	0.642 ± 0.012	**0.622 ± 0.044**

Table 4
Predictive performance for the diagnosis task (micro- and macro-averaged F1 scores). We report the mean and standard deviation across five folds. Higher values are better. The use of data of each model is indicated in columns 2–3 (I = the use of image data; T = use of tabular data). L/NL denote the linearity of the model for the tabular data transform (linear/non-linear).

			Micro-averaged F1 score ↑		Macro-averaged F1 score ↑	
	I	T	Validation	Testing	Validation	Testing
Linear Model	✗	L	0.594 ± 0.014	0.583 ± 0.021	0.571 ± 0.020	0.557 ± 0.022
ResNet	√	✗	0.561 ± 0.017	0.496 ± 0.031	0.549 ± 0.018	0.485 ± 0.025
Linear Model /w ResNet Features	√	L	0.603 ± 0.043	0.575 ± 0.054	0.588 ± 0.048	0.565 ± 0.052
Concat-1FC	√	L	0.636 ± 0.039	0.607 ± 0.045	0.629 ± 0.044	0.588 ± 0.049
Concat-2FC	√	NL	0.631 ± 0.029	0.579 ± 0.029	0.616 ± 0.031	0.567 ± 0.024
1FC-Concat-1FC	√	NL	0.636 ± 0.043	0.600 ± 0.034	0.623 ± 0.034	0.583 ± 0.027
Duanmu et al. (2020)	√	NL	0.651 ± 0.033	0.596 ± 0.020	0.628 ± 0.023	0.573 ± 0.021
FiLM (Perez et al., 2018)	√	NL	0.636 ± 0.039	0.589 ± 0.037	0.620 ± 0.044	0.572 ± 0.034
DAFT	√	NL	0.637 ± 0.025	**0.617 ± 0.040**	0.619 ± 0.037	**0.600 ± 0.045**

Table 5
Predictive performance for the time-to-dementia task. We report the mean and standard deviation across five folds. For the concordance index, higher values are better. For the integrated Brier score, lower values are better. The use of data of each model is indicated in columns 2–3 (I = the use of image data; T = use of tabular data). L/NL denote the linearity of the model for the tabular data transform (linear/non-linear).

			Concordance index ↑		Integrated brier score ↓	
	I	T	Validation	Testing	Validation	Testing
Kaplan–Meier	✗	✗	N/A	N/A	0.144 ± 0.015	0.148 ± 0.007
Linear Model	✗	L	0.726 ± 0.040	0.719 ± 0.077	0.120 ± 0.012	**0.122 ± 0.013**
ResNet	√	✗	0.669 ± 0.032	0.599 ± 0.054	0.137 ± 0.010	0.145 ± 0.013
Linear Model /w ResNet Features	√	L	0.743 ± 0.026	0.693 ± 0.044	0.133 ± 0.021	0.135 ± 0.011
Concat-1FC	√	L	0.755 ± 0.025	0.729 ± 0.086	0.116 ± 0.013	**0.122 ± 0.011**
Concat-2FC	√	NL	0.769 ± 0.026	0.725 ± 0.039	0.119 ± 0.015	0.130 ± 0.011
1FC-Concat-1FC	√	NL	0.759 ± 0.035	0.723 ± 0.056	0.120 ± 0.017	0.125 ± 0.008
Duanmu et al. (2020)	√	NL	0.733 ± 0.031	0.706 ± 0.086	0.125 ± 0.014	0.128 ± 0.017
FiLM (Perez et al., 2018)	√	NL	0.750 ± 0.025	0.712 ± 0.060	0.121 ± 0.014	0.131 ± 0.022
DAFT	√	NL	0.753 ± 0.024	**0.748 ± 0.045**	0.129 ± 0.023	**0.122 ± 0.015**

and Concat-2FC with respect to micro- and macro-averaged F1 scores. The FiLM-based model, which scales and shifts feature maps only based on the tabular data, is the best performing baseline model, but merely achieves a 0.011 higher mean bACC compared to the concatenation-based networks. While it is the network with the highest TPF_{AD}, the class with the lowest frequency, it performs below average for the F1 scores. In contrast, the proposed DAFT network outperforms all competing methods by a large margin of at least +0.021 bACC, +0.01 micro- and +0.012 and macro-averaged F1 score. The performance drop from validation to test set (−0.02 bACC, −0.02 micro- and −0.019 macro-averaged F1 score, −0.005 *c*-index, −0.007 IBS) is the lowest compared to other deep learned-based models, which highlights the generalizability of DAFT.

Table 6 depicts the confusion matrix of DAFT and compares it to the runner-up FiLM-based network. It shows that the improvement due to DAFT can be attributed to better identifying the MCI cohort. In particular, DAFT reduces the amount of MCI patients that are misclassified as AD by 19% (31 patients), and CN patients misclassified as MCI or AD

Table 6
Confusion matrix across all test sets for the diagnosis task for DAFT. Numbers in brackets denote the change relative to the FiLM-based network.

		Predicted					
		CN		MCI		AD	
Actual	CN	414	(+18)	107	(−10)	19	(−8)
	MCI	164	(+9)	242	(+22)	132	(−31)
	AD	22	(+6)	70	(−3)	171	(−3)

by 12.5% (18 patients). This is also evident from the increase in TPF_{MCI} by 0.033, and in TPF_{CN} by 0.039.

In the time-to-dementia task, the unimodal baselines show the same pattern as in the diagnosis task: the linear model (using only tabular data) is outperforming the ResNet (0.12 lower mean *c*-index), and the linear model with ResNet features (0.026 lower mean *c*-index). For the concatenation-based networks, the improvement in *c*-index is at most 0.01, which, when taking the variance into account, must be considered insignificant. As above, the network by Duanmu et al. (2020) is performing worst and is even outperformed by the linear model (+0.013 mean *c*-index). In contrast to the diagnosis task, the FiLM-based network is falling behind the concatenation-based networks on the time-to-dementia task (−0.013 mean *c*-index). Here, the best performing baseline model is Concat1-FC. Overall, the proposed DAFT is outperforming all models by at least +0.019 *c*-index. This demonstrates that a two-way exchange of information between the image and tabular information, which only DAFT facilitates, is crucial for time-to-dementia prediction.

Next, we focus on the IBS, which measures a model's discriminative ability and calibration. Here, we include the Kaplan–Meier estimator, which estimates the time to dementia solely based on observed conversion times without considering tabular or image information. Therefore, it can be considered as worst-case upper bound on the IBS (Graf et al., 1999). The IBS indicates that many deep learning models are poorly calibrated. In particular, the ResNet performs very poorly with an IBS just 0.003 below that of the Kaplan–Meier estimator. The linear model achieved an IBS that is clearly below that of the Kaplan–Meier estimator. While the concatenation-based networks outperform the linear model in terms of *c*-index, only Concat-1FC can match it in terms of IBS, which indicates that Concat-2FC and 1FC-Concat-1FC sacrifice calibration for discriminative ability. In contrast, our proposed DAFT outperforms the linear model in terms of *c*-index (+0.029), while still matching the linear model's IBS of 0.122.

In summary, the results on the predictive performance demonstrate that concatenation-based approaches are unable to fully utilize the complementary nature of image and tabular information. Notably, we observed that integrating tabular data at different stages of image representation within a CNN, as done by Duanmu et al. (2020), can severely deteriorate performance. The sole approach, that excels at integrating image and tabular information for both tasks is the proposed DAFT network: DAFT outperforms all competing methods by a large margin (+0.021 bACC, +0.019 *c*-index).

4.2. Ablation study

To justify various design choices of the proposed DAFT, we perform an extensive ablation study on the diagnosis and time-to-dementia tasks. First, we evaluate the location of the DAFT within the last ResBlock, second, the activation function σ for the scale $\alpha_{i,c}$, and third, the impact of dynamically scaling and/or shifting each feature map. As proposed in Perez et al. (2018), parameters of batch normalization layers immediately preceding the DAFT are turned off. Results are summarized in Table 7.

They demonstrate that the DAFT works well regardless of its location, indicating a strong robustness against this design choice. Most notably, the DAFT outperforms all other models (with the exception of FiLM) on the diagnosis task, regardless of its location. For the progression task, the performance only decreases when placing DAFT before the first ReLU. Disabling either $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ results in a decrease for both task of at least 0.013 bACC and 0.002 *c*-index, respectively. For the diagnosis task, we can observe that scaling seems to be more important than shifting. In contrast, for the progression task, the capacity of the DAFT appears to be sufficient if either scaling or shifting is enabled, as the performance drop is within the variance of those configurations. Finally, applying a non-linear activation function σ to the scale diminishes the diagnosis performance, but increases the mean *c*-index for progression analysis for both activation functions sigmoid and tanh. With just two configurations being outperformed by concatenation-based networks, we can conclude that the DAFT is robust with respect to the design choices. Moreover, optimizing the configuration of the DAFT to a particular task can yield further performance gains over concatenation-based approaches.

Table 7
Results of the ablation study for DAFT. Values are mean test set performance and standard deviation across five cross-validation folds. Note that our proposed configuration (last row) uses DAFT before the first convolution with shift and scale predicted dynamically and the identity function $\sigma(x) = x$.

Configuration	Balanced accuracy	Concordance index
Before Last ResBlock	0.598 ± 0.038	0.749 ± 0.052
Before Identity-Conv	0.616 ± 0.018	0.745 ± 0.036
Before 1st ReLU	0.622 ± 0.024	0.713 ± 0.085
Before 2nd Conv	0.612 ± 0.034	0.759 ± 0.052
$\boldsymbol{\alpha}_i = \mathbf{1}$	0.581 ± 0.053	0.743 ± 0.015
$\boldsymbol{\beta}_i = \mathbf{0}$	0.609 ± 0.024	0.746 ± 0.057
$\sigma(x) = \text{sigmoid}(x)$	0.600 ± 0.025	0.756 ± 0.064
$\sigma(x) = \tanh(x)$	0.600 ± 0.025	0.770 ± 0.047
Proposed	0.622 ± 0.044	0.748 ± 0.045

4.3. Impact of scale and shift

To evaluate the benefits of dynamically scaling and shifting feature maps, rather than scaling/shifting by a (learned) constant, we perform a test time ablation study. First, we train a standard ResNet on the diagnosis tasks until convergence. Next, we initialize the weights of a ResNet with the FiLM or DAFT block with the weights of that network. Finally, we train both models for 20 epochs (with early stopping based on the validation set performance), but fix all weights with the exception of the last ResBlock and the FC layer. We note that this setup is important, because when learning the networks from scratch, the distribution over image feature maps that are the input to the FiLM or DAFT block could differ. By fixing the weights of layers preceding those blocks, the input feature maps fed to the FiLM and DAFT block are identical, resulting in a meaningful analysis of the behaviors of FiLM and DAFT.

Fig. 2 depicts the values for $\alpha_{i,c}$ and $\beta_{i,c}$ for all patients and feature maps. From this figure, it is evident that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ values produced by FiLM are often centered around zero, whereas the values produced by the DAFT are always consistently different from zero (except for $c = 9$). Moreover, we can observe that the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ values due to the DAFT block are more concentrated and evenly scattered in both dimensions, whereas the values due to FiLM are often characterized by a high variance in one dimension and low variance in the other dimension (e.g. $\text{Var}(\beta) \gg \text{Var}(\alpha)$ for $c = 4$).

Next, we replace either $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ with its mean across the training set to remove the respective conditioning information. The results in Fig. 3 suggest that the DAFT achieves a more effective integration of tabular information by shifting feature maps via $\boldsymbol{\beta}$, as indicated by the large performance loss when fixing $\boldsymbol{\beta}$ compared to $\boldsymbol{\alpha}$. In contrast, FiLM relies on both scaling and shifting and shows an overall decrease of performance irrespective of which vector is constrained. Moreover, our third test time ablation experiments supports this hypothesis, were we add Gaussian noise to $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$. Fig. 4 shows that DAFT is more sensitive to distortions of $\boldsymbol{\beta}$, while in FiLM the performance loss is equal for the two. Additionally, FiLM's performance deteriorates more quickly than that of DAFT as the variance σ of the noise increases, which highlights DAFT's overall robustness.

4.4. Impact of tabular data

Next, we want to analyze the marginal contribution individual tabular features have on the overall performance in the proposed DAFT network. Let $\mathbf{I}$ denote the 3D image information, $\mathbf{x}$ the tabular information,

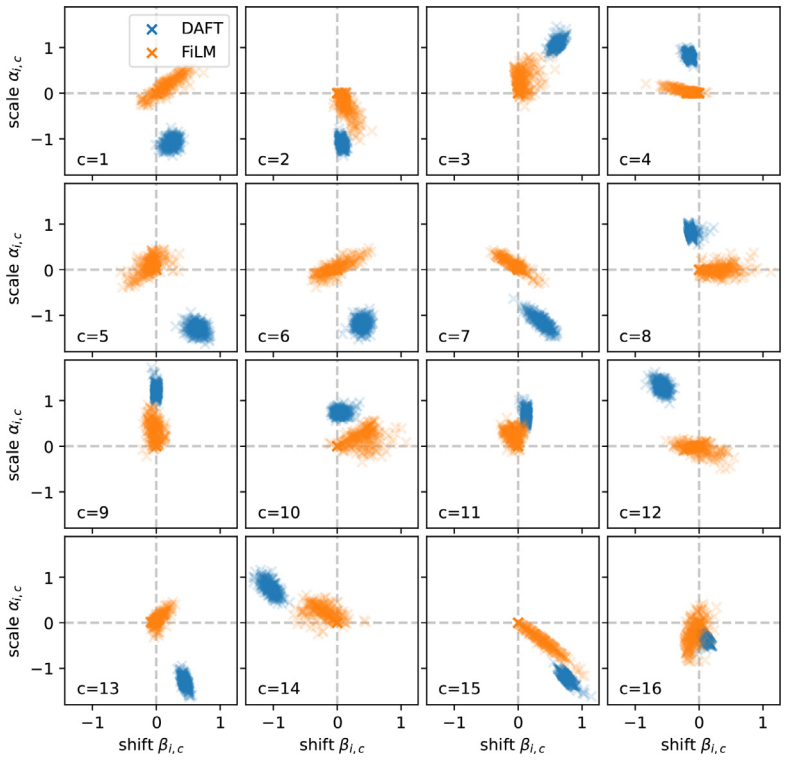

Fig. 2. Scatter plots for scale $\alpha_{i,c}$ and shift $\beta_{i,c}$ for all $C = 16$ feature maps for one model on the diagnosis task. Each sample i in the dataset is represented by a cross in each plot. The clusters of DAFT are compact and clearly tend towards non-zero values for each feature map. In contrast, for FiLM $\alpha_{i,c}$ and $\beta_{i,c}$ mostly remain close to the origin.

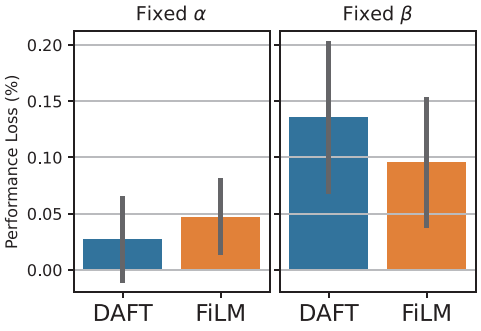

Fig. 3. Performance loss when setting either the scale $\boldsymbol{\alpha}$ or the shift $\boldsymbol{\beta}$ to the respective mean values in the training set.

y the actual label, and $\hat{M}(\mathbf{I}, \mathbf{x})$ the predicted class label by the DAFT network. We want to estimate the marginal contribution a single feature $j \in \{1, \ldots, P\} = \mathcal{F}$ has on the balanced accuracy $\text{bACC}(\{(y_i, \hat{M}(\mathbf{I}_i, \mathbf{x}_i))\}_{i=1}^{n})$. This can be cast as a problem from cooperative game theory, for which the Shapley Value is a suitable estimator (Covert et al., 2020; Shapley, 1953). The Shapley Value ϕ_j estimates the contribution of feature j by marginalizing over all possible subsets $\mathcal{S} \subseteq \mathcal{F}$:

$$\phi_j = \frac{1}{|\mathcal{F}|!} \sum_{\mathcal{S} \subseteq \mathcal{F} \setminus \{j\}} |\mathcal{S}|! \cdot (|\mathcal{F}| - |\mathcal{S}| - 1)! (\text{bACC}^{\mathcal{S} \cup \{j\}} - \text{bACC}^{\mathcal{S}}), \tag{4}$$

where $\text{bACC}^{\mathcal{S}}$ denotes the test set balanced accuracy of a model restricted to the features in the subset $\mathcal{S}$. The Shapley Values have the desirable property that they sum to the total improvement over the model using no tabular information: $\sum_{j=1}^{P} \phi_j = \text{bACC}^{\mathcal{F}} - \text{bACC}^{\emptyset}$. Hence, a negative Shapley Value would indicate that ignoring that feature would improve the overall performance. Note that the sum in (4) scales exponential in the number of tabular features, therefore we do not train a new model for each subset $\mathcal{S}$, but use a single pre-trained network where we mask the weights in the first FC layer of DAFT corresponding to the features in $\mathcal{S}$. This is also referred to as the Baseline Shapley approach (Sundararajan and Najmi, 2020).

Fig. 5 illustrates the estimated Shapley Values for the five DAFT-based models on the diagnosis task. We can observe that the three most important features across all folds are FDG-PET, T-tau, and Aβ_{42}. This result is reassuring as reduction of metabolic activity in cortical regions, as measured by FDG-PET, and high concentrations of CSF total tau and low

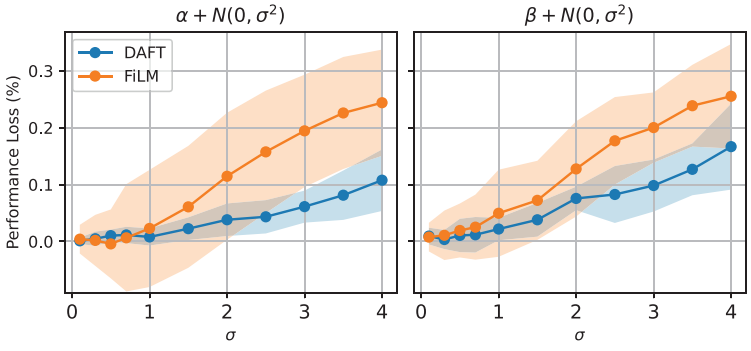

Fig. 4. Performance loss when distorting $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ with Gaussian noise. Lines represent the mean, shaded areas the standard deviation over 5 folds.

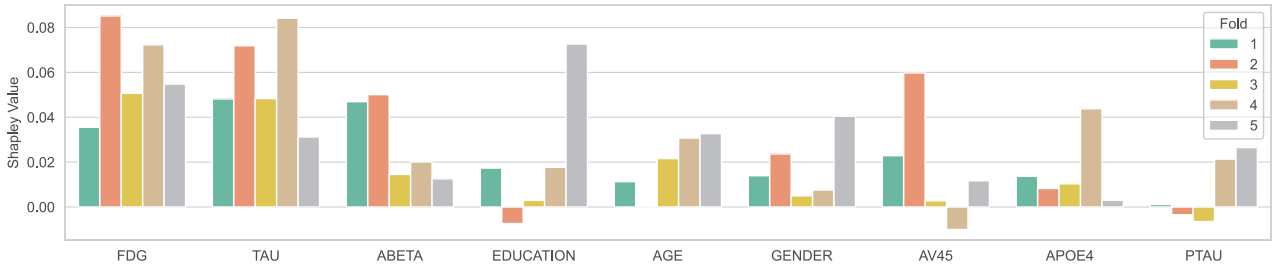

Fig. 5. Contribution of tabular features to the test set balanced accuracy of DAFT-based networks on the diagnosis task. The Shapley Value (y-axis) estimates the marginal contribution a single feature has on the balanced accuracy (see (4) for details). Positive (negative) Shapley Values indicate that a feature contributes to an increase (decrease) in performance.

Table 8
Predictive performance for the diagnosis task (balanced accuracy) of the repeatability study on ADNI and AIBL with a reduced set of tabular features. We report the mean and standard deviation across five folds and 15 random initializations of model weights. Higher values are better. The use of data of each model is indicated in columns 2–3 (I = the use of image data; T = use of tabular data). L/NL denote the linearity of the model for the tabular data transform (linear/non-linear).

			Balanced accuracy ↑		
			ADNI		AIBL
	I	T	Validation	Testing	Hold-out
Linear Model	✗	L	0.428 ± 0.037	0.417 ± 0.033	0.417 ± 0.009
ResNet	√	✗	0.554 ± 0.024	0.514 ± 0.036	0.493 ± 0.021
Linear Model /w ResNet Features	√	L	0.547 ± 0.039	0.536 ± 0.039	0.510 ± 0.021
Concat-1FC	√	L	0.577 ± 0.024	0.534 ± 0.041	0.515 ± 0.029
Concat-2FC	√	NL	0.521 ± 0.109	0.491 ± 0.098	0.475 ± 0.084
1FC-Concat-1FC	√	NL	0.570 ± 0.024	0.534 ± 0.042	0.515 ± 0.025
Duanmu et al. (2020)	√	NL	0.546 ± 0.052	0.513 ± 0.042	0.510 ± 0.039
FiLM (Perez et al., 2018)	√	NL	0.579 ± 0.022	0.541 ± 0.036	0.523 ± 0.028
DAFT	√	NL	0.581 ± 0.025	**0.550 ± 0.033**	**0.527 ± 0.024**

concentrations of CSF Aβ_{42} are well-known markers for AD (Blennow et al., 2001; Minoshima et al., 1997).

4.5. Generalization on AIBL and repeatability study

To validate our findings on hold-out data, we evaluate all models trained on ADNI on data from The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL; Ellis et al., 2009). As pointed out in Wen et al. (2020), models trained on ADNI typically do not generalize well on AIBL due deviating study protocols. In this experiment, we retrain all models for disease classification on a subset of tabular features, that are available for both ADNI and AIBL (age, gender, ApoE4 and a variable indicating if ApoE4 is missing). We follow the same pre-processing pipeline as for ADNI. Dataset statistics can be found in Table 1, third row. Additionally, we re-run every experiment, i.e. each hyper-parameter configuration, with 15 different random initializations of network weights, resulting in 225 trained models per test set. The results are summarized in Table 8.

DAFT performs best on ADNI and AIBL with respect to the reported bACC score. The gap to the runner-up model FiLM (Perez et al., 2018) is +0.009 bACC on ADNI and +0.004 bACC on AIBL. Compared to the initial experiment, the ranking of all models remains effectively the same, i.e. concatenation-based methods are outperformed by methods based on feature-merging on a channel-level (except for Duanmu et al., 2020). All models show a decrease in performance, and Concat-2FC performs worse than the unimodal ResNet and the Linear Model /w ResNet Features. The poor performance of Concat-2FC is due to training occasionally converging to a point at which the model only predicted one class. This highlights the importance of our work to promote robust deep learning architectures for heterogeneous feature fusion.

A t-test with Benjamini-Hochberg correction between DAFT and all competing methods on the test performance on ADNI and AIBL results in corrected p-values smaller than 0.018, except for FiLM: the corrected p-values are 0.12 and 0.297 on ADNI and AIBl respectively, strengthening our claim that feature-merging on a channel-level is superior to concatenation-based approaches. Compared to benchmark methods, the performance drop of DAFT compared to training on all available tabular data and the generalization of ADNI to AIBL is high (compare Tables 3 and 8). However, the only AD-specific biomarker in this experiment is ApoE4 and, thus, disease specific information readily available in the tabular data is limited. This manifests in the poor performance of the Linear Model. DAFT outperforms all other approaches, even across random initialization of network weights, and has the biggest relative increase in performance when meaningful tabular biomarkers are provided. This indicates that DAFT is able to incorporate tabular biomarkers more effectively than competing methods.

4.6. Generalization to other architectures

To validate our hypothesis that DAFT can be applied to any CNN to fuse image and tabular features, we change the backbone of all models to a ConvNet, which is essentially the ResNet in Fig. 1 without skip connections in ResBlocks. Again, we carry out a hyper-parameter search with the same search space as in previous experiments (see Section 3.2). The resulting bACC can be found in Table 9.

As in our main experiment, the performance of the Linear Model with ConvNet features increases marginally over the unimodal ConvNet, showing the existence of redundant high-level image and tabular features. DAFT outperforms all other models by more than +0.013 bACC, confirming its superior ability to integrate tabular data independent of the network's backbone architecture.

Table 9
Predictive performance for the diagnosis task (balanced accuracy) with a ConvNet as the backbone. We report the mean and standard deviation across five folds. Higher values are better. The use of data of each model is indicated in columns 2–3 (I = the use of image data; T = use of tabular data). L/NL denote the linearity of the model for the tabular data transform (linear/non-linear).

			Balanced accuracy ↑	
	I	T	Validation	Testing
Linear Model	✗	L	0.628 ± 0.034	0.582 ± 0.044
ConvNet	√	✗	0.587 ± 0.022	0.519 ± 0.027
Linear Model /w ConvNet Features	√	L	0.594 ± 0.018	0.534 ± 0.031
Concat-1FC	√	L	0.639 ± 0.011	0.604 ± 0.039
Concat-2FC	√	NL	0.652 ± 0.026	0.580 ± 0.018
1FC-Concat-1FC	√	NL	0.635 ± 0.030	0.579 ± 0.033
Duanmu et al. (2020)	√	NL	0.633 ± 0.032	0.571 ± 0.033
FiLM (Perez et al., 2018)	√	NL	0.644 ± 0.023	0.604 ± 0.018
DAFT	√	NL	0.643 ± 0.021	$\mathbf{0.617 \pm 0.018}$

Table 10
Runtime comparison. Training time is reported per epoch in seconds, inference time in milliseconds per forward pass. Times are reported for models with the ResNet backbone.

Model	Training	Inference
ResNet	8.91 s	1.80 ms
Concat1FC	8.94 s	1.91 ms
Concat2FC	8.92 s	1.97 ms
1FC-Concat-1FC	8.95 s	1.95 ms
Duanmu et al. (2020)	9.03 s	2.15 ms
FiLM (Perez et al., 2018)	8.74 s	2.02 ms
DAFT	8.96 s	2.15 ms

4.7. Runtime comparison

Since the models differ in terms of architecture and number of parameters (see Table 2), we further evaluated how these differences affect training and inference time. For each network, we measured the training time for one epoch of training, and for inference the time required for a forward pass of a batch of 256 samples (excluding time for I/O). All measures were carried out using PyTorch 1.5.1 and an NVIDIA GeForce GTX 1080 Ti graphics card.

Unsurprisingly, fully fitting a linear model can be carried out very efficiently (235 ms), whereas training one of the deep neural networks requires between 8.74 and 9.03 s per epoch (see Table 10). The wall time during inference scales with the number of weights and lies between 1.8 and 2.15 ms. While the differences for training are negligible, the differences in inference time are relatively high: With DAFT the inference time increases about 19% relative to ResNet. Nevertheless, for practical purposes this difference will remain unnoticeable. Therefore, we can conclude that the performance increase due to DAFT comes only at a minor increase in runtime.

5. Conclusion

The underlying changes that cause dementia can only be partially captured by brain MRI. While many previous deep learning methods focus purely on MRI, we have demonstrated that information on patient demographics, laboratory measurements, and genetics, all commonly encoded as tabular data, are required to put brain MRI into the right context and improve prediction accuracy. Other methods, that incorporate tabular data and brain MRI in deep learning frameworks, typically implement a naive concatenation mechanism, resulting in minimal exchange of information between the image- and tabular-related branches of those networks.

We proposed the Dynamic Affine Feature Map Transform (DAFT) to facilitate an improved two-way exchange of information between sources within a single CNN. DAFT is able to effectively incite or repress high-level concepts learned from a 3D image by conditioning feature maps of a convolutional layer on both image and tabular information. We compared DAFT against five state-of-the-art approaches on Alzheimer's disease diagnosis and time-to-dementia prediction. Our results showed that DAFT outperformed all previous deep learning approaches that combine image and tabular data by a large margin. Our experiments show that the features extracted by a CNN are, indeed, partially redundant to the discriminative tabular features readily available in clinical practice. Moreover, our exhaustive ablation study, generalization experiments, and repeatability study indicate that the DAFT is generally robust with respect to design choices. As a general concept to integrating image and tabular data, DAFT is applicable to many CNN architectures and medical data analysis tasks outside of dementia too.

Data/code availability statement

Alzheimer's Disease Neuroimaging Initiative (ADNI) and Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) used in this study are available at http://adni.loni.usc.edu and https://aibl.csiro.au, respectively, upon registration and compliance with the data usage agreement. The code used in this study is available at https://github.com/ai-med/DAFT.

Declaration of Competing Interests

Authors declare that they have no conflict of interest.

Credit authorship contribution statement

Tom Nuno Wolf: Methodology, Software, Validation, Writing – original draft. **Sebastian Pölsterl:** Data curation, Methodology, Software, Writing – original draft. **Christian Wachinger:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Acknowledgments

This research was supported by the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation, and the Federal Ministry of Education and Research in the call for Computational Life Sciences (DeepMentia, 031L0200A). The authors thank the Leibniz Supercomputing Centre for funding this project by providing computing time on its Linux-Cluster.

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

San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P6000 GPU used for this research.

References

Barnes, J., et al., 2010. Head size, age and gender adjustment in MRI studies: a necessary nuisance? NeuroImage 53 (4), 1244–1255. doi:10.1016/j.neuroimage.2010.06.025.

Blennow, K., Vanmechelen, E., Hampel, H., 2001. CSF total tau, Aβ42 and phosphorylated tau protein as biomarkers for Alzheimer's disease. Mol. Neurobiol. 24 (1–3), 087–098.

Braman, N., Gordon, J.W.H., Goossens, E.T., Willis, C., Stumpe, M.C., Venkataraman, J. (2021). Deep Orthogonal Fusion: Multimodal Prognostic Biomarker Discovery Integrating Radiology, Pathology, Genomic, and Clinical Data. In: Medical Image Computing and Computer Assisted Intervention MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12905. Springer, Cham. doi:10.1007/978-3-030-87240-3_64

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. ICPR.

Bron, E.E., et al., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage 111, 562–579.

Covert, I., Lundberg, S.M., Lee, S.-I., 2020. Understanding global feature contributions with additive importance measures. In: NeurIPS, vol. 33, pp. 17212–17223.

Cox, D.R., 1972. Regression models and life tables (with discussion). J. R. Stat. Soc. Ser. B (Statistical Methodology) 34, 187–220.

Duanmu, H., et al., 2020. Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data. In: MICCAI, pp. 242–252.

Ebrahimighahnavieh, M.A., Luo, S., Chiong, R., 2020. Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. Comput. Methods Programs Biomed. 187, 105242.

El-Sappagh, S., Abuhmed, T., Islam, S.M.R., Kwak, K.S., 2020. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. Neurocomputing 412, 197–215.

Ellis, K., Bush, A., Darby, D., et al., 2009. The Australian imaging, biomarkers and lifestyle (AIBL) study of aging. Int. Psychogeriatr. 21 (04), 672–687.

Esmaeilzadeh, S., Belivanis, D.I., Pohl, K.M., Adeli, E., 2018. End-to-end Alzheimer's disease diagnosis and biomarker identification. In: MLMI, pp. 337–345.

Faraggi, D., Simon, R., 1995. A neural network model for survival data. Stat. Med. 14 (1), 73–82.

Fischl, B., 2012. FreeSurfer. NeuroImage 62 (2), 774–781.

Frisoni, G.B., et al., 2008. Mapping local hippocampal changes in Alzheimer's disease and normal ageing with MRI at 3 Tesla. Brain 131 (12), 3266–3276.

Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M., 1999. Assessment and comparison of prognostic classification schemes for survival data. Stat. Med. 18 (17–18), 2529–2545.

Hao, J., Kosaraju, S.C., Tsaku, N.Z., Song, D.H., Kang, M., 2019. PAGE-Net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. In: Biocomputing 2020, pp. 355–366.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. CVPR.

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

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2020. Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. 42 (8), 2011–2023.

Jacenków, G., O'Neil, A.Q., Mohr, B., Tsaftaris, S.A., 2020. INSIDE: steering spatial attention with non-imaging information in CNNs. In: MICCAI, pp. 385–395.

Jack, C.R., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.

Jack, C.R., et al., 2013. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 12 (2), 207–216.

Jarrett, D., Yoon, J., van der Schaar, M., 2020. Dynamic prediction in clinical survival analysis using temporal convolutional networks. IEEE J. Biomed. Health Inform. 24 (2), 424–436.

Kopper, P., Pölsterl, S., Wachinger, C., Bischl, B., Bender, A., Rügamer, D., 2021. Semi-structured deep piecewise exponential models. In: Proc. AAAI Spring Symposium on Survival Prediction, vol. 146, pp. 40–53.

Li, H., Habes, M., Wolk, D.A., Fan, Y., 2019. A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimer's Dementia 15 (8), 1059–1070.

Li, S., Shi, H., Sui, D., Hao, A., Qin, H., 2020. A novel pathological images and genomic data fusion framework for breast cancer survival prediction. EMBC.

Liu, M., Zhang, J., Adeli, E., Shen, D., 2019. Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. IEEE Trans. Biomed. Eng. 66 (5), 1195–1206.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. ICLR.

Minoshima, S., Giordani, B., Berent, S., Frey, K.A., Foster, N.L., Kuhl, D.E., 1997. Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. Ann. Neurol. 42 (1), 85–94.

Mobadersany, P., et al., 2018. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc. Natl. Acad. Sci. USA 115 (13), E2970–E2979.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: ICML, pp. 807–814.

Pelka, O., et al., 2020. Sociodemographic data and APOE-e4 augmentation for MRI-based detection of amnestic mild cognitive impairment using deep learning systems. PLoS One 15 (9), e0236868.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A., 2018. FiLM: visual reasoning with a general conditioning layer. AAAI, 32.

Pölsterl, S., 2020. Scikit-survival: a library for time-to-event analysis built on top of scikit-learn. J. Mach. Learn. Res. 21 (212), 1–6.

Pölsterl, S., Sarasua, I., Gutiérrez-Becker, B., Wachinger, C., 2020. A wide and deep neural network for survival analysis from anatomical shape and tabular clinical data. In: Machine Learning and Knowledge Discovery in Databases, pp. 453–464.

Pölsterl, S., Wolf, T.N., Wachinger, C., 2021. Combining 3D Image and Tabular Data via the Dynamic Affine Feature Map Transform. In: Medical Image Computing and Computer Assisted Intervention MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12905. Springer, Cham doi:10.1007/978-3-030-87240-3_66.

Shapley, L.S., 1953. A value for *n*-person games. Contrib. Theory Games 2 (28), 307–317.

Spasov, S., Passamonti, L., Duggento, A., Liò, P., Toschi, N., 2019. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. NeuroImage 189, 276–287.

Stern, Y., et al., 2020. Whitepaper: defining and investigating cognitive reserve, brain reserve, and brain maintenance. Alzheimer's Dementia 16, 1305–1311. doi:10.1016/j.jalz.2018.07.219.

Sundararajan, M., Najmi, A., 2020. The many Shapley values for model explanation. In: ICML, vol. 119, pp. 9269–9278.

Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B., Wei, L.J., 2011. On the *C*-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat. Med. 30 (10), 1105–1117.

Wen, J., et al., 2020. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med. Image. Anal. 63, 101694.