



Vision’s Potential Unlocked: How Pretraining and Strategic Fine-tuning Improve Stroke Relapse Prediction

Christian Gapp¹ 

CHRISTIAN.GAPP@UMIT-TIROL.AT

Elias Tappeiner¹ 

ELIAS.TAPPEINER@UMIT-TIROL.AT

Martin Welk¹ 

MARTIN.WELK@UMIT-TIROL.AT

¹ *Institute of Biomedical Image Analysis, UMIT TIROL – Private University for Health Sciences and Health Technology, Eduard-Wallnöfer-Zentrum 1, 6060 Hall in Tirol, Austria*

Karl Fritscher² 

KARL.FRITSCHER@VASCAGE.AT

² *VASCage – Centre on Clinical Stroke Research, 6020 Innsbruck, Austria*

Stephanie Mangesius^{3,4} 


STEPHANIE.MANGESIUS@I-MED.AC.AT

Constantin Eisenschink^{3,4} 

CONSTANTIN.EISENSCHINK@I-MED.AC.AT

Astrid E. Grams^{3,4} 

ASTRID.GRAMS@I-MED.AC.AT

Elke R. Gizewski^{3,4} 

ELKE.GIZEWSKI@I-MED.AC.AT

³ *Department of Radiology, Medical University of Innsbruck*

⁴ *Neuroimaging Research Core Facility, Medical University of Innsbruck*

Philipp Deisl^{2,3,4} 

PHILIPP.DEISL@I-MED.AC.AT

Michael Knoflach^{2,5} 

MICHAEL.KNOFLACH@I-MED.AC.AT

⁵ *Department of Neurology, Medical University of Innsbruck*

Rainer Schubert¹ 

RAINER.SCHUBERT@UMIT-TIROL.AT

Editors: Under Review for MIDL 2026

Abstract

Unbalanced modality usage, especially modality collapse, remains a major limitation in multimodal learning, often preventing models from exploiting the full potential of multimodal datasets. Pretraining multimodal neural networks has been shown to enhance overall performance. Nevertheless, its effect on modality contribution remains largely unexplored. Moreover, freezing parts of the network during fine-tuning is crucial to mitigate catastrophic forgetting. Yet the impact of freezing strategies on modality contribution has also received little attention.

In this work, we explore how self-supervised image pretraining can mitigate modality contribution imbalance and enhance cross-modal integration for stroke relapse detection—a clinically critical task we recently addressed. To this end, two multimodal neural networks were pretrained in a self-supervised manner and subsequently fine-tuned under two distinct freezing strategies. Their performance was compared against both the baseline model from our previous work and models trained entirely from scratch in this work.

Our results demonstrate that pretraining enables a more comprehensive exploitation of the multimodal image-tabular dataset, outperforming both the prior baseline and all non-pretrained models. Furthermore, pretraining notably increased the vision’s modality contribution, while freezing strategies were found to significantly affect modality utilization as well. The overall best-performing model, based on a Vision Transformer, successfully overcame unimodal collapse through self-supervised pretraining.

These findings indicate that pretraining combined with strategic fine-tuning allows full use of multimodal medical datasets, supporting more balanced and effective models for tasks such as stroke relapse detection. The code for the pretraining step is publicly available at https://github.com/ChristianGappGit/SSL_Pretraining.

Keywords: Self-Supervised Image Pretraining, Modality Contribution, Multimodal Fusion, Stroke Relapse Prediction

1. Introduction

In medicine, patient data recorded for studies or just in clinical routine processes, is likely to be multimodal. Image data, such as X-rays, MRIs or CTs, tabular patient related data, such as demographic, histopathological data and text data representing clinical reports are quite common. However, applying multimodal, deep learning based neural networks for medical tasks like disease diagnosis remains a difficulty. Despite several multimodal fusion strategies (Xu et al., 2023), unbalanced modality contributions are still an issue. It is common for multimodal models to suffer from modality collapse, in which only a subset of modalities is effectively utilized (Javaloy et al., 2022). In this cases the unimodal trained networks are likely to outperform multimodal networks, as found by Wang et al. (2020). Reasons for collapses might be conflicting gradients between the modalities (Javaloy et al., 2022) or the dependence of the fusion strategy on the dataset, as first found by Ma et al. (2022).

Multimodal architectures built upon Vision Transformer (Dosovitskiy et al., 2021) backbones have been observed to exhibit unimodal collapse, particularly in vision-language tasks (Parcalabescu and Frank, 2023). Pretraining these ViT based models has a big impact on their performance in the downstream task, as results from Wang et al. (2022) demonstrate. Despite requiring extensive pretraining, ViTs have demonstrated superior generalization capabilities compared to ResNets (Chen et al., 2022). Pretrained models such as CLIP (Radford et al., 2021) and BiomedCLIP (Zhang et al., 2025)—trained on 15 million biomedical image-text pairs—achieved great results on several tasks. Still, multimodal pretraining is not straightforward. For non-image-text pair based medical datasets, these pretrained models can hardly be used. For image-tabular datasets, the modalities may be pretrained separately. (Ye et al., 2025) uses a pretraining method for tabular data that can be applied to several tasks even without being fine-tuned. For vision, self-supervised image pretraining (Tang et al., 2022) can aid the models in learning to extract image features. The subsequent fine-tuning stage can then be solved with knowledge learned in the pretraining step.

This study aims at analyzing the impact of self-supervised image pretraining on both the performance and modality contribution of different multimodal neural networks. For this, we take the image-tabular dataset from our previous published work (Gapp et al., 2025b), where a ResNet (He et al., 2016) based multimodal model for stroke relapse detection was trained. Therein we used 3D CTAs + tabular data, consisting of age, gender, CHD (coronary heart disease) and PAD (peripheral artery disease), and follow-up data on recurred stroke from 119 stroke patients. Importance-related analysis results indicate a fairly balanced contribution from both tabular and visual modalities, with the region of the arteria carotis communis appearing to be a relevant factor for stroke relapses. Here, with pretraining, we are fully utilizing the dataset containing 491 patients. Furthermore, besides retraining the ResNet based network from Gapp et al. (2025b), we pretrain and fine-tune a ViT (Dosovitskiy et al., 2021) based multimodal architecture. Performance AUC results are compared to results from Gapp et al. (2025b), as well as models trained from scratch here. In addition, with the modality contribution method from Gapp et al. (2025a) we quantify the usage of the modalities in the multimodal models. Overcoming imbalanced contributions is necessary to fully utilize the stroke dataset’s potential.

The goal of this work is to improve the performance for stroke relapse detection from [Gapp et al. \(2025b\)](#), thereby measuring the impact of pretraining and strategic fine-tuning. Specifically, we aim to quantify how both single-modality pretraining and the applied freezing strategy affect a modality’s contribution to a multimodal task—a topic that, to the best of our knowledge, has not yet been explored.

2. Multimodal Stroke Data

Data acquisition methods and study cohort composition for pretraining and fine-tuning are detailed below, including collection procedures and patient demographics.

2.1. Data Generation

As part of the project “Retrospective Pilot Project: Imaging Biomarkers for Vascular Diseases and Vascular Aging”, clinical and imaging data were collected from April 2023 onwards. The cohort included patients with at least one ischemic cerebral event (ICE) who had been admitted to the Stroke Unit of the Department of Neurology, Medical University of Innsbruck, since 2010. Anonymized imaging data was recorded on Siemens’ syngo.share platform (Version VA32C), while clinical data was collected in a custom database established by an external company. The anonymized clinical data included patient age, gender, and the occurrence of cardiac or peripheral events. The combined image-tabular dataset was labeled according to the occurrence of recurrent ICE, distinguishing between relapse and non-relapse cases. All CT angiography data from routine diagnostics were fully anonymized prior to analysis. The study was approved by the local institutional review board (IRB) of the Medical University of Innsbruck (EK-Nr: 1429/2021).

2.2. Study Population

After some data cleaning processes we could finally record fully usable vision and clinical tabular data from 491 patients. For the self-supervised pretraining task, 393 of them were used for training, 98 for validation, yielding an approximate 80:20 ratio.

For the fine-tuning task, which involves predicting RFS time and classifying patients into relapse and non-relapse groups, the dataset comprises 119 patients. Patients were selected based on their label (relapse vs. relapse-free) and RFS time, as they met the necessary criteria for this task (see [Gapp et al., 2025b](#)). Specifically, relapse patients with an RFS below 1,642 days and non-relapse patients with an RFS above 1,825 days were included, with an additional cut-off at 2,555 days applied for non-relapse cases.

The training dataset consists of 95 (including 32 relapses), the testing dataset of 24 (9 relapses) image-tabular pairs. Thereby it is ensured that the 24 patients used for evaluation were entirely part of the pretraining validation dataset, thus completely excluded from all training steps. A summary of the demographic and clinical characteristics of the study population for the fine-tuning task is provided in [Table 1](#).

characteristic	attribute	n	%
gender	men	79	66.4
	women	40	33.6
heart disease	CHD only	24	20.2
	PAD only	20	16.8
	CHD + PAD	6	5.0
	none	69	58.0
relapse status	no relapse	78	65.6
	relapse	41	34.4
		mean \pm SD	
age (years)		69.10 \pm 10.18	

Table 1: Baseline characteristics of the study population (N = 119).

2.3. Data Preprocessing

In the whole study, for both pretraining and fine-tuning tasks, we used 3D CTAs and tabular data recorded at the time of the first, initial stroke event. The labels (RFS, occurrence of relapse) were finalized at the end of the follow up.

Vision All images are registered to a fixed, representative image using affine transformations to ensure consistent alignment and comparable properties for downstream deep learning analyses. The images have a size of $224 \times 224 \times 320$ voxels with an isotropic spacing of $1 \times 1 \times 1$ mm.

Tabular The clinical tabular data includes information about the patients’ age, gender, and the heart diseases CHD and PAD. Heart diseases are encoded using a single bit each, while gender is represented with two bits. Age values are z-normalized across the entire dataset.

3. Training Configurations

Training is conducted in two stages: first, pretraining of the visual networks, and second, fine-tuning of the multimodal neural networks. We pretrain two visual nets—ResNetAutoEnc and VisionAutoEnc—and later fine-tune multimodal neural networks utilizing either a ResNet (He et al., 2016) or ViT (Dosovitskiy et al., 2021) backbone model. Given the inherently low dimensionality of the tabular data (four attributes), pretraining is unnecessary. Instead, a compact MLP with a single hidden layer is trained from scratch. The core focus of this work lies in advancing the visual modality.

3.1. Pretraining: Self-Supervised Learning (SSL)

Motivated from Tang et al. (2022), we apply a self-supervised learning approach as a pre-training step. To this end, the 3D CTAs are processed by an autoencoder (comprising

an encoder and decoder) that learns to extract features from the visual input and to reconstruct the same input image, optimized with an L1 loss. The encoder part is used to function as backbone model for the fine-tuning task later. Details to the pretraining architectures are presented below. Additional information on the pretraining step can be found at https://github.com/ChristianGappGit/SSL_Pretraining.

ResNetAutoEnc The encoder is a ResNet34 (see also vision backbone model in Figure 1). In the decoder, four convolutional layers without skip connections are applied to the encoded features to reconstruct the image.

ViTAutoEnc A Vision Transformer architecture with eight layers and eight attention heads is representing the encoder part (see also backbone model in Figure 2). The decoder, built with two convolutional layers and no skip connections, reconstructs the image.

Computation Time Pretraining for both models required approximately four days for 300 epochs on a NVIDIA L40S 48GB GPU, without computational optimizations.

3.2. Fine-tuning Task: Stroke Relapse Prediction

As fine-tuning task the stroke relapse prediction is done the same way as in (Gapp et al., 2025b). Therein, RFS prediction is performed as a regression task, followed by classification into relapse versus non-relapse groups. Classification is based on the predicted RFS time, using a threshold κ in days: predictions below this threshold are considered relapses, while predictions above it are considered non-relapses. Thresholds within the range $\kappa_{\text{low}} \leq \kappa \leq \kappa_{\text{high}}$ were explored in (Gapp et al., 2025b). However, in this work we keep κ fixed at $\kappa_{\text{low}} = 1,642$. With this setting, we retrain the XSRD-net architecture used in that study, a ResNetMLP. In addition, we fine-tune a ViTMLP employing a ViT (eight heads, eight layers) as vision backbone model. Architectural details are depicted in Figures 1 and 2.

Model Complexity The ResNet and ViT backbones have around 63.5M and 59.8M parameters, respectively. As the vision head for ViT has around 5.75M parameters, compared to only 2.56k for the the ResNet34, the vision models are comparable in size. Including tabular MLP (225), and fusion MLP (850), the total model sizes amount to roughly 63.60M for ResNetMLP and 65.55M for ViTMLP. Even though the MLPs for the tabular and fusion components are comparatively small, they are still sufficient and play an essential role in the multimodal task. The tabular MLP processes five inputs: one continuous feature for age and four binary features for gender (2 bits) and heart disease (2 bits for CHD and PAD). The fusion MLP combines two 5×1 outputs from the individual models to a 10×1 representation and learns joint features prior to the final classification head.

Computation Time Fine-tuning was performed over 250 epochs, requiring approximately 1 hour and 50 minutes per model on the same hardware.

3.3. Experiments

To assess the effect of pretraining on the fine-tuning task, we fine-tune pretrained multimodal models for each backbone, ResNet and ViT, and compare them to models trained from scratch. For this the AUC is computed as performance metric. The modality contribution is measured with the method from Gapp et al. (2025a). This occlusion sensitivity

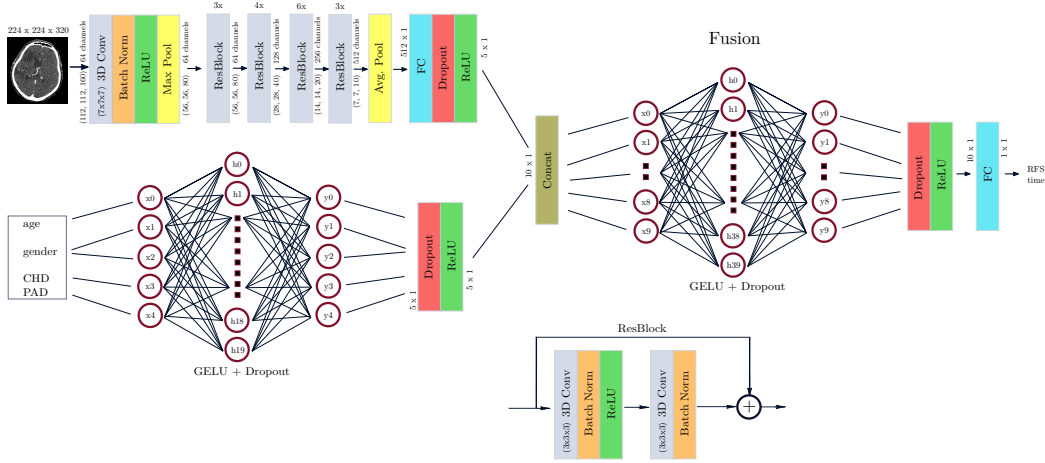


Figure 1: ResNetMLP: Vision model: ResNet34, Tabular model: MLP with one hidden layer. Fusion model: MLP with one hidden layer.

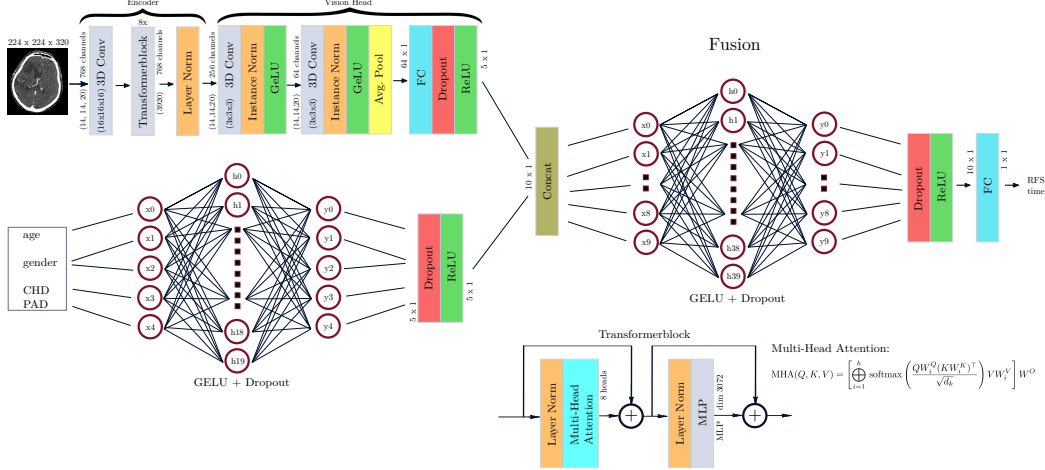


Figure 2: ViTMLP: Vision model: ViT with eight heads and eight layers, Tabular model: MLP with one hidden layer, Fusion model: MLP with one hidden layer.

based method includes a hyper-parameter h_{modality} , which defines how many sequences are occluded at a single model forward pass. We use $h_{\text{vision}} = 1$, occluding the entire image at once, and $h_{\text{tabular}} = 4$, where each attribute—age, gender (two bits occluded at once), CHD, and PAD—is occluded separately.

For each of the two pretrained multimodal models we fine-tune one multimodal model with a fully frozen backbone encoder, and another one following the freezing strategy illustrated in Figure 3, which successively unfreezes later backbone layers after 50, 100 and 150 epochs. A large part of the (pretrained) encoder weights are thus frozen throughout the entire 250 epochs. Weights from the last layer are trained 200, weights from the second last layer 150 and weights from the third last layer 100 epochs. Both strategies of either fully freezing or partly freezing encoder weights, rather than to just fine-tune the whole

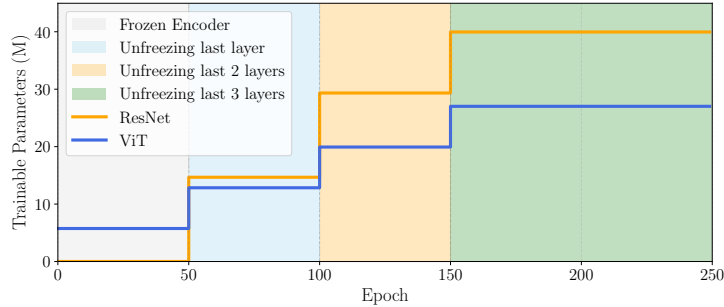


Figure 3: Visualization of parameter schedule during progressive layer unfreezing over 250 fine-tuning epochs for ResNet and ViT backbones. The training is divided into four stages, indicated by shaded background regions. Step curves show the cumulative number of trainable parameters (in millions) over time. ViT starts with more trainable parameters due to its larger vision head (see stage 1), whereas ResNet releases more parameters in stages 2-4, as larger layers are affected there.

pretrained model without freezing, are used to mitigate the risk of catastrophic forgetting, a well-known challenge in sequential learning (French, 1999; Goodfellow et al., 2015). By applying freezing approaches, previously learned features are preserved, allowing the model to effectively adapt to new data or, in our case, new tasks. All in all, thus six multimodal neural networks (for both ViT and ResNet backbones: two pretrained + one from scratch) are trained.

4. Results

This chapter presents the experimental results, organized into two sections: pretraining and fine-tuning.

4.1. Pretraining

In the Figure 4, we report the pretraining L1-losses of the two autoencoder backbones: ResNetAutoEnc and ViTAutoEnc.

4.2. Fine-tuning Task

Tables 2 and 3 summarize the effects of pretraining on multimodal disease classification performance and modality contributions. Table 2 compares the overall AUC and the relative impact of the vision modality (mc_v) between ResNetMLP and ViTMLP architectures under different training setups. Table 3 provides a more detailed breakdown of individual modality contributions, highlighting how vision and tabular features (age, gender, CHD, PAD) are utilized by the model when using pretrained weights compared to training from scratch.

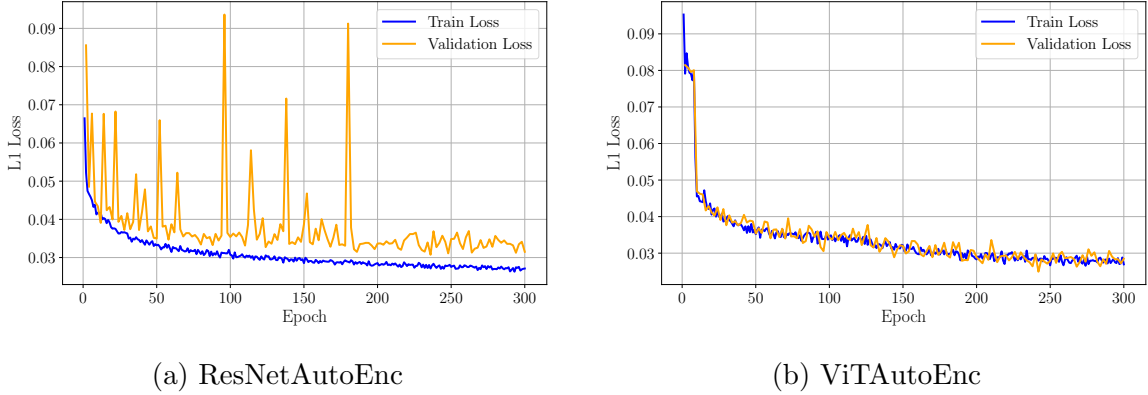


Figure 4: Pretraining L1 loss curves over epochs for training and validation across both architectures.

XSRD-Net	Training Setup	train AUC	test AUC	mc_v
ResNetMLP	from scratch	0.570	0.549	79.46%
	pretrained frozen	0.409	0.653	93.31%
	pretrained temp. frozen	0.695	0.715	92.20%
ViTMLP	from scratch	0.617	0.722	0.13%
	pretrained frozen	0.805	0.743	28.39%
	pretrained temp. frozen	0.626	0.729	5.53%

Table 2: Pretraining effects on performance for ResNetMLP and ViTMLP. The last column shows the modality contribution of vision mc_v . The whole analysis is done with the test dataset. Highlighted in green: overall best model.

XSRD-Net	Training Setup	vision		tabular		
		3D CTA	age	gender	CHD	PAD
ResNetMLP	from scratch	79.46%	8.09%	6.96%	0.68%	4.81%
	pretrained frozen	93.31%	2.21%	3.13%	0.31%	1.04%
	pretrained temp. frozen	92.20%	2.57%	3.91%	0.40%	0.92%
ViTMLP	from scratch	0.13%	12.95%	75.68%	4.59%	6.65%
	pretrained frozen	28.39%	8.73%	55.60%	3.03%	4.25%
	pretrained temp. frozen	5.53%	11.97%	71.84%	4.37%	6.29%

Table 3: Detailed modality contributions for RFS predictions using the test dataset. Green highlighted: modality contributions for overall best model from Table 2. Marked in red: unimodal collapse of ViTMLP trained from scratch.

5. Discussion

Self-supervised visual pretraining enables the usage of all summarized patient data within a study cohort, regardless of any disease specific labels or exclusion-criteria applied in the fine-tuning task. Despite fully utilizing the cohort, computation time aspects are promising. Although pretraining the model takes a few days, the fine-tuning step is realized in less than two hours. Moreover, the fine-tuned models can be efficiently retrained with additional data collected in the future. Results from the pretraining step (Figure 4) show that the ViTAutoEnc can extract visual features and reconstruct images more effectively than the ResAutoEnc. The ViTAutoEnc exhibits similar L1 losses on both the training and validation datasets, whereas the ResAutoEnc shows a tendency to slightly overfit on the training set. These results are also reflected in the fine-tuning stage.

First of all, with both architecture specific best neural networks we could increase the AUC performance on the stroke relapse detection task from Gapp et al. (2025b) (AUC: 0.71). The ResNetMLP with pretrained ResNet weights following the freezing strategy in Figure 3 during the fine-tuning reached an AUC of 0.72 on the test dataset. The pretrained ViTMLP with fully frozen encoder got the overall best performance with an AUC of 0.74 (Table 2). Furthermore, the two best models outperform the models trained from scratch. For the ResNetMLP both the model trained from scratch and the pretrained model with fully frozen encoder could not solve the fine-tuning task sufficiently. The ViTMLP model trained from scratch underwent a unimodal collapse, relying only on the tabular data while ignoring the vision input. Although the AUC performance on the test dataset is quite good, the model has not learned enough (train AUC only 0.62) to solve the task. The separation into the classes, relapse and non-relapse, is more or less gender and age specific (see also Table 3), which is interesting, but not medically relevant here.

The poor performance of the pretrained ResNetMLP with frozen encoder (AUC 0.65 on the test set and only 0.41 on the train set) may be attributed, on the one hand, to limited pretraining, and on the other, to the low number of training parameters in the subsequent vision head (consisting only of the classification layer). That is exactly why the progressive layer unfreezing strategy works here. The big layers are successively unfrozen and thus enable the network to learn important patterns in later epochs. By contrast, the ViTMLP’s results are quite the opposite. Since the pretraining performance was convincing, the pre-trained model with fully frozen encoder, and additional, learnable weights in the vision head (convolutional layers, classifier) was the overall best model (see Table 2). Applying the freezing strategy here resulted in a loss of performance (esp. training AUC) and probably generalizability. The encoder part of the model tried to adapt to the fine-tuning task at the cost of previously learned representations from pretraining.

The performance of the models can also be explained with modality contributions measured with the method from Gapp et al. (2025a). These results highlight the effectiveness of pretraining one modality for subsequent integration in a multimodal task. The ResNetMLP processed visual data with ca. 79.5%, while the pretrained versions used 93.31% (frozen encoder) and 92.20% (partly frozen encoder) of the image data (Table 3). Even the unimodal collapse that occurred for the ViTMLP trained from scratch could be tackled with the pretraining step. The modality contribution for vision increased from 0.13% to ca. 28.39% when using the overall best model: the pretrained ViTMLP with frozen encoder.

Although this seems to be still small compared to best version of ResNetMLP (92.20%), it must be much better balanced for the stroke prediction task as performance results confirm. The pretrained ViTMLP with frozen encoder could focus on training the fusion model more efficient and thus find cross modality patterns, such as gender, age, or heart disease specific vision features. Especially gender specific vision patterns may be the main finding by the ViTMLP as modality contributions in Table 3 demonstrate (55.6% contribution of gender). The 28.26% (=28.39%-0.13%) increase in the visual modality’s contribution, observed when comparing ViTMLP from scratch to ViTMLP with a pretrained frozen encoder, corresponded mainly to a decrease in gender (-20.08%), with additional decreases in age (-4.22%) and heart diseases (-3.96%).

6. Conclusion

Through utilization of self-supervised pretraining with 3D CTAs we could improve our previous results from Gapp et al. (2025b) for early detection of stroke relapses. While the ResNetMLP architecture worked out best with the pretrained model, which has been fine tuned by using progressive layer unfreezing, the pretrained ViTMLP with frozen encoder performed overall best with an AUC of 0.74 on the testing dataset. The contribution of the visual modality increased substantially throughout the pretraining stage across all models, provided that freezing was applied judiciously to specific parts of the network. Especially for ViTMLP, we could even tackle an unimodal collapse that occurred for the model trained from scratch. Vision features were used considerably more effective in the multimodal task (0.13% before vs. 28.39% after pretraining). Hence, pretraining unlocked a broader range of discriminative features for the stroke relapse detection task. By enhancing the contribution of individual modalities, multimodal datasets can be fully leveraged to reveal vision patterns linked to attributes such as age, gender, or cardiovascular disease.

With our contribution to the field of multimodal, medical, explainable AI, we aim to inspire future research in this area to use pretraining methods to reveal otherwise hidden information embedded in the data. Self-supervised pretraining can definitely lead to improvements across multiple fine-tuning tasks. While this may involve leveraging combined multimodal features, we are, to the best of our knowledge, the first to demonstrate the effects of both pretraining and strategic fine-tuning on modality contributions. Specifically, we analyze the modality contribution gap between pretrained models and those trained from scratch, revealing novel insights at the modality level. Due to legal restrictions, the dataset cannot be publicly released. Nevertheless, the full pretrain code, partial downstream-task components, and supplementary materials are available at https://github.com/ChristianGappGit/SSL_Pretraining.

Acknowledgments

This study is partly supported by VASCage – Centre on Clinical Stroke Research.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

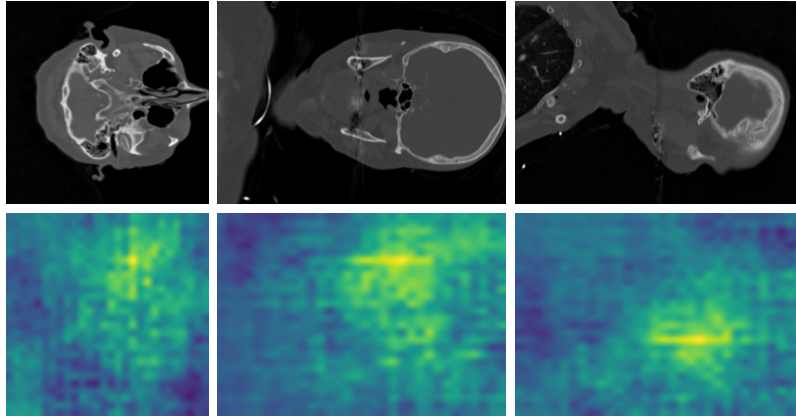
References

- Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=LtKcMgG0eLt>.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. doi: 10.48550/arxiv.2010.11929.
- Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, Apr 1999. ISSN 1364-6613. doi: 10.1016/S1364-6613(99)01294-2. URL [https://doi.org/10.1016/S1364-6613\(99\)01294-2](https://doi.org/10.1016/S1364-6613(99)01294-2).
- Christian Gapp, Elias Tappeiner, Martin Welk, Karl Fritscher, Elke R. Gizewski, and Rainer Schubert. What are you looking at? modality contribution in multimodal medical deep learning. *International Journal of Computer Assisted Radiology and Surgery*, Oct 2025a. ISSN 1861-6429. doi: 10.1007/s11548-025-03523-w. URL <https://doi.org/10.1007/s11548-025-03523-w>.
- Christian Gapp, Elias Tappeiner, Martin Welk, Karl Fritscher, Stephanie Mangesius, Constantin Eisenschink, Philipp Deisl, Michael Knoflach, Astrid E. Grams, Elke R. Gizewski, and Rainer Schubert. Xsrd-net: Explainable stroke relapse detection, 2025b. URL <https://arxiv.org/abs/2509.07772>.
- Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2015. URL <https://arxiv.org/abs/1312.6211>.
- 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, 06 2016. doi: 10.1109/CVPR.2016.90.
- Adrian Javaloy, Maryam Meghdadi, and Isabel Valera. Mitigating modality collapse in multimodal VAEs via impartial optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9938–9964. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/javaloy22a.html>.
- Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multi-modal transformers robust to missing modality? In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18156–18165, 2022. doi: 10.1109/CVPR52688.2022.01764.

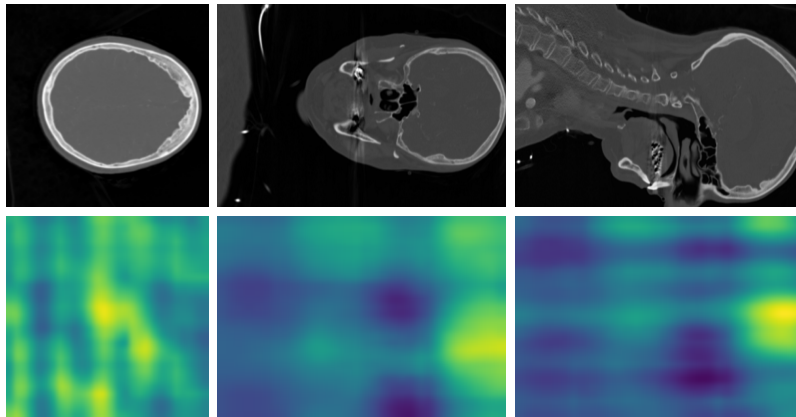
- Letitia Parcalabescu and Anette Frank. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.223.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.
- Luya Wang, Feng Liang, Yangguang Li, Honggang Zhang, Wanli Ouyang, and Jing Shao. Repre: Improving self-supervised vision transformer with reconstructive pre-training. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1437–1443. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/200. URL <https://doi.org/10.24963/ijcai.2022/200>. Main Track.
- Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12692–12702, 2020. doi: 10.1109/CVPR42600.2020.01271.
- Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(10):12113–12132, 10 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3275156.
- Han-Jia Ye, Qi-Le Zhou, Huai-Hong Yin, De-Chuan Zhan, and Wei-Lun Chao. Rethinking pre-training in tabular data: A neighborhood embedding perspective, 2025. URL <https://arxiv.org/abs/2311.00055>.
- Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025. URL <https://arxiv.org/abs/2303.00915>.

Appendix A. Interpretability Analysis for one Relapse

For the best-performing ReNetMLP and the best-performing ViTMLP an analysis for one example true positive training item (TP_1), that is a true predicted relapse, is provided. In Figure 5 occlusion sensitivity results are depicted. The maps are created using an occlusion mask size of $[8, 8, 10]$, resulting in overall 25.088 patches (= model forward passes for inference). Table 4 shows the detailed modality contributions for the item using the two networks.



(a) Occlusion Sensitivity with best ResNetMLP.



(b) Occlusion Sensitivity with best ViTMLP.

Figure 5: Occlusion sensitivity on 3D CTAs from the true positive predicted relapse example TP_1 (RFS = 107 days). (a) Top: original 3D CTAs. Bottom: corresponding saliency maps for the slices with ResNetMLP pretrained temp. frozen (visual importance is 82.40% for this example). (b) Top: original 3D CTAs. Bottom: Saliency maps for the slices with ViTMLP pretrained frozen (visual importance is 54.28%). From left to right: transversal slice (viewed from top), coronal slice (viewed from back), sagittal slice (viewed from left). Yellow regions mark high importance.

XSRD-Net	Training Setup	vision		tabular		
		3D CTA	age	gender	CHD	PAD
ResNetMLP	from scratch	84.65%	1.29%	5.96%	0.00%	8.10%
	pretrained temp. frozen	82.40%	1.42%	8.50%	0.00%	7.68%
ViTMLP	from scratch	0.14%	4.66%	73.72%	0.00%	21.48%
	pretrained frozen	54.28%	1.87%	35.94%	0.00%	7.91%

Table 4: Comparison of modality-level contributions for the true positive relapse example TP_1 (RFS = 107 days), evaluated using models trained from scratch and the best-performing ResNetMLP and ViTMLP architectures.

For this particular case (item TP_1), the best-performing ResNetMLP identifies the left side of the neck as the most influential region for relapse prediction. This finding aligns with the results from (Gapp et al., 2025b), where the carotid arteries were reported as highly relevant for RFS prediction. The visual analysis shows a thin spike along the left side of the neck, most likely corresponding to the arteria carotis communis (see Figure 5 (a), bottom row). By contrast, the ViTMLP model highlights the upper-central, slightly right intracranial area as the primary region of importance for this example item TP_1 (see Figure 5 (b)).

However, it should be noted that the regions of interest exhibit different absolute importance, as the contribution of the vision modality is higher for the ResNetMLP, whereas the gender feature has greater influence in the ViTMLP (see Table 4). The distinct region of interest identified by the ViTMLP may reflect a gender-related, high-dimensional visual feature. To explore this further, deeper synergy analysis between clinical attributes and vision patches is required.