ExpLIMEable: An exploratory framework for LIME

Sonia Laguna[*] ETH Zurich
Julian N. Heidenreich[*] ETH Zurich
Jiugeng Sun [*] ETH Zurich
Nilüfer Cetin [*] ETH Zurich

Ibrahim Al-Hazwani University of Zurich
Udo Schlegel University of Konstanz
Furui Cheng ETH Zurich
Mennatallah El-Assady ETH Zurich

Abstract

ExpLIMEable is a tool to enhance the comprehension of Local Interpretable Model-Agnostic Explanations (LIME), particularly within the realm of medical image analysis. LIME explanations often lack robustness due to variances in perturbation techniques and interpretable function choices. Powered by a convolutional neural network for brain MRI tumor classification, *ExpLIMEable* seeks to mitigate these issues. This explainability tool allows users to tailor and explore the explanation space generated post hoc by different LIME parameters to gain deeper insights into the model's decision-making process, its sensitivity, and limitations. We introduce a novel dimension reduction step on the perturbations seeking to find more informative neighborhood spaces and extensive provenance tracking to support the user. This contribution ultimately aims to enhance the robustness of explanations, key in high-risk domains like healthcare.

1 Introduction

The increasing use of machine learning necessitates explainable artificial intelligence (XAI) techniques to bridge the gap between the complexity of algorithms and the need for human comprehensibility, especially in high-stakes applications like healthcare Gunning et al. (2019). Local Interpretable Model-Agnostic Explanations (LIME) Ribeiro et al. (2016) is a post hoc local perturbation-based XAI technique that generates explanations based on the following steps: Identifying interpretable features, i.e., superpixels in the image domain; sampling the neighborhood of the input data by generating perturbations; fitting an interpretable model, to the predictions in the sampled neighborhood and estimating the importance of each feature from the predicted coefficients. Figure 1 shows an example of the described process in brain MRI images, the focus of this work. LIME is widely adopted for explaining complex models, especially in the image and medical domains Garreau & Mardaoui (2021); Ahsan et al. (2021). Yet, its explanations vary with factors like perturbation techniques, leading to research on its stability Visani et al. (2022); Slack et al. (2020).

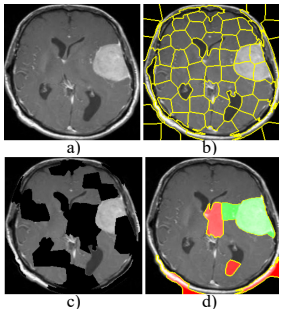

Figure 1: LIME overview: a) original image; b) segmented image; c) perturbed image; d) final explanation (green: adds to classification, red: supports other prediction).

To enhance LIME's understanding, we developed *ExpLIMEable*, an interactive dashboard built around the steps in Figure 1. It aids machine learning developers in assessing LIME's limitations, examining its sensitivity to different parameters, and streamlining prediction explanations in tumor classification. We analyze segmentation approaches and introduce a novel perturbation dimension reduction step to study LIME's robustness.

[*]Equal contribution. Correspondence to {slaguna, jheiden, jiusun, ncetin}@ethz.ch

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

2 Related work

Stability of LIME Research has intensified to assess the robustness of XAI techniques and works have highlighted LIME's challenges, such as sensitivity to input variations and hyper-parameters Alvarez-Melis & Jaakkola (2018); Bansal et al. (2020). Recent extensions aim to address these issues, including S-LIME Zhou et al. (2021), which determines the number of local perturbations required to guarantee stability, and B-LIME Abdullah et al. (2023) that incorporates bootstrap sampling to improve stability and local fidelity, but no definitive solution exists.

Visual diagnostics tools for LIME Given these challenges, works have focused on using visual diagnostics to assess explanations. For instance, Goode & Hofmann (2021) uses feature heatmaps and explanation scatter plots to analyze consistency and fidelity. Moreover, interactivity has been investigated with ExplainExplore Collaris & van Wijk (2020) providing interactive explanations, and other efforts adapt to specific subgroups Wang et al. (2019); Chan et al. (2020). Our work concentrates on the parameters and user interaction, introducing a novel sampling strategy. Unlike similar efforts Saito et al. (2020); Visani et al. (2020), we stay faithful to LIME's original implementation and focus on its inherent steps studying segmentation and dimension reduction. Inspired by the explAIner framework Spinner et al. (2020), we emphasize provenance tracking.

3 Implementation: LIME and tumor classification

Users This application is catered to machine learning experts with a basic understanding of explainability and LIME, specifically, who are interested in investigating its robustness with respect to changing modeling assumptions. With a user-in-the-loop approach, the dashboard allows for exploration, customization, and provenance tracking.

LIME implementation refinements This tool helps explore various configurations of the LIME algorithm interacting with the steps described in section 1. Firstly, the segmentation of the images that generate the superpixels is investigated with algorithms that include Felzenszwalb Felzenszwalb & Huttenlocher (2004), Slic Achanta et al. (2012), Quickshift Vedaldi & Soatto (2008) and Watershed Neubert & Protzel (2014). Secondly, we sample the perturbations by introducing a dimension reduction step that projects the perturbed images and clusters them. We explore three well-known dimensionality reduction techniques: UMAP McInnes et al. (2018), t-SNE Van der Maaten & Hinton (2008), and PCA Ringnér (2008). This way, we aim to select the most informative local perturbations, i.e., those within the cluster of the image under study.

Dataset and classification model We used a Brain Tumor MRI Dataset from Kaggle MRI of 3285 samples with four classes: three types of tumors and healthy brains. Five samples per class were kept as a test set and an 80%-20% train-validation split was used. The model architecture was EfficientNet-B1 Tan & Le (2019) pretrained on ImageNet Russakovsky et al. (2015). Further details of the dataset, pre-processing, and predictive model are provided in Appendix A.

4 ExpLIMEable: Visual analysis workspace

The interactive dashboard consists of two distinct and separate pipelines, "Preselected image" and "External upload", where users progress through steps using a top stepper and can move forward or backward. All throughout, onboarding explanations and ⓘ buttons clarify steps and algorithms and the left panel displays pipeline history. The dashboard uses calming blue tones and optimized colors for accessibility Wong (2011). The developed platform is available at http://b1-dimensionality-reduction-for-lime.course-xai-iml23.isginf.ch/.

Pipeline 1: Preselected images The main pipeline, as laid out in Figure 2, comprises four steps: *Image Selection*, *Explanation*, *Segmentation*, and *Reduction*, with a loop from Reduction to Explanation, initiated by the user. Further and detailed visualizations of the interface described in this section can be found in Appendix C. In the first *Image Selection* step, a machine learning expert chooses one of a set of suggested images and selects "next" in the navigation. Secondly, on the *Explanation* page, the user is invited to a tour of different LIME explanations for MRI classification, which were precomputed for the set of suggested images. Each explanation is

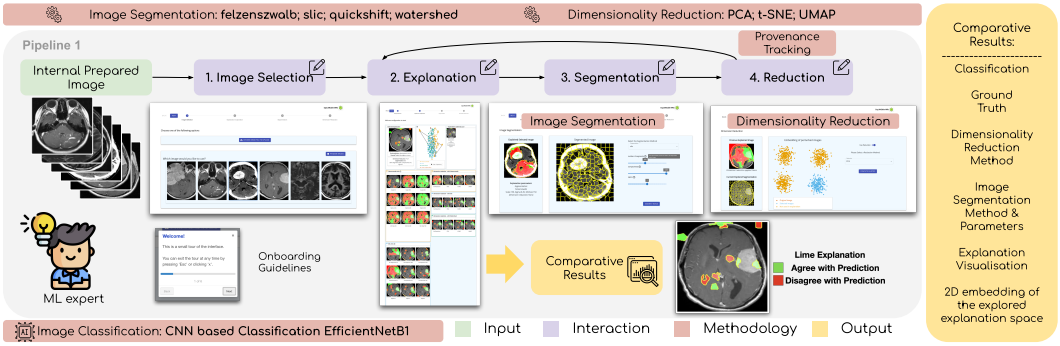

Figure 2: ExpLIMEable workflow, detailed in section 4. Depiction of the four steps in purple: 1. *Image Selection*, 2. *Explanation*, 3. *Segmentation*, 4. *Reduction*, and a, potentially endless, loop back to step 2. The algorithms used and methodology are highlighted in red, the user input in green, and the pipeline outputs in yellow. The ML expert is portrayed as the user of this pipeline and visual examples of the interface at the different steps are included for reference, with further detailed visualizations in Appendix C.

based on different segmentation and dimension reduction methods with adjustable parameters, mentioned in section 3. Explanations are presented in a grid of rows, each allowing one parameter variation, to depict its influence. In the same page, a total of 500 explanations, including those displayed in the interface, are projected with UMAP onto a 2D scatter-plot for enhanced comparison, as seen in Step 2 in Figure 2. Explanations can be selected by the user from the external image grid and a trajectory is generated within this embedding indicating the point from the latest selection. This provenance tracking allows monitoring and navigating the remaining, unexplored explanation space and facilitates an understanding of the explanation's sensitivity toward the parameters, each labeled with a specific color. To prevent overplotting, the trajectory map can be discarded. Users can customize the interface layout with a movable grid, and investigate parameters, while selecting a reference configuration to generate new explanations.

In the next *Segmentation* step, the user chooses a segmentation algorithm, adjusts parameters using sliders, and iterates until they achieve the desired result. Step 3 in Figure 2 shows an example of a segmented MRI with the selection of the segmentation method and parameters. The final step is an optional dimension *Reduction*, our novel methodological contribution. After segmentation, perturbed images are projected with a dimension reduction method chosen by the user, and a scatter plot (Step 4 in Figure 2) visually indicates the embedding space and which perturbation cluster will be used in the subsequent explanation. After dimension reduction, the user loops back to the *Explanation* step with the new computed explanation projected to the scatter plot with the pre-computed examples for comparison. All newly computed explanations are cataloged in a dedicated panel (See Step 2 in Figure 2) for better provenance tracking and guidance. The user can engage in a continuous loop of Explanation, Segmentation, and Reduction to thoroughly explore all possible explanations in the scatter plot space, eventually finding a satisfactory configuration. Overall, the integration of provenance tracking enhances the narrative by preserving a record of past explorations. Users also have the option to redo or change the image returning to image selection or switch to Pipeline 2, the "user upload" branch.

Pipeline 2: External upload The "External upload" branch, an alternative to Pipeline 1, starts with a user image upload. Hereafter, it does not include an exploratory explanation map as pre-computations are not possible. This branch involves *Image Upload Selection*, *Segmentation*, dimension *Reduction*, and a final *Explanation* page. This pipeline can be used by previous users of Pipeline 1 to try tailored parameter configurations on their own data samples and by a second user group, application domain experts, corresponding to clinicians in this medical data scenario, to obtain individual explanations. The latter group uses the default setting to obtain the explanations for simplicity. Further details of this secondary pipeline can be found in Appendix B, visualized in Figure 3. In using this tool, there exists the potential for testing samples that fall outside the distribution of data used for training. It is important for the user to exercise caution and remain aware of this possibility. Note that the current deployed version using the provided link is a preliminary prototype and only Pipeline 1 works reliably.

5 Discussion and conclusion

In this study, we introduce a tool to enhance LIME's understanding, despite its inherent complexity. Our efforts have been directed toward enhancing user experience, with improvements in visual aesthetics, the incorporation of distinctive color schemes, and the integration of features such as multiple provenance tracking paths and onboarding guides. We argue that incorporating a potentially endless loop within the interface yields substantial advantages, enabling users to navigate the system more effectively to understand the explanation space. The introduction of dimension reduction of the perturbations represents a novel contribution for controlled sampling. This platform holds promise for machine learning experts to enhance their understanding of explainability methods. Given the fast-paced deployment of models in everyday tasks and the growing regulatory requirements, we regard our platform as a valuable asset. Particularly, this tool tackles a medical imaging problem, a high-risk domain where explainability is key. Overall, we believe that the platform will provide clearer results when explaining robust models, as they have been shown to highlight discriminate features better Shah et al. (2021).

Limitations The main pipeline of this tool, while promising for clinical decision support, necessitates prior machine learning expertise to fully harness its potential, and therefore only machine learning experts are intended to use it. This limitation is surpassed by Pipeline 2, where a user can directly generate an explanation. Hereafter, this extends the potential users to clinicians but a simplification of the platform would be required to make it fully translatable to the clinics. Currently, quantitatively evaluating XAI methods remains challenging due to a lack of consensus and standardized metrics. Presently, we do not provide a quantitative metric for explanation comparisons, which is desired to assess the robustness and the efficacy of dimensionality reduction. In addressing these challenges, existing frameworks for classification evaluation can serve as foundational building blocks. For instance, comparing predictions after removing the most influential features based on different XAI methods Hooker et al. (2019); Rong et al. (2022) provides a viable approach and our next efforts will tackle this problem.

Future work As for future work, our aspirations encompass expanding the interactivity during the dimension reduction phase and conducting comprehensive user studies with questionnaires and surveys to gain valuable insights into user interactions with explanations and the use of the tool. Moreover, extending the pre-computed dataset will facilitate generalization of Pipeline 1. Additionally, implementing parameter exploration for Pipeline 2 is a next step, offering users full autonomy in their exploratory endeavors. However, this is accompanied by the challenge of addressing out-of-domain scenarios, which may need retraining the model with the appropriate distribution. While the current dashboard predominantly focuses on image data, it can be adapted to other data types, such as tabular data or signals. The segmentation step would be substituted with alternative perturbation methods, such as introducing random noise or removing data points, while the dimension reduction step would remain unchanged. Visualizations for other data types would be replaced with tables or plots. Lastly, our focus has been on LIME as one of the most prevalent XAI methods. However, we posit that a similar platform could be extended to other methods, such as SHAP Lundberg & Lee (2017), which considers all possible feature combinations for attribution. We could accommodate its most used variation, KernelSHAP Lundberg & Lee (2017), equivalent to weighted linear regression in LIME with a specific kernel. This extension would enable the investigation of perturbation effects and dimension reduction within our framework, with the optimization tailored to kernel-weighted linear regression. Moreover, using this framework with GradCAM Selvaraju et al. (2017) or other gradient based approaches widely used for imaging data is also possible to study their variability. However, we focused on LIME as the control of its intermediate steps is a more relevant contribution to understanding XAI, as gradient based approaches lack these steps.

Conclusion In summary, the growing reliance on machine learning in decision-making, particularly in critical domains like healthcare, necessitates the development of XAI. One widely discussed explainable technique is LIME, which is model-agnostic and provides locally interpretable explanations. However, LIME is not robust and its explanations depend on various parameters and perturbation methods. To tackle this challenge, we propose a tool designed to enhance comprehension of LIME and allow users to explore diverse explanations to understand which parameters the method is most sensitive or robust to. Our focus lies in the specific context of

brain MRI tumor classification, providing a structured workflow for machine learning developers to gain insights into model behavior. We introduce a novel approach to sampling in the perturbation space. Applying dimensionality reduction to such perturbations, we aim to select more reliable instances for a more robust explanation. In essence, *ExpLIMEable* facilitates the exploration of explanations generated by different LIME instances, providing a practical solution to shed light on system limitations, sensitivity, and robustness with a coherent narrative with provenance tracking.

Acknowledgments and disclosure of funding SL is supported by the Swiss State Secretariat for Education, Research, and Innovation (SERI) under contract number MB22.00047. ME is supported by the ETH AI Center.

References

`https://github.com/Ashish-Arya-CS/Coursera-Content`.

Abdullah, T. A. A., Zahid, M. S. M., Ali, W., and Hassan, S. U. B-LIME: An Improvement of LIME for Interpretable Deep Learning Classification of Cardiac Arrhythmia from ECG Signals. *Processes*, 11(2), 2023. doi: $10.3390/\mathrm{pr11020595}$.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Ahsan, M. M., Nazim, R., Siddique, Z., and Huebner, P. Detection of covid-19 patients from ct scan and chest x-ray data using modified mobilenetv2 and lime. In *Healthcare*, volume 9, pp. 1099. MDPI, 2021.

Alvarez-Melis, D. and Jaakkola, T. S. On the Robustness of Interpretability Methods. In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, 2018.

Bansal, N., Agarwal, C., and Nguyen, A. SAM: The Sensitivity of Attribution Methods to Hyperparameters. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8670–8680, 2020. doi: $10.1109/\mathrm{CVPR42600.2020.00870}$.

Chan, G. Y.-Y., Yuan, J., Overton, K., Barr, B., Rees, K., Nonato, L. G., Bertini, E., and Silva, C. T. Subplex: Towards a better understanding of black box model explanations at the subpopulation level. *arXiv preprint arXiv:2007.10609*, 2020.

Collaris, D. and van Wijk, J. J. Explainexplore: Visual exploration of machine learning explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 26–35. IEEE, 2020.

Felzenszwalb, P. F. and Huttenlocher, D. P. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004.

Garreau, D. and Mardaoui, D. What does lime really see in images? In *International conference on machine learning*, pp. 3620–3629. PMLR, 2021.

Goode, K. and Hofmann, H. Visual diagnostics of an explainer model: Tools for the assessment of lime explanations. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(2):185–200, 2021.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. XAI-Explainable artificial intelligence. *Science Robotics*, 4(37), 2019. doi: $10.1126/\mathrm{scirobotics.aay7120}$. URL `https://www.science.org/doi/abs/10.1126/scirobotics.aay7120`.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Lundberg, S. M. and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, 2017.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Neubert, P. and Protzel, P. Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In *2014 22nd international conference on pattern recognition*, pp. 996–1001. IEEE, 2014.

Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Ringnér, M. What is principal component analysis? *Nature biotechnology*, 26(3):303–304, 2008.

Rong, Y., Leemann, T., Borisov, V., Kasneci, G., and Kasneci, E. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*, 2022.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Saito, S., Chua, E., Capel, N., and Hu, R. Improving lime robustness with smarter locality sampling. *arXiv preprint arXiv:2006.12302*, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and 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.

Shah, H., Jain, P., and Netrapalli, P. Do input gradients highlight discriminative features? *Advances in Neural Information Processing Systems*, 34:2046–2059, 2021.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Spinner, T., Schlegel, U., Schäfer, H., and El-Assady, M. explainer: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1064–1074, 2020. doi: 10.1109/TVCG.2019.2934629.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for Convolutional Neural Networks. *Proceedings of the ICML 2019*, pp. 6105–6114, 2019. doi: 10.48550/arXiv.1905.11946.

Van der Maaten, L. and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

Vedaldi, A. and Soatto, S. Quick shift and kernel methods for mode seeking. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*, pp. 705–718. Springer, 2008.

Visani, G., Bagli, E., and Chesani, F. Optilime: Optimized lime explanations for diagnostic computer algorithms. *arXiv preprint arXiv:2006.05714*, 2020.

Visani, G., Bagli, E., and Chesani, F. Optilime: Optimized lime explanations for diagnostic computer algorithms, 2022.

Wang, J., Gou, L., Zhang, W., Yang, H., and Shen, H.-W. Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *IEEE transactions on visualization and computer graphics*, 25(6):2168–2180, 2019.

Wong, B. Points of view: Color blindness. *Nature Methods*, 8(6), 2011.

Zhou, Z., Hooker, G., and Wang, F. S-LIME: Stabilized-LIME for Model Explanation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 2429–2438, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3447548.3467274.

.

Appendix

A Implementation details

Dataset and pre-processing The dataset consists of 2D MRI frames for three types of tumors: glioma, meningioma, and pituitary, as well as MRIs of healthy brains, with 931, 942, 906, and 506 images respectively. During the pre-processing of 2D frames of brains, we have filtered duplicate images and detected the outline of the skull with the Python library OpenCV. The images have been cropped based on this contour and then resized to 240x240 pixels. We used 80% of the images for training the ML model and the other 20% for validation. Finally, the test set comprises five images per class. The example image in Figure 1 is part of the test set.

Classification model The predictive model we use in the backend is based on the EfficientNet-B1 architecture Tan & Le (2019). The model was pre-trained on ImageNet Russakovsky et al. (2015) and later fine-tuned for 30 epochs on our dataset using TensorFlow. The final model achieves a balanced accuracy of 97.9% on the validation dataset and all test images are classified correctly. In order to ensure meaningful interactions with the visual interface, the predictive model has to be reasonably accurate to yield sound explanations in combination with LIME.

B Pipeline 2: External Upload

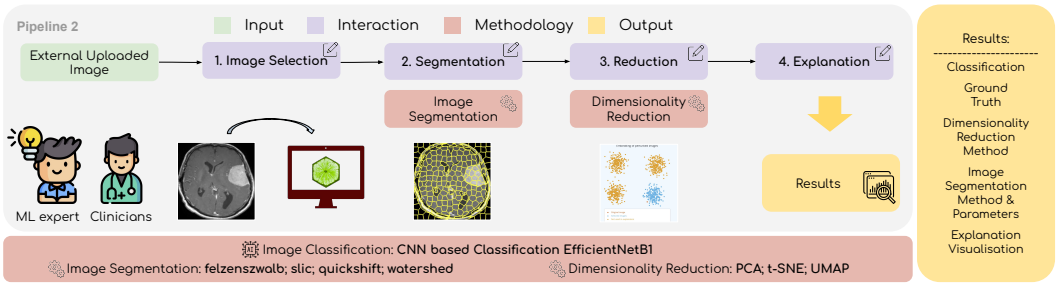

Figure 3: ExpLIMEable workflow, detailed in Section 4. Visualization of the four steps of the interactive workflow in purple: 1. *Image Upload Selection*, 2. *Segmentation*, 3. *Reduction*, and 4. *Explanation*. The methodology used is highlighted in red, user input in green, and outputs in yellow. The machine learning expert and the clinicians as application domain experts are portrayed as the users of this pipeline.

This secondary pipeline is envisioned as a tool for a user to directly generate an explanation of an uploaded image using either default or customized settings. The "External upload" branch starts with the user uploading their own image. This branch is an extension to the main Pipeline 1 (See Section 4) and does not include the exploratory map. The workflow in this branch consists of *Image Selection*, *Segmentation*, dimension *Reduction*, and *Explanation*, as depicted in Figure 3). In this branch there are no pre-computed explanations as user-uploaded data cannot be anticipated in advance, hence there is no comparative map among different parameters. Hereafter, a clinician can make use of this simplified Pipeline 2 in the default setting to acquire an explanation for a prediction made by a machine learning model. Moreover, focusing on the main user group, machine learning experts, we enable users to compute new explanations using the different segmentation techniques and dimension reduction approaches proposed in section 3, granting them the freedom to explore. For now, we do not perform any sanity checks concerning the validity of the uploaded data and consider it the user's responsibility to upload sensible data close to the model's original training data. The user can walk through and explore the steps of segmentation and dimension reduction and investigate the respective influence by reviewing their final explanations, displayed as in Figure 8 in Appendix C. The user can upload new images to continue to explore. This allows the user to verify, on their own data, findings, and results from interacting with the main branch on pre-computed samples. In future versions, we plan to compute an explanation overview in the same manner as the current main branch. This could be done by simply informing the user about the long waiting time, and allowing them to switch

tasks or work in parallel sessions until the computation is complete. Once finalized, this branch would be able to guide the user through the same exploration as the main branch. Note that the current deployed version using the provided link is a preliminary prototype and only Pipeline 1 works reliably.

C ExpLIMEable Interface

In this chapter, we include a detailed overview of the visual interface of *ExpLIMEable*, to clarify the described steps in Section 4 and the inclusion of provenance discussed throughout the manuscript. Firstly, Figure 4 shows the first step of both pipelines, *Image Selection* and the components within this page, such as the stepper bar, that is present throughout the workflow. From here, the user can upload an image to proceed to Pipeline 2 or select an image out of a suggested set to get started with Pipeline 1. Secondly, Figure 5 shows the *Segmentation* step and the range of parameters and algorithms that can be studied. On the left, a panel with the selected explanation to be used as a reference in Pipeline 1 is displayed for user guidance. Subsequently, Figure 6 visualizes the dimension *Reduction* step, the possible configurations that can be studied, and the resulting embedding space. Recall that each point corresponds to a perturbation and those belonging to the cluster of the original image are the ones used in the subsequent explanation. As a final step, Figure 7 shows how the *Explanation* page looks like in Pipeline 1. The components for provenance tracking are highlighted and the scatter-plot with the embedding space of the explanation is displayed with a sample trajectory that results from exploring multiple explanations. Finally, an example of the grid of pre-computed explanations is displayed together with the newly generated explanation. Note that the disposition of this page includes a movable grid that the user can rearrange at any point in time to aid their mental model. Finally, Figure 8 displays the *Explanation* step in Pipeline 2 where the newly generated explanation is shown along with the final prediction and parameter configuration. Recall that this second pipeline does not include an explanation exploration because new user uploaded data is used at all times.

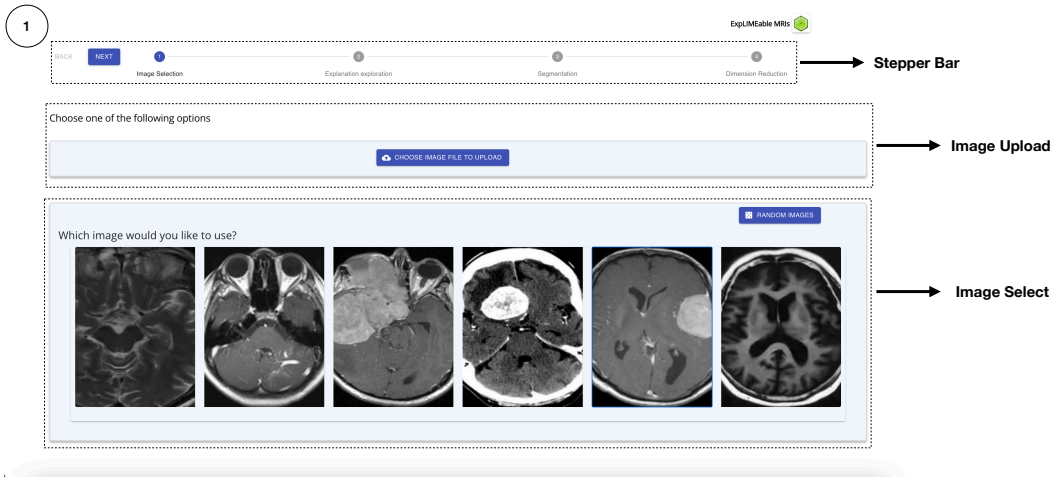

Figure 4: Overview of the interactive components (section 4): ① *Image Selection*: Including the stepper to guide through the pipeline and entrance point to both pipelines.

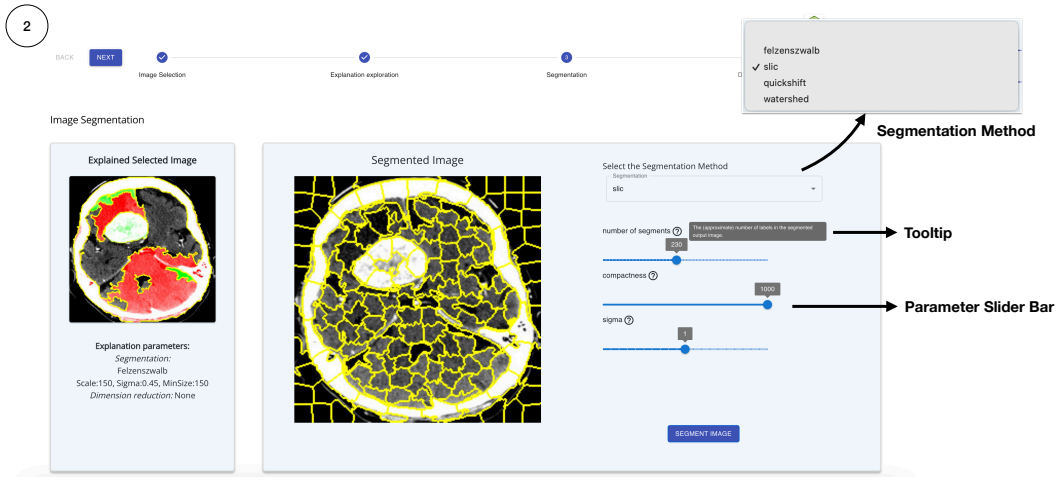

Figure 5: Overview of the interactive components (section 4): ② *Segmentation*: Method selection and resulting segmentation that will generate the perturbations.

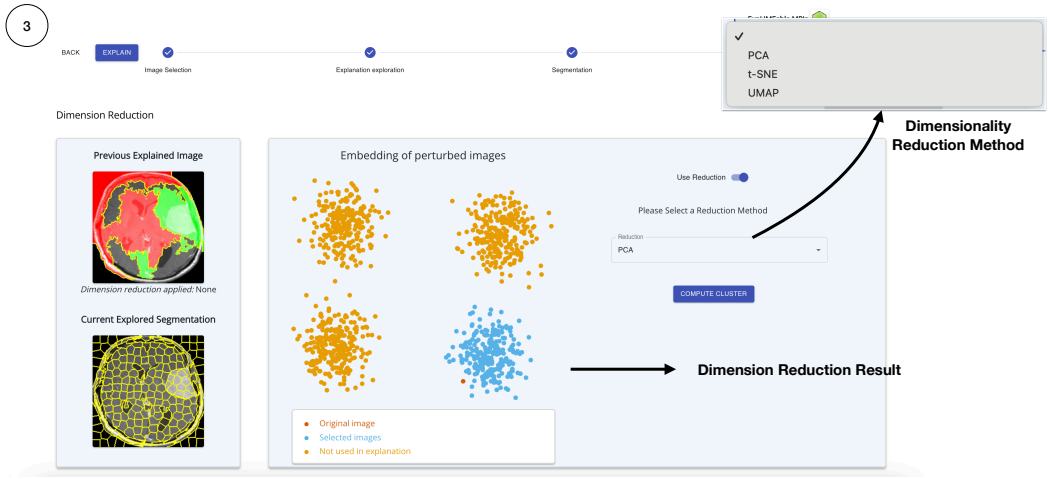

Figure 6: Overview of the interactive components (section 4): ③ *Reduction*: Method selection and clustered embeddings of the local perturbations.

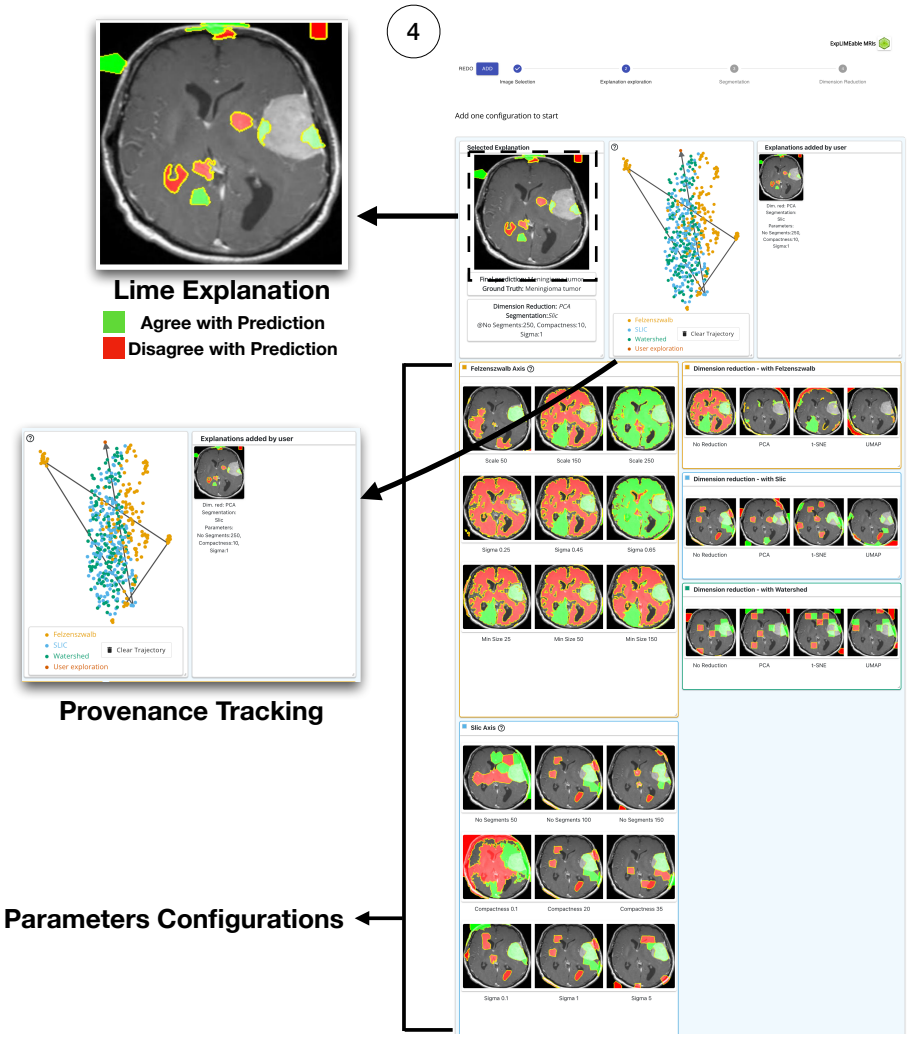

Figure 7: Overview of the interactive components (section 4): ④ *Explanation* page of Pipeline 1: Model prediction, pre-computed explanations to explore parameters, the 2D embedding of the explanations with user trajectory for provenance tracking and user-added explanation.

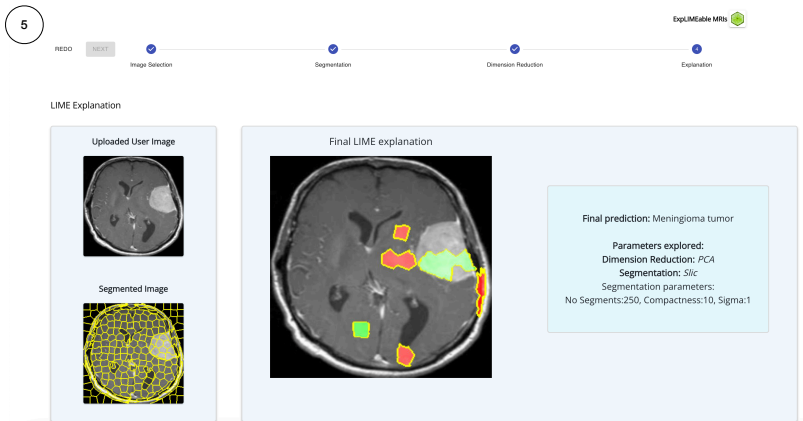

Figure 8: Overview of the interactive components (section 4): ⑤ Explanation page of Pipeline 2.