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

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

A. UK Biobank

A.1. Clinical Cardiac Features

The included clinical features for the cardiac prediction tasks are listed along with their UK Biobank field IDs in tables 3 and 4. The features labeled *extracted* were calculated using the pipeline outlined in [2–4, 10] with public code from [3].

A.2. ICD Codes

The ICD10 codes for cardiac infarction are I210, I211, I212, I213, I214, I219, and I252.

The ICD10 codes used for CAD labels are those within the categories I20-I25 - Ischemic heart diseases. The full list is I200, I201, I208, I209, I220, I221, I228, I229, I210, I211, I212, I213, I214, I219, I240, I248, I249, I250, I251, I252, I253, I254, I255, I256, I258, and I259.

B. Implementation Details

B.1. Pretraining

The augmentations used during pretraining on cardiac tasks were

- RandomHorizontalFlip (probability = 0.5)
- RandomRotation (degrees = 45)
- ColorJitter (brightness = 0.5, contrast = 0.5, saturation = 0.5)
- RandomResizedCrop (size = 128, scale = (0.2, 1.0))

The lower bound of the RandomResizedCrop was increased to 0.2 to ensure that a portion of the heart was in every view.

The augmentations used during pretraining on DVM car tasks were

- ColorJitter (brightness = 0.8, contrast = 0.8, saturation = 0.8, probability = 0.8)
- RandomGrayScale (probability = 0.2)
- GaussianBlur (kernel_size = 29, sigma = (0.1, 2.0, probability = 0.5))
- RandomResizedCrop (size = 128, scale = (0.08, 1))
- RandomHorizontalFlip (probability = 0.5)

The torchvision library was used for all augmentations. Each image was augmented during pretraining 95% of the time. The other 5% of the time the image was merely resized to 128x128. All validation and test images were simply resized to 128x128.

To effectively augment the tabular data, a fraction of a subject's features are randomly selected to be "corrupted" (i.e. augmented), following [1]. Each corrupted feature's value is sampled with replacement from all values for that feature seen in the dataset. This is also called sampling from the empirical marginal distribution. Categorical data was only one-hot encoded after corruption to ensure that each semantic feature had equal chance of being selected and categorical fields were not split up over multiple columns.

All contrastive pretraining models were trained for 500 epochs with a cosine annealing scheduler with warmup of 10 epochs. The learning rate and weight decay were chosen based on validation performance with possible values being $3e^{-3}$, $3e^{-4}$, and $3e^{-5}$ for the learning rate and either $1e^{-4}$ or $1.5e^{-6}$ for the weight decay. The Adam optimizer [8] was used with a batch size of 512.

Multimodal CLIP Loss [11] The standard CLIP loss was used as outlined in the methods section of this paper. In our experiments, a temperature of 0.1 worked best, which follows [5]. The corruption rate of the tabular augmentation was set to 0.3 after sweeping in increments of 0.1. A learning rate of $3e^{-3}$ was used for the cardiac pretraining and $3e^{-4}$ for the DVM pretraining. A weight decay of $1e^{-4}$ was used for the cardiac pretraining and $1.5e^{-6}$ for the DVM pretraining.

SimCLR [5] We used the NTXent loss which is based on the InfoNCE [9] loss and compares the embedding of a view against all other views in a batch. The temperature was kept at 0.1 as outlined in the original paper. A learning rate of $3e^{-3}$ was used for the cardiac pretraining and $3e^{-4}$ for the DVM pretraining. A weight decay of $1e^{-4}$ was used for the cardiac pretraining and $1.5e^{-6}$ for the DVM pretraining.

Bootstrap Your Own Latent (BYOL) [7] We used the pytorch lightning bolts implementation of BYOL. BYOL uses an online network and a target network. The online network has a prediction head ontop of a projection head. The target networks weights are an exponential moving average of the online network's weights. The loss is the cosine similarity between the prediction of the online network and the projection of the target network. The output of the target network has a stop gradient applied so no weight updates are made outside of the exponential moving average, which is tempered by τ_{base}. The projector hidden dimension was 4096 and projector out dimension was 256. The predictor hidden dimension was also 4096 and predictor out dimension was also 256. τ_{base} was set to 0.9995 as we used a smaller batch size (512 vs 4096) as recommended in the original paper. A learning rate of $3e^{-4}$ was used for the cardiac pretraining and $3e^{-4}$ for the DVM pretraining. A weight decay of $1.5e^{-6}$ was used for all pretrainings.

Simple Siamese Network (SimSiam) [6] SimSiam is similar to BYOL in that it also has a predictor on top of a projector, but it only uses a single encoder. The prediction of the one view is compared to the projection of the other view using the cosine similarity loss with a stop gradient again being used on the side of the projection. The projector hidden dimension was 2048 and projector out dimension was 2048. The predictor hidden dimension was 512 and predictor out dimension was 2048, creating a bottleneck as recommended in the original paper. A learning rate of $3e^{-4}$ was used for cardiac pretraining and $3e^{-5}$ for DVM pretraining. A weight decay of $1.5e^{-6}$ was used for all pretrainings.

Barlow Twins [12] Barlow Twins calculates the cross correlation matrix between the embeddings of the two views and pushes it towards the identity matrix, effectively maximizing similarity between views from the same subject and minimizing similarity between all other views. The advantage of barlow twins is that it does not require a predictor head, large batches, gradient stopping or a moving average of the weights, unlike previous methods. We use projector hidden dimensions and projector out dimensions of 8192, as recommended in the original paper. A learning rate of $3e^{-3}$ and a weight decay of $1e^{-4}$ was used for all pretrainings.

B.2. Finetuning

A learning rate sweep covering 6 learning rates, ($3e^{-2}$, $1e^{-2}$, $3e^{-3}$, $1e^{-3}$, $3e^{-4}$, $1e^{-4}$) was undertaken for every model during finetuning. The supervised models were swept in their entirety, and the contrastive models were swept during both finetuning settings, frozen and trainable. The optimal learning rate based on validation metric performance was selected. Early stopping based on the validation metric was used with a minimum delta of 0.0002 and a patience of 10 epochs. The Adam optimizer without weight decay and a batch size of 512 were used.

C. Low Data Contrastive Pretraining

As seen in figure 1, our multimodal pretraining strategy excels at all data quantities and outperforms the imaging only contrastive baselines by even larger margins. This underscores the strength of the learned representations that need minimal examples to perform on downstream classification tasks. This makes our framework particularly well suited to rare disease classification where few positive samples are available.

D. Tabular Results

As shown in table 1, we found that the tabular encoder trained with our multimodal framework remains competitive with SCARF. On the cardiac tasks, when freezing the encoder, the multimodal pretrained model rivaled SCARF. SCARF proved slightly stronger when allowing the entire network to be trainable. On the DVM car model prediction task our multimodal framework improved upon SCARF in the frozen setting and was slightly stronger in the trainable setting.

E. Low Data Training with Label as a Feature

As shown in table 2, our label as a feature (LaaF) strategy for supervised contrastive learning is particularly effective in the low data regime. LaaF consistently outperforms both supervised contrastive learning and false negative elimination, either alone or in combination with the aforementioned loss modifications. In the very low data regime (1% or 700 samples), LaaF by itself surpasses both SupCon and FN Elimination.

F. Explainability

F.1. Integrated Gradients

Figure 3 shows the integrated gradient attribution scores for tabular embeddings with respect to all cardiac features. Here the importance of the morphometric features, shown in orange, is underscored through their increased frequency towards the most important features.

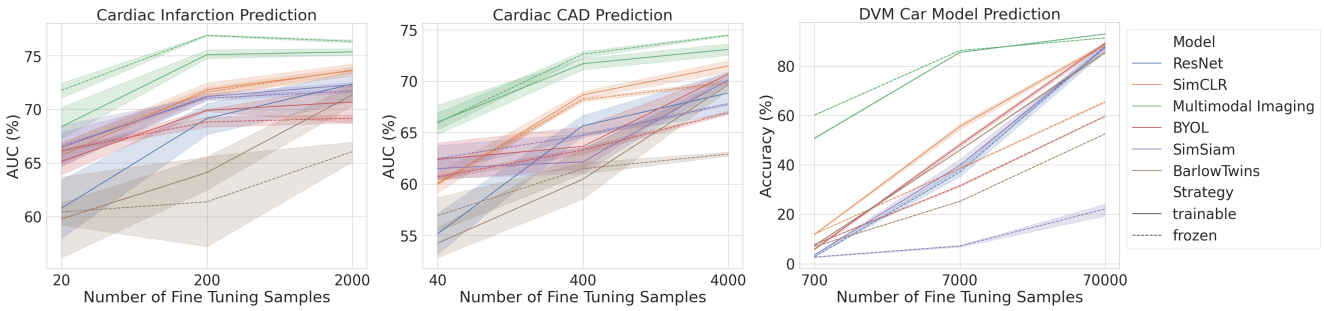

Figure 1. Performance of the imaging models with different number of finetuning training samples. Shaded regions indicate 95% confidence intervals. Pretraining with both images and tabular data excels at all data quantities and is well suited for rare disease identification when only tens or hundreds of labels are available.

Table 1. Performance of our framework using tabular input on the tasks of cardiac infarction, coronary artery disease (CAD), and DVM car model prediction. Our model performs similarly to SCARF on the cardiac tasks and is stronger on the DVM task. The best performing model for every input type is displayed in **bold** font and the second best is underlined. 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 MLP	83.35±0.29	83.35±0.29	79.61±7.19	79.61±7.19	91.99±0.10	91.99±0.10
SCARF	85.16±0.60	**86.01±0.39**	**83.21±0.42**	**84.23±0.25**	88.29±0.40	92.64±0.29
Multimodal Tabular	**85.76±0.27**	85.20±0.14	83.15±0.20	83.44±0.67	**92.88±0.30**	**93.08±0.16**

Table 2. Frozen eval results when incorporating labels into the contrastive pretraining process at 100%, 10%, and 1% training dataset sizes on the DVM task. Our label-as-a-feature 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.

Model	Top-1 Acc. (%) DVM (100%)	Top-1 Acc. (%) DVM (10%)	Top-1 Acc. (%) DVM (1%)
Multimodal Baseline	91.43±0.13	86.30±0.08	60.18±0.21
Supervised ResNet50	87.97±2.20	30.69±14.02	2.84±0.00
Label-as-a-Feature (LaaF)	93.56±0.08	89.87±0.03	**67.50±0.10**
FN Elim.	92.39±0.18	87.61±0.07	63.95±0.14
FN Elim. + LaaF	94.07±0.05	89.99±0.05	63.37±0.70
SupCon	93.82±0.11	89.75±0.08	63.29±0.33
SupCon + LaaF	**94.40±0.04**	**90.37±0.05**	64.01±0.77

F.2. Morphometric Features Impact on Pretraining

As shown in figure 2, despite comprising less than a fourth of all features, pretraining with only morphometric features converges to almost the exact same loss as training with all 117 features. Training without any morphometric features converged to a final loss almost twice as high. This emphasizes the importance of morphometric features in the minimization of the CLIP loss, as they are easily extracted from images and facilitate the learning of useful features.

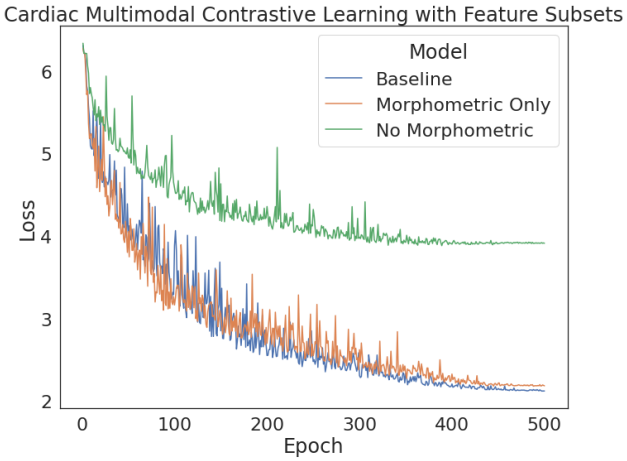

Figure 2. Contrastive loss during multimodal pretraining. Training with only morphometric features converged to a similar loss of the baseline which included all 120 features. Training with no morphometric features had markedly less similarity between the projected embeddings of the same subject, showing the importance of the morphometric features for the multimodal training process.

References

[1] 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

[2] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):1–12, 2018. 1, 5

[3] Wenjia Bai, Hideaki Suzuki, Jian Huang, Catherine Francis, Shuo Wang, Giacomo Tarroni, Florian Guitton, Nay Aung, Kenneth Fung, Steffen E Petersen, et al. A population-based phenome-wide association study of cardiac and aortic structure and function. *Nature medicine*, 26(10):1654–1662, 2020. 1, 5

[4] Wenjia Bai, Hideaki Suzuki, Chen Qin, Giacomo Tarroni, Ozan Oktay, Paul M Matthews, and Daniel Rueckert. Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In *International conference on medical image computing and computer-assisted intervention*, pages 586–594. Springer, 2018. 1, 5

[5] 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

[6] 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

[7] 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

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[9] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[10] Steffen E Petersen, Nay Aung, Mihir M Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Jane M Francis, Mohammed Y Khanji, Elena Lukaschuk, Aaron M Lee, et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (cmr) in caucasians from the uk biobank population cohort. *Journal of Cardiovascular Magnetic Resonance*, 19(1):1–19, 2017. 1, 5

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pages 8748–8763, 2021. 1

[12] 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

Table 3. Included cardiac features and their UK Biobank Field ID. Features labeled *extracted* were calculated using the pipeline outlined in [2–4, 10].

Tabular Feature	UK Biobank Field ID
Alcohol drinker status	20117
Alcohol intake frequency.	1558
Alcohol usually taken with meals	1618
Amount of alcohol drunk on a typical drinking day	20403
Pulse rate	95, 102
Angina diagnosed by doctor	3627, 6150
Augmentation index for PWA	12681
Average heart rate	22426
Basal metabolic rate	23105
Beef intake	1369
Blood pressure medication regularly taken	6153, 6177
Body fat percentage	23099
Body mass index (BMI)	23104, 21001
Body surface area	22427
Cardiac index	22425
Cardiac index during PWA	12702
Cardiac operations performed	20004
Cardiac output	22424
Cardiac output during PWA	12682
Central augmentation pressure during PWA	12680
Central pulse pressure during PWA	12678
Central systolic blood pressure during PWA	12677
Cholesterol lowering medication regularly taken	6177
Cooked vegetable intake	1289
Current tobacco smoking	1239
Diabetes diagnosis	2443, 120007, 23104, 21001
Diastolic blood pressure	4079, 94
Diastolic brachial blood pressure during PWA	12675
Duration of heavy DIY	2634
Duration of light DIY	1021
Duration of moderate activity	894
Duration of other exercises	3647
Duration of strenuous sports	1001
Duration of vigorous activity	914
Duration of walks	874
Duration walking for pleasure	981
End systolic pressure during PWA	12683
End systolic pressure index during PWA	12684
Ever had diabetes (Type I or Type II)	120007
Ever smoked	20160
Exposure to tobacco smoke at home	1269
Exposure to tobacco smoke outside home	1279
Falls in the last year	2296
Frequency of consuming six or more units of alcohol	20416
Frequency of drinking alcohol	20414
Frequency of heavy DIY in last 4 weeks	2624
Frequency of other exercises in last 4 weeks	3637
Frequency of stair climbing in last 4 weeks	943
Frequency of strenuous sports in last 4 weeks	991
Frequency of walking for pleasure in last 4 weeks	971
Heart rate during PWA	12673
Height	12144
High blood pressure diagnosed by doctor	2966, 6150
Hip circumference	49

Table 4. Included cardiac features and their UK Biobank Field ID. Features labeled *extracted* were calculated using the pipeline outlined in [2–4, 10].

Tabular Feature	UK Biobank Field ID
Hormone replacement therapy medication regularly taken	6153
Impedance of whole body	23106
Insulin medication regularly taken	6153, 6177
Lamb/mutton intake	1379
LVCO (L/min)	*extracted*
LVEDV (mL)	*extracted*
LVEF (%)	*extracted*
LVESV (mL)	*extracted*
LVM (g)	*extracted*
LVSV (mL)	*extracted*
Mean arterial pressure during PWA	12687
Number of beats in waveform average for PWA	12679
Number of days/week of moderate physical activity 10+ minutes	884
Number of days/week of vigorous physical activity 10+ minutes	904
Number of days/week walked 10+ minutes	864
Oral contraceptive pill or minipill medication regularly taken	6153
Overall health rating	2178
P duration	12338
Pace	3079
Pack years adult smoking as proportion of life span exposed to smoking	20162
Pack years of smoking	20161
Past tobacco smoking	1249
Peripheral pulse pressure during PWA	12676
Pork intake	1389
PP interval	22334
PQ interval	22330
Processed meat intake	1349
Pulse wave Arterial Stiffness index	21021
QRS duration	12340
RR interval	22333
RVEDV (mL)	*extracted*
RVEF (%)	*extracted*
RVESV (mL)	*extracted*
RVSV (mL)	*extracted*
Salad / raw vegetable intake	1299
Sex	31
Shortness of breath walking on level ground	4717
Sitting height	20015
Sleep duration	1160
Sleeplessness / insomnia	1200
Smoking status	20116
Smoking/smokers in household	1259
Standing height	50
Stroke diagnosed by doctor	6150
Stroke volume during PWA	12686
Systolic blood pressure	4080, 93
Systolic brachial blood pressure during PWA	12674
Tense / highly strung	1990
Time spent driving	1090
Time spent using computer	1080
Time spent watching television (TV)	1070
Total mass	23283
Total peripheral resistance during PWA	12685
Usual walking pace	924
Ventricular rate	12336
Waist circumference	48
Weight	23098, 21002
Weight change compared with 1 year ago	2306
Whole body fat-free mass	23101
Whole body fat mass	23100
Whole body water mass	23102
Worrier / anxious feelings	1980

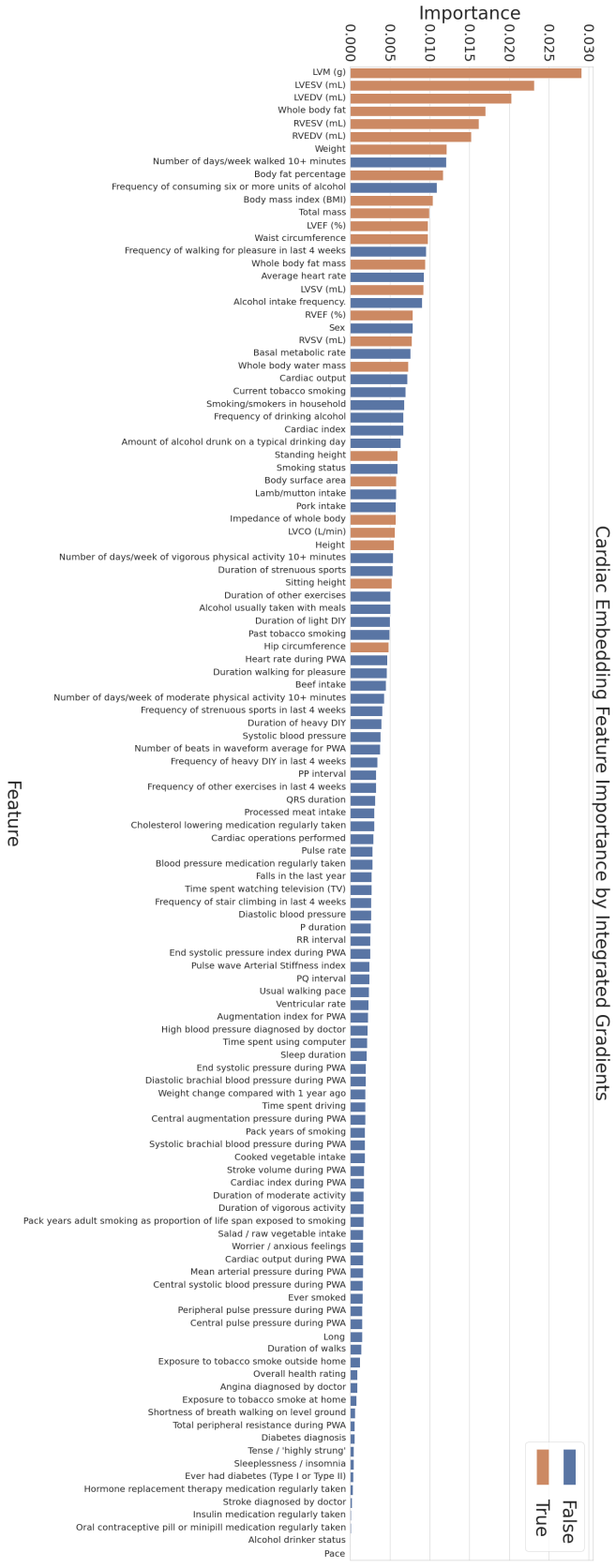

Figure 3. The integrated gradient attribution scores for tabular embeddings with respect to cardiac features. Orange color indicates morphometric feature.