Explainability in mulimodal deep transformation models for stroke outcome prediction

Lisa Herzog[1], Jonas Brändli[2], Maurice Schneeberger[3], Loran Avci[2], Nordin Dari[2], Martin Hänsel[1], Hakim Baazaoui[1], Pascal Bühler[2,4], Susanne Wegener[1], and Beate Sick[2,3,4]

[1] University Hospital Zurich, Department of Neurology, Frauenklinikstrasse 26, 8091 Zurich, Switzerland
[2] Zurich University of Applied Sciences, Institute of Data Analysis and Process Design (IDP), Technikumstrasse 9, 8401 Winterthur, Switzerland
[3] University of Zurich, Epidemiology, Biostatistics and Prevention Institute (EBPI), Hirschengraben 84, 8001 Zurich, Switzerland
[4] Thurgau Institute for Digital Transformation (TIDIT), Hauptstrasse 21a, 8280 Kreuzlingen, Switzerland beate.sick@uzh.ch, beate.sick@tidit.ch

Abstract. Multimodal prediction models based on imaging and clinical data are increasingly used for clinical decision support, yet their interpretability remains limited. We present multimodal Deep Transformation Models (DTMs), which combine statistical approaches and neural networks to achieve strong predictive performance while preserving interpretability for tabular data. A key contribution of this work is the adaption of the xAI methods Grad-CAM and Occlusion to DTMs relying on 3D CNNs, enabling interpretation of the image branch through the generation of explanation maps. We developed DTMs to predict functional independence three months after stroke using diffusion-weighted imaging and clinical data from 407 patients. In a ten-fold cross-validation, the models achieved state-of-the-art predictive performance (AUC 0.81 [0.75, 0.87]) while maintaining interpretability for tabular features, with functional independence before stroke and stroke severity on admission emerging as the strongest predictors. Explanation maps from both xAI methods highlighted consistent regions, including frontal lobe areas which are known to be associated with age, a strong predictor of functional outcome. Notably, these regions disappeared once age was included as an explicit tabular predictor. Similarity analyses of explanation maps revealed distinct spatial patterns, providing meaningful insights into stroke pathophysiology, systematic error analysis and hypothesis generation.

Keywords: Explanation maps · multimodal neural network models · deep transformation models · stroke.

1 Introduction

Accurate and trustworthy outcome prediction models are increasingly important for clinical decision support, yet building such models remains challenging

due to the multimodal nature of patient data and strict requirements on model transparency. Clinically suitable models should jointly integrate heterogeneous sources such as neuroimaging and structured clinical data, achieve strong predictive performance, and provide interpretable results that allow clinicians to understand how predictions arise [14]. Classical statistical models offer interpretability through interpretable parameter estimates but struggle with high-dimensional inputs such as medical images, whereas modern deep learning (DL) achieves strong performance at the cost of limited explainability.

Deep Transformation Models (DTMs) combine statistical modeling with neural networks (NNs) to enable interpretable multimodal prediction [9, 1]. They estimate a conditional probability distribution of the outcome while integrating different input modalities with (potentially deep) NNs. Their design preserves interpretability through interpretable parameter estimates for tabular features while allowing complex feature extraction from imaging data. Prior work has demonstrated strong predictive performance of DTMs in medical applications [6, 2]. However, the image branch remains difficult to interpret.

Post-hoc explainable artificial intelligence (xAI) methods provide a practical solution to the black box problem. Without altering the architecture, they highlight the contribution of image regions to a model's output. While such explanations do not render models inherently transparent and may have limitations regarding robustness [14, 16], they currently represent the most practical approach for explainability. However, their application to probabilistic DTMs has not been established yet.

In this work, we extended Occlusion [17] and Grad-CAM [15] to multimodal DTMs with 3D CNN image branches. Using a cohort of 407 stroke and transient ischemic attack (TIA) patients, we evaluated predictive performance, interpretability of clinical parameters and systematically analyzed explanation maps to identify characteristic spatial patterns associated with functional outcome prediction. This establishes a general framework for explainability in multimodal transformation models and supports hypothesis generation and error analysis in clinical prediction tasks.

2 Material and methods

2.1 Data

We analyzed a retrospectively collected cohort of 407 patients with ischemic stroke (n=295) or TIA (n=112) admitted to the University Hospital Zurich between 2014 and 2018. TIA and stroke patients frequently present with similar neurological symptoms, making neuroimaging essential for diagnosis and characterization of tissue injury. All patients underwent diffusion-weighted imaging (DWI) within three days after symptom onset, with most scans acquired within the first 24 hours. Brain volumes were preprocessed to a size of 128x128x28 and normalized to have zero mean and unit standard deviation. Tabular data included demographics, vascular risk factors and admission information (see

Fig. 5). Categorical variables were dummy-encoded and numerical variables standardized to ensure comparability of parameter estimates. Functional outcome was assessed using the modified Rankin Scale (mRS) at three months, dichotomized into favorable (mRS 0-2, n=332) and unfavorable (mRS 3-6, n=75) outcome. Ethical approval was obtained from the local ethics committee and all participants have provided written informed consent.

2.2 Deep Transformation Models

DTMs are interpretable probabilistic models for multimodal prediction estimating a conditional probability distribution $F_Y(y|\mathsf{B}, x)$ of an outcome Y given imaging data B and tabular features x [9]. Instead of estimating the outcome distribution directly, DTMs estimate a transformation function h that maps a predefined latent distribution F_Z to the targeted outcome distribution, such that

$$F_Y(y|\mathsf{B}, x) = F_Z(h(y|\mathsf{B}, x)). \tag{1}$$

The parameter-free probability distribution F_Z defines the interpretational scale of the model parameters. For instance, using the standard logistic distribution $\sigma(z) = \frac{1}{1+\exp(-z)}$ yields a formulation analogous to logistic regression, allowing additive model components to be interpreted on the log-odds scale [7, 9].

For binary outcomes, the transformation function represents a threshold on the z-scale, partitioning $F_Z(z)$ into two segments which determine the class probabilities $p_0 = P(y_0|\mathsf{B}, x) = \sigma(h(y_0|\mathsf{B}, x))$ and $p_1 = P(y_1|\mathsf{B}, x) = 1 - p_0$ (see Fig. 2). Here, y_0 represents a favorable, y_1 an unfavorable outcome. We define

$$h(y_0|\mathsf{B}, x) = \vartheta_0(\mathsf{B}) - x^\top \beta, \tag{2}$$

where $\vartheta_0(\mathsf{B})$ is estimated by a 3D CNN processing imaging data and $x^\top\beta$ with a NN that represents a linear shift for tabular features. The NNs are jointly fitted by minimizing the negative log-likelihood

$$\text{NLL} = -\frac{1}{n}\sum_{i=1}^{n} \log(F_Z(y_k|\mathsf{B}, x) - F_Z(y_{k-1}|\mathsf{B}, x)) \tag{3}$$

using standard stochastic optimization. With $F_Z(z) = \sigma(z)$, $\exp(\beta_k)$ describes the multiplicative change in the odds of a favorable outcome associated with a one unit increase in x_k, when holding the remaining inputs constant. For a detailed discussion we refer to [9].

2.3 Explanation maps

To obtain visual explanations of the image branch, we adapted Occlusion [17] and Grad-CAM [15] to the DTM framework (see Fig. 1). Both methods operate on the trained CNN and account for the transformation function output. Grad-CAM provides insight into spatial features represented in the network, whereas Occlusion directly evaluates the effect of localized perturbations on the probability for the predicted class, therefore, providing complementary perspectives.

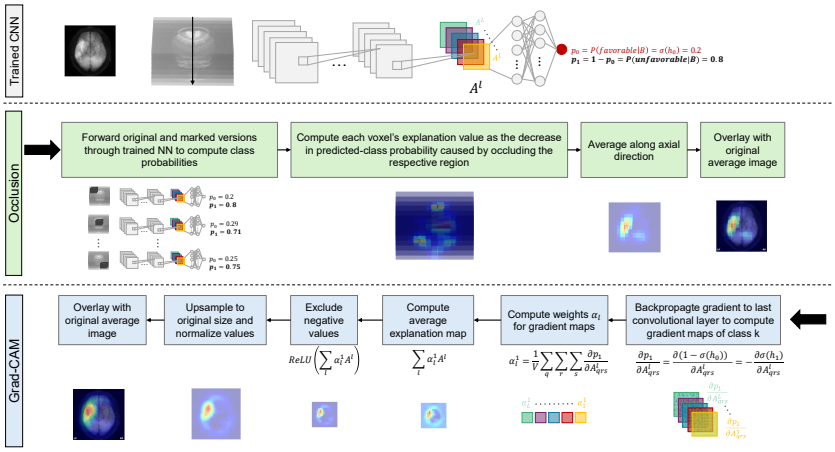

Fig. 1. Overview of the Grad-CAM and Occlusion method for DTMs. Both approaches build on the trained CNN. Occlusion identifies relevant regions by systematically masking parts of the input and measuring the resulting change in the probability of the predicted class. Grad-CAM derives voxel importance from the gradients flowing from the predicted class back to the last convolutional layer.

Occlusion Occlusion was performed by systematically masking regions of the 3D input and measuring changes in the probability of the predicted class. Regions whose masking decreased the predicted probability were considered important for the decision. For the 3D images, we used an occlusion size of 18x18x4 and a stride width of 10x10x3.

Grad-CAM In standard 2D CNN classifiers, Grad-CAM leverages gradient information flowing from the output node for the predicted class k into the last convolutional layer. In DTMs, however, the 3D CNN contributes to the transformation function rather than directly producing class probabilities. Considering $h_0 = h(y_0|B, x) = \vartheta_0(B) - x^T\beta$, we therefore quantify the importance of each 3D activation map A^l, $l = 1, \ldots, L$ in the last convolutional layer by computing the gradient for each voxel $A^l_{(q,r,s)}$

$$\frac{\partial p_1}{\partial A^l_{(q,r,s)}} = \frac{\partial \sigma(\vartheta_0(\mathsf{B}) - x^T\beta)}{\partial A^l_{(q,r,s)}} = \sigma(\vartheta_0(\mathsf{B}))(1 - \sigma(\vartheta_0(\mathsf{B}))) \frac{\partial \vartheta_0(\mathsf{B})}{\partial A^l_{(q,r,s)}}. \tag{4}$$

The last expression only involves differentiable components while the gradient of $\vartheta_0(\mathsf{B})$ is calculated automatically via backpropagation. Pooling these gradients yields map weights

$$\alpha^1_l = \frac{1}{R \cdot S} \sum_q \sum_r \sum_s \frac{\partial p_1}{\partial A^l_{qrs}}. \tag{5}$$

The final explanation map is obtained via a weighted combination of activation maps followed by ReLU and upsampling

$$L_1^{\text{Grad-CAM}} = \text{ReLU}\left(\sum_l \alpha_l^1 \cdot A^l\right). \tag{6}$$

This formulation enables gradient-based explanation maps consistent with the probabilistic DTM structure. Derivations for other transformation functions are equivalent.

2.4 Experiments

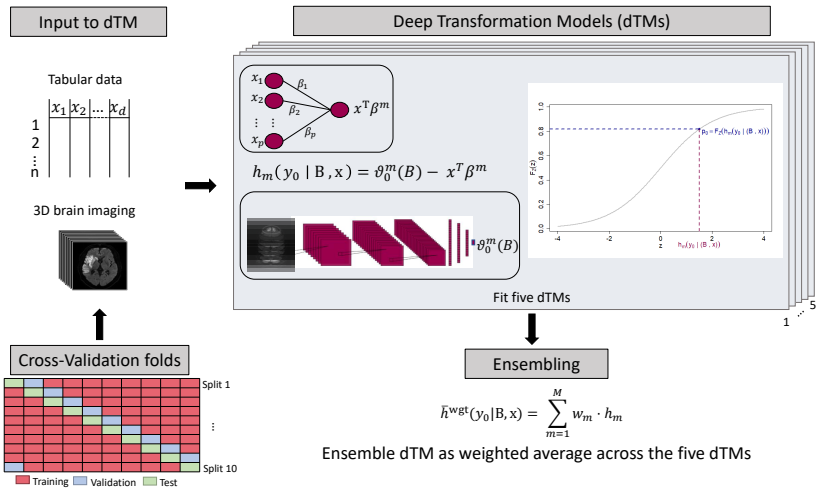

Fig. 2. DTM training and evaluation setup. In each CV fold, we fit $M = 5$ DTMs on the training data and determine a threshold on the validation set to separate the standard logistic distribution $F_Z(z) = \sigma(z)$ into two segments to yield a probability for favorable $p_0 = \sigma(h_m(y_0|\mathsf{B}, x))$ vs. unfavorable $(1 - p_0)$ outcome. The final ensemble DTM threshold is a weighted average of the M thresholds.

We compared an imaging-only model with a complex intercept ($\mathbf{CI_B}$; $h(y_0|\mathsf{B}) = \vartheta_0(\mathsf{B})$), a tabular-only model with simple intercept and linear shift ($\mathbf{SI\text{-}LS_x}$; $h(y_0|x) = \vartheta_0 + x^T\beta$; logistic regression), and a multimodal model ($\mathbf{CI_B\text{-}LS_x}$; $h(y_0|\mathsf{B}, x) = \vartheta_0(\mathsf{B}) + x^T\beta$). In addition, we fitted a null model to assess if the models have learned meaningful features to enable interpretability of explanation maps. The linear shift terms were implemented with a fully connected, single-layer NN. The image branch of the models used a 3D CNN with four convolutional blocks (3x3x3 kernels, 32-32-64-64 filters), global average pooling and two fully connected layers (128 units, dropout 0.3). During training, data augmentation (zooming, flipping, shifting, rotating) was applied.

All models were evaluated using stratified ten-fold cross-validation while ensuring that test data was unseen and untouched during the training procedure (see Fig. 2). For the models including imaging data, deep ensembling ($M = 5$) [10] was applied, resulting in five transformation functions per fold that were averaged with weights w_m optimized to minimize the NLL on the validation data with $\bar{h}^{wgt} = \sum_{m=1}^{M} w_m h_m$. The five 3D explanation maps resulting from Occlusion and Grad-CAM were averaged using the same weights w_m. To produce a final, interpretable 2D explanation map, the 3D ensemble map was overlayed with the original 3D brain volume and averaged along the axial direction (see Fig. 1).

Test performance was assessed using NLL, AUC, accuracy, specificity and sensitivity. For specificity, sensitivity, and accuracy, we report 95% Wilson confidence intervals (CIs). CIs for NLL and AUC were obtained via bootstrapping. Class labels were derived from predicted probabilities using thresholds determined to maximize the geometric mean of the true-positive and true-negative rates on the validation set of each fold. All code is publicly available at `https://github.com/liherz/xAI_paper`.

Similarity plots of Grad-CAM explanation maps were used to identify spatial patterns, understand stroke pathophysiology and perform error analysis and hypothesis generation. Therefore, we applied t-SNE on the features of the explanation maps extracted from the last convolutional layer of a ResNet-50 trained on Imagnet [4]. An interactive plot is available under `https://liherz.shinyapps.io/tsne_xAI_shinyio/`.

3 Results

Table 1. Predictive performance of models using diffusion weighted imaging (DWI) data, tabular clinical data (Clinical) and a combination of both. NLL = Negativ log-likelihood, AUC = Area under the Receiver Operating Characteristic Curve.

	DWI	**Clinical**	**DWI + Clinical**
NLL	0.452 [0.379, 0.528]	0.368 [0.308, 0.433]	0.384 [0.306, 0.470]
AUC	0.714 [0.649, 0.777]	0.809 [0.747, 0.867]	0.812 [0.752, 0.868]
Specificity	0.792 [0.745, 0.832]	0.783 [0.736, 0.824]	0.795 [0.749, 0.835]
Sensitivity	0.480 [0.371, 0.591]	0.627 [0.514, 0.727]	0.640 [0.527, 0.739]
Accuracy	0.735 [0.690, 0.775]	0.754 [0.710, 0.794]	0.767 [0.723, 0.805]

The multimodal model achieved the highest predictive performance (AUC 0.812 [0.752, 0.868], see Table 1). However, the model using clinical data alone performed similar (AUC 0.809 [0.747, 0.867]), indicating that clinical features contain most of the predictive information. All models showed higher specificity (correctly predicted favorable outcomes), likely reflecting the predominance of favorable outcomes in the cohort. Comparison with the null model (AUC 0.498

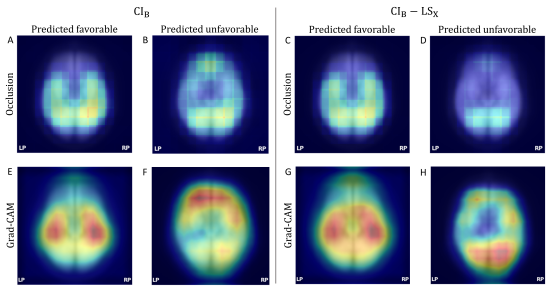

Fig. 3. Average explanation maps. The figure shows the average across the explanation maps for patients predicted having a favorable or unfavorable outcome, produced with the CI_B and CI_B-LS_X model when applying Occlusion and Grad-CAM.

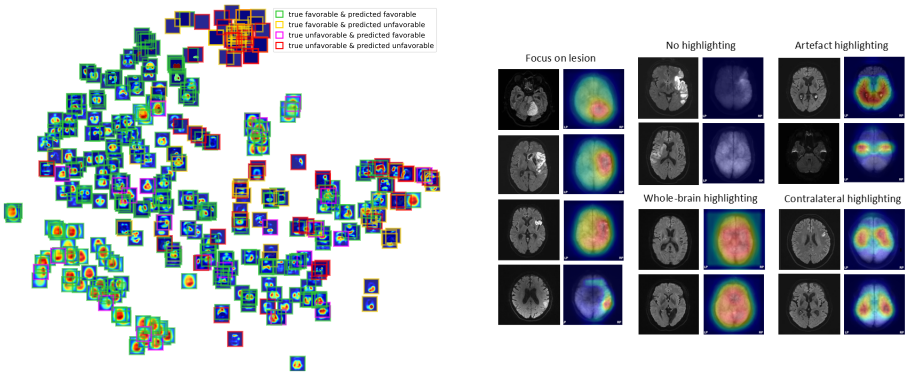

Fig. 4. Grad-CAM explanation analysis. (A) Similarity structure of explanation maps visualized using t-SNE, illustrating clustering behavior across correctly and incorrectly predicted favorable and unfavorable outcome groups. (B) Distinct spatial patterns of averaged explanation maps with representative volumetric examples.

[0.430, 0.570]) confirms that all models have learned meaningful patterns, supporting a cautious but informative interpretation of the explanation maps.

Occlusion and Grad-CAM highlighted similar regions (see Fig. 3). However, Grad-CAM yielded smoother and more distinct average explanation maps while being less computationally expensive. For unfavorable outcome prediction, the imaging-only model $\mathbf{CI_B}$ focused on frontal lobe regions, which are known to be associated with age, a strong predictor of functional outcome [3]. These regions disappeared in the multimodal $\mathbf{CI_B\text{-}LS_x}$ model, indicating that the imaging-only model had captured age-related patterns to predict unfavorable outcomes that became redundant when age was included explicitly as a tabular predictor.

Cluster analysis revealed several consistent patterns (see Fig. 4). In most stroke cases, the model focused on lesion areas appearing as hyperintensities on DWI. Occasionally, imaging artifacts were highlighted (e.g. ventricular hyperintensities). Frequently, the model highlighted bilateral regions, suggesting

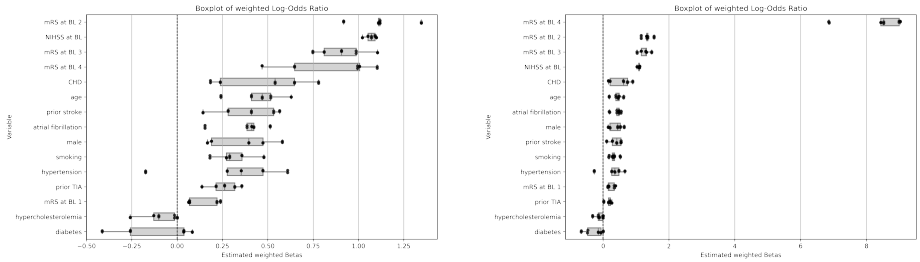

Fig. 5. Estimated log-odds ratios with 95% bootstrap confidence intervals for tabular predictors derived from (A) the SI-LS_X model and (B) the tabular component of the CI_B-LS_X model. Estimates were obtained from 10-fold cross-validation and averaged across ensemble members. Categorical features (sex, smoking, hypertension, prior stroke or TIA, atrial fibrillation, coronary heart disease (CHD), diabetes, hypercholesterolemia, pre-stroke mRS (mRS at BL)) are shown with respect to the reference category. Parameter estimates of numeric variables (age, National Institutes of Health Stroke Scale on admission (NIHSS at BL)) are based on the standardized measures.

assessment of hemispheric/contralateral symmetry. When strong clinical predictors (e.g. high NIHSS scores) were available, Grad-CAM maps often showed minimal activation, indicating reduced reliability on images. In patients with TIA or very small lesions, when only a few slices show visible abnormalities, activation was on the entire image, suggesting that the image branch contributed global rather than localized information.

Beyond imaging explanations, DTMs provide interpretable parameters for tabular predictors when being included as linear shift terms (see Fig. 5). Here, pre-stroke mRS and admission NIHSS emerged as the strongest predictors in models with and without imaging data.

4 Discussion

In this work, we extended Occlusion and Grad-CAM to DTMs with multimodal input, enabling interpretability for models integrating 3D imaging alongside clinical data. Our models for functional outcome prediction achieved state-of-the-art performance in acute ischemic stroke and enabled to identify tabular features and imaging patterns relevant for outcome prediction. Tabular data contained most of the predictive information with NIHSS on admission and pre-stroke mRS being most important. Both Grad-CAM and Occlusion offered meaningful insights into stroke pathophysiology and highlighted characteristic patterns such as lesion areas. Interestingly, the imaging-only model focused on the frontal lobe area for predicting unfavorable outcomes which is associated with age, a well-known predictor of functional outcome. Please note that this does not contradict the finding that NIHSS and pre-stroke mRS were the most important tabular predictors as these two measures are not directly visible on the image. Overall, Grad-CAM and Occlusion maps highlighted similar regions. However,

Grad-CAM produced smoother, more visually coherent maps while requiring less computational power making it the preferred approach.

Our work has several limitations. Although our goal was on explainability, external validation of the models remains necessary and generalizability is limited due to the single-center cohort. This will be addressed in future work. Further, 10-fold CV allowed us to evaluate test performance on the 407 patients and yielded robust estimates, but prediction performance is limited with an AUC of around 0.8. Nonetheless, this test performance is in line with the literature and comparable to other multimodal approaches for outcome prediction in acute ischemic stroke reporting AUC values between 0.65 and 0.8 [8, 11, 12]. We assume that the performance is not a model limitation but rather a clinical one as three-month mRS is influenced by post-stroke events (e.g. rehabilitation, recurrent strokes), which are unpredictable in the acute setting. Currently, we additionally test other potential outcome measures (like 24 hour NIHSS) for improved prediction. Moreover, we collected more data and will extend the approaches to vision transformers, which represent state-of-the-art architectures. In addition, we integrated novel 3D CNN foundation models such as [13] to potentially improve predictive performance [5].

In conclusion, DTMs combine the strengths of deep learning and statistical modeling, providing reliable and interpretable predictions for multimodal data. By adapting explanation maps to this framework, the image branch becomes more transparent, supporting hypothesis generation and assessment of model trustworthiness, which is critical for clinical translation.

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

References

1. Baumann, P.F., Hothorn, T., Rügamer, D.: Deep conditional transformation models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 3–18. Springer (2021)
2. Campanella, G., Kook, L., Häggström, I., Hothorn, T., Fuchs, T.J.: Deep conditional transformation models for survival analysis. arXiv:2210.11366 (2022)
3. Fujita, S., Mori, S., Onda, K., Hanaoka, S., Nomura, Y., Nakao, T., Yoshikawa, T., Takao, H., Hayashi, N., Abe, O.: Characterization of brain volume changes in aging individuals with normal cognition using serial magnetic resonance imaging. JAMA Network Open **6**(6), e2318153–e2318153 (2023)
4. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. IEEE International Conference on Computer Vision (2015). https://doi.org/10.1109/ICCV.2015.123
5. Herzog, L., Bühler, P., de la Rosa, E., Sick, B., Wegener, S.: Outcome prediction and individualized treatment effect estimation in patients with large vessel occlusion stroke. In: Image Analysis in Stroke Diagnosis and Interventions: 5th International Workshop, SWITCH 2025, Held in Conjunction with MICCAI 2025, Daejeon, South Korea, September 23, 2025, Proceedings. pp. 73–82 (2025). https://doi.org/10.1007/978-3-032-07945-9_8, https://doi.org/10.1007/978-3-032-07945-9_8

6. Herzog, L., Kook, L., Götschi, A., Petermann, K., Hänsel, M., Hamann, J., Dürr, O., Wegener, S., Sick, B.: Deep transformation models for functional outcome prediction after acute ischemic stroke. Biometrical Journal **65**(6), 2100379 (2023)
7. Hothorn, T., Kneib, T., Bühlmann, P.: Conditional transformation models. Journal of the Royal Statistical Society. Series B: Statistical Methodology **76**(1), 3–27 (2014). https://doi.org/10.1111/rssb.12017
8. Klug, J., Leclerc, G., Dirren, E., Carrera, E.: Machine learning for early dynamic prediction of functional outcome after stroke. Communications Medicine **4**(1), 232 (2024). https://doi.org/10.1038/s43856-024-00666-w, `https://doi.org/10.1038/s43856-024-00666-w`, published online: 13 November 2024
9. Kook & Herzog, L.L., Hothorn, T., Dürr, O., Sick, B.: Deep and interpretable regression models for ordinal outcomes. Pattern Recognition **122**, 108263 (2022)
10. Kook, L., Götschi, A., Baumann, P., Hothorn, T., Sick, B.: Deep interpretable ensembles. Neurocomputing **682** (Jun 2026). https://doi.org/10.1016/j.neucom.2026.133394, `https://doi.org/10.1016/j.neucom.2026.133394`
11. Liu, Y., Yu, Y., Ouyang, J., Jiang, B., Yang, G., Ostmeier, S., Wintermark, M., Michel, P., Liebeskind, D.S., Lansberg, M.G., Albers, G.W., Zaharchuk, G.: Functional outcome prediction in acute ischemic stroke using a fused imaging and clinical deep learning model. Stroke **54**(9) (2023). https://doi.org/10.1161/STROKEAHA.123.044072, `https://doi.org/10.1161/STROKEAHA.123.044072`, originally published 24 July 2023
12. Oliveira, G., Fonseca, A., Ferro, J., Oliveira, A.: Deep learning-based extraction of biomarkers for the prediction of the functional outcome of ischemic stroke patients. Diagnostics **13**(24), 3604 (2023). https://doi.org/10.3390/diagnostics13243604, `https://doi.org/10.3390/diagnostics13243604`
13. Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H., Aerts, H.J.W.L.: Vision foundation models for computed tomography (2025), `https://arxiv.org/abs/2501.09001`
14. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**, 206–215 (2019)
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
16. van der Velden, B.H.M., Kuijf, H.J., Gilhuijs, K.G.A., Viergever, M.A.: Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis **79** (Jul 2022). https://doi.org/10.1016/j.media.2022.102470
17. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)