ExpLIMEable: A Visual Analytics Approach for Exploring 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

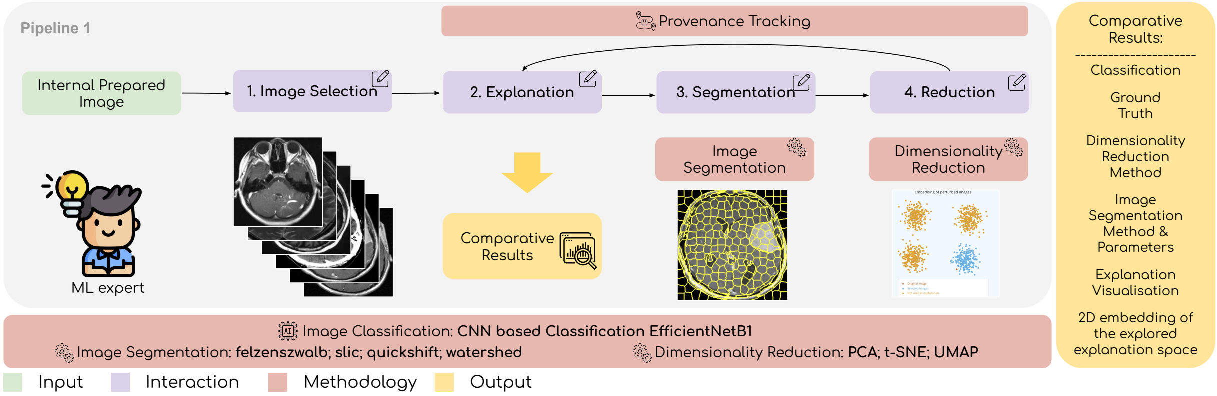
Pipeline 1
Provenance Tracking
Internal Prepared Image
1. Image Selection
2. Explanation
3. Segmentation
4. Reduction
ML expert
Comparative Results
Image Segmentation
Dimensionality Reduction
Comparative Results:
Classification
Ground Truth
Dimensionality Reduction Method
Image Segmentation Method & Parameters
Explanation Visualisation
2D embedding of the explored explanation space
Image Classification: CNN based Classification EfficientNetB1
Image Segmentation: felzenszwalb; slic; quickshift; watershed
Dimensionality Reduction: PCA; t-SNE; UMAP
Input
Interaction
Methodology
Output

Figure 1: Main pipeline of *ExpLIMEable*, detailed in Sect. 5.1. Including the four steps of the interactive workflow in purple: 1. image selection, 2. explanation, 3. segmentation, and 4. reduction. The methodology is highlighted in red, the user input in green, and the corresponding pipeline outputs in yellow. The ML expert is portrayed as the user of this pipeline.

Abstract

We introduce *ExpLIMEable* for enhancing the understanding of Local Interpretable Model-Agnostic Explanations (LIME), with a focus on medical image analysis. LIME is a popular and widely used method in explainable artificial intelligence (XAI) that provides locally faithful and interpretable post-hoc explanations for black box models. However, LIME explanations are not always robust due to variations in perturbation techniques and the selection of interpretable functions. The proposed visual analytics application aims to address these concerns by enabling the users to freely explore and compare the explanations generated by different LIME parameter instances. The application utilizes a convolutional neural network (CNN) for brain MRI tumor classification and allows users to customize post-hoc LIME parameters to gain insights into the model's decision-making process. The developed application assists machine learning developers in understanding the limitations of LIME and its sensitivity to different parameters, as well as the doctors in providing an explanation to machine learning models, enabling more informed decision-making, with the ultimate goal of improving its robustness and explanation quality.

Index Terms: Explainable AI—Visualization—LIME—Healthcare

1 Introduction

With the increasing use of machine learning (ML) methods for decision-making and problem-solving, there is a growing need for interpretable and explainable ML techniques. These techniques emerge from a necessity to design intelligible ML systems comprehensible to humans and provide explanations for predictions from opaque models, i.e. CNNs. The terms interpretability and explainability are still commonly used interchangeably in the literature [22]. However, we would like to make the following distinction: Interpretable ML focuses on developing models that are inherently interpretable, while explainable artificial intelligence (XAI) aims to provide explanations for existing black-box models [19, 22]. These research areas aim to establish trust, causality, fairness, and privacy to facilitate a comprehensive understanding of models' decision process and mitigate biases and errors [19, 22, 30]. However, the lack of consensus among different XAI methods creates confusion for users, as the explanations are typically method-dependant [24].

One widely used technique in the field of explainability is the Local Interpretable Model-Agnostic Explanations (LIME) algorithm [26]. It is a widely-used XAI method that has been extensively discussed in the literature and that finds rich application to image data [12], including medical ones [4]. LIME is a post-hoc method that provides local explanations using human-interpretable representations, even for complex models that are otherwise non-intelligible to humans. However, LIME itself is not always robust, as the explanations can depend on factors such as the perturbation techniques or the type of interpretable function employed. Extensive research has been conducted to examine the stability of LIME and explore pipelines that enhance our comprehension of its behavior [39], as well as study its biases [32].

In high-stakes applications like healthcare, where machine learning plays a crucial role [18], explanations of ML models' decisions become vital due to the associated risks. Given the wide usage of LIME across high-risk disciplines and its lack of robustness and consensus, we propose *ExpLIMEable* to enhance the understanding of LIME. For the design of the interactive visual interface, we draw inspiration from previous works dedicated to improving the understanding of explanations [33]. Specifically, we introduce *ExpLIMEable* to address a medical image analysis challenge, brain MRI tumor classification. Within this context, we aim to provide a workflow that allows users to freely explore various explanations generated by different LIME parameter instances for the predictions of a large, black-box CNN model. *ExpLIMEable* is designed to serve two distinct purposes. Firstly, it supports machine learning developers in comprehending the limitations of LIME, as well as its sensitivity and robustness to various parameters. This understanding can guide developers toward configuring LIME in a more robust manner. Secondly, it supports clinicians in their decision-making process by providing an explanation for machine learning predictions.

[*] e-mail: {slaguna, jheiden, jiusun, ncetin}@ethz.ch

[†] Equal contribution

The main contribution of this work is the development of an interactive framework that allows machine learning experts to explore and understand the pitfalls and robustness of LIME in healthcare applications, particularly, in MRI brain tumor classification. Explainability holds great importance in healthcare, as the predictions of ML models heavily influence medical decisions. We extensively analyze various segmentation approaches within LIME and introduce a novel dimension reduction step to the local perturbations of LIME to evaluate the impact on the robustness of the method. Additionally, we provide a simplified pipeline for clinicians to obtain an explanation for their machine learning prediction.

2 Background

2.1 Explainable AI

Explainable AI refers to the field of research focused on developing AI systems that can provide understandable and transparent explanations for their decisions [14]. It aims to bridge the gap between the complexity of AI algorithms and the need for human comprehensibility. Various methods have been proposed in the literature to achieve explainability. Rule-based approaches, such as decision trees, generate explicit rules that can be easily interpreted. Model-agnostic techniques, such as LIME [26] and SHAP [20], approximate the behavior of black-box models by assigning feature importance scores, enabling post-hoc explanations [15]. Gradient-based methods measure how much each feature contributes to the final prediction based on their gradient. Finally, counterfactual explanations, measure how much individual feature values would have to be altered to flip the overall model's prediction [19, 22, 30]. Specifically in the healthcare domain, XAI has found applications in diagnostic systems, where it can provide clinicians with explanations for their predictions, aiding in decision-making and fostering trust [21,35]. Despite their benefits, XAI models have limitations. The pursuit of comprehensive explanations leads to a trade-off between accuracy and interpretability. At the same time, the lack of standardized evaluation metrics and consistent definitions of explainability hampers unified frameworks and poses additional challenges. Additionally, ensuring that explanations are tailored to different user backgrounds and contexts remains an ongoing research area as there is wide variability between explanation methods. Nonetheless, XAI holds the potential to enhance human-AI collaboration, increase trustworthiness, and facilitate the adoption of AI systems in critical domains. Numerous tools have been developed to take care of visualization approaches to represent explained data and architectures and improve the interactions with XAI users. For a comprehensive overview of visualization solutions to XAI methods, we refer the reader to [5] and [17].

2.2 Local Interpretable Model-agnostic Explanations

2.2.1 LIME algorithm

LIME is a perturbation-based local XAI technique that explains the behavior of a predictive model for a given input image [26]. Thus, the technique is model-agnostic but data-dependent. LIME works based on interpretable representations of the input, e.g. features/columns in tabular data. Likewise, for image classification, it is common practice to use superpixels, larger image patches of similar pixels, as interpretable features. LIME consists of the following steps to obtain the explanations for a target model:

1 *Identify interpretable features*: For a given input image Fig. 2 a) the superpixels are computed via segmentation (compare Fig. 2 b)).

2 *Sampling for Local Exploration*: The neighborhood of the input image is sampled and perturbation of the original image is generated. There exist multiple alternatives for this generation and in this work, we will focus on zero replacement, i.e. for each perturbed image the color of randomly drawn superpixels is turned to black (Fig. 2 c)).

3 *Approximate target model*: For each perturbed image, the output of the target model is computed. Later, an interpretable model is fitted locally on these samples. In this work, we fit a ridge regression model to approximate these outputs.

4 *LIME explanations*: Based on the coefficients of the locally fitted interpretable model, the importance of each superpixel is estimated for the prediction of the model. In Fig. 2 d) the green regions have the strongest positive influence on the prediction. Red regions on the other hand support a different classification.

2.2.2 LIME sensitivity

The steps described in Sect. 2.2.1 are sensitive to the design choices and can introduce biases at different levels. In step 1, the choice of segmentation will influence the final superpixels, hence, the space of features to explain the method. e.g., if a superpixel covers a large region, sub-regions that potentially have different semantic meanings are therefore indistinguishable. The choice of superpixel replacement in step 2 will affect the final explanation as the generated local perturbations will be at different distances in space from the image of interest. At the same time, the effect of certain replacement methods might differ when treating almost uniform superpixels compared to superpixels with largely heterogeneous textures. Additionally, if perturbations are sampled uniformly, possible correlations between features are disregarded. Finally, step 3 assumes that there exists a linear model that can fit the local decision boundary, which can bias the final performance of the explanation method.

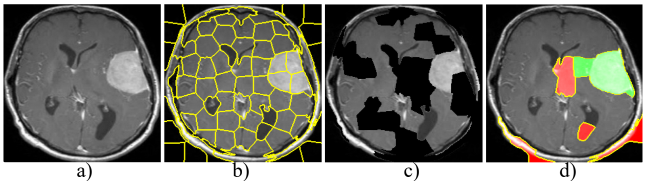

Figure 2: LIME explanation: a) input image; b) segmented image; c) example of a perturbed image in the local neighborhood; d) final explanation (green segments: contribute towards the classification, red segments: support a different prediction)

3 Related work

There are numerous examples in literature that explore the use of XAI techniques in an interactive fashion, focusing on providing explanations to machine learning models or on understanding the explanation methods themselves. For example, the explAIner framework [33], provides a comprehensive pipeline for model understanding, diagnosis, and refinement. It incorporates global monitoring and steering mechanisms, such as provenance tracking and reporting, to enhance trust and confidence in the explanations. The system includes various explainers, both low- and high-abstraction, to cater to different user groups. Through its iterative workflow, this framework enables users to gain insights into the models, diagnose issues, and refine them effectively using a wide range of XAI techniques.

In recent years, there has been a significant amount of effort dedicated to investigating the robustness of XAI techniques. Many works report the lack or insufficient coherence in these methods. For example, Alvarez-Melis and Jaakkola [6] highlight the negative effect of small input variations on the stability of the LIME explanations. On the same notion, Bansal et al. [7] report a high sensitivity of LIME to its hyperparameters. Finally, Dieber and Kirrane [10] mention a lack of global interpretability of local LIME explanations.

Based on these findings and criticism, various algorithmic changes to LIME have been proposed. Some of the most recent extensions include S-LIME [42] that aims at determining the number of local perturbations required to guarantee stability and B-LIME [2] that incorporates bootstrap sampling in an effort to improve stability and local fidelity. Despite the vast array of new ideas and extensions, no comprehensive solution has yet been found.

As many open challenges continue to stay unanswered and many relationships remain unintelligible to humans, recent works advocate the use of visual diagnostics to assess XAI explanations. For instance, Goode and Hofmann [13] develop and discuss various visualizations, including feature heatmaps, explanation scatter plots, and assessment metric plots, to analyze the consistency in LIME explanations across different observations and investigate the fidelity of the local model approximation. Based on these visual diagnostic tools Goode and Hofmann [13] encourage further research to improve the effectiveness and reliability of LIME as an explanatory method, which will be the focus of our study. Related work has focused on interacting with LIME. ExplainExplore [9] revised the LIME method by utilizing both linear models and shallow tree-based models as surrogate models. It enables local explanations for specific instances through interactive display of generated samples and direct manipulations. Moreover, [40] enhanced the process of generating training samples by utilizing deep generative networks, and [8] modified LIME to calculate average feature contributions for providing explanations specifically tailored to a selected subgroup.

In our pipeline, we focus on different parameters of the LIME algorithm and the user interaction with these, including a novel sampling strategy. Previous works have explored sampling alternatives before, such as [31] which proposes a generative adversarial network to generate the sampled perturbations for the locally fitted model for explanations. Following a similar line, Visania et al. [38] propose OptiLIME, a framework that provides freedom to choose the best adherence-stability trade-off level, basing the sampling on geometrical properties and a new optimization scheme on tabular data. However, we focus on the original implementation of LIME and its application to medical data, extending the OptiLIME idea of assessing the robustness, but based on different segmentation and dimension reduction steps, without modifying the overall optimization scheme and staying faithful to LIME. Overall, we address a similar problem as previous works trying to find a stable LIME configuration. However, we provide the user with an interactive framework with an enhanced understanding of the method's parameters and provenance tracking for better comprehension, following a similar line as the explAIner framework [33], in our case targeting solely LIME to solve healthcare applications.

4 Tumor Classification for Decision Support

4.1 Users and Tasks

The application is catered to machine learning experts with a basic understanding of LIME who are interested in investigating its sensitivity to different modeling assumptions and who are keen on improving the robustness of LIME explanations.

The *ExpLIMEable* pipeline offers a flexible visual interface that allows extensive exploration, analysis, and innovation. Thereby, it combines the central notions of the explAIner framework [33]: understanding, diagnosis, and refinement. In particular, the dashboard allows for:

1 **Exploration**: The user can analyze different explanation results with a comprehensive overview and comparative dashboard.

2 **Customization**: The user has the freedom to include new explanations based on user-desired configurations and modify modeling assumptions to steer explanations.

3 **Provenance tracking**: The user can keep track of the exploration of parameters and of all explanations in time.

We address the challenges of LIME and the parameters by keeping humans in the loop, with a more comprehensive user scenario described in Sect. 6.

4.2 MRI data, processing and guidance

The dataset used in the proposed XAI pipeline is the Brain Tumor MRI Dataset from Kaggle, which was also part of a Coursera course [1]. The dataset includes 2D MRI frames for three types of tumors: glioma, meningioma, and pituitary. It includes MRIs of healthy brains, with 931, 942, 906, and 506 images, respectively.

During the pre-processing phase, we filter duplicate images and detect the outline of the skull with the Python OpenCV library. We crop the images based on this contour and resize them to 240x240 pixels. We use 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 2 is part of the test set.

We pre-compute selected test cases to facilitate a seamless user experience and responsive explorations. These computations include segmentations of MRI images, lower dimensional embeddings of the perturbed images, and corresponding LIME explanations. However, due to resource constraints, we limited the pre-computations to six images only: three meningioma tumors and three examples of healthy brains. For each image, we computed a total of 36 different explanations using LIME, which will be shown in the dashboard to serve as user's guidance for parameter exploration. Furthermore, we pre-compute a total amount of 500 parameter combinations per image, which are mapped to a lower dimensional embedding using UMAP [23] to provide an overview of possible explanations. Overall, this structure allows the user to freely interact with the data and generate new examples based on their own input. The designed interface improves user guidance by providing a flexible dashboard with extensive data visualization and provenance tracking, detailed in Sect. 5.2.

4.3 Machine learning model

The predictive model we use in the backend is based on the EfficientNet-B1 architecture [34]. The model was pre-trained on ImageNet [29] 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.

4.4 LIME implementation refinements

In this work, we explore various configurations of the LIME algorithm based on the four steps described in Sect. 2.2.1. Firstly in Step 1, we investigate different segmentation algorithms and their influence on the resulting superpixel geometries to uncover the potential biases that the segmentation introduces. These algorithms include: Felzenszwalb [11], Slic [3], Quickshift [37] and Watershed [25]. Secondly, between steps 2 and 3, we introduce a novel dimension reduction approach to select specific perturbations of LIME. Here, all perturbed images are transformed to a lower dimensional space and clustered based on embedding distance. Later, only samples in the cluster of the original image are chosen to carry out the explanations in Step 3. This way, we aim to select the most informative local perturbations for a more stable, knowledgeable explanation. We explore three well-known dimension reduction techniques: UMAP [23], t-SNE [36], and PCA [27]. With this approach, we avoid a uniform sample selection and hinder potential biases in the correlation of the data.

5 Visual Analysis Workspace

The developed platform is available at http://b1-dimensionality-reduction-for-lime.course-xai-iml23.isginf.ch/.

5.1 Layout of the Interactive Dashboard

The interactive dashboard consists of two distinct and separate pipelines. The so-called ”Preselected image” and an ”External up-

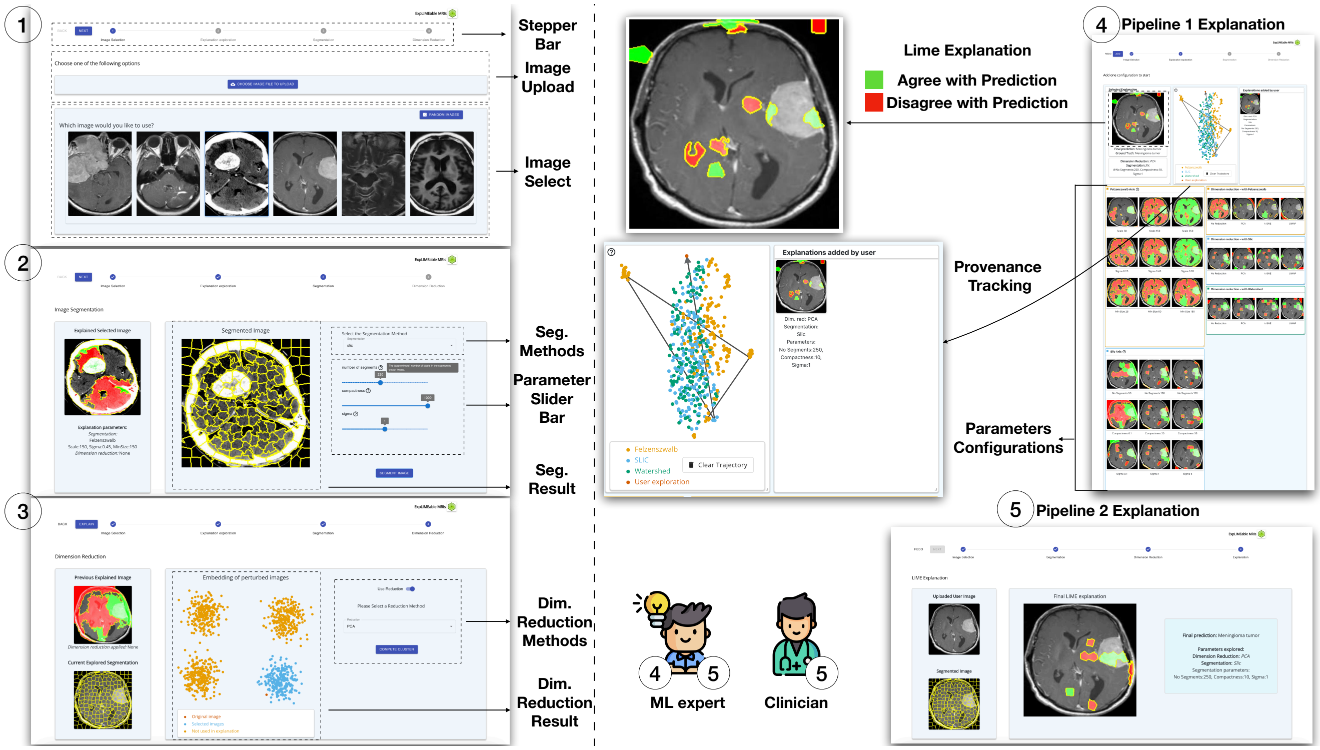

Figure 3: Overview of the interactive components (Sect. 5): ① Image selection: including the stepper to guide through the pipeline and entrance point to both pipelines, ② Segmentation: method selection and resulting segmentation, ③ Reduction: method selection and clustered embeddings of the local perturbations, ④ Exploration page: Model prediction, pre-computed explanations to explore parameters, the 2D embedding of the explanations with user trajectory for provenance tracking and user-added explanation, ⑤ Explanation page of external upload branch.

load" branch. In both scenarios, users can navigate through various stages, and at the top of the page, there is a stepper that serves as a visual guide indicating past, present, and future steps. Throughout the dashboard, there are two buttons to the left of the stepper that allow to navigate forward or backward within the pipeline. Fig. 3 ① shows the stepper as well as the entrance points for both pipelines. For all phases of the pipeline, an initial onboarding setup introduces the user to the functionalities of the page and provides brief explanations of each step. Additional ⓘ buttons are scattered throughout the interface to explain selected algorithms. In all intermediate steps, the panel on the left side shows the history of the current pipeline, e.g. the currently chosen MRI and its segmentation (compare Fig. 3 ② or ③ left). Please note that the background colors throughout the dashboard are within a range of calming and desaturated blue tones to avoid an over-saturation of the interface. This choice is made considering the interface's inherent complexity. Other colors are selected based on an optimized color palette by Wong [41] considering color vision deficiency.

5.1.1 Pipeline 1: Preselected images

This first branch is the main pipeline and it includes four different steps: image selection, explanation, segmentation, and reduction, with a potentially endless loop from reduction back to explanation, as laid out in Fig. 1. The user, an ML expert, starts with image selection, which currently allows choosing one out of six pre-processed images, and selects "next" in the navigation panel.

Explanation Exploration

On the Explanation page, the user is invited to an exploration of different LIME explanations for the classification of the selected MRI. The explanations are based on different combinations of segmentation and dimension reduction methods coupled with different parameter settings, described in Sect. 4.4. The display is arranged in different rows, also called axes. Each of these rows varies only one parameter at a time to depict its influence on the final explanations. Fig. 3 ④ shows the segmentation axes on the left and the axes for the dimension reduction methods on the right. The user can freely select and investigate each explanation. In order to improve the navigation through the numerous provided examples, each explanation is mapped with UMAP from the image space onto a point in the 2D space and summarized in a scatter-plot (see Fig. 3 ④ left). For more details please refer to the provenance tracking in Sect. 5.2. This allows the user to keep track of the remaining, unexplored space. In order to improve the user experience, the user can rearrange the visual interface by moving and resizing all axis components within the dashboard to match individual preferences. It is intended that the user investigates and compares different configurations and analyses their proximity via the scatter plot. After exploration, the user can move on to the next step of the pipeline to compute new, custom explanations based on the last selected reference explanation which is always displayed in a designated panel (see Fig. 3 ④).

Segmentation

During this step, various segmentation algorithms can be selected. The corresponding parameter values can be adjusted via sliders. The user can continuously change it until reaching the desired outcomes. Fig. 3 ② shows an example of a segmented MRI with the selection of the segmentation method and parameters.

Reduction

The last step of the workflow is the optional step of dimension reduction. The user can decide to skip this step and continue with the standard LIME pipeline. However, it is encouraged to explore the influence of dimension reduction. Once the user proceeds with a segmented image from the previous step, the perturbed images are transformed into a lower dimensional space based on the user's choice. This embedding is used to select only the closest perturbations for the explanation, as described in 4.4. The central component of the reduction page shows the result of the dimension reduction in the form of a scatter plot. The color coding highlights which perturbations would be part of the following explanation (Fig. 3 ③).

Explanation loop

After finishing the dimension reduction step, the user is taken back to the explanation and exploration step where the newly added explanation is now displayed. Additionally, this explanation is added to the scatter plot for the user to compare the new explanation with the pre-computed examples and to reason about its reliability. Ideally, the user stays in a continuous loop of Explanation, Exploration, Segmentation, and Reduction until the entire space of all possible explanations, depicted in the scatter plot, is fully explored and until the user is satisfied with a specific configuration or neighborhood of configurations. All of the computed new explanations are added in a new panel on the explanation page for improved provenance tracking and guidance (see Fig. 3 ④ top right). At the explanation/exploration step, the user always has the option to re-do (button on the top left) and potentially choose a new image during image selection or change to the "user upload" branch.

5.1.2 Pipeline 2: External upload

The "external upload" branch starts with the user uploading their own image. This branch is an extension to the main Pipeline 1 and does not yet include the exploratory map. The workflow in this branch consists of image selection, segmentation, dimension reduction, and explanation (see Pipeline 2 in Fig. 4). 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, the clinician can make use of this simplified Pipeline 2 in the default setting to acquire an explanation for a prediction made by an ML model. Moreover, focusing on the main user group, ML experts, we enable users to compute new explanations using the different segmentation techniques and dimension reduction approaches proposed in Sect. 4.4, 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 (see Fig. 3 ⑤). The user can upload new images to continue to explore. This allows the user to verify, 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 potential waiting times, and allowing to 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. The current deployed version using the provided link is a preliminary prototype and only Pipeline 1 works reliably.

5.2 Provenance tracking

Provenance tracking in our study is accomplished using two approaches. Firstly, as briefly introduced in Sect. 4.2, we employ UMAP to map the set of pre-computed explanations to a lower dimensional embedding. Within this embedding, we introduce a trajectory view that enables users to select explanations from an external image grid and navigate through the explanation space using an arrow as a guide (see Fig. 3 ④ detail). This trajectory view facilitates an understanding of the explanation's sensitivity toward each parameter. As the user traverses through new explanations in the pipeline, each iteration adds a new data point to the 2D embedding, and the trajectory follows this new point. To prevent overplotting, the trajectory map can be discarded at any point in time.

In addition, we incorporate provenance tracking in the visualization of the explanations. Whenever users generate a new explanation, it is automatically displayed in a dedicated panel in the interface and can be selected as a reference for subsequent explorations. While users can explore in an infinite loop, for user-friendliness, only up to 9 images are displayed at a time to avoid overcrowding. Once the limit is reached, each new image replaces the oldest one, yet all images are still included in the 2D exploration map. Thanks to the flexible panel layout, the user can rearrange the components to bring the past history closer to any other desired component. The interface is designed to support users in exploring the space of possible explanations, while the integration of provenance tracking enhances the narrative by preserving a record of past explorations.

6 User scenarios

We highlight the practical use of *ExpLIMEable* based on two user scenarios that are developed around different analysis tasks, focusing on pipeline 1, the main platform in *ExpLIMEable*. We refer to Fig. 3 for visual examples of the described use cases.

Understanding the sensitivity of LIME through Exploration This scenario describes a user who experienced ambiguous explanations by LIME and now wants to understand why the explanations do not align with their mental model and how the different hyperparameters influence the final explanation. At the start, the user receives onboarding prompts that ease the first contact with the dashboard. The user wants to explore different explanations and selects one of the pre-computed images (Fig. 3 ①). Arriving at the exploration page (Fig. 3 ④), the user follows a second set of onboarding information and reads it carefully. At any moment, the user can go back and forth between the onboarding information in case some components remain unclear. After familiarizing them with the exploration page the user now has a good overview of the capabilities of the dashboard and starts to explore different pre-computed explanations. While going through the different explanations the user keeps track of the trajectory via the scatter plot (Fig. 3 ④). After a while, the user notices a small distinct cluster of explanations that they have not explored yet. Using the color coding in the scatter plot and tracing the trajectory the user can find the corresponding parameter set and starts to investigate. In this case, this cluster contained images segmented into large superpixels each spanning considerable portions of the brain. As a result, LIME indicated that the majority of the brain scans were positively contributing to the tumor classification. To the user, this is less intuitive since they also seem to include healthy tissue. Thus, these explanations were skipped by the user initially. Before moving on, the user notes down the critical parameter range. The user now wonders about the influence of the segmentation around the tumor region. Therefore, they start to add custom explanations with different segmentation methods (Fig. 3 ②). Some of these combine the entire tumor in one superpixel and others are designed to divide the tumor into several superpixels. One tentative takeaway is that LIME seems to perform more robustly if the tumor is captured in one large superpixel rather than several smaller ones. The user notes their findings and continues with another pre-computed image.

Improving LIME's robustness by customized explanations In the second scenario, the user already went through the first exploration stage and is now driven to improve the robustness of LIME. The user is aware of challenges arising from out-of-domain predictions and thus downloads the original training dataset. Afterward, the user begins to upload (Fig. 3 ①) new images to *ExpLIMEable* to probe if the findings translate and generalize to other examples. Curious about the dimension reduction feature (Fig. 3 ③) the user

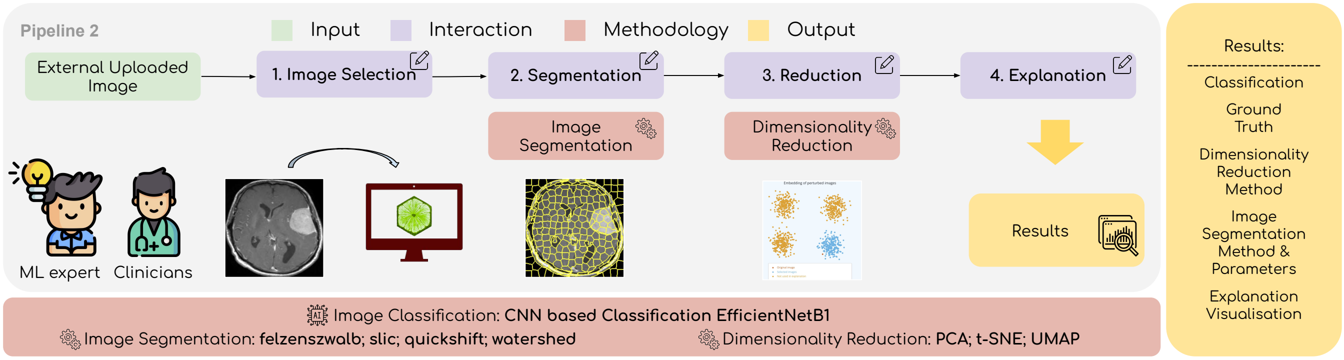
Pipeline 2
Input
Interaction
Methodology
Output
External Uploaded Image
1. Image Selection
2. Segmentation
3. Reduction
4. Explanation
Image Segmentation
Dimensionality Reduction
Results
ML expert
Clinicians
Image Classification: CNN based Classification EfficientNetB1
Image Segmentation: felzenszwalb; slic; quickshift; watershed
Dimensionality Reduction: PCA; t-SNE; UMAP
Results:
Classification
Ground Truth
Dimensionality Reduction Method
Image Segmentation Method & Parameters
Explanation Visualisation

Figure 4: Pipeline 2 of *ExpLIMEable*, detailed in Sect. 5.1.2. Including the four steps of the interactive workflow in purple: 1. image selection, 2. segmentation, 3. reduction, and 4. explanation. The methodology is highlighted in red, user input in green, and outputs in yellow. The ML experts and clinicians are portrayed as the users of this pipeline.

extensively explores the influence of the selected local perturbations on the resulting explanations (Fig. 3 ⑤). Based on these findings, it is envisioned that the ML expert then implements changes to LIME in an effort to make the algorithm more robust.

7 Discussion

In this study, we have developed a tool that we believe offers significant value to understanding the use of LIME, despite its inherent complexity. To enhance user experience, we have invested efforts into improving the tool's aesthetics, incorporating visually discernible colors, and implementing features such as multiple provenance tracking paths and on-boarding guides. We argue that including an unlimited loop within the interface offers substantial benefits, enabling users to navigate the system more effectively and derive greater value as their learning curve in terms of usability can be extended indefinitely. The addition of dimension reduction represents a novel contribution, and we intend to further explore this aspect to enhance the functionality of LIME. This platform will be of great use to machine learning experts, as it facilitates their comprehension of explainability, critical in our medical application and beyond. Given the increasing deployment of models in everyday tasks and the growing regulatory requirements on mandatory explanations, we consider our platform a valuable asset in this evolving landscape.

Regarding limitations, we believe that this tool can provide great value for clinical decision support but Pipeline 1 still requires vast simplifications to make the full system translatable. Currently, only Pipeline 2 can be used by clinicians. In general, the quantitative evaluation of XAI methods is challenging due to a lack of agreement and standardized metrics. At this stage, we do not provide a quantitative metric to compare different evaluations. However, this would be needed to conclude the absolute effect on robustness and whether the dimensionality reduction is beneficial. Existing frameworks to evaluate classification could serve as the basis for further steps, i.e., comparing the predictions after removing the features that contributed most based on different XAI methods [16, 28]. During segmentation and dimension reduction, the user is shown a range of recommended parameters. For now, these are constant for all images. Providing individual parameter ranges for each image, e.g. based on histogram data, could improve user guidance.

In terms of future work, we would like to expand the interactions during the dimension reduction step. Moreover, conducting a comprehensive evaluation user study, with questionnaires and surveys, will provide valuable feedback and insights, to understand how users handle the explanations. Additionally, expanding the pre-computed dataset beyond the current six images will be beneficial for generalization. As the next steps, incorporating parameter exploration for the ”upload branch” will offer users complete freedom in their explorations. However, this comes with the challenge of handling out-of-domain scenarios, which could be circumvented by retraining the CNN with the correct distribution. Regarding the stakeholders in the future, we aim to assist clinicians by providing a better, more robust understanding of the model, to make more informed decisions with a potential optimal configuration. This user will not need to understand the machine learning components but rather select the parameters that yield a faithful explanation aligned with their mental model. We are aware of the challenge of adapting the system but we believe that having this tool will help to incentivize the transition of ML to the clinics. The current dashboard focuses on image data as it is a relevant use case of LIME. However, it could be adapted to other data types, i.e., tabular data or signals. Instead of the segmentation step, we would have an alternative perturbation method to study, i.e. adding random noise or removing data points. However, the dimension reduction step would remain unchanged. Visualization of the new data types would be replaced with tables or plots. Lastly, we have focused on LIME, as one of the most used XAI methods. However, a similar platform could be developed for other XAI models like SHAP [20]. At its core, it considers all possible combinations of features to allocate contributions. We believe that this platform can be extended to its most used variation, KernelSHAP [20], which is equivalent to using weighted linear regression in LIME with a specific kernel. Hereafter, one could study the effect of perturbations and dimension reduction with our framework, modifying the optimization to a kernel-weighted linear regression problem.

8 Conclusion

In conclusion, the increasing use of machine learning techniques in decision-making has created a demand for explainable ML models, especially in high-stakes applications like healthcare. One widely discussed explainable technique is LIME, which is model-agnostic and provides locally interpretable explanations. However, LIME itself is not always robust, and the quality of explanations can vary based on the parameters and perturbation techniques used. To address this issue, we propose a tool that enhances the understanding of LIME and allows users to explore different explanations. We focus on the specific application of brain MRI tumor classification and provide a workflow that enables machine learning developers to gain insights into the model's behavior. Additionally, we introduce a novel approach to sampling in the perturbation space. By applying dimensionality reduction to such perturbations, we aim to select more reliable instances for a more robust explanation. With the help of this tool, users can explore the explanations generated by different LIME instances, providing insights into the system's limitations, sensitivity, and robustness with a smooth narrative keeping track of provenance and explanations of tasks at all steps. With a comprehensive overview, the tool guides users toward an optimal, faithful explanation aligned with their mental model. Overall, our study contributes to the field of XAI by addressing the lack of robustness in LIME explanations and providing a practical tool for enhanced understanding and decision-making in medical image analysis.

REFERENCES

[1] https://github.com/Ashish-Arya-CS/Coursera-Content.

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

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

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

[5] G. Alicioglu and B. Sun. A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, 102:502–520, 2022.

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

[7] N. Bansal, C. Agarwal, and A. Nguyen. 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/CVPR42600.2020.00870

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

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

[10] J. Dieber and S. Kirrane. Why model why? assessing the strengths and limitations of lime. *arXiv preprint arXiv:2012.00093*, 2020.

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

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

[13] K. Goode and H. Hofmann. 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.

[14] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang. XAI-Explainable artificial intelligence. *Science Robotics*, 4(37), 2019. doi: 10.1126/scirobotics.aay7120

[15] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek. *Explainable AI Methods - A Brief Overview*, pp. 13–38. Springer International Publishing, Cham, 2022. doi: 10.1007/978-3-031-04083-2_2

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

[17] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini. State of the art of visual analytics for explainable deep learning. In *Computer Graphics Forum*, vol. 42, pp. 319–355. Wiley Online Library, 2023.

[18] X. Liu, L. Song, S. Liu, and Y. Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021.

[19] P. Love, W. Fang, J. Matthews, S. Porter, H. Luo, and L. Ding. Explainable Artificial Intelligence: Precepts, Methods, and Opportunities for Research in Construction. 11 2022. doi: 10.48550/arXiv.2211.06579

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

[21] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

[22] R. Marcinkevičs and J. E. Vogt. Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *WIREs Data Mining and Knowledge Discovery*, 13(3):e1493, 2023. doi: 10.1002/widm.1493

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

[24] S. Mohseni, N. Zarei, and E. D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

[25] P. Neubert and P. Protzel. 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.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin. "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.

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

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

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

[30] W. Saeed and C. Omlin. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, p. 110273, 2023.

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

[32] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. 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.

[33] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady. 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

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

[35] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pp. 359–380. PMLR, 2019.

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

[37] A. Vedaldi and S. Soatto. 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.

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

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

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

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

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