Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Foreground and Background Separation In Mouse Brain Tissues Using Statistical Tests. Can Tests Substitute For No Prior Knowledge ?

Sotirios Raptis

sotiris.raptis.n@gmail.com

NTUA https://orcid.org/0009-0000-8828-8173

Research Article

Keywords: Background, foreground, Kolmogorov Smirnov test, Kullback Leibler divergence, ROC analysis, Hierarchical Clustering, SOM, Logistic Regression

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9269152/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Example Journal, 2026, xx: 1–41

Foreground and Background Separation In Mouse Brain Tissues Using Statistical Tests. Can Tests Substitute For No Prior Knowledge ?

Sotirios N. Raptis[1*]

[1] *Alumnus with Department of Informatics, National Technical University of Athens(NTUA)*

Disclaimer: Department of Informatics, National Technical University of Athens(NTUA), Athens, Greece.

ABSTRACT

The separation of foreground from background regions has long been a central challenge in computer vision, particularly in situations where no predefined model is available. Nevertheless, large image regions that are likely to contain salient subjects can be identified as foreground through blind modeling approaches. Objective evaluation can be achieved using self-organizing techniques—such as Self-Organizing Maps (SOMs), clustering, and predictive methods—and by assessing their results through self-consistency checks using statistical tests across multiple segmentation trials with varying sampling and partitioning conditions.

Conventional machine-learning approaches, including logistic regression, can further provide insight into image differentiation by partitioning pixels into foreground and background clusters. These methods were applied to a mouse

*Corresponding author: sotiris.raptis.n@gmail.com.

Received 27 xxxxx 2026; Revised xx xxxxx 2026
ISSN xxxx-xxxx; DOI 10.1561/xxx.xxxxxxxx
© 2026 Sotirios N. Raptis

brain scan. Foreground and background were treated as the two primary pixel classes and were assigned using: (1) logistic regression, which provides a binary classification (foreground versus non-foreground); (2) SOM-based neural networks, which assign each image block to one of two output nodes corresponding to foreground or background; and (3) hierarchical clustering, which further groups and subdivides blocks within each category, and (4) KM (K-Means) that is operated with 2 main classes and no other information.

The studied case does not need to fall into the category of immediate DL (Deep Learning) applications. One practical reason is that the data cannot adequately support a DL scheme (many layers capturing very fine details, such as synapses and their structure). At the stage where the paper can be used, there is no need to track such scans from a series of "NanoZoomer" brain scans. Instead, one can use the results from the paper (maps) to train a DL scheme and also use the performance results as additional parameters. Possibly, at a much higher resolution of 10000 $pixels \times 10000 pixels$, DL would be a one-way solution. Examples include automated artifact detection in brain images, small tumor detection, automatic image tagging (what is a lesion), and the analysis of MRI, CT, and EEG data. An example of the use of weakly supervised segmentation of brain images using DL can be found in **Gila2025**, where the weak supervision is applied to bounding boxes around seed-pixels or entire scans (slices) selected from a scan. The work also uses blocks around pixels (seed-pixels), but only when the LGR method is applied. Reportedly, annotated masked (e.g., BG) slices or boxes of labeled BG pixels are tracked across a scanned brain's volume. The degree of supervision is estimated to be greater than the one adopted here, since LGR is used in only one of the three implemented methods and not in KM, HCL, or the tests.

The resulting pixel-level maps indicating foreground and background regions were evaluated for self-consistency by varying block size, number of nodes, seed-point placement,

and the number of clusters. This methodology may serve as a useful preprocessing step prior to the application of more specialized image analysis techniques.

Keywords: Background, foreground, Kolmogorov Smirnov test, Kullback Leibler divergence, ROC analysis, Hierarchical Clustering, SOM, Logistic Regression

JEL Codes: Comma-separated JEL codes go here

1 Introduction

An image can generally be divided into engaging and unengaging regions, commonly referred to as foreground and background, which may contain objects belonging to multiple image classes. This work addresses the problem of distinguishing foreground from background as a dual-class segmentation task and argues that the lack of prior information can be compensated for by statistical methods capable of effectively discriminating image regions.

Defining foreground and background largely involves resolving the ambiguity associated with certain pixels or blocks, a subject extensively discussed in the literature. Such ambiguity often arises from overlap between background (background) and foreground (foreground). In many cases, salient regions are not sharply defined. The present study is based on the self-consistency of dual-class segmentation results. In particular, it evaluates whether maps of the two fundamental classes formed using foreground (foreground) pixels are mutually consistent. A preliminary approach to assessing self-consistency involves generating alternative maps with additional clusters and varying block sizes. In addition, seed pixels may be employed as prior background patterns, followed by segmentation based on similarity to these patterns. This approach is referred to as pixel annotation; however, it does not constitute a true Ground Truth (GT) bias, since the number of GT pixels used is very limited, in this case not exceeding ten. The paper argues that generating statistical foreground/backgroundmaps that can be compared against one another may lead to more reliable segmentation outcomes.

The core idea is to apply statistical testing methods, such as hypothesis testing, to evaluate the generated maps and the spatially adjacent classified blocks after applying segmentation algorithms that rely minimally on prior annotations. The KL and KS tests are representative examples of such methods, with the hypothesis under examination being the similarity between neighboring blocks. Standard statistical models used to approximate similarity statistics typically assume a normal distribution, i.e., $\sim N(0,1)$. It is assumed that the statistics of the resulting segments do not significantly deviate from some variant of a normal distribution. This assumption allows the computation of p-values or significance measures associated with block similarity, whether classified as foreground or background. Such assumptions are widely adopted in statistical learning and are considered realistic in this context.

The separation of foreground from background using no prior information has not been addressed in this context. That is, one adopts an iterative approach that involves segmentation trials using different sampling and classification parameters or block sizes, as well as agreement across the resulting foreground/background maps, rather than relying on a single model for foreground or background pixels. Moreover, to the best of one's knowledge, the NULL hypothesis approach that comes into play is not widely applied in this context and replaces the Gound Truth missing.

In terms of choosing the optimal classification , this would eventually apply either to the optimal number of clusters or to the optimal block size. There is no formal way to determine this in a generic setting. As discussed in this work, a large block size would likely suppress the small periodicity of textures, which would otherwise produce a larger number of smaller clusters that might be more accurate. On the other hand, a small block size would better capture fine structures that would be missed when absorbed into larger ones associated with a large block size. This work does not apply pixel-based segmentation, and the minimal block dimension is 2. However, as also discussed, small block sizes introduce other challenges, such as those encountered with the HCL methods and the large number of blocks that must be compared within the hierarchy of clusters . In terms of the number of blocks, this was chosen to be 2 (FG/BG), and less often 3 or 4, in order to examine the iterative aspect of the approach—that is, whether FG or BG areas

further split when more clusters are used. This direction was not pursued extensively, given the many other challenges faced in this work as well as practical limitations regarding space. The seed points used with the LGR method constitute the only form of minimal prior knowledge. They cannot be considered typical prior knowledge because (a) they were chosen randomly from visually perceived FG/BG areas that were not explicitly modeled, (b) the FG/BG regions were inherently complex even when considered in isolation, and (c) as discussed in the present work, the ratio of seed points (from 2 up to a maximum of 4) was quite small relative to the overall amount of data that were classified. Moreover, the only factor uniting (grouping) FG/BG blocks was the statistical significance of their similarity, rather than spatial proximity, which would otherwise provide explicit prior knowledge via the seed points. Self-consistency testing and multiclass merging are employed to address the absence of a strict theoretical distinction between foreground and background. In this framework, two primary pixel classes are defined: foreground and background **Dong2018SingleLabelMI**. Their distinction largely depends on the specific application. The effectiveness of this definition depends on how consistently these image regions are reproduced when acquisition parameters change, whether for the foreground or the background **DebaniPrasadMishra**. As a result, the foreground and background often encompass more precise content-related image classes, leading naturally to a multiclass problem.

Although the present work combines the "weakly supervised" LGR method with clearly unsupervised ones (K-Means , Hierarchical Clustering, tests), the assessment here is that it does not obviously fall into the category of weakly supervised approaches. Again, given the extent to which Gound Truth is used (seed points in LGR), the proportion of data used for training is very low (pixels used relative to image size). Without further resorting to additional references, hybrid approaches would typically involve specific methods that require extensive parameter setup. These include, for example, *'Mixture of Gaussians'* (GMM), which reports a success rate from 85% to over 95% **Moch2020**.

Further, one can also report *'binary pattern texture analysis'* (BPA) to distinguish what is FG and what is BG. This approach reports success rates of (90˘95)%, as implied, for example, in **Aligholi2022**. Another category includes video analysis methods (identifying mov-

ing objects), which can achieve success rates above 95% in background subtraction. However, these rates can drop to 52% or 75% in more complex images, such as those considered here. Another possible approach would be the *'class imbalance'* method between FG or BG clusters , although no reported success rates were found. Similarly, for methods based on *'reference background images'* for BG modeling, no specific success rates were identified. The use of areas around Gound Truth or seed points can be viewed as an approach related to this category.

A wide range of other methods can also be found in the literature. Among them, many implicitly or explicitly rely on some form of Gound Truth. The implicit reliance refers to comparisons with a known case (paradigm), whereas the explicit reliance involves fine-tuning a specific model or related parameters.

The approach employs regression (i.e., averaging) to simplify and smooth the background. The Linear Regression models are known to predict a pixel's value through linear extrapolation based on neighboring pixels, resulting in a smoother image where fine signal variations are suppressed. Robust edge detection combined with the removal of low-intensity edges (primarily noise) further supports background smoothing. A rich background tends to produce many low-intensity edges following edge detection. By adjusting edge-detection parameters—such as pixel-difference thresholds and the spatial extent of local averaging—sparse, weak edges can be eliminated. Consequently, background variations that generate uninformative edges and do not correspond to actual structural boundaries can be largely suppressed.

Edge detection is inherently a point-based method that depends on single-pixel values. In contrast, the present work applies the segmentation to blocks rather than individual pixels, and subsequently uses pixel-level maps to evaluate the segmentation. Thus, comparisons are performed at the pixel level, even though classification is carried out at the block level. The evaluation metrics are therefore applied per pixel, meaning that each pixel inherits the class label of the block to which it belongs, rather than being classified directly by its own gray-level value.

Various feature types can be used to construct masks. For example, the work in **RyanThiermann** employs image responses to structure-removal filters to assign pixel labels. These may include edge micro-orientation detectors, where convolution responses or neighborhood re-

sponses to matched filters sensitive to relevant or irrelevant structures are used. Such features can capture approximate background periodicity across multiple orientations, after which the filtered image is subtracted from the original to produce a mask that removes large uninformative regions.

This approach is applied in cell-imaging applications, where it is used to suppress cytoplasmic noise.

The partitioning of an image into two regions is closely related to occlusion detection, as discussed in **JYLee**, namely the identification of background and foreground (foreground–background separation). This work examines how block size influences this process. Small blocks (*e.g.* 2×2 or 4×4) tend to expose fine foreground/background transitions or repetitive transition patterns, as illustrated in Fig. 1a. However, such configurations often make foreground/background separation difficult due to strong texture.

Alternative approaches, such as those in **LiShengli**, employ the segmentation -based Region Growing method, where pixels are progressively added to regions based on similarity criteria. Although region growing avoids rigid block-based analysis and allows greater flexibility, it requires careful selection of seed pixels and growth rules. The shifted growing strategy mitigates local interruptions by systematically relocating the seed pixel. When the background is more homogeneous than the foreground, this method can provide robust separation.

Motion-analysis techniques can also assess continuity across segmentation results by modeling background variations. In **HWei**, changes induced by motion are contrasted with spatial variations, assuming a static spatial background model and a time-varying motion-dependent one. The approach used here differs in that it relies on global classification parameters—such as class centers and confidence values of block memberships —rather than local changes alone, offering greater robustness. Instead of explicit motion modeling, the membership criterion is derived from KL or KS tests, complemented by binary agreement of pixel-wise classification maps.

Combined foreground / background separation and motion-based methods have been applied to medical imaging, as shown in **ChaochaoZhou**, where motion artifacts in cranial imagery are suppressed. In Digital Subtraction Angiography, accurate vessel delineation is essential for detecting abnormalities. Minor motion during CT acquisition degrades

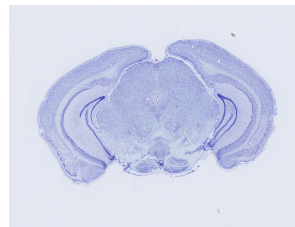
this accuracy, while block-based registration and masking—using a *'precontrast fluoroscopic background'*—help isolate true background regions. Since vessels exhibit tube-like structures with Gaussian cross-sections, they span multiple pixels; the minimum block size of four pixels used here is therefore suitable. In this context, the Logistic Regression classification is effective, as it allows training from a small number of manually selected healthy tissue blocks.

In Cryo-EM tomography, the background can be modeled as additive Gaussian noise **Mansournia2018SeparationIL**, with molecular structures treated as foreground. Additionally, as noted in **Xuebin**, foreground / background separation may depend on image resolution: blurred regions are classified as background, while sharper regions correspond to foreground. Although this criterion is not directly applicable to Fig. 1a, modifying the classification parameters or block size reveals alternative background definitions that reflect the scale of target structures. Resolution changes also aid in identifying fiber connectivity and anatomical boundaries **HamptonCM**; **JeongM**.

More generally, foreground / background separation can be extended to multiclass settings **DebaniPrasadMishra** by designating clusters of interest as foreground and assigning remaining regions to the background. Cropping and annotation can then be used to train classifiers such as SVMs (*'Support Vector Machine'*) for foreground identification. Similarly, Logistic Regression has been used for face extraction **SuraSabahKeiteb**. In contrast to feature-driven approaches, the present work relies on statistical similarity testing against a limited set of reference blocks.

2 The data

In Fig. 1a, one can see the image that is the target of this work. One can see the presence of a complex foreground creates ambiguity. To address this, the Hierarchical Clustering method was applied using multiple block sizes to obtain statistically separable regions, assuming low $p - value$ values for blocks belonging to different classes. Increasing block size effectively "flattens" the foreground or the background **JYLee** and alleviates overlap cases in which both foreground and background pixels appear within the same block. In such situations, all pixels inherit the



(a) FF3 [*1], MBA

Figure 1: GRP1 The foreground and background are statistically separated in a mouse brain's CT scan. ([*1] `All images are assigned an ID ("FFXX") for review purposes that is to be removed`)

block's label.

Moreover, both foreground and background may consist of complex structures, an issue discussed in **Phalempin**. In many cases, the background corresponds largely to noise, and modeling this noise enables its removal through an appropriate noise model. This situation is exemplified in Fig. 1a. The cited work highlights that a highly detailed background, encoded through large pixel depths (i.e., many gray levels per pixel), can give rise to a complex background that is difficult to analyze. High pixel depth may degrade edge clarity by producing numerous small pixel differences that reduce the sharpness of dominant edges. Conversely, it may also introduce fine details that generate new edges capable of revealing hidden structures.

3 Methodology

The initial idea was to generate maps of foreground and background pixels using Logistic Regression for foreground/background separation. Specifically, image blocks of varying sizes were classified into two fundamental image classes using block dimensions in the set $2, 4, 8, 16, 32, 64$. The Logistic Regression method operates with minimal training, requiring only a small number of mouse clicks to designate foreground or background blocks on the image plane. A typical image of size (900×730) pixels contains a number of blocks in the range $[160, 10000]$ for block sizes in $[8, 64]$. Consequently, using up to ten clicks does not constitute a supervised approach. In practice, Logistic Regression was applied using at most five annotated blocks per category (ten in to-

tal for both foreground and background), selected in a largely random manner. The KL and KS tests were then used to assess how effectively the remaining (non-seed) blocks were statistically classified as foreground or background, based on the hypothesis of belonging to either a foreground or a background region. This setup resembles a multi-class problem, as foreground and background may not be adequately represented by a single block category.

The second method employed the HCL (Hierarchical Clustering) classification and required no annotated blocks, relying instead on a predefined number of clusters to complete the merging of block classes inherent to HCL. For a final number of clusters ($n = 2$) and the same range of block sizes, HCL produced alternative maps of foreground and background pixels. The KL and KS tests were again applied to examine the hypothesis that blocks belong, in statistical terms, to the same generic cluster or not (i.e., foreground or background).

A third method employed was the well-known Self-Organized-Maps clustering . It is well established **Pavel2025** that SOM-type Neural Networks can map data from an arbitrary initial dimension to a one- or two-dimensional grid of output nodes, which can be interpreted as final clusters . It should be noted that the similarity measure—or the $p - value$ in KS testing—depends on the variance of gray-level values within each block category inside a given cluster . Consequently, the number of blocks collectively attributed to a foreground or background region depends on how these regions are defined. Thus, except for the Logistic Regression method, the remaining approaches—requiring no manual feedback—primarily assess the validity of using the $p - value$ as a criterion for effective cluster assignment across different metrics (KM, HCL).

In the approach adopted here, pixel blocks of at least four pixels per dimension (i.e., $4^2 = 16$ pixels) were processed. This choice provided effective smoothing while preserving the notion of block-based classes rather than individual pixels. The objective of the study is to capture foreground and background using statistical discrepancies between pixel blocks, rather than relying on neighborhood information. Consequently, even neighboring blocks were independently classified as foreground or background based on statistical information derived from distant blocks.

Classification results were obtained using Logistic Regression, Hier-

archical Clustering, or K-Means across a range of block sizes. Blocks were classified first, and individual pixels inherited the class label of their corresponding block, producing classification maps that represent different versions of the original image. These maps were then compared both against the original image and among themselves. The outcome was a binary pixel map, where pixels were assigned a value of one only when there was agreement across all classification maps regarding their foreground or background membership. Additional maps were also generated for low and high $(p-value)$s, rather than agreement alone. In these $(p-value)$s maps, pixels assigned high $p-value$ values indicate blocks consistently classified into the same class. To generate KS or KL maps, gray-level histograms of the corresponding blocks were extracted and used in KL or KS testing to evaluate similarity.

To reduce bias arising from enforcing block-level class assignments, as in LGR, different block sizes were used. Additionally, for methods such as LGR that require minimal training, slightly different pixel locations were selected for the same block size. Training clicks ranged from two to ten pixels for foreground and background blocks, corresponding to one to two clicks per class when using five seed pixels.

Overall, the goal was to benchmark foreground/background separation with minimal reliance on prior knowledge. This was achieved by comparing three classification methods: (1) HCL, (2) Logistic Regression, and (3) K-Means . None of these methods uses explicit prior models for foreground or background , except for Logistic Regression. The selected training points were chosen randomly from both evident and less evident foreground and background regions, including areas with low or high overlap between the two classes.

Two types of images were examined. The first consisted of mouse brain images acquired using a nanozoomer scanner, where it is not always clear which regions should be considered foreground or background. The objective was not to segment specific neurological regions, but rather to evaluate classification consistency under changes in seed-point location (for Logistic Regression) and block size. The second type comprised Cryo-EM images of molecular (protein) volume projections. In this case, the foreground corresponds to single-particle protein projections that must be tracked across tilts (i.e., successive rotations), while the background corresponds to cytoplasmic content to be treated as a single removable class.

Self-consistency was evaluated by measuring changes in classification results as three key parameters varied: (a) block size, (b) seed-point location (when applicable), and (c) the number of clusters . As discussed earlier, block size plays a crucial role in defining the scale of the regions of interest (foreground or background). Larger blocks may increase overlap between foreground and background regions, whereas smaller blocks improve classification accuracy but may introduce fragmented foreground/background patterns within larger homogeneous regions.

For each block size or seed-point configuration, a new foreground/background pi map was generated, with values corresponding to cluster IDs set to 0 for background and 1 for foreground. This map is represented by the binary matrix $M \in 0, 1^{m \times n}$, as defined in (1):

$$M(i,j) = \begin{cases} 1, M(i,j) \in FG \\ 0, \ M(i,j) \in BG \end{cases} \tag{1}$$

Each binary matrix corresponds to a specific classification method ($CLS_i, \ i \in [1,2,3]$) and block size ($D_i \in 2,4,8,16,32,64 = DR, \ i \in [1,6]$), or to a specific seed-point configuration. Accordingly, a set of classification maps was obtained for each image examined:

$MM_{image_l.CLS_i,D_k} = \{M_{image_l,D_{k_1},CLS_{i_1}}, ...M_{image_l,D_6,CLS_3}\}]$

The number of clusters for Hierarchical Clustering and K-Means was fixed at two, and seed-point locations (relevant only for Logistic Regression) were not treated as tunable parameters, since the approach aims to minimize prior knowledge. The Logistic Regression method, a standard binary classification technique, was used with minimal training. For example, an image of size (728×928) pixels yields approximately 320,000 blocks for block size (8×8), from which no more than ten blocks were used for training—approximately 0.003% of the data. These values are indicative averages and need not satisfy $N_{blocks} \times BlockSize = ImageSize$. Training clicks were selected using different visual interpretations of foreground and background regions, with the total number of clicks remaining below ten.

Indicative training blocks for the image in Fig. 1a are shown in Fig. 2a through Fig. 2e for *'MBA'*. To avoid overloading the paper with intermediate results, not all block maps or images are presented. Instead, representative cases were selected to cover a range of conditions, including mixed foreground/background, clean background, and

clean foreground regions. The Logistic Regression method follows a probabilistic formulation for foreground/background classification and belongs to a well-established class of such approaches.

$$p_{FG/BG}(i,j) = \frac{1}{1+exp^{-B(\hat{i},j)_i^T * \hat{w}_i}} = \begin{cases} 1 & M(i,j) \in B_(i_0,j_0) \; and \; P(i_0,j_0) \in FG \\ 0 & M(i,j) \in B_(i_0,j_0) \; and \; P(i_0,j_0) \in BG \end{cases}$$

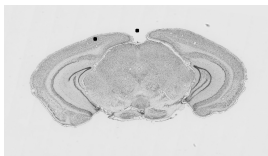
where $\hat{w}_i$ is a vector of weights resulting from the training of the Logistic Regression classifier using training pairs of the type $((BLKtraining, BG/FG\ 1(for\ i \in [2,5], GTi_{training}))$, with $(GTX_{training})$ being a ready annotation (FG,BG) associated with the clicked pixel $(P_0 = (i_0, j_0))$. Of course, any variable/data that is denoted $X_{training}$ indicates that it was used for training. The term $(BLK_{i,k})$ denotes an image area of a given dimension and block size (D), centered around the selected point, that is, $(B_{i_0,j_0} = [I(i,j)], for\ i \in [i_0 - D/2, i_0 + D/2],\ j \in [j_0 - D/2, j_0 + D/2])$, where the full image under examination is $(I_{m \times n})$.

Hence, the Logistic Regression -based method constructs a classification map using at most five (and typically only two) pre-annotated pixels, which may belong to either the foreground or the background class, with at least one pixel provided per category. All remaining non-annotated blocks are then classified using the trained model.

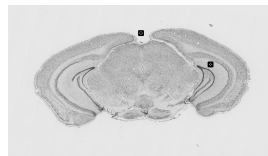
Then, one slightly varies the (i_0, j_0)s, obtaining different models and corresponding maps for six (D)s and two or three $((i_0, j_0))$s. These maps are used to mask the original image, and the resulting image blocks (B) are statistically examined by extracting their histograms, which are then used as inputs to (a) the KL divergence test and (b) the KS similarity test.

The tests are performed using, as a reference histogram (the model of the test), the average histogram per image category (FG, BG). For the KL metric, the KL divergence is computed as in (2), while the blocks are assigned an FG-KL value according to (3). The final decision is then made as described in (2, 3, 4).

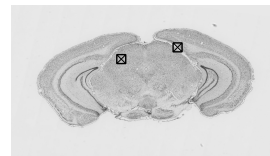
$$KL : KL(H_1, H_2) = -\sum\nolimits_i H_{1_i} * log\left(\frac{H_{1i}}{H_{2i}}\right) \tag{2}$$



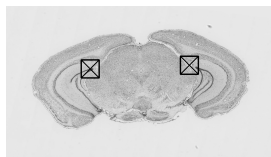
(a) (FF116, MBA, KM, 2 seeds , SZ=4*[1])



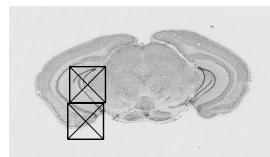
(b) (FF117, MBA, KM, 2 seeds, SZ=8)



(c) FF118, MBA, KM, 2 seeds, SZ=16



(d) FF119, MBA, KM, 2 seeds, SZ = 32)



(e) FF120, MBA, KM, 2 seeds, SZ=64)

Figure 2: GRP2 Indicative FG and BG training blocks for image MBA (1a) in all 5 block-sizes and different seed points. The main seed if where the two diagonals cross. *[1]: "SZ" indicates the dimension of the block, hence $SZ = 2$ is 4 pixels as number of pixels

$$
\begin{array}{l}
N_{FG}: \quad Number\ of\ FG\ blocks \\
nB: \quad Number\ of\ intensity\ buckets\ per\ block\ of\ size\ D \\
H(B_i) = H_i \\
Histogram(B, nB): \quad Histogram of B_i, using (nB) buckets \\
KL(B_1, B_2) = KL(H(B_1), H(B_2)) \\
\quad : KL divergence\ between\ blocks\ B_1\ and\ B_2 \\
Average\ H\ of\ FG\ blocks: \quad (\hat{H_{FG}} = \hat{H(B_i)}, B_i \in FG \\
Average\ of\ FG\ blocks\ (\hat{B_{FG}}) \quad = \hat{B}_i, B_i \in FG \\
Average\ KL\ of\ FG\ blocks (\hat{KL_{FG}}) = \\
\quad \hat{KL_i}, B_i \in FG \rightarrow "KlFgBlkAve" \\
KL\ of\ B_i\ block: \\
\quad (KL_i) = KL(BLK_i, \hat{HstFgBlkAve}) \\
"HstFgBlkStd" \leftarrow STD\ of\ H\ for\ FG\ blocks
\end{array}
\tag{3}
$$

Then, the KL evaluation is:

$$
\begin{array}{l}
KL_{EVAL(}i_0.j_0) \in \left\{ \begin{array}{ll} FG & Cond\ 1 \\ BG & Cond\ 2 \end{array} \right. \\
where: \\
Cond\ 1 \\
\quad |KL_{FG}(H_i, \hat{H_{FG}}) - \hat{KL_{FG}}| <= 3 * STD(KL_{FG}) \\
Cond\ 2 \leftarrow \\
\quad |KL(H_i, \hat{H_{BG}}) - \hat{KL_{BG}}| <= 3 * STD(KL_{BG})
\end{array}
\tag{4}
$$

The KS test is non-parametric and only considers counts in raw data (the empirical CDFs of data counts' sums). Given two data sets represented as block histograms, either $H(\mathrm{FG}_i), i \in \mathrm{FG}$ or $H(\mathrm{BG}_i), i \in \mathrm{BG}$, the KS distance can be defined as:

$$D_{KS} = \sup(H_{\mathrm{BG}_I}, \hat{H}_{\mathrm{BG}})$$

, used for BG-inclusion, and

$$D_{KS} = \sup(H_{\mathrm{FG}_I}, \hat{H}_{\mathrm{FG}}$$

used for FG-inclusion. Here, $\sup(A, B)$ is the instant maximum of the pairs of CDFs, $\mathrm{CDF}\big(i \in 1, \ldots, |P_A|\big)$ and $\mathrm{CDF}\big(i \in 1, \ldots, |P_B|\big)$, where

i runs through the length of each array. $|X|$ denotes the length of the set X. One obtains an instant CDF subset

$$\mathrm{CDF}_{A/B}\big[i \leq |P_{A/B}|\big] \subset \mathrm{CDF}_{A/B},$$

for each pair being compared. That is, one histogram H corresponds to the newly examined block, and the other to the current average histogram H over all FG or BG blocks.

3.1 Use of Hierarchical Clustering for classification

The Hierarchical Clustering is an unsupervised classification method that, for (N) data pieces (blocks), initially assumes the same number of initial clusters . This number is gradually reduced by merging similar data blocks into larger clusters, while a distance measure is used to separate blocks or split larger clusters, depending on the hierarchy direction. The distance metric is typically Euclidean, and the data consist of all blocks of the examined image.

As with Logistic Regression, the analysis was conducted over the same range of block sizes (DR); however, the location (seed point) was not a parameter. The Hierarchical Clustering method is less biased than Logistic Regression since no prior knowledge is used. The only prior information involved can be treated as a parameter and concerns the number of clusters to be extracted. In this case, this number is fixed at 2 (FG, BG). Different (D)s are still applied, and the resulting HCL maps are evaluated using KL and KS tests.

In the absence of a specific prior-knowledge setting, this was defined only at the end, after all trials had been conducted. That is, once all pixel maps were created, one could then examine (1) their self-similarity and consistency across pixel maps, and (2) the fact that, over a series of T trials, the Gound Truth would emerge as the most likely form of prior knowledge. In case (2), this would correspond to the class (FG/BG) that dominated the set of pixel labels across the trials, or alternatively to the "majority" class. The "majority" class would be the dominant FG/BG class assigned to the pixels surrounding the examined pixel.

3.2 The Self-Organized-Maps (SOM) clustering

Due to the complexity of the images and the similarity of foreground and background patterns, a 1-D (2-D) grid of output nodes (10×1) was

selected to map the image blocks to output SOM nodes. The resulting 10 clusters were subsequently merged into two fundamental groups, namely the foreground and the background ones. The selection was based on the best KS or KL test outcomes, that is, clusters yielding minimal Kullback Leibler divergence and Kolmogorov Smirnov test metrics were merged so that the blocks within each merged SOM cluster were as similar as possible. Owing to space limitations, further experimental details are not reported here.

After the SOM clusters were assigned to either the foreground or the background clusters , the image blocks associated with these SOM clusters were finally classified as foreground or background blocks, respectively.

Low-overlap FG/BG segments (as in Fig. 1a) and low texture simplify FG BG rendering. Nevertheless, even after isolating the brain areas from the rest of the image (for example, in Fig. 9d), it remains difficult to unambiguously determine whether a region belongs to FG or BG.

3.3 Evaluation

3.3.1 Use of KS for to value classification

Both the KL divergence test and the KS (Kolmogorov Smirnov test) test assess the similarity between FG and BG blocks in different ways. The KS metric yields a probability that the intensity histograms (or distributions) of two blocks belong to the same class or are drawn from the same unknown underlying distribution. The $p - value$ used in the KS metric is the probability that the NULL hypothesis—namely, similarity—holds under the χ^2 test for the two blocks.

Here, similarity refers to the membership (membership) of both blocks to either the FG or the BG. The NULL hypothesis states that the intensity distributions of the two blocks both belong to either the FG or the BG class. Consequently, a low $p - value$ ($p.value \leq 0.05$) indicates that one block is BG and the other is FG, as classified by any of the three methods (HCL, LGR, KM).

If two blocks are classified into the same (FG/BG) category but the KS test yields a high $p - value$ (indicating dissimilarity), this results in a conflict between the KS test and the examined classification method and may trigger further merging or splitting. Conversely, a low $p -$

value indicates agreement with the assigned (FG/BG) category. The relationship between the KS test and a given classification method may change as the block size (D) varies.

Using the Kullback Leibler divergence method, one obtains a divergence map that can be transformed into a probability map by assuming that the divergence values for foreground and for background follow normal distributions (N_{FG} or N_{BG}), with corresponding parameters ($\mu FG, \sigma FG$) estimated from the sets of FG or BG blocks.

3.3.2 Evaluation by comparing clusters maps

Instead of comparing classified blocks using KL or KS tests, one can superpose the classification maps produced across a range of block sizes or by varying the number of clusters and seed points (for LGR). As discussed, multiple label-maps can be generated per image, as described in Section (3). More formally, one can define at least two such maps, (M_1) and (M_2), obtained from different trial parameters, and compare them by examining the level of disagreement between pixel class assignments.

A disagreement occurs when, for a given block size and number of clusters , a method assigns a label ($L_1(i,j) = L_1 \rightarrow M_1$) to a pixel ($(i,j)$), and under a different block size or number of clusters , the assigned label changes to ($L_2(i,j) = L_2 \rightarrow M_2$).

As discussed, three clustering methods were employed, namely HCL, KM, and LGR, using between 2 and 4 seed pixels (for LGR) and six block sizes on average. HCL did not perform well for small block sizes (2,4) or for a small number of clusters (2 to 4). Additional clusters (3 or 4) were examined to determine whether further segmentation of already labeled (FG or BG) pixels would result in additional merging or splitting into sub-FG or sub-BG regions (iterative HCL). Thus, multiple label-maps could be produced using the three methods and all feasible combinations of segmentation parameters (block size, number of clusters , seed point locations) for a given image.

Let the maps obtained from (k) segmentation trials be denoted as ($L_k(i,j), k \geq 3$). For each pixel, one can then define a metric corresponding to the number of distinct labels encountered across trials. This metric indicates how many different (L_k)s are observed at pixel location ((i,j)). A higher number of unique labels implies greater vari-

ability and increased ambiguity, making such pixels likely candidates for frontier regions or intermediate FG/BG areas.

Maps depicting the number of unique classes per pixel were generated, and their histogram was computed to assess how frequently label changes occurred per segmentation passage. This process was applied to both input images, all three segmentation methods (KM, LGR, Hierarchical Clustering), five block sizes, and between two and four seed points. Overall, 3D volumes were constructed per image and per segmentation method, where each plane corresponded to a different sampling or segmentation configuration (that is, number of clusters , number of seed points), or block-wise (five cases), or seed-point-wise. The number of distinct labels per pixel was then computed.

Due to space limitations, intermediate results from the pixel label-change evaluation are not shown. Using this criterion, the most effective method was identified as the one for which the uniqueness histogram exhibited high bins for few label changes, indicating that only a small number of pixels experienced frequent label changes, while the majority remained stable.

The above procedure can also be applied separately to each of the three segmentation methods to define two pixel categories: *'vague'* (likely frontier pixels) and *'non vague'* (unlikely frontier pixels). The underlying rationale is that pixels located at boundaries between valid FG/BG regions are more likely to exhibit label changes across segmentation settings, whereas pixels firmly belonging to FG or BG regions are expected to remain stable.

One can construct a label-change array per pixel as in equation below:

$$\overline{L_(i,j)} = L_1(i,j), ..., L_n(i,j)$$

where (n) denotes the number of distinct labels observed (not the number of segmentation trials). For $(L_{(i,j)} \in 0, 1)$, the pixels correspond to (FG, non-FG) classes. Consequently, FG/BG classification evaluation can also be expressed in terms of changes in $(L(i,j))$ or label differences. This concept is formalized mathematically using 3D maps or volumes, as defined in (5).

$$
\begin{aligned}
V_{image_l,k}(i,j,k) = \quad & L(i,j,k) \\
& \ni \; L(i,j,k) \in \{0,1\} \\
& \ni \; (i,j) \in [1...m, 1...n] \\
& \ni \; k \in T \in [1, NoTrials]
\end{aligned}
\tag{5}
$$

The (T)s differed for each segmentation method and spanned the range of block sizes, number of clusters , and seed points. For example, HCL had fewer (T)s due to the volume of intermediate data produced by its hierarchical nature, particularly for small block sizes, where a large number of very small clusters are generated.

To examine changes in the classification maps across the (T)s, the uniqueness of labels per pixel was evaluated. As discussed, the first criterion involved computing the number of unique cluster labels assigned to each pixel across all block sizes, seed points, and numbers of clusters . Pixels exhibiting low values are more likely to be reliable candidates for FG or BG membership, since this indicates agreement across all classification methods.

Let $N_L(i,j)$ denote the number of distinct values of $(M(i,j,k))$ as k varies. One can then construct a uniqueness matrix (N_L), which highlights pixels that are more likely to belong to FG or BG regions. The quantity $N_L(i,j)$ does not directly indicate FG or BG , but rather reflects the confidence associated with the classification decision. Low values of $N_L(i,j)$ correspond to high confidence, whereas high values indicate low confidence.

There are several possible approaches to assessing how local cluster assignments change, either by comparing different labeling methods or by analyzing changes within a pixel's local neighborhood. For example, one could examine whether changes at a given pixel propagate to neighboring pixels. This approach was not adopted here, as the focus of this work is on single-pixel labeling consistency across different segmentation passages. As such, the proposed approach is not a segmentation method in itself—since it does not attempt to merge pixels into larger cluster segments —but rather a metric for evaluating classification stability. One can state that, under the map-evaluation approach (that is, using classification maps rather than Kullback Leibler divergence evaluation), the "true" $(h_{FG/BG})$ map is derived from the set of L maps by selecting the most likely pixel labels (the dominant

ones). Accordingly, a segmentation method can be evaluated using the histogram of the (N_L) matrix, which reflects the distribution of likely false and correct classifications.

High values of (N_L) indicate a higher likelihood of false classification , whereas low values of (N_L) suggest correct classification for $D_i \in DR$, evaluated across the entire DR range (and not across individual trials).

A radical change indicates that the method tends to produce highly variable classification maps and, therefore, lacks stable performance. Indicative results for both cases are shown in Fig. 11.

One can further extend the analysis by considering additional block sizes or a larger number of clusters . In this case, multi-dimensional maps can be constructed. Specifically, in the maps $(M(i, j, k))$ defined above, the index k spans the range $k \in [3, 10]$.

3.3.3 ROC analysis

In addition to using significance tests and self-consistency measures, ROC analysis (*'Receiver Operating Characteristic'*) was employed to further assess self-consistency. ROC analysis evaluates the probability of incorrect prediction or pixel misclassification as foreground or background , and it operates at the pixel level. That is, unlike the KS and KL evaluations used earlier, ROC analysis does not consider the region to which a pixel belongs. Instead, it relies solely on the so-called *'Gound Truth'*.

The present work aims, in principle, to investigate an approach that uses minimal prior knowledge, which can be regarded as another form of Gound Truth. To connect this work with the mainstream ROC framework, all intermediate classifications from the trials were used as Gound Truth. This choice was motivated by the absence of a proven Gound Truth and by the objective of the paper itself. Thus, for each pixel, one may have up to N_T most likely classifications as foreground or background. This is reflected in the results, where multiple ROC curves are presented per plot, with ROC parameters corresponding to specific combinations of block size and number of clusters, each yielding a different segmentation.

ROC analysis involves computing prediction performance for two pixel categories: foreground and background. Two-class pixel maps were obtained using Logistic Regression, K-Means , and Hierarchical

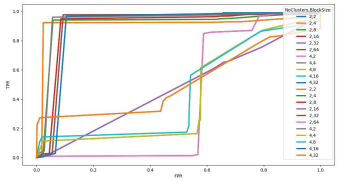
Clustering methods, across all block sizes and seed-point configurations over N_T trials. The ROC curve's true-positive parameter for a pixel at location (x, y) is denoted as $GT(x, y)$, where GT may correspond to either a foreground or a background pixel. In this formulation, a true positive (TP) corresponds to a correctly identified foreground pixel, while a true negative (TN) corresponds to a correctly identified background pixel.

The true positive rate (TPR) and false positive rate (FPR) are defined as $TPR = N_{TP}/Number\ Of\ FGs\ In\ A\ Trial$ and $FPR = Number\ Of\ FPs\ In\ A\ Trial/N_{BG}$, where $N_{TP} = N_{FG}$ and $N_{TN} = N_{BG}$. The TP and TN pixels are taken from the GT image, while the false positives (FP) and false negatives (FN) are derived from the maps produced by the segmentation methods.

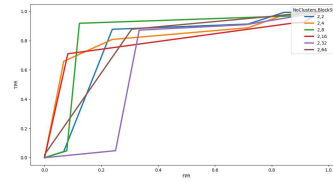
Representative ROC curves are shown in Fig. 3b for the MBA–LGR case, where each "case" corresponds to an (image, segmentation method) pair. Each plot contains between four and six ROC curves, corresponding to distinct (Gound Truth, NoCluster, BlockSize) combinations. Additionally, each plot may include five to six ROC curves associated with different execution parameters, such as the number of clicks (for LGR) and seed-point configurations.

The AV ROC curves exhibit superior performance: all AV ROCs lie above the naive classifier line ($TPR = FPR$), display steeper slopes, and reach TPR values above 80% by the second or third trial. In contrast, several MBA ROC curves fall below the naive classifier and exhibit a more gradual increase. This behavior is attributed to the structural complexity of MBA images, where the distinction between foreground and background is less clear due to the rich and heterogeneous content of brain imagery. As a result, multiple regions may be interpreted as foreground or background , rather than a simple binary separation, leading to a slower increase in TPR.

Furthermore, boundaries in AV images are generally less complex, whereas MBA images exhibit more intricate yet well-defined boundaries. This complexity also affects the KS and KL maps discussed earlier (Table 1), where perfect self-consistency is not observed (all values remain below 0.45). Overall, self-consistency appears to depend on whether a larger number of expected classes—beyond a strict binary separation—is allowed, enabling better representation of content variability and improved internal consistency.



(a) FF102, MBA, KM



(b) FF104, MBA, LGR

Figure 3: GRP3 Indicative ROC analysis plots on image MBA for sample block-sizes, clusters , seed points and numbers of seeds

3.4 Comparing the KS, KL and the maps for classification evaluation

The evaluation strategy was largely based on self-consistency metrics derived from the application of similarity measures across the different segmentation methods. That is, trials using slightly different sampling schemes (block sizes) as well as varying classification parameters produced foreground or background pixel maps of the same underlying image. From these maps, (a) the ratios of foreground to background pixels were computed as the sampling or segmentation conditions changed, and (b) the resulting binary maps ($FG \leftarrow 1$, $BG \leftarrow 0$) were compared pixel-wise as binary flags. Specifically, the binary pixels were treated as strings of length-T flags, where T is defined in Section 3.3.2.

The Kullback Leibler divergence divergence was used to compute block similarity when blocks were produced using LGR. Similarity was assessed based on the binary membership of a block to a foreground or background area and on whether this membership was supported by the NULL hypothesis. The NULL hypothesis indicated non-resemblance when the KL divergence between two compared blocks (that is, their equivalent full-range histograms) deviated significantly from the average KL divergence observed when comparing arbitrary pairs of blocks (either FG or BG).

Specifically, for all blocks classified as FG or BG using LGR, the algorithm first generated *'equivalent'* full-range histograms with 255 bins (buckets). These histograms were then used to compute the KL divergence, and a $p-value$ was obtained under the assumption that the

KL divergence values follow a normal distribution, with mean equal to the average KL divergence and standard deviation equal to its standard deviation.

If the $p-value$ of the KL test was found to be below 0.05, resemblance was rejected, using the $p-value$ as the significance criterion. The BG blocks were identified as those whose KL divergence values lay in the tails of the resulting distribution, that is, below the 0.05 threshold. In practice, neighboring blocks exhibiting large Kullback Leibler divergence deviations in their histograms were assigned to the background class, allowing the background area to encompass blocks with higher variability.

The typical threshold applied for FG/BG block similarity was the standard one, namely 0.05, as commonly used in NULL hypothesis testing. The threshold used with KL, defined as $3 \times KL_{STD}$ away from the mean, is also typical and is applied when the $p-value$ standard is not directly applicable. Assuming that the KL metric follows a normal distribution, the cumulative probability corresponding to $3 \times STD$ is approximately 99.7%. In six-sigma quality control (QC) theory, or in general outlier identification, this region is considered sufficient **Diana2025**. In the present context, QC concepts can be applied within the framework of cluster quality.

Table 1 presents the comparative evaluation results for the examined images (Fig. 1a) (MBA). The table is organized into separate columns for each segmentation method (namely HCL, KM, and LGR) and, subsequently, for each evaluation method (namely Kolmogorov Smirnov test and Kullback Leibler divergence). The rows correspond to different trials, defined by specific combinations of seed points and block sizes.

Table 1 shows that, for both images, FG and BG blocks evaluated using Kolmogorov Smirnov test or Kullback Leibler divergence metrics lie predominantly within the interval $[0.25, , 0.45]$. The table presents the comparison results by dataset, with AV results spanning columns 2 to 7 and MBA results spanning columns 8 to 13. The comparison covers the five block sizes discussed earlier. Although the table is exhaustive, the main conclusion is that the KS and KL metrics yield broadly comparable results.

For the KS metric, this range corresponds to an average confidence level, or $p-value$, indicating that FG and BG blocks within the

Size	Image MBA											
	This work								Found from similar cases (not all are referenced)			
	HCL	FG BG	SOM	FG BG	LGR	FG BG	KM	FG BG	GMM	FG BG	BPA	FG BG
	KL	KS	KL	KS	KL	KS	KL	KS	KL	KS	KL	KS
4	0.41 0.24	0.21 0.14	0.22 0.19	0.32 0.39	0.14 0.16	0.15 0.19	NR*4	NR	(0.357-0.223)*3	(0.357-0.223)	0.004	0.004
8	0.71 0.44	0.23 0.30	0.03 0.39	0.23 0.30	0.23 0.30	0.43 0.86	NR	NR	0.294	0.295	0.014	0.014
16	0.33 0.88	0.38 0.37	0.21 0.37	0.43 0.17	0.32 0.52	0.39 0.21	NR	NR	NR	NR	0.04	0.04
32	0.68 0.78	0.27 0.26	0.37 0.90	0.27 0.26	0.38 0.19	0.37 0.26	NR	NR	NR	NR	0.034	0.034
64	0.26 0.26	0.27 0.21	0.22 0.24	0.25 0.25	0.30 0.76	0.22 0.25	NR	NR	NR	NR	(0.105-0.01)	(0.105-0.01)

*2 FG/BG blocks KL $p-value$, *3 using the relationship ($PRC = F(KL)$) in section 5, *4 No report found in the literature5

Table 1: Kolmogorov Smirnov test, Kullback Leibler divergence comparisons for the image (MBA) (1a) and the three segmentation methods (HCL, SOM, LGR, KM, GMM, BPA)

same segmentation are neither highly similar nor strongly dissimilar. A similar observation holds for the KL metric, where the probability of falling within $3 \times STD$s from the mean KL value (as defined in 4) also lies within approximately the same interval. These results suggest that the KM, HCL, and LGR methods exhibit comparable levels of self-consistency.

The restriction to only two clusters (FG/BG) may limit the achievable similarity. Nevertheless, some extreme cases can be observed, including very high KL/KS values (for example, 0.68/0.78 for AV/HCL/KL(KS)/4, corresponding to image AV, block size 4, method HCL, and KL/KS comparison) as well as very low values (for example, 0.14/0.15 for MBA/LGR/KL(KS)). In general, smaller block sizes (for example, size 4) tend to favor higher similarity. Among the examined methods, KM appears to produce more self-consistent segmentation results across block sizes.

4 Results

4.1 Results using Hierarchical Clustering segmentation

Larger block sizes facilitate image "averaging," since a greater number of pixels are examined collectively as either background or foreground. It can also be observed that the above process of varying block sizes

and cluster numbers effectively operates as a spatially iterated Hierarchical Clustering. In this sense, large background or foreground clusters produced using a fixed (or small) number of 2-cluster HCL are progressively decomposed into smaller-range HCL segments . Conceptually, this is equivalent to applying separate foreground / background separation operations using a 2-cluster segmentation , not over the entire image, but within previously identified segments classified as either background or foreground.

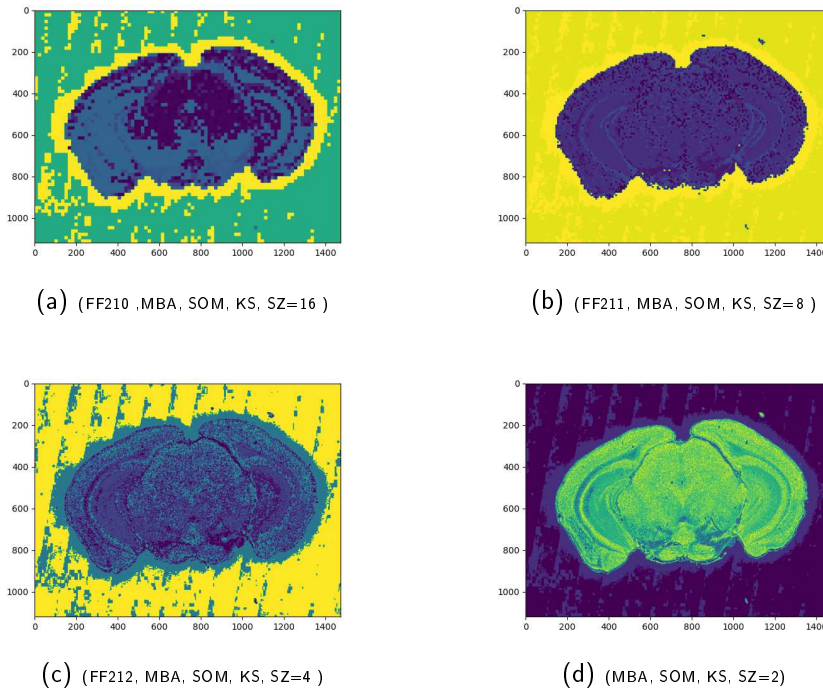
Equivalently, a known background or foreground cluster is further subdivided into smaller background or foreground clusters , applied only to blocks with a known class label (background or foreground). This process introduces a range of $(p-value)$s (and thus variable similarity) within a larger initial background or foreground region.

The results of applying HCL to the image shown in Fig. 1a are presented in Figs. 10a through 8c. The analysis again employs triplets of parameters $((NoOfClusters, BlockSize, FigureID))$: (2, 16, MBA Fig. 8f *'two (1-FG, 1-BG) initial point HCL segmentation with 2 pixels block size segmentation'*), (2, 32, MBA Fig. 8e *'two (2-FG, 2-BG) initial point segmentation with 4 pixels block size'*), (3, 16, MBA Fig. 8a *'three (1-FG, 2-BG) initial point segmentation with 2 pixels block size segmentation'*), (3, 32, MBA Fig. 8b *'three (2-FG, 1-BG) initial point segmentation with 4 pixels block size segmentation'*), and (3, 64, MBA Fig. 8c *'two (2-FG, 2-BG) initial point segmentation with 4 pixels block size segmentation'*).

In figures such as Fig. 8f, the brain areas appear less textured, allowing clearer visualization of the foreround / background segmentation results as the number of clusters changes. For a low number of clusters (for example, 2), the segmentation essentially corresponds to a brain versus non-brain separation. As the number of clusters increases, the brain tissue begins to separate into subregions. For larger block sizes, the brain areas become more coherent, but spatial granularity decreases, and finer brain sub-areas tend to vanish.

4.2 Results from applying SOM on input image

The KS self-consistency values for the MBA images are illustrated in Figs. 4c through 4a, corresponding to block sizes in the array $(2, 4, 8, 16)$. For brevity, this collection of results is referred to as the (MBA-SOM-



(a) (FF210 ,MBA, SOM, KS, SZ=16)

(b) (FF211, MBA, SOM, KS, SZ=8)

(c) (FF212, MBA, SOM, KS, SZ=4)

(d) (MBA, SOM, KS, SZ=2)

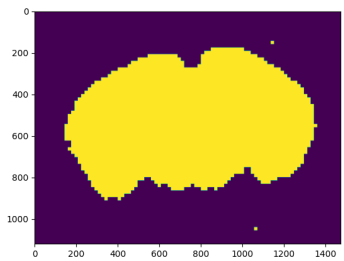
Figure 4: GRP4 KS test consistency for SOM segmentation on MBA image 1a for different block-sizes

KS) set of maps.

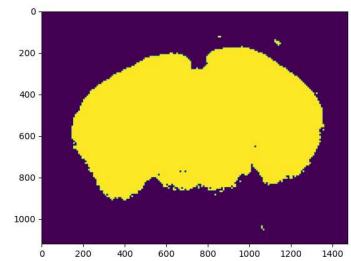
Similarly, the KL self-consistency values for the MBA images are shown in Figs. 6a through 6d, for block sizes in the same array $(2, 4, 8, 16)$. These results are denoted as the (MBA-SOM-KL) set of maps.

By examining the (MBA-SOM-KS) and (MBA-SOM-KL) sets, it becomes evident that the multi-class nature of SOM segmentation is more clearly captured by the KS tests. Since the KS test quantifies the confidence that blocks originate from different underlying data sources (or not), it provides richer segmentation detail, whereas the KL-based evaluation conveys substantially less structural information.

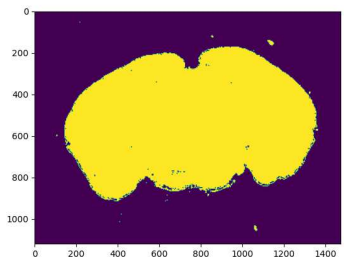
For clearer comparison, the results are also presented in a compact form as pairs of $(NoOfCusters,\ BlockSize)$, complemented by the corresponding figure ID, thus forming triples. More specifically, for Fig. 1a (MBA), one obtains the triple (4, 4, Fig. 9c), which corresponds



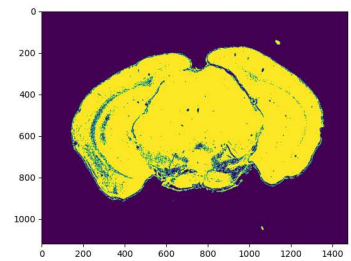
(a) FF214 (MBA, SOM, KL, SZ=16)



(b) FF217 (MBA, SOM, KL, SZ=8)

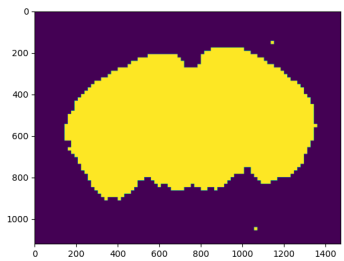


(c) FF216 (MBA, SOM, KL, SZ=4)

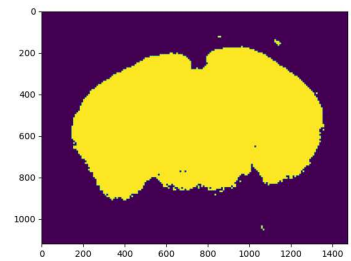


(d) FF215 (MBA, SOM, SZ=2

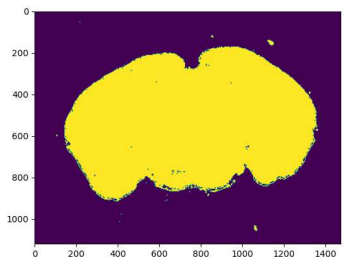
Figure 5: GRP5 KL test consistency for SOM segmentation on MBA image 1a for different block-sizes



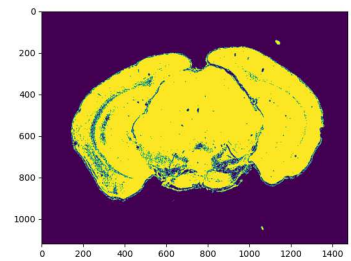
(a) FF214 (MBA, SOM, KL, SZ=16)



(b) FF217 (MBA, SOM, KL, SZ=8)



(c) FF216 (MBA, SOM, KL, SZ=4)



(d) FF215 (MBA, SOM, SZ=2)

Figure 6: GRP6 KL test consistency for SOM segmentation on MBA image 1a for different block-sizes

to a *'4-cluster KM segmentation with 4 pixels block size'*.

4.3 Results using the KL-LGR method

The KL resemblance method using LGR classification was computed, ideally, over the defined range of block sizes (that is, the (D_i)s), in order to cover as wide a range of FG/BG overlaps as possible. The term ideally refers to the fact that the Hierarchical Clustering method does not support the full range of (D_i)s, due to its hierarchical nature, which leads to excessive computational cost and an explosion in cluster data for very small block sizes. In addition, very fine blocks are not always compatible with very large block sizes (for example, 32) in foreground / background separation applications, since strong FG/BG overlap may cause the KL match to deteriorate.

Another important parameter is the selection of the initial points, both in terms of their spatial location and whether they are sampled from known FG or BG regions. Representative combinations of FG/BG seed pixels are illustrated, for example, in Fig. 7a.

The results of this analysis are presented using triplets of execution parameters of the form $((NoOfClusters,\ BlockSize,\ FigureID))$, namely: (2, 2, MBA Fig. 7a *'two (1-FG, 1-BG) initial point segmentation with 2 pixels block size segmentation for the image in Fig. 1a'*), (4, 4, AV Fig. 7b *'two (2-FG, 2-BG) initial point segmentation with 4 pixels block size for the image in Fig. 1a'*), and (4, 4, MBA Fig. 7b).

The results obtained by applying LGR segmentation to the image in Fig. 1a (MBA), as shown in Fig. 7a, indicate that a finer segmentation using four clusters (Fig. 7b) reveals more FG details than the corresponding two-cluster segmentation in Fig. 7a. This behavior can be interpreted as an *'iterative Hierarchical Clustering'* segmentation process, in which larger FG or BG regions are progressively decomposed into smaller ones by applying the LGR method locally, that is, solely within regions already identified as FG or BG pixels.

In the AV image, there are no obvious foreround / background regions except for the single particle or disc located near the top right, corresponding to the darker part of the image. As a result, the image content is dominated by rich texture variations that depend strongly on block size. Here, "following" means that texture exhibits more periodic behavior than large homogeneous regions, and this periodicity

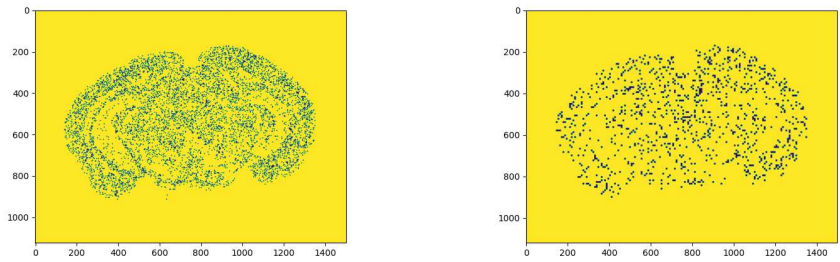


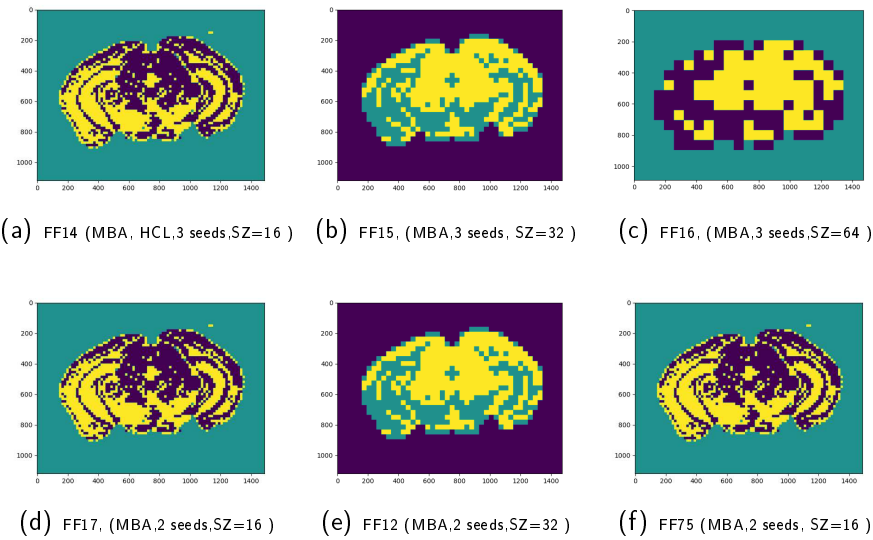
Figure 7: GRP7 Results using the LGR segmentation on image 1a for different block-sizes and LGR clusters

can be captured at different scales depending on the size of the kernel used to analyze it. Thus, different block sizes reveal different periodic structures.

Various combinations of initial (seed) points were selected, with at most two pixels belonging to either the FG or the BG. This procedure does not constitute supervised or Gound Truth -based classification , given that the image size is approximately (1500×1125 pixels) for Fig. 1a (MBA).

The results of applying the KL metric to LGR classification can be observed in the figures ranging from Fig. 7a to Fig. 7a. The LGR-based results exhibit substantial merging of FG and BG pixels, as shown in Fig. 7a. Similar merging behavior is observed when increasing the block size from 2 to 4, as in Fig. 7b, which is derived from Fig. 1a. The image "averaging" effect associated with larger block sizes is clearly visible in Fig. 7b.

If one compares the figures in Fig. 7b and Fig. 7b, which correspond to different methods (LGR and HCL) applied to the same input image shown in Fig. 1a, one can observe that the HCL method produces more homogeneous and spatially coherent regions, both for low and for high numbers of clusters . One reason for this behavior is that LGR is essentially a local foreground / background separation probability model, in which class assignment is performed locally using from 2 to 4 seed pixels, as discussed earlier. Consequently, some blocks or pixels conform



(a) FF14 (MBA, HCL,3 seeds,SZ=16) (b) FF15, (MBA,3 seeds, SZ=32) (c) FF16, (MBA,3 seeds,SZ=64)

(d) FF17, (MBA,2 seeds,SZ=16) (e) FF12 (MBA,2 seeds,SZ=32) (f) FF75 (MBA,2 seeds, SZ=16)

Figure 8: GRP8 Results using the HCL segmentation on both image MBA for different block-sizes and numbers of clusters

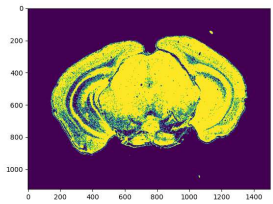
well to this local model, while others do not, leading to more sparse or scattered segmentation results.

In contrast, HCL relies on more global, image-wide criteria to compute distances among blocks before deciding whether to merge or split groups of blocks into foreground or background regions. This global perspective promotes the formation of larger, more uniform regions and explains the more solid appearance of the resulting segmentation maps.

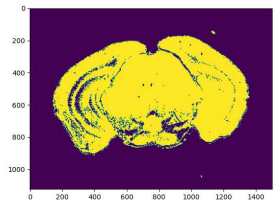
In the figures from Fig. 9a to Fig. 9b, the results of applying the K-Means algorithm to the input image shown in Fig. 1a are presented. The KM FG/BG separation method also tends to produce larger and less sparse segments . In particular, one can observe the limited penetration of isolated FG spots into larger FG segments , and vice versa.

Therefore, improved image segmentation is better supported by foreground / background separation, as adopted in this work.

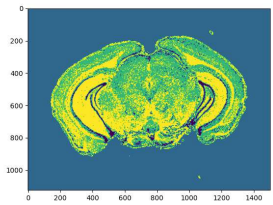
The results from the analysis of Kullback Leibler divergence divergence evaluation on HCL segmentation is outlined as triplets of execution parameters of the type $((NoOfClusters,\ BlockSize,\ FigureID))$ and are: (2, 8,MBA Fig.10a *'2-bin/bucket histogramsfor the KL metric*



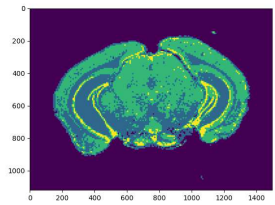
(a) FF37, MBA-KM, (2 seeds,SZ=2)



(b) FF38, MBA-KM, (2 seeds,SZ=4)

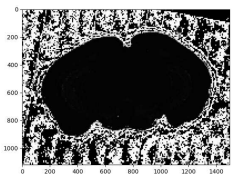


(c) FF34 , MBA-KM, (4 seeds,SZ=4)

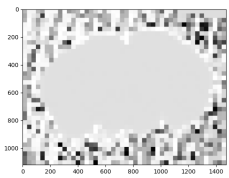


(d) FF36, MBA-KM, (4 seeds,SZ=8)

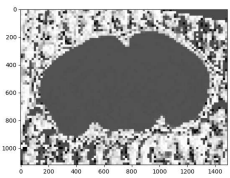
Figure 9: GRP9 Results using the SOM segmentation on image 1a for different block-sizes and SOM clusters



(a) FF73, MBA-HCL-KL,(2 seeds,SZ=8)

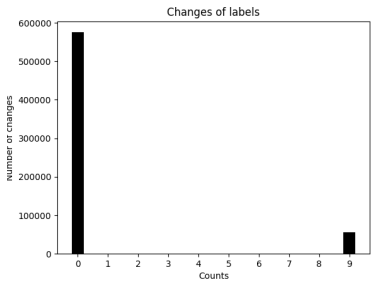


(b) FF74, MBA-HCL-KL,(2 seeds,SZ=32)

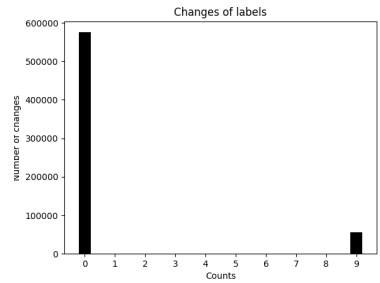


(c) FF75, MBA-HCL-KL,(2 seeds,SZ=16)

Figure 10: GRP10 Results using the the KL similarity on HCL segmentation on image 1a for different block-sizes and Hierarchical Clustering clusters



(a) FF88, MBA Label changes counts using using Logistic Regression for image Fig. 1a



(b) FF98, MBA Label changes counts using using KM for image Fig. 1a

Figure 11: GRP11 Evaluation results using the labels change method and 3 classification methods. Large changes frequencies show low segmentation performance. The parameters are ($Number\ Of\ Changes\ (x-axis), Frequency\ OF\ Number\ Of\ Changes\ (y-axis)$)

applied on blocks from HCL 2cluster segmentation with 8 pixels block size for image in Fig. 1a‘),(2, 32, MBA Fig.10b ‘*2-bin histogramson blocks from HCL 2-cluster segmentation with 32 pixels block size*‘), (2, 16, MBA Fig.10c ‘*2-bin histogramson blocks from HCL 2cluster segmentation with 16 pixels block size*‘).

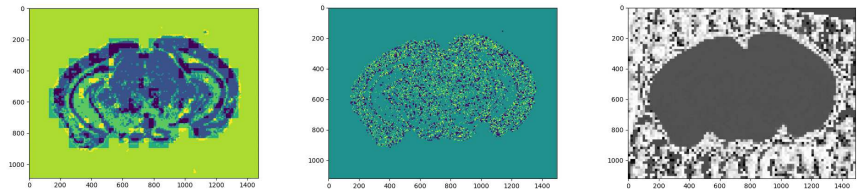
Using the Logistic Regression method on Fig. 1a (AV) with only two clusters (simple foreround / background) and a variable number of block sizes one gets the segmentation as seen in Fig. 7a.

4.4 Evaluation of results using label changes

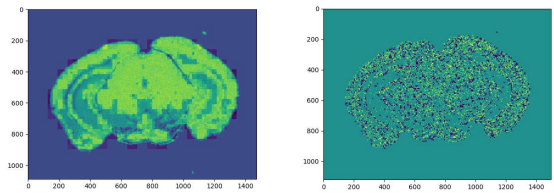
Evaluation maps produced, as in section 3.3.2, can be seen in the figure Fig.12a when HCL is applied on the MBA image or Fig. 1a, and in Fig. 12e when LGR is applied on MBA, or Fig. 1a.

Relevant histograms for label change maps (the (N_L)s of section 3.3.2) that show the analogy of pixels with varying IDs over those with stable labels can be seen in Fig. 11b.

Overall, the KM classification proved better for foreground / background separation taking into account the standard deviationof the pixel-label changes, as well as the maps seen in Fig. 13c that represent label-change frequencies over the series of trials.

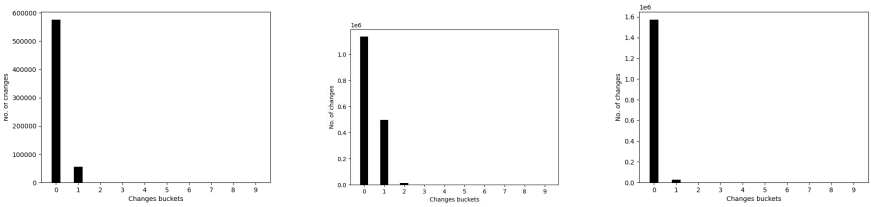


(a) F78, MBA (b) F80, MBA (c) FF90, MBA



(d) FF84, MBA (e) FF85, KM, MBA

Figure 12: GRP12 Evaluation results using the labels change map method. Light areas are more "vague"(many label changes) pixels while darker ones are known fixed label ones (either FG orBG)



(a) FF95, MBA-KM, Fig.1a (b) FF97, MBA-HCL, Fig.1a (c) FF100, MBA-LGR, Fig.1a

Figure 13: GRP13 Evaluation results using the labels change per image and classification method used

5 Limitations

The present work makes no use, or very limited use, of prior knowledge. This entails both risks and be$(KL = 0)$ or $(KS \quad PV = 0.01 \quad for \quad over-fit)$ benefits. While the benefits have already been discussed, the main limitations are that the method can be used primarily as a clustering schema combining statistics and ML, and that most parameters cannot be tuned in a systematic way, as previously noted. These are only two examples. Still, in most programming languages, evaluation modules can be implemented to run the entire schema over a highly variable set of hundreds of images and produce global statistics for maps and self-consistency. The methods (LGR, LR, HCL, KM, significance tests, KL, KS) are not directly comparable with other approaches due to their generic nature. For example, the present work already compares the above mainstream methods using the maps presented. However, methods that rely on weakly supervised schemes do not fall into the same category as the approach proposed here. For instance, one cannot directly compare a method that adopts an explicit background model, such as MRF (Markov Random Fields), with the present one, which only evaluates inner similarity among BG/FG blocks. The former can achieve higher accuracy due to BG/FG learning and training, whereas the latter is generally less accurate but more robust, as it can be applied to a wide variety of backgrounds. Another limitation is the strong dependence on the image statistics. Typically, a model is designed to address very specific needs and applications and can therefore yield more meaningful results. For example, identifying neuronal synapses in brain regions would benefit from a trained model. The present work instead offers elasticity (flexibility), allowing the detection of multiple clusters that may belong to a single synapse region, without explicitly identifying that these clusters represent the same structure. A trained model could be used to merge them into the same cluster . Different success metrics for GMM, KM, or BAP methods have been reported in the literature, since the approach presented here is not found elsewhere. Table 1 provides a rough conversion between KS and KL segmentation evaluation metrics and the conventional % success rate used in other computer vision applications, mainly in the video analysis domain. This conversion is based on the analogy that a perfect KL resemblance, corresponding to $(KL = 0)$ or $(KS' \quad PV = 0.01 \quad for \quad over - fit)$,

maps to a success rate above 95%. The idea pursued was to use an empirical relationship of the form $PRC(\%) = F(KL)$, linking success rates (*'PRC'*) expressed as percentages and reported in the literature to the KL metric used here, which is expressed as a value in $\in [0, \infty]$. This relationship can be expressed as $PRC = a * exp^{-b*KL}$. By imposing the conditions that when $KL = 0$ then $PRC = 100$, and when $KL \to \infty$ then $PRC = 0$, one obtains $a = 100$ and b as an arbitrary positive constant ($b = 1$). Applying these values back to the empirical relationship yields $PRC = 100 * exp^{-KL}$, or equivalently $KL = log(\frac{100}{PRC})$.

References

[Dong2018SingleLabelMI] Qi Dong and Xiatian Zhu and Shaogang Gong, *Single-Label Multi-Class Image Classification by Deep Logistic Regression*. ArXiv. [Online]. Vol. abs/1811.08400. pp. 086001, 2018. doi = 10.1117/1.2338565. Available: https://arxiv.org/abs/1811.08400.

[Diana2025] Diana Bratić, Petar Miljković, Denis Jurečić, and Tvrtko Grabarić, *AI-Driven Random Forest Model and the Six Sigma Approach for Enhancing Offset Printing Process and Product Quality*. Appl. Sci. 2025, 15(5), 2266. DOI = https://doi.org/10.3390/app15052266.

[DebaniPrasadMishra] Debani Prasad Mishra, Sanhita Mishra, and et al., "Image classification using machine learning". *Indonesian Journal of Electrical Engineering and Computer Science*, [Online], vol. 31, no. 3, pp. 1551-1558, 2018. doi = 10.11591/ijeecs.v31.i3.pp1551-1558. Available: https://www.ijeecs.iaescore.com/index.php/IJEECS/article/view/1551.

[HamptonCM] Hampton CM, Liu J, Taylor DW, and et al., "The 3D structure of villin as an unusual F-Actin crosslinker". *Structure*, vol. 10, no. 16, pp. 1882-1891, 2008. doi = 10.1016/j.str.2008.09.015. [PMID: 19081064, PMCID: PMC2782859]. Available: https://www.sciencedirect.com/science/article/pii/S0969212608002787.

[Hampton2008] Hampton CM, Liu J, Taylor DW, et al., "The 3D structure of villin as an unusual F-Actin crosslinker", *Structure*, vol. 10, no. 16, pp. 1882-1891, 2008. doi = 10.1016/j.str.2008.09.015. Available: https://www.researchgate.net/figure/Electron-Micrographs-of-F-ActinVillin-Crosslinks-A-Electron-micrograph-of-fig4_23661890.

[Bohland2009] Bohland JW, Wu C, et al., "A Proposal for a Coordinated Effort for the Determination of Brainwide Neuroanatomical Connectivity in Model Organisms at a Mesoscopic Scale", *PLoS Computational Biology*, vol. 5, no. 3, pp. e1000334, 2009. doi = https://doi.org/10.1371/journal.pcbi.1000334.

[JeongM] Jeong M, Kim Y, et al., "Comparative three-dimensional connectome map of motor cortical projections in the mouse brain", *Scientific Reports*, vol. 2, no. 6, p. 20072, 2016. doi = https://doi.org/10.1038/srep20072.

[Meng2018] Meng Kuan Lin, Yeonsook Shin Takahashi, et al., "A High-throughput Neurohistological Pipeline for Brain-Wide Mesoscale Connectivity Mapping of the Common Marmoset", *bioRxiv*, vol. 1, pp. 123-130, 2018. doi = https://doi.org/10.1101/315804.

[Phalempin] Phalempin M., Lippold E., Vetterlein D., et al., "An improved method for the segmentation of roots from X-ray computed tomography 3D images: Rootine v.2", *Plant Methods*, vol. 17, p. 39, 2021. doi = https://doi.org/10.1186/s13007-021-00735-4.

[RyanThiermann] Ryan Thiermann, Michael Sandler, et al., "Tools and methods for high-throughput single-cell imaging with the mother machine", *eLife*, vol. 12, article RP88463, 2024. doi = https://doi.org/10.7554/eLife.88463.4.

[JYLee] J. Y. Lee, R. H. Park, J. Kim, "Occlusion Handling by Successively Excluding Foregrounds for Light Field Depth Estimation Based on *Foreground-Background Separation*", *IEEE Access*, vol. 9, pp. 103927-103936, 2021. doi = https://doi.org/10.1109/ACCESS.2021.3098819, Available: https://www.researchgate.net/publication/353390462_Occlusion_Handling_by_Successively_Excluding_

Foregrounds_for_Light_Field_Depth_Estimation_Based_on_Foreground-Background_Separation.

[LiShengli] Li Shengli, Cao Zihao Li, et al., "Complex Background Removal Method for Video of an Inclined Cable Based on UAV Detection of Structural Vibration Frequencies", *Structural Control and Health Monitoring*, vol. 1630928, 25 pages, 2024. doi = https://doi.org/10.1155/2024/1630928, Available: https://onlinelibrary.wiley.com/doi/epdf/10.1155/2024/1630928.

[HWei] H. Wei and Q. Peng, "A block-wise Frame Difference method for real-time video motion detection", *International Journal of Advanced Robotic Systems*, vol. 15, no. 4, pp. 123-134, 2018. doi = 10.1234/abcd5678, Available: https://journals.sagepub.com/doi/10.1177/1729881418783633.

[ChaochaoZhou] Chaochao Zhou, Ramez N Abdalla, Dayeong An, Syed H A Faruqui, Teymour Sadrieh, Mohayad Alzein, Rayan Nehme, Ali Shaibani, Sameer A Ansari, Donald R Cantrell, "Reducing motion artifacts in craniocervical background subtraction angiography with deformable registration and unsupervised deep learning", *Radiology Advances*, vol. 1, no. 3, pages N/A, 2024. doi = https://doi.org/10.1093/radadv/umae020, Available: https://academic.oup.com/radadv/article/1/3/umae020/7727676?login=false.

[Mansournia2018SeparationIL] Mohammad Ali Mansournia, Angelika Geroldinger, Sander Greenland, Georg Heinze, "Separation in Logistic Regression: Causes, Consequences, and Control", *American Journal of Epidemiology*, vol. 187, no. 4, pp. 864-870, 2018. doi = https://doi.org/10.1093/aje/kwx319, Available: https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2024.1393564.

[JoseLuisVilas] Jose Luis Vilas, Jose Maria Carazo, Carlos Oscar Sorzano, *Emerging Themes in CryoEM - Single Particle Analysis Image Processing*, *Chemical Reviews*, vol. 122, no. 17, pp. 13915-13951, 2022. doi = https://doi.org/10.1021/acs.chemrev.1c00850.

[Xuebin] Xuebin Liu, Yuang Chen, et al., *Foreground-background separation and deblurring super-resolution method*, *Optics and Lasers in Engineering*, vol. 184, no. 1, p. 108629, 2025. doi = https://doi.org/10.1016/j.optlaseng.2024.108629, Available: https://www.sciencedirect.com/science/article/pii/S0143816624006079.

[DebaniPrasadMishra] Debani Prasad Mishra, Sanhita Mishra, et al., *Image classification using machine learning*, *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 3, pp. 1551-1558, 2018. doi = https://doi.org/10.11591/ijeecs.v31.i3.pp1551-1558.

[SuraSabahKeiteb] Sura Sabah Keiteb, Entsar Arebe Fadam, "Face Image Recognition by Using Some Dimension Reduction Algorithms and Logistic Regression", *Journal of Economics and Administrative Sciences*, vol. 30, no. 143, pp. 405-419, 2024. doi = https://doi.org/10.33095/2k5d7267, Available: https://www.researchgate.net/publication/385058088_Face_Image_Recognition_by_using_some_Dimension_Reduction_Algorithms_and_Logistic_Regression.

[Pavel2025] Pavel Stefanovič, Olga Kurasova, "An advanced visualization of self-organizing maps by determining data clusters", *Nonlinear Analysis Modelling and Control*, vol. 31, no. 3, p. 1551, 2025. DOI = https://doi.org/10.15388/namc.2025.30.43803, Available: https://www.researchgate.net/publication/396521445_An_advanced_visualization_of_self-organizing_maps_by_determining_data_clusters.

[Gull89a] S. F. Gull, *Developments in maximum-entropy data analysis*, in *Maximum Entropy and Bayesian Methods*, J. Skilling, Ed., pp. 53-71, Kluwer Academic, Dordrecht, 1989.

[Hanson93c] K. M. Hanson, *Introduction to Bayesian image analysis*, in *Medical Imaging: Image Processing*, M. H. Loew, Ed., Proc. SPIE, vol. 1898, pp. 716-731, Feb. 1993. doi = https://doi.org/10.1117/12.154577.

[Moch2020] Moch. Arief Soeleman, Shreyak Yogi, Aris Nurhindarto, *An Improvement for Background Modelling using a Mixture of*

Gaussian and Region Growing in Moving Objects Detection, *Journal of Physics: Conference Series*, vol. 1430, no. 1, p. 012032, 2020. doi = https://doi.org/10.1088/1742-6596/1430/1/012032.

[Aligholi2022] S. Aligholi, R. Khajavi, M. Khandelwal, D. J. Armaghani, *Mineral Texture Identification Using Local Binary Patterns Equipped with a Classification and Recognition Updating System (CARUS)*, *Sustainability*, 2022, vol. 14, p. 11291. doi = https://doi.org/10.3390/su141811291.

[Gila2025] V. Gila, A. Marinovszki, L. Ruskó, L. Ferenczi, I. Megyeri, *Weakly-Supervised Brain Segmentation for MR Images*, *International Journal of Radiation Oncology*Biology*Physics*, vol. 123, no. 1, Supplement, pp. e761-e762, 2025. doi = https://doi.org/10.1016/j.ijrobp.2025.06.3213, Available: https://www.sciencedirect.com/science/article/pii/S0360301625038180.

5.1 Acknowledgements

The author is thankful to the source that made the input image open for research. The source is also properly referenced in the references section.