A Self-Supervised Model for Multi-modal Stroke Risk Prediction

Camille Delgrange
Signal Processing Institute
EPFL University
Lausanne, Switzerland
camille.delgrange@alumni.epfl.ch

Olga Demler
Brigham and Women's Hospital
Harvard Medical School
Boston, Massachusetts, USA
odemler@bwh.harvard.edu

Samia Mora
Brigham and Women's Hospital
Harvard Medical School
Boston, Massachusetts, USA
smora@bwh.harvard.edu

Bjoern Menze
Department of Quantitative Biomedicine
University of Zurich
Zurich, Switzerland
bjoern.menze@uzh.ch

Ezequiel de la Rosa
Department of Quantitative Biomedicine
University of Zurich
Zurich, Switzerland
ezequiel.delarosa@uzh.ch

Neda Davoudi[*]
ETH AI Center, Department of Computer Science
Department of Quantitative Biomedicine
University of Zurich
Zürich, Switzerland
neda.davoudi@ai.ethz.ch

Abstract

Predicting stroke risk is a complex challenge that can be enhanced by integrating diverse clinically available data modalities. This study introduces a self-supervised multimodal framework that combines 3D brain imaging, clinical data, and image-derived features to improve stroke risk prediction prior to onset. By leveraging large unannotated clinical datasets, the framework captures complementary and synergistic information across image and tabular data modalities. Our approach is based on a contrastive learning framework that couples contrastive language-image pretraining with an image-tabular matching module, to better align multimodal data representations in a shared latent space. The model is trained on the UK Biobank, which includes structural brain MRI and clinical data. We benchmark its performance against state-of-the-art unimodal and multimodal methods using tabular, image, and image-tabular combinations under diverse frozen and trainable model settings. The proposed model outperformed self-supervised tabular (image) methods by 2.6% (2.6%) in ROC-AUC and by 3.3% (5.6%) in balanced accuracy. Additionally, it showed a 7.6% increase in balanced accuracy compared to the best multimodal supervised model. Through interpretable tools, our approach demonstrated better integration of tabular and image data, providing richer and more aligned embeddings. Gradient-weighted Class Activation Mapping heatmaps further revealed activated brain regions commonly associated in the literature with brain aging, stroke risk, and clinical outcomes. This robust self-supervised multimodal framework surpasses state-of-the-art methods for stroke risk prediction and offers a strong foundation for future studies integrating diverse data modalities to advance clinical predictive modeling.

[*]Corresponding author

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

1 Introduction

Stroke ranks as the second leading cause of death worldwide, responsible for 11.6% of global fatalities in 2019. It often results in neurological damage and long-term disability in adults, imposing significant health and economic challenges [1, 2]. Early detection through predictive models is crucial in preventing severe outcomes, as cerebrovascular events can cause irreversible brain damage within hours [3]. The complexity of stroke, driven by multiple risk factors, highlights the importance of integrating multi-modal data to improve diagnostic accuracy and treatment strategies. Among the various imaging techniques, Magnetic Resonance Imaging (MRI) stands out as a highly effective tool, offering high-resolution, non-invasive assessments of structural abnormalities and detailed visualization of the brain's vascular network [4].

Uni-modal predictive models Prior works mainly use convolutional neural networks (CNN) that can leverage the high-dimensional imaging information for diagnosing patients [5]. Yu et al. applied deep learning algorithms to extract meaningful imaging features in an increasing order of hierarchical complexity to make predictions of the infarct volume [6]. Other models that use only clinical data, often assume linear relationships between traditional risk factors such as age, gender, smoking status, blood pressure, diabetes, cholesterol levels, and body mass index [7, 8, 9]. Alaa et al. used AutoPrognosis, an ensemble machine learning approach, to outperform conventional models like the Framingham score and Cox models [10]. A major limitation of these models is that they don't integrate complementary information from other modalities, similar to how clinicians diagnose using multiple data sources. Biobanks like the UK Biobank (UKB) have become invaluable in this context, providing vast datasets integrating imaging and clinical information to train machine learning models for disease prediction [11, 12].

Multi-modal predictive models Several studies have employed multi-modal data to improve diagnostic capabilities by integrating diverse data types [13]. For example, MultiSurv model has shown success by fusing image and tabular data for cancer survival prediction [14]. multi-modal models combining image and clinical data have demonstrated better prediction performance for disability prediction in stroke patients [15, 16]. However, CNNs tend to prioritize image features, and simple image-tabular CNN concatenation fails to enhance predictive models due to insufficient cross-modal interactions. To address this, Wolf et al. developed the Dynamic Affine Feature Map Transform (DAFT), which conditions convolutional feature maps on both image and tabular data, enabling a two-way information exchange via an auxiliary neural network [17]. While DAFT reduces issues related to the large number of trainable parameters in standard 3D CNNs and the curse of dimensionality, it may sacrifice some predictive power compared to deeper models like ResNet. Although recent models show promise in biomedical prediction tasks, their clinical translation is hindered by limited annotated datasets, low disease prevalence, and the risk of overfitting. Self-supervised learning (SSL) is a powerful technique for extracting representative features from unlabeled data, making it valuable for early disease risk identification.

Self-supervised models Unlike traditional supervised learning, SSL defines pretext tasks that allow models to learn meaningful representations from raw data [18]. One prominent SSL technique is contrastive learning, which trains encoders to generate augmented views of a sample, maximizing similarity between these views while minimizing similarity with other samples [18]. Popular methods such as SimCLR [19], BYOL [20], and MOCO [21] have demonstrated success in imaging tasks, while VIME [22] and SCARF [23] are leading approaches for tabular data. Emerging approaches, like contrastive language-image pre-training (CLIP) strategy, have evolved from unimodal methods to integrate diverse modalities. While there was an extensive work done for cardiovascular diseases prediction [24, 25, 26], stroke risk prediction through volumetric brain images and clinical health records remains underexplored.

We present for the first time, to the best of our knowledge, a self-supervised multi-modal approach integrating 3D brain MRIs with clinical tabular data for stroke risk prediction. As depicted in Figure 1, our methodology incorporates cross-modal interactions via CLIP loss [27] and image-tabular matching (ITM) loss [28, 25]. We demonstrate that our learning strategy outperforms leading (self-)supervised unimodal methods and that multi-modal image-tabular pre-training leads to better representations and improved downstream performance. Lastly, we validate the model's learned features through visual activation maps, which align with established clinical and neurological findings on stroke-related brain pathology. Code is available at `https://github.com/CamilleDelgrange/SSMSRPM`.

2 Materials and Methods

2.1 Dataset

Our analyses are performed on T2-Fluid Attenuation Inversion Recovery (FLAIR) brain volumes, and over a subset of clinical information spanning across five categories, extracted from the UKB: demographics, lifestyle, biomarkers, comorbidities, and medication. The complete list of features is available in the supplementary materials. Continuous features are standardized using z-score normalization, while categorical data is one-hot encoded. For our experiments, we use 5000 and 500 samples for training and validation sets respectively for the model pre-training stage. Train, validation, and test subset for the downstream fine-tuning stage use 278, 93, and 93 samples respectively. The fine-tuning sets are stratified according to sex, age and stroke diagnosis to account for confounders to avoid spurious correlations and class imbalance. To handle missing tabular data, we use an iterative multivariate imputer based on Multivariate Imputation by Chained Equations (MICE) [29], modelling missing features as a function of existing features over multiple imputation rounds. Missing categorical data is replaced by the most frequent category. This step is performed after data normalization, to ensure that the means and standard deviations are calculated only from recorded values. The 3D brain images are registered to Montreal Neurological Institute brain template (MNI) space, have uniform dimensions of $182 \times 218 \times 182$ and a voxel size of $1mm^3$ and are processed using the UKB imaging pipeline [30]. Key image-derived phenotypes (IDPs), such as segmented brain tissue volumes and white matter hyperintensity (WMH) volumes, are extracted and used as brain IDPs. Brain lesion segmentation is performed using the BIANCA tool to produce 3D binary lesion masks [31]. Furthermore, lesion segmentation masks are characterized by pyradiomics [32] through radiomic features such as volume, area, elongation, and sphericity and these features are used as lesion IDPs.

2.2 Multi-modal self-supervised framework

Our pipeline is split into two sequential steps. First, we pre-train the tabular and imaging encoders (Figure 1 A) and then we fine-tune them with labels from downstream task (Figure 1 B). Each batch of data contains pairs of imaging x_{j_i} and tabular x_{j_t} samples. These samples are augmented by random transformations $t \sim \tau$ from a set of parametric transforms τ, such as random cropping and affine transforms for the images, or random feature corruption for the tabular data. We use an image augmentation rate of 95% for the model to still occasionally see unaltered data to capture the original data distribution for transfering the learnt features to the downstream task. The corruption rate of the tabular data is set to 0.3 as in the original tabular method SCARF [23]. For a given reference point, known as anchor x, the positive samples are the ones derived from x transformations while other samples in the batch are considered as negative samples. Augmented images x_{j_i} and tabular data x_{j_t} are passed through the imaging encoder f_{θ_I} and tabular encoder f_{θ_T} to generate the embeddings. These embeddings are propagated through the 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 L2-normalized onto a unit hypersphere. The projections are pulled and pushed in the shared latent space according to the CLIP loss [27], which maximizes the cosine similarity of projections from the positive samples and minimizes the similarity of projections from the negative samples in the batch. In contrast to the original InfoNCE loss used in SimCLR [19], and following the CLIP loss, the projected embeddings similarities are contrasted between data modalities. An image projection is therefore defined as :

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

Considering all N subjects in a batch, the loss for the imaging modality is defined as follows:

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

where τ is the temperature parameter. In our experiments, a temperature of 0.1 is selected to work best, following [19]. $l_{t,i}$ is computed analogously and CLIP loss is defined as follows:

$$\mathcal{L}_{clip} = \lambda l_{i,t} + (1 - \lambda) l_{t,i} \tag{3}$$

We choose value of 0.5 for the λ as regularization parameter. The aim is to learn patient-wise representations invariant to the variation of the image-tabular pairs. Hard negative samples are crucial in contrastive learning as they help the model distinguish between similar samples, preventing trivial solutions and enhancing its robustness. We implement a hard negative mining strategy to predict whether image-tabular data pairs are positive or negative, using image-tabular matching (ITM) loss. In this approach, for each image or tabular representation, we identify an unmatched tabular or image representation from the mini-batch. This selection is based on similarity scores computed using the CLIP method, which serves as the sampling weight for the negative pairs [25, 28]. A multi-modal interaction module is introduced, as shown in Figure 1, which takes the output of the projector heads to perform inter-modality learning and generates a multi-modal representation. It uses a cross-attention mechanism [33], enabling tabular embeddings to attend to relevant image embeddings. The multi-modal interaction module contains two transformer layers, with four attention heads and a hidden dimension of 256, each including self-attention, cross-modal attention, an MLP feed-forward module and layer normalization [25]. The output of the multi-modal module is a [CLS] token, aggregating the information from the entire sequence, used for downstream classification task, where the model needs a single feature vector representing the entire input [34]. The [CLS] embedding is capturing a joint representation of the image-tabular pair that is fed into the ITM predictor (a linear layer) to match the prediction based on a binary cross-entropy loss $\mathcal{L}_{ITM}$. Therefore, the complete loss is expressed as:

$$\mathcal{L} = (\mathcal{L}_{CLIP} + \mathcal{L}_{ITM})/2 \tag{4}$$

Downstream task predictions After pre-training, the projection heads were replaced by fully connected layers. Extracting the representation before the projector has been shown to improve downstream tasks performance [18]. For downstream fine-tuning and binary classification of healthy versus stroke (Figure 1 B), we employ ensemble learning to improve model generalization and performance by leveraging the rich representations from the image encoder, tabular encoder, and the multi-modal transformer interaction module. All pre-trained models are evaluated using linear probing (frozen) and fine-tuning (trainable). The frozen models use tuned linear classifiers after the feature extractors. The datasets used for model fine-tuning are balanced in each batch of training, validation, and test subset. This way, we reduce potential bias due to class-imbalance, as well as unstable and slow training due to imbalance batch distributions.

3 Experiments

3.1 Benchmarking

The herein proposed solution is compared against supervised and SSL strategies, each of them using imaging, tabular, and integrated imaging-tabular methodologies.

3.1.1 Supervised learning methods

To benchmark our proposed method, we implement two state-of-the-art, supervised image-based models, namely ResNet50 [21] and DenseNet121 [35], two supervised tabular data approaches, namely a two-layer tabular MLP model and a tabular transformer encoder inspired from Du et. al. (2024) [25]. We conduct an ablation study using a supervised MLP model with various feature combinations to identify the optimal feature set. This process helps us to select the final combination of features for improved model performance. The combinations include: i) clinical tabular data only, spanning the previously mentioned categories (clinical), ii) clinical data with brain extracted IDPs (clinical + brain IDPs), and iii) clinical data, brain IDPs and lesion IDPs (clinical + brain IDPs + lesion IDPs). Furthermore, we implement three supervised, multi-modal (imaging-tabular) learning models, namely a simple concatenation fusion model (CF) [36], a CF model integrated with the tabular transformer encoder inspired by the work from Du et. al. (2024) [25] (CF + Transformer), and DAFT model [17]. All models employing an imaging encoder are implemented with ResNet50 as backbone. DAFT block is integrated within ResNet50 from the third stage onwards. To alleviate over-fitting, an early stopping strategy is adopted, with a minimal delta (divergence threshold) of 1×10^{-4}, a maximal number of epochs of 50, and a patience of 15 epochs.

3.1.2 Self-supervised learning methods

Our model is compared against leading, self-supervised contrastive solutions, including: i) the unimodal, image-based SimCLR[19] approach, ii) the unimodal, tabular data-based SCARF [23]

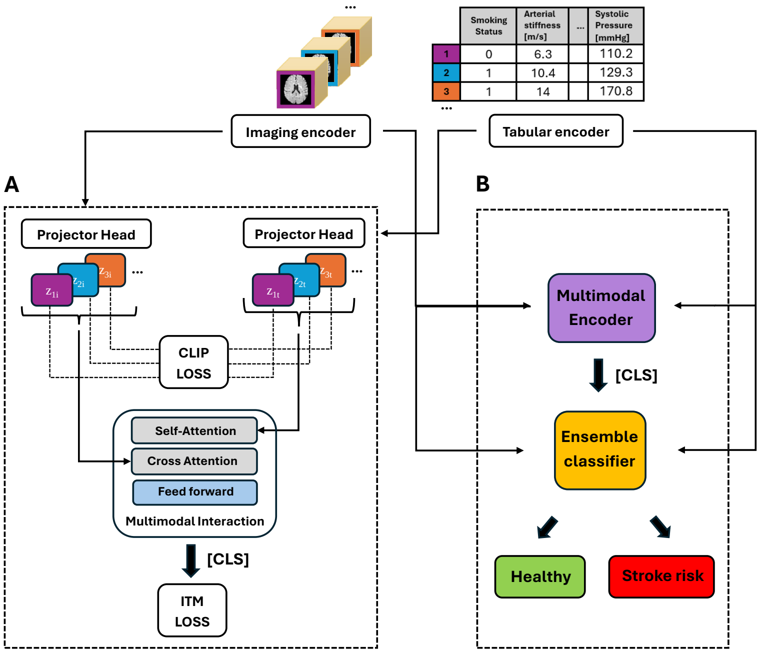

Figure 1: Pipeline for joint imaging and tabular data pre-training (A) and downstream fine tuning (B). CLIP loss is applied on projected data to align the image and tabular representations. Hard negative pairs are mined through CLIP similarities within the batch. A transformer block with self-attention and cross-attention layers is used to cross-attend both modalities, resulting in a multi-modal [CLS] token fed to a classifier and used for further downstream fine tuning. Image-Tabular Matching (ITM) loss evaluates the image-tabular pair matching. In the downstream task, an ensemble classifier is fine-tuned to predict healthy versus stroke from pre-trained imaging, tabular and multimodal encoders.

approach, and *iii*) the multi-modal, CLIP method (without ITM loss). The hyperparameters and training configurations for all SSL pre-training approaches are adapted for our specific dataset and task, and are obtained through hyper-parameter search. All models are pre-trained for 100 epochs using an Adam optimizer [37]. The learning rate is warmed up linearly for 10 epochs and decayed following a cosine annealing scheduler. For all methods, as in the CLIP+ITM method, the image augmentation rate is 95% and the tabular corruption rate 0.3 during pre-training and 80% and 0.3 during downstream fine tuning. SimCLR is trained using the NTXent objective [19] and the temperature parameter is kept to 0.1. Hidden and projected dimensions are respectively 2048 and 128 for both modalities [19]. The same parameters are used for SCARF [23]. Learning rates are chosen with a sweep through a range of learning rates, by tracking the validation loss. Weight decay and dropout rate are added depending on the level of overfitting observed at the validation loss. The same early stopping strategy is employed as in the supervised learning methods. The downstream fine-tuning is using the same parameters as the presented method, using only a single modality classifier for unimodal SSL pre-trained methods and a fused representation vector with a single linear classifier for the CLIP only model. Trainable models in each SSL unimodal method means that the other modality is incorporated during fine-tuning as a full trainable model. The employed batch size for all methods is 6. All SSL models are pretrained on a Tesla V100-SXM2 (32GB, 42 CPUs), and inference is performed on an NVIDIA GeForce RTX 4090 (24GB, 62 CPUs). Pretraining took $\sim$ 24 hours, while fine-tuning took less than 1 hour.

3.2 Interpretability and qualitative analysis

Embeddings visualization is done using a two-dimension Uniform Manifold Approximation and Projection (UMAP) technique [38], to evaluate the quality of the generated latent space embedding after pre-training, using validation samples. Such approach allows to qualitatively assess the latent space representation and data distribution after encoding each modality with either a unimodal or

Table 1: Feature selection with supervised MLP results using different combinations of tabular data. ROC-AUC: area under the receiver operating characteristic curve; bAcc (%): balanced Accuracy; Se (%): sensitivity. For each metric, the best-performing method is highlighted in **bold** and the second-best is underlined.

Model	Metrics (%)			
	AUC	**bAcc**	**F1**	**Se**
MLP (clinical)	65.36	61.29	60.87	60.87
MLP (clinical + brain IDPs)	**71.37**	60.31	65.96	67.39
MLP (clinical + brain IDPs + lesion IDPs)	65.63	**62.58**	**68.69**	**73.91**

multi-modal pre-trained model, giving some hints about the successfullness of the learning strategies. A latent space embedding size of 2048 dimensions is produced.

For qualitative analysis, we generated imaging heatmaps using the Gradient-weighted Class Activation Mapping (GradCAM) technique [39] to visualize the regions in each slice that contributed most significantly to the model's predictions across given brain MRI volume. GradCAM [39] heatmaps are normalized in the range of 0 to 1 and are upsampled with trilinear interpolation to match the original image space. The 7th layer of the ResNet50 encoder is used to allow capturing high-level features and spatial structure that is suitable for visualization. The most informative slice (defined as the one accounting with the highest heatmap activation scores) for each view in axial, sagittal, or coronal plane is generated. We use the 3D GradCAM implementation from MONAI.

3.3 Performance assessment

All models are evaluated through the area under the Receiver Operating Characteristic (ROC) curve. Binary classification metrics, namely balanced accuracy, F1-Score, and sentitivty, are included. Classification metrics are reported at the Youden-index operating point ($J = \text{Sensitivity} + \text{Specificity} - 1$) retrieved from the (validation set) ROC curve. The metrics are chosen bearing in mind that potential clinical applications of this study could serve as screening and risk stratification tools, where the models sensitivity plays an important role to avoid missing positive stroke cases.

4 Results and Discussion

4.1 Benchmarking

To determine which tabular features to include, we conducted a supervised-learning ablation analysis using various combinations of tabular data subgroups. As shown in Table 1, the models that incorporate clinical and brain IDPs achieve the highest ROC-AUC scores. However, the method that also includes lesion IDPs outperforms in binary classification metrics, such as F1-score and sensitivity. To prioritize model robustness while maintaining a smaller feature set, the subsequent benchmarking of models using tabular data is performed using only clinical and brain IDPs.

A summary of the different models performance is shown in Table 2. It is observed that the proposed multi-modal learning strategy outperforms all other methodologies across all considered metrics, with the *trainable* model setting performing slightly better than the *frozen* one.

When comparing models based on learning approach and data modality, it can be observed that the best performing imaging supervised learning strategy is DenseNet121 (ROC-AUC 66.79%). In DenseNet architectures, layers are densely connected, which improves feature reuse and gradient flow, leading to richer feature representations. However, this dense connectivity increases memory overhead during training, particularly with the large inputs used in this study. To optimize the trade-off between efficiency and memory usage, we selected ResNet50 as the encoder for SSL pre-training, accepting a minor reduction in performance.

When comparing SSL strategies, it is evident that fine-tuning both data modalities in multi-modal approaches significantly boosts performance. The performance gap is considerable when comparing these multi-modal models with unimodal image-based models, showing that image data alone is insufficient for effectively addressing the task. The best performing method is the CLIP+ITM model, performing better than all unimodal (tabular and imaging) SSL methods. Interestingly, DAFT performs similar to the multi-modal SSL methods in terms of ROC-AUC and balanced accuracy, although exhibits poor F1-score and sensitivity results. There is no clear difference in performance between trainable and frozen settings across all models. We hypothesize this is because the pre-trained models have already developed robust, transferable representations, making fine-tuning less impactful. Additionally, the small size of the fine-tuning dataset may limit the effectiveness of further learning beyond what was achieved during pre-training. Besides, it could be hypothesized that freezing the model may serve as a form of regularization, helping to mitigate overfitting, particularly in this setting with limited labeled data.

Table 2: Benchmarking performance results. F and T denote frozen and trainable pre-trained encoders. ROC-AUC: area under the receiver operating characteristic curve; bAcc (%): balanced Accuracy; Se (%): Sensitivity. For each metric, the best-performing method is highlighted in **bold** and the second-best is underlined. The overall best performing method is highlighted in gray.

Model	Tabular	Image	Metrics (%)			
			AUC	**bAcc**	**F1**	**Se**
(a) Supervised Image						
ResNet-50 [40]	-	T	63.25	57.08	60.01	65.22
DenseNet121 [35]	-	T	66.79	66.79	69.90	78.26
(b) Supervised Tabular						
MLP	T	-	71.37	60.31	65.96	67.39
Transformer [25]	T	-	64.38	62.21	47.62	32.61
(c) Supervised multi-modal						
Concat Fuse (CF) [36]	T	T	65.26	60.29	62.63	67.39
Concat Fuse (CF) [w/ Transformer] [36, 25]	T	T	63.48	52.08	66.17	95.65
DAFT [17]	T	T	73.82	63.51	69.57	65.01
(d) SSL Image						
SimCLR [19]	-	F	64.99	52.38	33.33	28.91
SimCLR [19]	-	T	65.59	55.67	43.83	34.78
SimCLR [19]	T	F	72.02	65.56	64.44	63.04
SimCLR [19]	T	T	72.11	65.56	64.44	63.04
(e) SSL Tabular						
SCARF [23]	F	-	71.18	62.42	63.91	67.39
SCARF [23]	T	-	70.35	64.48	62.92	60.87
SCARF [23]	F	T	72.16	62.16	43.48	53.34
SCARF [23]	T	T	72.02	67.85	73.08	78.26
(f) SSL multi-modal						
CLIP [26]	T	T	73.41	61.5	67.24	80.78
CLIP [26]	F	F	73.54	71.00	70.97	71.74
CLIP+ITM [25, 28]	T	T	74.42	**71.11**	**74.22**	**84.78**
CLIP+ITM [25, 28]	F	F	**74.75**	62.77	67.29	76.60

4.2 Interpretability and qualitative analysis

4.2.1 Embeddings visualization

Figure 2 shows the UMAP embeddings distribution for unimodal and multi-modal data models. On one hand, it can be observed that in Fig. 2 A, there is a clear distinction between (unimodal learnt) tabular and imaging data modalities, with data samples clustered by data-type. In this case, the

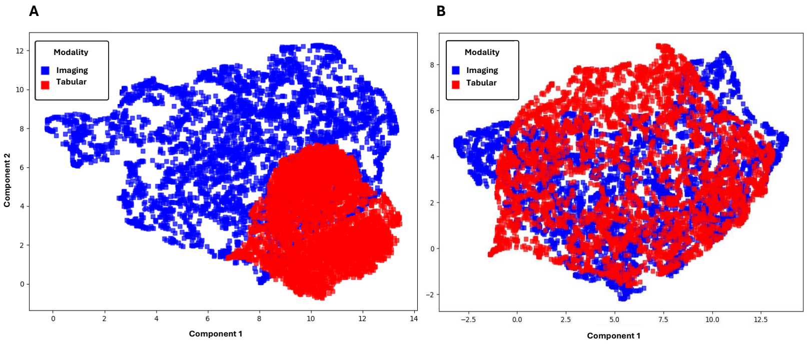

Figure 2: 2D UMAP projections of tabular and imaging embeddings from the validation set, using (A) unimodal pre-trained tabular and imaging encoders and (B) multi-modal pre-trained tabular and imaging encoders.

embeddings generated from imaging data and tabular data are significantly different from each other in the feature space when generated with a unimodal pre-trained model (i.e., either SCARF or SimCLR). The tight clustering of red points suggests that the tabular data embeddings are more homogeneous and possibly more concentrated in the feature space compared to the broad representation of brain MRI images. Therefore, unimodal-data encoders have learned modality-specific features, without capturing interactions between them. On the other hand, in Fig. 2 B the UMAP plot obtained for the best performing multi-modal model (CLIP+ITM) is shown. In this case, there is significant overlap between the tabular and imaging embeddings, suggesting that the model has found common representations for the two different data types, either via shared visual features or via learning associated clinical patterns in tabular and brain MRIs. Thus, CLIP+ITM is able to encode the underlying patient representation in a common latent space by reducing data augmentation noise. Still, there are data-points in the plot having distinct representations within each modality, suggesting that the model could not project them to the modality-shared latent space. The broad distribution of points across the entire UMAP space suggests that the embeddings capture a wide variety of features from both imaging and tabular data, rather than collapsing all data points into a narrow cluster. These results expose the enhanced performance of the multi-modal SSL strategy by projecting diverse data modalities into a shared embedding space, and thus suggesting a better model starting point for downstream analysis.

4.2.2 Imaging heatmaps

Figure 3 shows results from the GradCAM experiment obtained over predicted samples. When inspecting the positive predicted scans (True Positives and False Positives), the model tends to highlight anatomical regions surrounding the lateral ventricles and (periventricular) white matter areas. Such patterns could be associated to white matter hyperintensities, which are known predictors of brain atrophy and age-related brain alterations [41] and also stroke risk predictors in elderly individuals [42]. In different studies, correlations have been observed between common age-related structural brain changes and brain pathologies [41, 43, 44]. When assessing scans #2 and #4 of the true positive patients in Fig. 3, the activation maps are also showing anatomical regions distant from the lateral ventricles, showing high activations. Supported from literature, those activations could be related to white matter hyperintensities (deep white matter, in this case), often appearing in regions of the brain that are not immediately adjacent to the cortical surface, but commonly located in subcortical white matter or in deep white matter tracts [41]. Such deep white matter hyperintensities are associated with chronic vascular disease and other chronic pathologies (e.g. multiple sclerosis) [41]. When evaluating negatively predicted patients (True Negatives and False Negatives), the scans are showing less emphasis on the (periventricular) white matter region but instead highlight areas of the lower brain (cerebellum, posterior brain) and the cortex. We hypothesize that these areas may

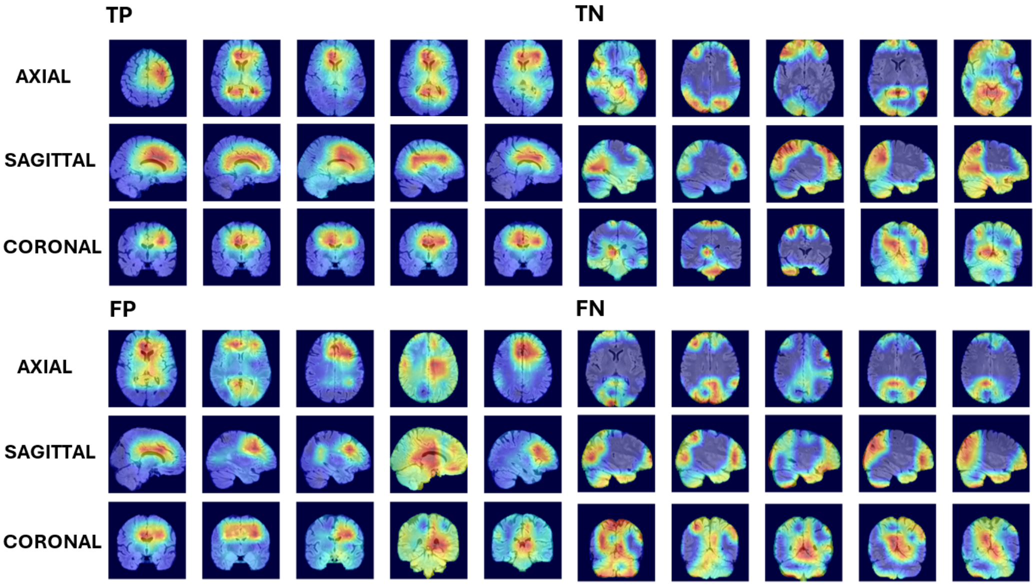

Figure 3: GradCAM-activated brain regions for five patients, categorized as TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). Red (blue) indicates higher (lower) activations.

reflect patterns related to normal aging or normal brain atrophy processes, rather than anomalous brain conditions. Overall, we can hypothesise from these visualizations that the multi-modal SSL stroke risk predictor model focuses on abnormal brain aging patterns for its predictions. We therefore believe that future experiments including brain-age and brain structure-age biomarkers could help enhancing the models predictability, since they have been shown to be associated with overall cardiovascular risk [45], clinical outcome in stroke [46] and overall risk of mortality [47].

Limitations. Our study is limited by the use of the UK Biobank, whose demographic characteristics may not fully represent the diversity of global populations, potentially impacting the model's generalizability and clinical utility. Future research should validate our approach using more diverse external datasets to improve applicability. Additionally, our test set was constrained by the limited availability of pre-stroke imaging samples, as most stroke datasets focus on post-onset cases. Finally, the heterogeneity in the time between imaging and stroke onset in the UK Biobank could influence model performance, necessitating further experiments to disentangle these effects. Future work could also include improving model efficiency by testing further architectures and techniques to reduce model parameters (e.g. network pruning).

5 Conclusion

We hereby present an SSL model integrating diverse data modalities for predicting stroke risk. The model's performance is compared against state-of-the-art (self-)supervised models employing both unimodal and multi-modal data, including tabular and imaging datasets. A comprehensive set of experimental settings is utilized, encompassing different subgroupings of tabular features—such as clinical data, brain IDPs, and lesion IDPs—as well as various training regimes that combine pre-training and fine-tuning based on data modality.

Our results demonstrate that the CLIP model on multi-modal data, combined with an ITM loss, outperforms single-modality alternatives.The CLIP+ITM model surpasses the self-supervised tabular (image) data SCARF (SimCLR) model by 2.6% (2.6%) in ROC-AUC, and by 3.3% (5.6%) in balanced accuracy terms. Our framework also demonstrated an AUROC improvement of 0.93% and 7.6% balanced accuracy from the best multi-modal supervised method. Additionally, the proposed model produces well-aligned multi-modal representations in a common, data modality-independent space, which is unattainable with unimodal tabular or imaging data models. Thus, CLIP-ITM effectively leverages complementary and synergistic information from diverse data modalities.

Using interpretable GradCAM heatmaps, we identified activated brain regions commonly associated with brain aging, stroke risk, and clinical outcomes. On one hand, the activated areas indicate that the model primarily focuses on deep and periventricular white matter hyperintensities for predicting positive samples, which may be more common and extensive in patients identified as at risk for stroke. On the other hand, the prediction of negative samples highlights the cerebellum, posterior brain regions and cortical areas. These results demonstrate the model's capacity to extract task-specific features linked to stroke risk, which are well-supported by existing literature.

In conclusion, we propose a robust self-supervised multi-modal learning approach for stroke risk prediction. Our model offers a strong foundation for future studies that aim to integrate multiple data modalities into prediction models.

Acknowledgement

This research has been conducted using the UK Biobank resource under application number 81959 and grants NHLBIR21HL156174, NHLBIK24HL136852, and K24HL136852. This project was supported by the grant #2023-N-306 of the 1st Joint Call of the Swiss Data Science Center (SDSC) and the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology). N.D. is partially supported by the ETH AI Center postdoctoral fellowship. B.M., N.D., and E.D. acknowledge support of the Helmut-Horten-Foundation.

References

[1] Liyuan Pu et al. "Projected Global Trends in Ischemic Stroke Incidence, Deaths and Disability-Adjusted Life Years From 2020 to 2030". In: *Stroke* 54.5 (May 2023), pp. 1330–1339. ISSN: 15244628. DOI: 10.1161/STROKEAHA.122.040073. URL: https://www.ahajournals.org/doi/10.1161/STROKEAHA.122.040073.

[2] Valery L. Feigin et al. "Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019". In: *The Lancet. Neurology* 20.10 (2021), pp. 1–26. ISSN: 1474-4465. DOI: 10.1016/S1474-4422(21)00252-0. URL: https://pubmed.ncbi.nlm.nih.gov/34487721/.

[3] Gagan D. Flora and Manasa K. Nayak. "A Brief Review of Cardiovascular Diseases, Associated Risk Factors and Current Treatment Regimes". In: *Current Pharmaceutical Design* 25.38 (Sept. 2019), pp. 4063–4084. ISSN: 13816128. DOI: 10.2174/1381612825666190925163827. URL: https://pubmed.ncbi.nlm.nih.gov/31553287/.

[4] Valentina Hartwig et al. *Biological effects and safety in magnetic resonance imaging: A review*. 2009. DOI: 10.3390/ijerph6061778. URL: https://pubmed.ncbi.nlm.nih.gov/19578460/.

[5] Huanhuan Zhang and Yufei Qie. *Applying Deep Learning to Medical Imaging: A Review*. Sept. 2023. DOI: 10.3390/app131810521. URL: https://www.mdpi.com/2076-3417/13/18/10521/htm%20https://www.mdpi.com/2076-3417/13/18/10521.

[6] Yannan Yu et al. "Use of deep learning to predict final ischemic stroke lesions from initial magnetic resonance imaging". In: *JAMA network open* 3.3 (2020), e200772–e200772.

[7] Julia Hippisley-Cox et al. "Development and validation of a new algorithm for improved cardiovascular risk prediction". In: *Nature Medicine* 30.5 (Apr. 2024), pp. 1440–1447. ISSN: 1546170X. DOI: 10.1038/s41591-024-02905-y. URL: https://www.nature.com/articles/s41591-024-02905-y.

[8] Jaejin An et al. "Recurrent atherosclerotic cardiovascular event rates differ among patients meeting the very high risk definition according to age, sex, race/ethnicity, and socioeconomic status". In: *Journal of the American Heart Association* 9.23 (Dec. 2020). ISSN: 20479980. DOI: 10.1161/JAHA.120.017310. URL: https://www.ahajournals.org/doi/10.1161/JAHA.120.017310.

[9] Jia You et al. "Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study". In: *Stroke and vascular neurology* 8.6 (Dec. 2023), pp. 475–485. ISSN: 2059-8696. DOI: 10.1136/SVN-2023-002332. URL: https://pubmed.ncbi.nlm.nih.gov/37105576/.

[10] Ahmed M. Alaa et al. "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants". In: *PLoS ONE* 14.5 (May 2019). ISSN: 19326203. DOI: 10.1371/journal.pone.0213653. URL: https://pubmed.ncbi.nlm.nih.gov/31091238/.

[11] Thomas J. Littlejohns et al. *The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions*. May 2020. DOI: 10.1038/s41467-020-15948-9. URL: https://www.nature.com/articles/s41467-020-15948-9.

[12] UK BioBank. "UK Biobank - UK Biobank". In: (2021). URL: https://www.ukbiobank.ac.uk/.

[13] Sidong Liu et al. *Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders*. Sept. 2015. DOI: 10.1007/s40708-015-0019-x. URL: https://braininformatics.springeropen.com/articles/10.1007/s40708-015-0019-x.

[14] Luís A. Vale-Silva and Karl Rohr. "Long-term cancer survival prediction using multimodal deep learning". In: *Scientific Reports* 11.1 (June 2021), pp. 1–12. ISSN: 20452322. DOI: 10.1038/s41598-021-92799-4. URL: https://www.nature.com/articles/s41598-021-92799-4.

[15] Adam White et al. "Predicting recovery following stroke: deep learning, multimodal data and feature selection using explainable AI". In: *arXiv preprint arXiv ...* (Oct. 2023). arXiv: 2310.19174. URL: https://arxiv.org/abs/2310.19174v1%20https://arxiv.org/abs/2310.19174%0Ahttps://arxiv.org/pdf/2310.19174.

[16] Yongkai Liu et al. “Functional Outcome Prediction in Acute Ischemic Stroke Using a Fused Imaging and Clinical Deep Learning Model”. In: *Stroke* 54.9 (Sept. 2023), pp. 2316–2327. ISSN: 15244628. DOI: 10.1161/STROKEAHA.123.044072. URL: https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.123.044072.

[17] Tom Nuno Wolf, Sebastian Pölsterl, and Christian Wachinger. “DAFT: A universal module to interweave tabular data and 3D images in CNNs”. In: *NeuroImage* 260 (Oct. 2022), p. 119505. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2022.119505.

[18] Randall Balestriero et al. “A Cookbook of Self-Supervised Learning”. In: (2023). arXiv: 2304.12210. URL: http://arxiv.org/abs/2304.12210.

[19] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *37th International Conference on Machine Learning, ICML 2020* PartF168147-3.Figure 1 (2020), pp. 1575–1585. arXiv: 2002.05709.

[20] Jean Bastien Grill et al. “Bootstrap your own latent a new approach to self-supervised learning”. In: *Advances in Neural Information Processing Systems*. Vol. 2020-Decem. Neural information processing systems foundation, June 2020. ISBN: 2006.07733v3. arXiv: 2006.07733. URL: https://arxiv.org/abs/2006.07733v3.

[21] Kaiming He et al. “Momentum Contrast for Unsupervised Visual Representation Learning”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Nov. 2020, pp. 9726–9735. DOI: 10.1109/CVPR42600.2020.00975. arXiv: 1911.05722. URL: https://arxiv.org/abs/1911.05722v3.

[22] Rein Houthooft et al. “VIME: Variational information maximizing exploration”. In: *Advances in Neural Information Processing Systems*. Vol. 0. Neural information processing systems foundation, May 2016, pp. 1117–1125. arXiv: 1605.09674. URL: https://arxiv.org/abs/1605.09674v4.

[23] Dara Bahri et al. “Scarf: Self-Supervised Contrastive Learning Using Random Feature Corruption”. In: *ICLR 2022 - 10th International Conference on Learning Representations* (2022), pp. 1–24. arXiv: 2106.15147.

[24] Adityanarayanan Radhakrishnan et al. “Cross-modal autoencoder framework learns holistic representations of cardiovascular state”. In: *Nature Communications* 14.1 (Dec. 2023). ISSN: 20411723. DOI: 10.1038/s41467-023-38125-0. URL: /pmc/articles/PMC10140057/%20/pmc/articles/PMC10140057/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10140057/.

[25] Siyi Du et al. “TIP: Tabular-Image Pre-training for Multimodal Classification with Incomplete Data”. In: *arxiv* (July 2024). arXiv: 2407.07582. URL: https://arxiv.org/abs/2407.07582v1%20http://arxiv.org/abs/2407.07582.

[26] Paul Hager, Martin J. Menten, and Daniel Rueckert. “Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data”. In: *arXiv* (2023), pp. 23924–23935. DOI: 10.1109/cvpr52729.2023.02291. arXiv: 2303.14080.

[27] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of Machine Learning Research*. Vol. 139. ML Research Press, Feb. 2021, pp. 8748–8763. ISBN: 9781713845065. arXiv: 2103.00020. URL: https://arxiv.org/abs/2103.00020v1.

[28] Junnan Li et al. “Align before Fuse: Vision and Language Representation Learning with Momentum Distillation”. In: *Advances in Neural Information Processing Systems*. Vol. 12. Neural information processing systems foundation, July 2021, pp. 9694–9705. ISBN: 9781713845393. arXiv: 2107.07651. URL: https://arxiv.org/abs/2107.07651v2.

[29] Stef van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate imputation by chained equations in R”. In: *Journal of Statistical Software* 45.3 (Dec. 2011), pp. 1–67. ISSN: 15487660. DOI: 10.18637/jss.v045.i03. URL: https://www.jstatsoft.org/index.php/jss/article/view/v045i03.

[30] Fidel Alfaro-Almagro et al. “Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank”. In: *Neuroimage* 166 (2018), pp. 400–424.

[31] Fidel Alfaro-Almagro et al. “Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank”. In: *NeuroImage* 166 (Feb. 2018), pp. 400–424. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2017.10.034.

[32] Joost JM Van Griethuysen et al. “Computational radiomics system to decode the radiographic phenotype”. In: *Cancer research* 77.21 (2017), e104–e107.

[33] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. Vol. 2017-Decem. Neural information processing systems foundation, June 2017, pp. 5999–6009. ISBN: 1706.03762v7. arXiv: 1706.03762. URL: https://arxiv.org/abs/1706.03762v7.

[34] Chao Ye et al. “CT-BERT: Learning Better Tabular Representations Through Cross-Table Pre-training”. In: *Proceedings of the ACM Web Conference 2024 (WWW ’24), May 13â•fi17, 2024, Singapore, Singapore* 1 (July 2023). DOI: 10.1145/XXXXXXX.XXXXXXX. arXiv: 2307.04308. URL: https://arxiv.org/abs/2307.04308v1.

[35] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Vol. 2017-Janua. Institute of Electrical and Electronics Engineers Inc., Aug. 2017, pp. 2261–2269. ISBN: 9781538604571. DOI: 10.1109/CVPR.2017.243. arXiv: 1608.06993. URL: https://arxiv.org/abs/1608.06993v5.

[36] Simeon Spasov et al. “A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer’s disease”. In: *NeuroImage* 189 (Apr. 2019), pp. 276–287. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2019.01.031. URL: https://pubmed.ncbi.nlm.nih.gov/30654174/.

[37] Diederik P. Kingma and Jimmy Lei Ba. “Adam: A method for stochastic optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, Dec. 2015. arXiv: 1412.6980. URL: https://arxiv.org/abs/1412.6980v9.

[38] Leland McInnes, John Healy, and James Melville. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: (2018). arXiv: 1802.03426. URL: http://arxiv.org/abs/1802.03426.

[39] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2017-Octob. Institute of Electrical and Electronics Engineers Inc., Dec. 2017, pp. 618–626. ISBN: 9781538610329. DOI: 10.1109/ICCV.2017.74.

[40] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2016-Decem. IEEE Computer Society, Dec. 2016, pp. 770–778. ISBN: 9781467388504. DOI: 10.1109/CVPR.2016.90. arXiv: 1512.03385. URL: https://arxiv.org/abs/1512.03385v1.

[41] Lukas A. Grajauskas et al. *MRI-based evaluation of structural degeneration in the ageing brain: Pathophysiology and assessment*. Jan. 2019. DOI: 10.1016/j.arr.2018.11.004.

[42] J-H Park et al. “White matter hyperintensities and recurrent stroke risk in patients with stroke with small-vessel disease”. In: *European Journal of Neurology* 26.6 (2019), pp. 911–918.

[43] Hui Guo et al. “MRI assessment of whole-brain structural changes in aging”. In: *Clinical Interventions in Aging* 12 (Aug. 2017), pp. 1251–1270. ISSN: 11781998. DOI: 10.2147/CIA.S139515. URL: http://dx.doi.org/10.2147/CIA.S139515.

[44] Yong Soo Shim et al. “Pathological correlates of white matter hyperintensities on magnetic resonance imaging”. In: *Dementia and Geriatric Cognitive Disorders* 39 (Feb. 2015), pp. 92–104. ISSN: 14219824. DOI: 10.1159/000366411.

[45] Ann-Marie G De Lange et al. “Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study”. In: *NeuroImage* 222 (2020), p. 117292.

[46] Sook-Lei Liew et al. “Association of brain age, lesion volume, and functional outcome in patients with stroke”. In: *Neurology* 100.20 (2023), e2103–e2113.

[47] James H Cole et al. “Brain age predicts mortality”. In: *Molecular psychiatry* 23.5 (2018), pp. 1385–1392.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes- both quantitative and qualitative experiments were conducted to support the claims of this work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the limitations are discussed after the results in page 8.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There is no theoretical assumption or proof in this work.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes, all the experimental parameters as well as all the models architecture details are given in sections 2 and 3. The code will be published open source for full reproducibility in the camera-ready version.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be released in the camera-ready version to fully reproduce all the experimental results as described. The data is from the UK Biobank and is therefore not publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, every detail is shared concerning the pretraining and training splits, hyperparameters and optimizers used in sections 2 and 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Error bars or confidence intervals are not really applicable in this context/case/experiments. Train/test splits and model initialization are precised.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details for compute resources are elaborated in section 3.1.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, data anonymisation was preserved according to UKB guidelines and societal impact was maximized in the context of this research work. No participant was harmed and they were all volunteers to the UKB.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, the discussion and conclusion elaborate on the potential benefits of anticipating stroke events in healthy populations.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All references to data (UK Biobank), codes, and models are referenced accordingly.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Code, documentations and pre-trained models checkpoints will be released for camera-ready version.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing. We use the data from human participants that were already collected by UK Biobank.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing. We use the data from human participants that were already collected by UK Biobank.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.