Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data

Paul Hager[1,2] Martin J. Menten[1,2,3] Daniel Rueckert[1,2,3]
[1]Technical University of Munich, [2]Klinikum Rechts der Isar, [3]Imperial College London
{paul.hager, martin.menten, daniel.rueckert}@tum.de

Abstract

Medical datasets and especially biobanks, often contain extensive tabular data with rich clinical information in addition to images. In practice, clinicians typically have less data, both in terms of diversity and scale, but still wish to deploy deep learning solutions. Combined with increasing medical dataset sizes and expensive annotation costs, the necessity for unsupervised methods that can pretrain multimodally and predict unimodally has risen.

To address these needs, we propose the first self-supervised contrastive learning framework that takes advantage of images and tabular data to train unimodal encoders. Our solution combines SimCLR and SCARF, two leading contrastive learning strategies, and is simple and effective. In our experiments, we demonstrate the strength of our framework by predicting risks of myocardial infarction and coronary artery disease (CAD) using cardiac MR images and 120 clinical features from 40,000 UK Biobank subjects. Furthermore, we show the generalizability of our approach to natural images using the DVM car advertisement dataset.

We take advantage of the high interpretability of tabular data and through attribution and ablation experiments find that morphometric tabular features, describing size and shape, have outsized importance during the contrastive learning process and improve the quality of the learned embeddings. Finally, we introduce a novel form of supervised contrastive learning, label as a feature (LaaF), by appending the ground truth label as a tabular feature during multimodal pretraining, outperforming all supervised contrastive baselines.[1]

1. Introduction

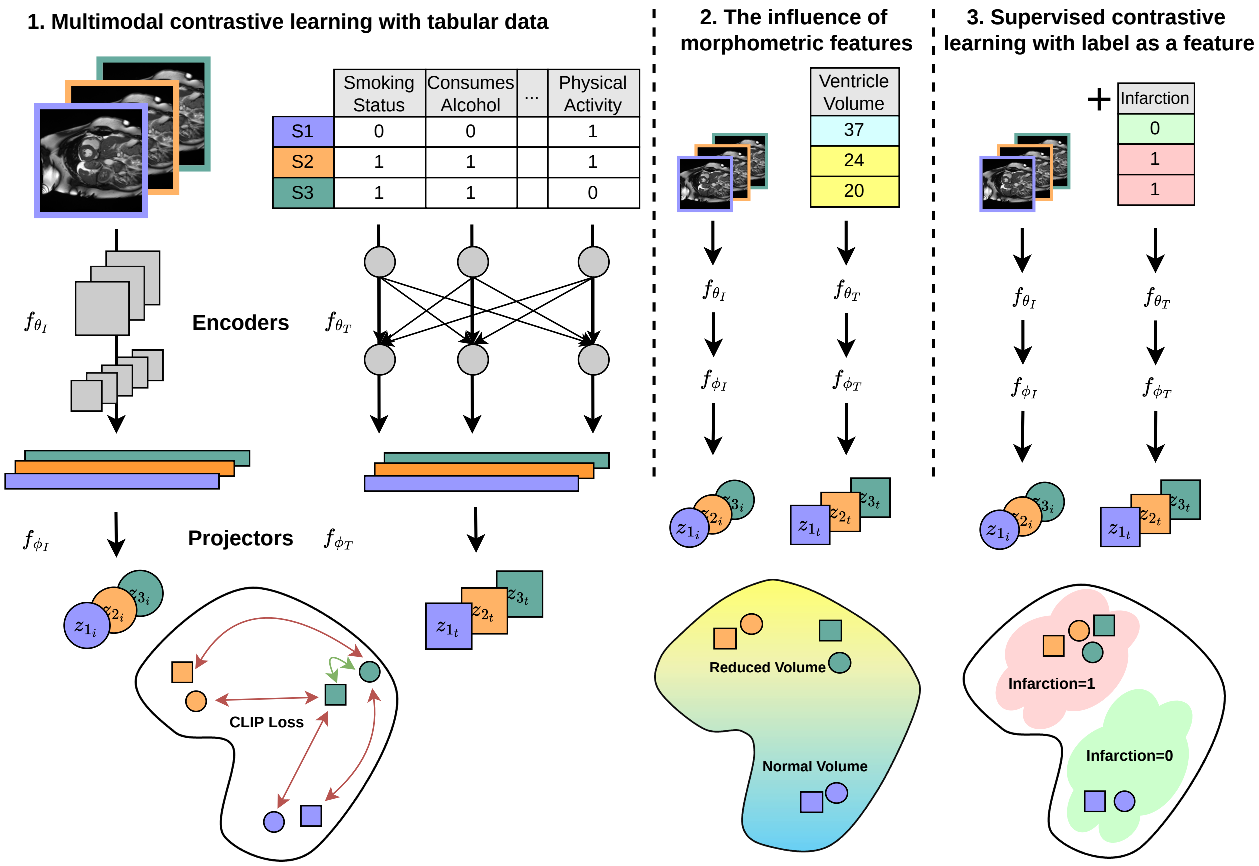
Modern medical datasets are increasingly multimodal, often incorporating both imaging and tabular data. Images can be acquired by computed tomography, ultrasound, or magnetic resonance scanners, while tabular data commonly originates from laboratory tests, medical history and patient lifestyle questionnaires. Clinicians have the responsibility to combine and interpret this tabular and imaging data to diagnose, treat, and monitor patients. For example, cardiologists may ask about a patients' family history and record their weight, cholesterol levels, and blood pressure to better inform diagnoses when examining images of their heart.

Beyond diagnostics, multimodal data is also crucial to advance the understanding of diseases motivating the creation of biobanks. Going far beyond the scale of typical datasets in hospitals, biobanks pool vast amount of information from large populations. Multimodal biobanks include the German National Cohort [21] with 200,000 subjects, Lifelines [52] with 167,000 subjects, and the UK Biobank [54] with 500,000 subjects. The UK Biobank includes thousands of data fields from patient questionnaires, laboratory tests, and medical examinations, in addition to imaging and genotyping information. Biobanks have already proven useful in the training of machine learning models to predict many diseases such as anaemia [39], early brain aging [32] and cardiovascular disease [1, 51].

There is a substantial interest in deploying algorithms that have been developed using these large-scale population studies in clinical practice. However, acquiring the same quality of data, both in terms of diversity of modalities and number of features, is not feasible in a busy clinical workflow [20]. Furthermore, low disease frequencies make supervised solutions hard to train. Consequently, there is a clear need for unsupervised strategies that can learn from biobank scale datasets and be applied in the clinic where considerably less data, in size and dimension, is available.

Our contribution To address these needs, we propose the first contrastive framework that utilizes imaging and tabular data, shown in figure 1. Our framework is based on SimCLR [13] and SCARF [6], two leading contrastive learning solutions, and is simple and effective. We demonstrate the utility of our pretraining strategy on the challenging task of

[1] https://github.com/paulhager/MMCL-Tabular-Imaging



	Smoking Status	Consumes Alcohol	...	Physical Activity
S1	0	0		1
S2	1	1		1
S3	1	1		0

Ventricle Volume
37
24
20

Infarction
0
1
1

Figure 1. We combine imaging and tabular data in a contrastive learning framework. We observe that morphometric features, describing shape and size, are of outsized importance in multimodal contrastive training and their inclusion boosts downstream task performance. By simply adding the label as a tabular feature we introduce a novel form of supervised contrastive learning that outperforms all other supervised contrastive strategies.

predicting cardiac health from MR images. Beyond medical imaging, we show that our framework can also be applied when combining natural images and tabular data using the DVM car advertisement dataset [29].

Experimentally, we observe that our tool leverages morphometric features during contrastive learning. Morphometric features describe the size and shape of an object and therefore correlate with extractable imaging features. We quantitatively demonstrate the importance of these features in the contrastive learning process using attribution methods, such as integrated gradients [55], and ablation experiments.

Finally, we introduce a new supervised contrastive learning method called label as a feature (LaaF). By appending the target label as a tabular feature, our method outperforms previously published strategies that incorporate labels into the contrastive framework. Our method is also highly flexible and can be combined with the aforementioned strategies to further improve performance.

2. Related Work

Self-supervised learning with images aims to extract useful features from unlabeled data. Historically, this was attempted by solving hand-crafted pretext tasks such as jigsaw puzzles [44, 58, 59, 72], colorization [36, 63, 71], image inpainting [45], and context prediction [7, 12, 19]. The major difficulties with using these methods is that they tend to overfit on the specifics of their pretext task, limiting their utility for downstream tasks.

Contrastive learning has emerged as a popular and performant successor to pretext tasks. Contrastive learning trains encoders by generating augmented views of a sample and maximizing their projected embedding similarity while minimizing the similarity between the projected embeddings of other samples [24]. It has been popularized recently by implementations such as SimCLR [13], MOCO [25], BYOL, [23] and others [9, 10, 14, 16, 70]. We use the contrastive framework of SimCLR as the basis for our work.

Deep learning with tabular data has recently begun to yield results that are competitive with classical machine learning methods [4, 8, 28], though for many applications they still underperform simpler algorithms [8, 53]. Self-supervised learning is being explored in the tabular domain with frameworks such as VIME [66] and contrastive methods such as SubTab [61] and SCARF [6]. We base our tabular augmentations on those used in SCARF.

Multimodal contrastive learning with images is becoming more important as the number of multimodal datasets increases and multimodal training strategies become more effective. Approaches such as CLIP [49], which combines images and text, are general-purpose vision models that are able to solve new tasks in a zero-shot manner. Some of these models use internet-size datasets and are described as foundational models, such as UniCL [65], Florence [68], ALIGN [31], and Wu Dao 2.0 [18]. Outside of the image-language domain, there has also been progress on multimodal contrastive learning using two different imaging modalities [47, 58], audio and video [37], video and text [64, 73], and imaging and genetic data [57]. While literature on generative self-supervised tabular and imaging models [3] [34] exists, it is limited in scope, using only two or four clinical features. To the best of our knowledge, there is no implementation of a contrastive self-supervised framework that incorporates both images and tabular data, which we aim to address with this work.

Supervised learning within contrastive frameworks has been shown to outperform the binary cross entropy loss in some cases and create more robust embeddings [33]. Supervised contrastive learning [33] maximizes the similarity of the projected embeddings of all views in a batch from the same class. This also addresses the problem of false negatives in contrastive learning, which is that the contrastive loss minimizes projected embedding similarity between different samples even if they are part of the same class according to a downstream task (i.e. false negatives). By utilizing the available labels, supervised contrastive learning is able to circumvent this problem and outperforms other methods that heuristically identify and eliminate false negatives [15,30]. We propose a solution for supervised learning in our multimodal contrastive framework that takes advantage of the unique strengths of tabular data by appending the label as a tabular feature.

3. Methods

3.1. Contrastive Framework for Tabular and Imaging Data

We base our multimodal framework on SimCLR [13]. Let our dataset be x and a unique sample be j. Each batch contains pairs of imaging x_{j_i} and tabular x_{j_t} samples which are augmented. Each augmented imaging sample x_{j_i} in the batch is passed through an imaging encoder f_{θ_I} to generate the embedding $\widetilde{x_{j_i}}$. Each augmented tabular sample x_{j_t} in the batch is passed through a tabular encoder f_{θ_T} to generate the embedding $\widetilde{x_{j_t}}$. The embeddings are propagated through separate projection heads f_{ϕ_I} and f_{ϕ_T} and brought into a shared latent space as projections z_{j_i} and z_{j_t} which are then L2 normalized onto a unit hypersphere. The projections are pulled and pushed in the shared latent space according to the “CLIP” loss [49], which maximizes the cosine similarity of projections from the same sample and minimizes the similarity of projections from different samples. In contrast to the original InfoNCE [43] loss and following CLIP, we only contrast projections between modalities, never within one modality.

i and t can be used interchangeably and so, without loss of generality, the projection of an image is defined as

$$z_{j_i} = f_{\phi_I}(f_{\theta_I}(x_{j_i})) \tag{1}$$

Considering all subjects $\mathcal{N}$ in a batch, the loss for the imaging modality is

$$\ell_{i,t} = -\sum_{j\in\mathcal{N}} log \frac{\exp(\cos(z_{j_i}, z_{j_t})/\tau)}{\sum\limits_{k\in\mathcal{N},k\neq j} \exp(\cos(z_{j_i}, z_{k_t})/\tau)}. \tag{2}$$

$\ell_{t,i}$ is calculated analagously and the total loss is thus

$$\mathcal{L} = \lambda\ell_{i,t} + (1-\lambda)\ell_{t,i}. \tag{3}$$

The images in the batch are augmented based on the standard contrastive augmentations specified in [13]: horizontal flips, rotations, color jitter, and resized crop. We do not use Gaussian blurring on the cardiac dataset in order to preserve fine-grained features in the MR images [5]. To effectively augment the tabular data, a fraction of a subject’s features are randomly selected to be “corrupted” (i.e. augmented), following [6]. Each corrupted feature’s value is sampled with replacement from all values for that feature seen in the dataset. Full implementation details are in the supplementary materials.

3.2. Explainability using Integrated Gradients

To improve our understanding of the dynamics of the multimodal training, we analyze the importance of the individual tabular features in generating the embeddings. Using test samples, we take the pretrained tabular encoder of our multimodal model and calculate the integrated gradients [55] of each dimension of the embeddings. This integrates the gradients of the encoder along the straightline path from a baseline sample, in our case a zero vector, to the test sample in question. This yields the importance value of each tabular feature in generating the downstream prediction for that sample. We then take the absolute value and calculate the mean importance of each feature across all embedding dimensions. Categorical features have their means summed over all choices. We use these results to categorize features and better understand how training in a multimodal setting influences unimodal performance.

3.3. Contrastive Learning with Labels

Incorporating labels into the contrastive learning process is typically done by modifying the loss function [15,33]. We propose to take advantage of the unique structure of tabular data and directly append the downstream class label as a tabular feature. We explore the benefits of combining our method with existing strategies for incorporating labels into the training process, such as supervised contrastive learning and false negative elimination.

4. Experiments and Results

4.1. Datasets

As a first dataset, we used cardiac MR images and clinical information from the UK Biobank population study. Our aim was to predict past and future risk of myocardial infarction and coronary artery disease (CAD). We used short axis cardiac MR images, which provide a cross-sectional view of the left and right ventricle of the heart. The images used are two-channel 2D images whose channels are the middle baso-apical slice of the short axis cardiac MRI image at end-systolic and end-diastolic phases. The short axis images were chosen as left ventricular function and morphometry are impacted by both CAD [69] and cardiac infarction [56]. Conversely, the left ventricle is a high-risk area in which early warning signs of cardiac dysfunction may be visible [2, 50, 60]. The images were zero-padded to 210x210 pixels and min-max normalized to a range of 0 to 1. After augmentations (see supplementary materials), final image size was 128x128 pixels.

A subset of demographic, lifestyle and physiological features out of 5,000 data fields included in the UK Biobank dataset were selected for the tabular data. These features were chosen based on published correlations with cardiac outcomes. They include information about the subjects' diet [67], physical activity [42], weight [48], alcohol consumption [46], smoking status [35], and anxiety [11]. Only features with at least 80% coverage were included. The full list of features can be found in the supplementary materials. The continuous tabular data fields were standardized using z-score normalization with a mean value of 0 and standard deviation of 1 while categorical data was one-hot encoded. The total size of the dataset was 40,874 unique subjects, split into 29,428 training, 7,358 validation, and 4,088 testing pairs of imaging and tabular data.

The first prediction target is myocardial infarction as defined by the International Classification of Diseases (ICD10) code. ICD10 codes are maintained by the World Health Organization, used to record diagnoses during hospital admissions, and made available through the UK Biobank. The second prediction target is CAD, also defined by ICD10 code. The ICD codes used for each class are listed in the supplementary materials. We combine past and future diagnoses since infarctions and CAD can go undiagnosed for many years and may only be recorded once a patient has to be treated for a severe cardiac event, making it difficult to establish when the disease began [41, 62]. As both diseases are low frequency in the dataset (3% for infarction and 6% for CAD), finetune train splits were balanced using all positive subjects and a static set of randomly chosen negative subjects. The test and validation sets were left untouched.

The second dataset is the Data Visual Marketing (DVM) dataset that was created from 335,562 used car advertisements [29]. The DVM dataset contains 1,451,784 images of cars from various angles (45 degree increments) as well as their sales and technical data. For our task we chose to predict the car model from images and the accompanying advertisement data. The images were all 300x300 pixels and after augmentations (see supplementary materials) final image size was 128x128. All fields that provided semantic information about the cars in question were included, such as width, length, height, wheelbase, price, advertisement year, miles driven, number of seats, number of doors, original price, engine size, body type, gearbox type, and fuel type. Unique target identifying information like brand and model year were excluded. The width, length, height and wheelbase values were randomly jittered by 50 millimeters so as not to be uniquely identifying. We pair this tabular data with a single random image from each advertisement, yielding a dataset of 70,565 train pairs, 17,642 validation pairs, and 88,207 test pairs. Car models with less than 100 samples were removed, resulting in 286 target classes.

To handle missing tabular data, we used an iterative multivariate imputer which models missing features as a function of existing features over multiple imputation rounds. This was done after normalization, to ensure that the means and standard deviations were calculated only from recorded values. The missing features were initialized with the mean and then entire columns were imputed in order from least amount of missing features, to most amount of missing features. A regressor was fit with all other features as input and the currently examined column as dependent variable. This process was repeated a maximum of n times or until $\frac{max(abs(X_t - X_{t-1}))}{max(abs(X))} < \epsilon$, where X_t is the feature vector being imputed at time point t and ϵ is a provided tolerance, typically $1e^{-3}$. Categorical data was then rounded to the nearest integer i.e. category.

Table 1. Performance of our framework on the tasks of myocardial infarction, coronary artery disease (CAD) and DVM car model prediction from images. Our multimodal pretrained model outperforms all other models on every task. The best performing model for every input type is displayed in **bold** font. Our method is highlighted gray.

Model	AUC (%) Frozen / Infarction	AUC (%) Trainable / Infarction	AUC (%) Frozen / CAD	AUC (%) Trainable / CAD	Top-1 Accuracy (%) Frozen / DVM	Top-1 Accuracy (%) Trainable / DVM
Supervised ResNet50	72.37±1.80	72.37±1.80	68.84±2.54	68.84±2.54	87.97±2.20	87.97±2.20
SimCLR	73.69±0.36	73.62±0.70	69.86±0.21	71.46±0.71	65.48±0.48	88.76±0.81
BYOL	69.18±0.43	70.69±2.09	66.91±0.19	70.66±0.22	59.73±0.28	89.18±0.90
SimSiam	71.72±0.18	72.31±0.26	67.79±0.12	70.13±0.35	22.11±2.83	87.43±0.88
BarlowTwins	66.06±1.11	71.35±1.23	62.90±0.23	69.63±0.58	52.57±0.08	85.47±0.82
Multimodal Imaging	**76.35±0.19**	**75.37±0.43**	**74.45±0.09**	**73.08±0.75**	**91.43±0.13**	**93.00±0.18**

4.2. Experimental Setup

All imaging encoders are ResNet50s [26] that generate embeddings of size 2048. Our multimodal model uses a tabular encoder which is a multilayer perceptron (MLP) with one hidden layer of size 2048 that generates embeddings of size 2048. All weights are randomly initialized. Our imaging projector is an MLP with one hidden layer of size 2048 and our tabular projector generates projections directly from the embeddings with a fully connected layer. Projection size is 128 following [13]. After pretraining, the projection head is removed and a fully connected layer to the output class nodes is added.

To evaluate the effectiveness of our model, we compare it to a fully supervised ResNet50 as well as multiple contrastive solutions such as SimCLR [13], BYOL [22], SimSiam [17], and BarlowTwins [70]. We use linear probing of frozen networks to evaluate the quality of the learned representations [13, 15, 33]. We also benchmark each network while leaving all weights trainable during finetuning, as this typically improves upon the frozen counterpart and would be used in practice. For the cardiac classification tasks we use area under the receiver operating characteristic curve (AUC) as our metric because the dataset is severely unbalanced for our targets. With only 3-6% positive labels, a model that always predicts the negative class would achieve an accuracy of 94-97%. For the DVM cars dataset we report top-1 accuracy as we have 280+ classes. Results are reported as a mean and standard deviation calculated over five different seeds set during finetuning. Both cardiac tasks were evaluated from a single pretrained model. Full experimental details including the setup of the baseline models can be found in the supplementary materials.

4.3. Multimodal Pretraining Improves Unimodal Prediction

Our main results showing the strength of our multimodal pretraining framework are found in table 1. All results shown use only images as input, as this is the clinically relevant task. Results using tabular inputs are shown in the supplementary materials. Our multimodal pretrained model substantially outperformed all other models on all three tasks and both finetuning strategies. SimCLR generally outperformed all other contrastive strategies, highlighting our decision to base our multimodal strategy on it.

On the cardiac tasks, the multimodal model achieved its best results when freezing the encoder. We hypothesize that this is due to overfitting on the imaging modality during finetuning. We suspect when provided with an imaging-only signal during finetuning that the encoder discarded features that were learned from tabular data.

When predicting the car model from images of the DVM dataset, our multimodal model outperformed other pretraining strategies by even larger margins during frozen linear probing. This shows that for a homogeneous dataset like the DVM cars, having an additional differentiating signal such as tabular data can better align the learned features to the downstream target classes.

4.4. Multimodal Pretraining is Beneficial in Low-Data Regimes

When investigating rare medical conditions with very low label frequencies, models must be performant when few positive samples are available. In order to test the performance of the learned encoders in this low-data regime, we sampled the finetuning training dataset to 10% and 1% of its original size, with each subset being wholly contained within each superset. Because of the low frequencies of the positive class, this resulted in balanced training set sizes of 200 (10%) and 20 (1%) for myocardial infarction and 400 (10%) and 40 (1%) for CAD. The test and validation set was kept identical to the full data regime. A graphical representation of the results is shown in figure 2. We find that that in low-data regimes our multimodal framework generally outperforms the imaging-only contrastive method by larger margins than with full data. This indicates improved representations that require less finetuning samples to achieve the same performance and higher utility when rare diseases are the target. We again see that our multimodal frozen encoder consistently outperforms our trainable encoder. We bench-

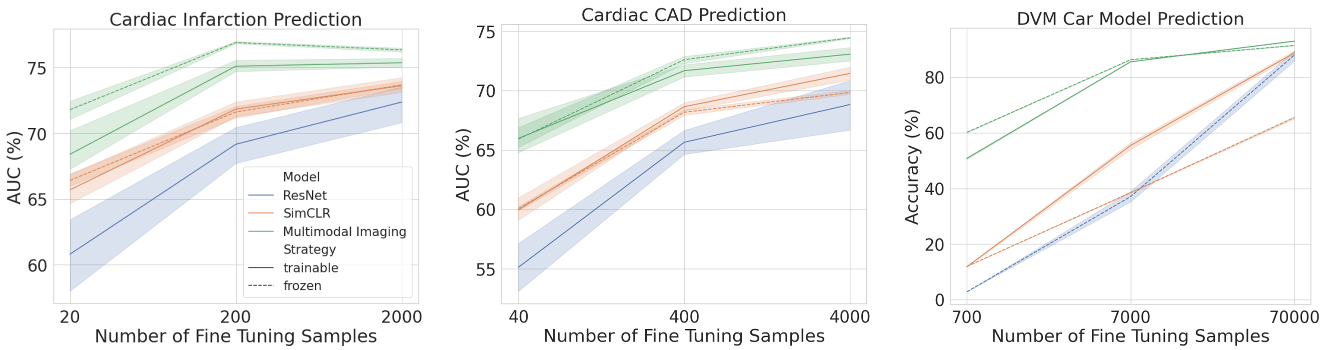


Figure 2. Performance of the imaging models with different number of finetuning training samples. Shaded regions indicate 95% confidence intervals. Pretraining with both images and tabular data excels at all data quantities and is well suited for rare disease identification when only tens or hundreds of labels are available.

Table 2. Frozen finetune performance of multimodal models pretrained with all features, morphometric features only, and no morphometric features. Even though the total importance of morphometric features was less than that of non-morphometric features on the cardiac task, their exclusion worsened or had equal impact on downstream performance. Best score is in **bold** font, second best <u>underlined</u>.

Experiment	Tabular Features	Importance Percentage (%)	AUC (%) Infarction	AUC (%) CAD	Tabular Features	Importance Percentage (%)	Top-1 Accuracy (%) DVM
MM Imaging Baseline	117	100.0	**76.35±0.19**	**74.45±0.09**	16	100.0	<u>91.43±0.13</u>
Morphometric Features	24	47.0	75.22±0.30	<u>73.71±0.09</u>	5	56.4	**92.33±0.05**
Non-Morphometric Features	93	53.0	<u>75.46±0.19</u>	72.18±0.25	11	43.6	89.14±0.24

mark against SimCLR as it was the strongest contrastive pretraining strategy. Comparisons against all pretraining strategies can be found in the supplementary materials.

4.5. Morphometric Features Improve Embedding Quality

To explore why training in a multimodal fashion improves the unimodal encoders, we analyzed the contributions of the tabular features to the improved embeddings. A unique strength of tabular data is that each of its input nodes corresponds to a single feature. We divided the features into two categories, morphometric and non-morphometric. Morphometric features are related to size and shape and have direct correlates in the images, such as ventricular volume, weight or car length. Using integrated gradients, we calculated the importance of each feature across the test samples.

To generate the cardiac embeddings, the seven most important features were all morphometric, even though they only represent one fifth of all features. Furthermore, all 24 morphometric features were found in the top half of the importance rankings. These results are shown in figure 3 and supplementary materials.

We hypothesize that the model focuses on morphometric tabular features because these have direct correlates in the images. Extracting these features in both modalities increases the projected embedding similarity and minimizes the contrastive loss, as shown in the supplementary materials.

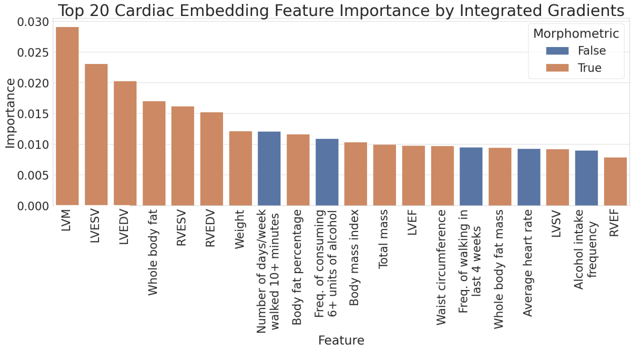


Figure 3. Top 20 most impactful features for calculating embeddings determined using integrated gradient feature attribution method. The morphometric features are colored orange and comprise 15 of the 20 most impactful features.

This is corroborated by the Guided GradCam results, shown in figure 5, where it is seen that the imaging model primarily focused on the left ventricle. Incidentally, the three most important features according to the integrated gradients are left ventricle mass (LVM), left ventricle end systolic volume (LVESV), and left ventricle end diastolic

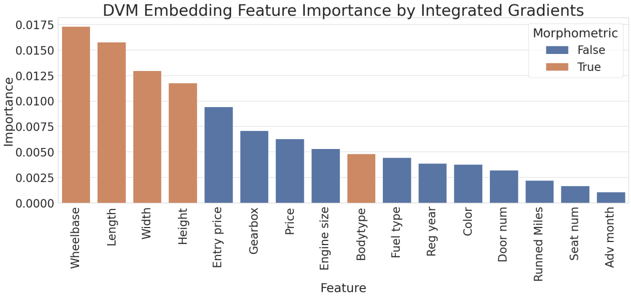


Figure 4. Impact of features for calculating DVM embeddings determined using integrated gradient feature attribution method. The morphometric features are colored orange and comprise the four most impactful features.

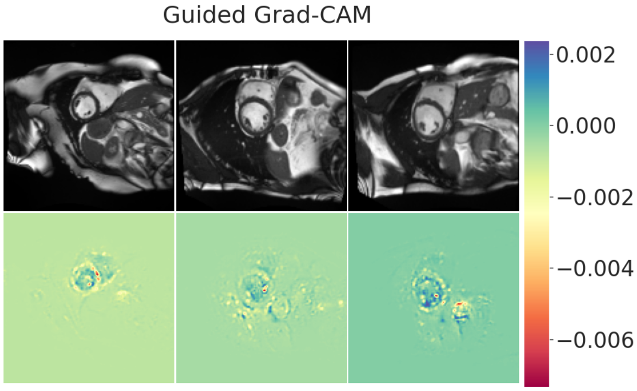


Figure 5. Guided Grad-CAM results for predicting CAD on test images. The most important features are centered around the left ventricle, matching the most important tabular features.

volume (LVEDV). To analyze the impact of these tabular features on downstream performance, we trained once with only morphometric features and once with only non-morphometric features. We observed that the morphometric features have an outsized impact on generating the embeddings. Table 2 shows that even though they only contribute 46.99% of the total importance and constitute 24 out of 117 features, their exclusion on CAD prediction degrades performance, and is equal to the exclusion of non-morphometric features on infarction prediction. In general, this shows that the multimodal pretraining process is fairly robust to feature selection, especially when the total feature set is so large and there exist collinearities within the data.

Similar results are seen on the DVM dataset where the top 4 most important features are all morphometric features, despite there only being 5 morphometric features in total, as seen in figure 4. Table 2 shows the effect of removing these features, which led to a substantial drop in accuracy. When training with only morphometric features the accuracy increased highlighting their importance on tasks that are shape driven, like car model identification.

4.6. Appending the Label as a Tabular Feature Boosts Supervised Contrastive Strategies

We introduce a novel form of supervised contrastive learning by including the ground truth label as a tabular feature (LaaF). We benchmark the effectiveness of this approach by comparing it to both supervised contrastive learning with full label information as well as false negative elimination with full label information. The results presented in table 3 show that LaaF outperforms or rivals both strategies.

On the cardiac binary classification tasks there is a sharp class imbalance. 97% of the subjects are negative for myocardial infarction, leading false negative elimination to remove large portions of each batch before calculating the contrastive loss. This leads to worse representations as batch sizes are drastically reduced during training. As contrastive learning, and especially SimCLR, is known to be sensitive to batch sizes [13] this degrades downstream performance. Supervised contrastive learning performed even worse as it did not converge during pretraining. Again, due to the class imbalance, the supervised contrastive loss function results in a degenerate solution as approximately 97% of the batch is projected to a single embedding. Analogous behaviour was seen on the CAD prediction task, where 94% of the samples are in the negative class.

On the cardiac task, LaaF performs better than false negative elimination and supervised contrastive learning, but does not offer substantial gains over the imaging baseline. We attribute this to the fact that the cardiac setting has 120 included features, which lessens the importance of any one feature. Additionally, imbalanced binary classification is a difficult task for supervised contrative learning as explained above. Increasing the importance of the ground truth label in the pretraining process and adapting supervised contrastive learning to the binary case is left to future work.

On the DVM task, where we have 286 classes, the trend follows established literature. False negative elimination improves upon the baseline and supervised contrastive learning improves upon false negative elimination [15]. Our method by itself, without modifying the loss function, surpasses false negative elimination and approaches supervised contrastive learning. As expected, labels have a higher impact when more classes are present as shown in [15].

Importantly, adding the label as a tabular feature can also be combined with false negative elimination and supervised contrastive learning. This highlights the flexibility of our method as it can be used with any supervised contrastive strategy. With LaaF, we improve upon both losses and achieve our best scores on the DVM car model prediction task. The effect was similarly pronounced in the low data regime as shown in the supplementary materials.

Table 3. Frozen evaluation when incorporating labels into the contrastive pretraining process. Our Label as a Feature (Laaf) strategy consistently outperforms supervised contrastive learning (SupCon) and false negative elimination (FN Elimination), either alone or in combination. Best score is in **bold** font, second best underlined. Our methods are highlighted gray. A dash indicates failure to converge.

Contrastive	Label Used	Model	AUC (%) Infarction	AUC (%) CAD	Top-1 Accuracy (%) DVM
✓		Multimodal Imaging Baseline	76.35±0.19	**74.45±0.09**	91.43±0.13
	✓	Supervised ResNet50	72.37±1.80	68.84±2.54	87.97±2.20
✓	✓	Label as a Feature (LaaF)	**76.60±0.42**	73.76±0.31	93.56±0.08
✓	✓	FN Elimination	75.38±0.06	72.45±0.09	92.39±0.18
✓	✓	FN Elimination + LaaF	75.30±0.05	72.39±0.08	94.07±0.05
✓	✓	SupCon	—	—	93.82±0.11
✓	✓	SupCon + LaaF	—	—	**94.40±0.04**

5. Discussion and Conclusion

In this work we presented the first contrastive framework that combines tabular and imaging data. Our contribution is motivated by rich clinical datasets available in biobanks that contain vast amounts of information on participants' medical history, lifestyle, and physiological measures in combination with medical images. However, it is unfeasible to gather such detailed tabular data in a clinical setting due to time and budget constraints. Our solution pretrains on large datasets of tabular and imaging data to boost performance during inference using only images as input. We demonstrated the utility of our tool on the challenging task of cardiac health prediction from MR images, beating all contrastive baselines and the fully supervised baseline. Our method also translates to the natural image domain where we showed its strength on the task of car model prediction from advertisement data.

Through attribution and ablation experiments we showed that morphometric tabular features have outsized importance for the multimodal learning process. We hypothesize that these features, which are related to size and shape, have direct correlates in the image and thus help minimize the multimodal self-supervised loss. This suggests that extracting morphometric features from the images or collecting them from another source, to include them as tabular features, improves the learned representations. Finally, we presented a simple and effective new supervised contrastive learning method when using tabular data. Simply appending the target label as a tabular feature outperformed loss modifying strategies such as contrastive learning with false negative elimination and approached supervised contrastive learning. This strategy can also be combined with any supervised contrastive loss modification to achieve state-of-the-art results, surpassing all other strategies.

Limitations In our study we examined the benefit of our framework only for classification tasks. Future work should aim to test the behavior of the framework with further tasks such as segmentation and regression. We hypothesize that segmentation could benefit from the framework if morphometric features such as the sizes of the to-be-segmented regions are included in the tabular data and regression if morphometric features are regressed.

A further shortcoming of this work is that we only included white subjects from the UK Biobank population dataset because other ethnicities were drastically underrepresented in the study, making up only 5% of the total population. Significant racial disparities in coronary infarction and CAD risk have been repeatedly found [27, 38, 40] and could lead to spurious correlations being learned. Future work could use balanced datasets or explore propagated biases learned with unbalanced datasets, to identify and counteract any learned biases.

Conclusion In conclusion, for the first time, our work presents an effective and simple strategy to take advantage of tabular and imaging data in self-supervised contrastive learning. Our method is particularly relevant in a clinical setting where we wish to take advantage of extensive, multimodal biobanks during pretraining and predict unimodal in practice. We believe tabular data is an understudied and underappreciated source of data for deep learning, which is easy to collect and ubiquitous, as any numerical or categorical feature can be represented. It is also highly interpretable due to the fact that each feature directly represents a semantic concept. We hope that this inspires future works to unlock this untapped potential.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 87802.

References

[1] Ahmed M Alaa, Thomas Bolton, Emanuele Di Angelantonio, James HF Rudd, and Mihaela Van der Schaar. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 uk biobank participants. *PloS one*, 14(5):e0213653, 2019. 1

[2] Fabio Angeli, Paolo Verdecchia, Monica Trapasso, and Gianpaolo Reboldi. Left ventricular hypertrophy and coronary artery calcifications: a dangerous duet? *American Journal of Hypertension*, 31(3):287–289, 2018. 4

[3] Luigi Antelmi, Nicholas Ayache, Philippe Robert, Federica Ribaldi, Valentina Garibotto, Giovanni Frisoni, and Marco Lorenzi. Combining multi-task learning and multi-channel variational auto-encoders to exploit datasets with missing observations-application to multi-modal neuroimaging studies in dementia. 2021. 3

[4] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021. 3

[5] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3458–3468. IEEE. 3

[6] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. *arXiv preprint arXiv:2106.15147*, 2021. 1, 3

[7] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E. Petersen, Yike Guo, Paul M. Matthews, and Daniel Rueckert. Self-supervised learning for cardiac MR image segmentation by anatomical position prediction. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 541–549. Springer International Publishing. 2

[8] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *arXiv preprint arXiv:2110.01889*, 2021. 3

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 2

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2

[11] Christopher M Celano, Daniel J Daunis, Hermioni N Lokko, Kirsti A Campbell, and Jeff C Huffman. Anxiety disorders and cardiovascular disease. *Current psychiatry reports*, 18(11):1–11, 2016. 4

[12] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019. 2

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR. ISSN: 2640-3498. 1, 2, 3, 5, 7

[14] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2

[15] Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive learning. *arXiv preprint arXiv:2106.03719*, 2021. 3, 4, 5, 7

[16] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753. IEEE. 2

[17] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 5

[18] GPT-3 Demo. Wu dao 2.0 | GPT-3 demo. 3

[19] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430. IEEE. 2

[20] David C Dugdale, Ronald Epstein, and Steven Z Pantilat. Time and the patient–physician relationship. *Journal of general internal medicine*, 14(Suppl 1):S34, 1999. 1

[21] German National Cohort (GNC) Consortium geschaeftsstelle@ nationale-kohorte. de. The german national cohort: aims, study design and organization. *European journal of epidemiology*, 29(5):371–382, 2014. 1

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 5

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. (arXiv:2006.07733). 2

[24] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, volume 2, pages 1735–1742. IEEE. 2

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735. IEEE. 2

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE. 5

[27] Frederick K Ho, Stuart R Gray, Paul Welsh, Jason MR Gill, Naveed Sattar, Jill P Pell, and Carlos Celis-Morales. Ethnic differences in cardiovascular risk: examining differential exposure and susceptibility to risk factors. *BMC medicine*, 20(1):1–10, 2022. 8

[28] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022. 3

[29] Jingming Huang, Bowei Chen, Lan Luo, Shigang Yue, and Iadh Ounis. Dvm-car: A large-scale automotive dataset for visual marketing research and applications. *arXiv preprint arXiv:2109.00881*, 2021. 2, 4

[30] Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 986–996. IEEE. 3

[31] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[32] Benedikt Atli Jónsson, Gyda Bjornsdottir, TE Thorgeirsson, Lotta María Ellingsen, G Bragi Walters, DF Gudbjartsson, Hreinn Stefansson, Kari Stefansson, and MO Ulfarsson. Brain age prediction using deep learning uncovers associated sequence variants. *Nature communications*, 10(1):1–10, 2019. 1

[33] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc. 3, 4, 5

[34] Wonjun Ko, Wonsik Jung, Eunjin Jeon, and Heung-Il Suk. A deep generative–discriminative learning for multimodal representation in imaging genetics. *IEEE Transactions on Medical Imaging*, 41(9):2348–2359, 2022. 3

[35] Jeffrey B Lakier. Smoking and cardiovascular disease. *The American journal of medicine*, 93(1):S8–S12, 1992. 4

[36] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016. 2

[37] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*, 2020. 3

[38] Telly A Meadows, Deepak L Bhatt, Christopher P Cannon, Bernard J Gersh, Joachim Röther, Shinya Goto, Chiau-Suong Liau, Peter WF Wilson, Genevieve Salette, Sidney C Smith, et al. Ethnic differences in cardiovascular risks and mortality in atherothrombotic disease: insights from the reduction of atherothrombosis for continued health (reach) registry. In *Mayo Clinic Proceedings*, volume 86, pages 960–967. Elsevier, 2011. 8

[39] Akinori Mitani, Abigail Huang, Subhashini Venugopalan, Greg S Corrado, Lily Peng, Dale R Webster, Naama Hammel, Yun Liu, and Avinash V Varadarajan. Detection of anaemia from retinal fundus images via deep learning. *Nature Biomedical Engineering*, 4(1):18–27, 2020. 1

[40] Aditi Nayak, Albert J Hicks, and Alanna A Morris. Understanding the complexity of heart failure risk and treatment in black patients. *Circulation: Heart Failure*, 13(8):e007264, 2020. 8

[41] Richard W Nesto. Screening for asymptomatic coronary artery disease in diabetes. *Diabetes Care*, 22(9):1393, 1999. 4

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Physical activity and reduced risk of cardiovascular events. *arXiv preprint arXiv:1807.03748*, 116, 2018. 4

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[44] Kaiyue Pang, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10344–10352. IEEE. 2

[45] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544. IEEE. 2

[46] Mariann R Piano. Alcohol's effects on the cardiovascular system. *Alcohol research: current reviews*, 38(2):219, 2017. 4

[47] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Natasa Sladoje. CoMIR: Contrastive multimodal image representation for registration. In *Advances in Neural Information Processing Systems*, volume 33, pages 18433–18444. Curran Associates, Inc. 3

[48] Tiffany M Powell-Wiley, Paul Poirier, Lora E Burke, Jean-Pierre Després, Penny Gordon-Larsen, Carl J Lavie, Scott A Lear, Chiadi E Ndumele, Ian J Neeland, Prashanthan Sanders, et al. Obesity and cardiovascular disease: a scientific statement from the american heart association. *Circulation*, 143(21):e984–e1010, 2021. 4

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learn-

ing transferable visual models from natural language supervision. pages 8748–8763, 2021. 3

[50] Zahra Raisi-Estabragh, Nicholas C Harvey, Stefan Neubauer, and Steffen E Petersen. Cardiovascular magnetic resonance imaging in the uk biobank: a major international health research resource. *European Heart Journal-Cardiovascular Imaging*, 22(3):251–258, 2021. 4

[51] Tyler Hyungtaek Rim, Chan Joo Lee, Yih-Chung Tham, Ning Cheung, Marco Yu, Geunyoung Lee, Youngnam Kim, Daniel SW Ting, Crystal Chun Yuen Chong, Yoon Seong Choi, et al. Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. *The Lancet Digital Health*, 3(5):e306–e316, 2021. 1

[52] Salome Scholtens, Nynke Smidt, Morris A Swertz, Stephan JL Bakker, Aafje Dotinga, Judith M Vonk, Freerk Van Dijk, Sander KR van Zon, Cisca Wijmenga, Bruce HR Wolffenbuttel, et al. Cohort profile: Lifelines, a three-generation cohort study and biobank. *International journal of epidemiology*, 44(4):1172–1180, 2015. 1

[53] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022. 3

[54] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015. 1

[55] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017. 2, 3

[56] Martin G St John Sutton and Norman Sharpe. Left ventricular remodeling after myocardial infarction: pathophysiology and therapy. *Circulation*, 101(25):2981–2988, 2000. 4

[57] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20921, 2022. 3

[58] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. In Aasa Feragen, Stefan Sommer, Julia Schnabel, and Mads Nielsen, editors, *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, pages 661–673. Springer International Publishing. 2, 3

[59] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. In *Advances in Neural Information Processing Systems*, volume 33, pages 18158–18172. Curran Associates, Inc. 2

[60] Connie W Tsao, Philimon N Gona, Carol J Salton, Michael L Chuang, Daniel Levy, Warren J Manning, and Christopher J O'Donnell. Left ventricular structure and risk of cardiovascular events: a framingham heart study cardiac magnetic resonance study. *Journal of the American Heart Association*, 4(9):e002188, 2015. 4

[61] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:18853–18865, 2021. 3

[62] Paul Valensi, Luc Lorgis, and Yves Cottin. Prevalence, incidence, predictive factors and prognosis of silent myocardial infarction: a review of the literature. *Archives of cardiovascular diseases*, 104(3):178–188, 2011. 4

[63] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018. 2

[64] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[65] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 3

[66] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. VIME: Extending the success of self- and semi-supervised learning to tabular domain. In *Advances in Neural Information Processing Systems*, volume 33, pages 11033–11043. Curran Associates, Inc. 3

[67] Edward Yu, Vasanti S Malik, and Frank B Hu. Cardiovascular disease prevention by diet modification: Jacc health promotion series. *Journal of the American College of Cardiology*, 72(8):914–926, 2018. 4

[68] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[69] Miguel Zabalgoitia, Jens Berning, Michael J Koren, Asbjørn Støylen, Markku S Nieminen, Björn Dahlöf, Richard B Devereux, LIFE Study Investigators, et al. Impact of coronary artery disease on left ventricular systolic function and geometry in hypertensive patients with left ventricular hypertrophy (the life study). *The American journal of cardiology*, 88(6):646–650, 2001. 4

[70] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2, 5

[71] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[72] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer

Science, pages 420–428. Springer International Publishing. 2

[73] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459, 2021. 3